Department of Mathematics, Washington University, November 6, 2007

Nonparametric Priors For Bayesian Social Science Models with an Application to Bureaucratic Politics

JEFF GILL
Center for Applied Statistics, Washington University

GEORGE CASELLA
Department of Statistics, University of Florida

Supported by NSF Grants: DMS-0631632 and SES-0631588.

Presidential Appointment Contradiction

Divergent Views of Political Appointees

- They add little to the president's already limited control over federal agencies (Fenno 1959, Noll 1971, Kaufman 1981, Beck 1982, Bill Clinton 1998).
- They are the single greatest source of presidential influence over the bureaucracy (Brigman 1981, Moe 1982, 1985, Stewart and Cromartie 1982, Hedge and Menzel 1983, Wood and Anderson 1993, Wood and Waterman 1991, 1993).
- The Agency's creation and legal role matters more for presidential control than who runs them (Howell and Lewis 2002, Kagan 2001, Cohen and Krause 2000, Waterman 1989, Cooper and West 1988) .

Presidential Appointment Contradiction

Maybe It Doesn't Matter

- These executives tend to remain on the job for very short periods of time.
- Two years on average,
  - recent maximum: 2.8 years (Johnson),
  - recent minimum: 1.9 (Ford).
- The famous "Government of Strangers" from Heclo's (1977).
- How is it then possible for political appointees to be a primary source of presidential power while simultaneously serving so briefly?

Presidential Appointment Contradiction

A Question of Loyalty?

- Classic principal-agent perspective: presidents can decrease the costs of monitoring the bureaucracy and offset bureaucratic informational advantages by using *loyal* agents to represent them.
- "Loyal" appointees tend to stay in office longer (Cohen 1986). Which means they are more trusted agents *and* they have more time to learn the intricacies of their jobs.
- However, there is evidence that presidents are not always able to get their preferred nominee (Mackenzie 1987), and often settle for a second or third choice.

Presidential Appointment Contradiction

Difficulty In Recruiting

- Some positions require specific technical knowledge.
- All positions require management and budgetary experience.
- Why Do Potential Nominees Reject the Offer?
  - long hours (73% reported working $>$ 61 hours),
  - low pay relative to the private sector,
  - *1978 Ethics in Government Act* requirements,
  - length and/or hostility of confirmation process,
  - cost and trouble of relocating families.

Bureaucratic Politics Data

- Contains every federal political appointee to full-time positions requiring Senate confirmation from November, 1964 through December, 1984, (*collected by Mackenzie and Light, ICPSR Study Number 8458, Spring 1987*). Biography cases: 1528, survey cases: 532.
- **Outcome Variable:** `stress` as a surrogate measure for self-perceived effectiveness and job-satisfaction, measured as a five-point scale from "not stressful at all" to "very stressful."
- **Explanatory Variables:**
  - ⟼ `Government Experience`,
  - ⟼ `Conservative`,
  - ⟼ `Committee Relationship`,
  - ⟼ `Career.Exec-Compet`,
  - ⟼ `Career.Exec-Liaison/Bur`,
  - ⟼ `Career.Exec-Liaison/Cong`,
  - ⟼ `Career.Exec-Day2day`,
  - ⟼ `Career.Exec-Diff`,
  - ⟼ `Confirmation Preparation`,
  - ⟼ `Hours/Week`,
  - ⟼ `President Orientation`.

Ordinal Model for Survey of Political Executives

- For the observed vector $\mathbf{Y}$: $$P(\mathbf{Y} \leq r|\mathbf{X}) = P(\mathbf{U} \leq \theta_r) = P(\mathbf{X}'\boldsymbol{\beta} + \mathbf{E} \leq \theta_r)$$ $$= P(\mathbf{E} \leq \theta_r - \mathbf{X}'\boldsymbol{\beta}) = F_{\mathbf{E}}(\theta_r - \mathbf{X}'\boldsymbol{\beta}).$$
- A normal assumption on the errors produces the ordered probit specification: $$F_{\mathbf{E}}(\theta_r - \mathbf{X}'\boldsymbol{\beta}) = P(\mathbf{Y} \leq r|\mathbf{X}) = \Phi(\theta_r - \mathbf{X}'\boldsymbol{\beta})$$
- *Conjugate priors* lead to the posterior distribution: $$\pi(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto \prod_{i=1}^{n} \prod_{j=1}^{C-1} \prod_{k=1}^{p} \Big[\Phi(\theta_j - \mathbf{X}_i'\boldsymbol{\beta}) - \Phi(\theta_{j-1} - \mathbf{X}_i'\boldsymbol{\beta})\Big]^{z_{ij}} \exp\left(-\frac{(\boldsymbol{\beta}_k - \mu_{\gamma_k})^2}{2\sigma_\gamma^2} - \frac{\theta_j^2}{2\sigma_\theta^2}\right)$$
- *Improper uniform priors* lead to the simpler posterior distribution: $$\pi(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto \prod_{i=1}^{n} \prod_{j=1}^{C-1} \prod_{k=1}^{p} \Big[\Phi(\theta_j - \mathbf{X}_i'\boldsymbol{\beta}) - \Phi(\theta_{j-1} - \mathbf{X}_i'\boldsymbol{\beta})\Big]^{z_{ij}}$$

The Use of Prior Distributions in Political Science

Subsequent Questions

- When do priors actually matter in political science research?
- What is the best way to specify known prior information?
- Why do Bayesian political scientists typically default to uninformed priors?
- Why are reviewers typically skeptical about informed priors, even with lots of supporting evidence?
- Topically:

Can priors that use the data help us recover latent hierarchical information in the data?

Mixtures as a Starting Point

- Consider first a discrete K-mixture form for the outcome y:

$$f(y|\phi) = \sum_{k=1}^{K} \omega_k \phi(y|\boldsymbol{\theta}_k),$$

  where all $\omega_k \geq 0$, and $\sum_{k=1}^{K} \omega_k = 1$.

- Suppose instead we make the mixture continuous (equivalent to letting $K \to \infty$):

$$f(y|g) = \int f(y|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

  where there are often further dependencies: $g(\boldsymbol{\theta}|\xi, \mathbf{Z} \ldots)$.

- Rather than specifying some parametric form here, a Dirichlet process prior is placed on the mixing distribution, now denoted G.

Mathematical Definitions

- Y is a random variable taking values on the measurable space $(\mathcal{Y}, \mathfrak{B})$, defined by the support of Y and an arbitrary (for now) abstract space.
- The "parameter" of interest here is P, the associated, but *unknown*, probability measure taking values in $\mathcal{P}$, the collection of *all* probabilities measures on $(\mathcal{Y}, \mathfrak{B})$.
- Define $\mathcal{C}$ as the smallest σ-field generated by sets of the form:

$$\{P : P(A) < r\}, \qquad \text{where:} \quad A \in \mathfrak{B},\ r \in [0:1]$$

- Now define ν as a probability measure on $(\mathcal{P}, \mathcal{C})$, which can be used as a prior distribution for the unknown P.
- We are interested in computing ν^*, the posterior distribution of $P|Y$.
- ν is called a Dirichlet measure if for every measurable partition $\{B_1, \ldots, B_K\}$ (and finite K) of the parameter space, the distribution of $P(B_1), \ldots, P(B_K)$ under ν is a Dirichlet: $f(\mathbf{y}|\alpha_1, \ldots, \alpha_K) \propto y_1^{\alpha_1 - 1} \cdots y_K^{\alpha_K - 1} \quad 0 \leq y_i \leq 1, \sum_{i=1}^{K} y_i = 1, 0 < \alpha_i, \forall i \in [1, 2, \ldots, K]$.

Putting Some Distributional Structure On These Definitions

- Ferguson (1973, 1974) and Antoniak (1974) introduced the Dirichlet process prior for nonparametric G, which is this random probability measure on the space of all measures.
- We notate this distribution over the space of distributions by:
  - G_0, a base distribution (finite non-null measure) which forms the expected value of the distributions,
  - $m > 0$, a concentration precision parameter (finite and non-negative scalar) giving the spread of distributions around G_0,
  - therefore mG_0 is a base measure,
  - leading to the prior specification $G \sim \mathcal{DP}(m, G_0) \in \mathcal{P}$.
- For *any* finite partition of the parameter space, $\{B_1, \ldots, B_K\}$, the joint distribution of the probabilities has the Dirichlet distribution:

$$(G(B_1), \ldots, G(B_K)) \sim \mathcal{D}(mG_0(B_1), \ldots, mG_0(B_K),$$

where for some *given* partition, these are just multinomial probabilities.

General Model Specification

- This works in the modeling sense as follows...
  - The data, y_i, are exchangeable from an unknown mixture of distributions with parameters $\boldsymbol{\phi}_i$:

$$y_i|\boldsymbol{\phi}_i \sim f_i(y_i|\boldsymbol{\phi}_i, \mathbf{X}).$$

  - The mixing distribution over $\boldsymbol{\phi}_i$ is given by:

$$\boldsymbol{\phi}_i|G \sim G.$$

  - The prior for this mixing distribution comes from a Dirichlet process influenced from the data with parameters m and G_0.

$$G|m, G_0 \sim \mathcal{DP}(m, G_0)$$

- Thus the joint density for $\mathbf{y}$ after integrating over $\boldsymbol{\phi}$ is $\prod_{i=1}^{n} f_i(y_i|G)$.
- All such models need to somehow bypass the infinite-dimension G and estimate the parameters of interest.

Restating the Model for Ordinal Outcomes

- The standard ordered probit model assumes first that there is a multinomial selection process:

$$Y_i \sim \text{Multinomial}(1, (p_1, p_2, \ldots, p_C)), \quad i = 1, \ldots n$$

where $\sum_j p_j = 1$, and $Y_i = (y_{i1}, \ldots, y_{iC})$ is a $C \times 1$ vector with a 1 in one position and 0 elsewhere. The placement of the 1 denotes the class that the observation falls into.

- The p_j are ordered by the probit model

$$p_j = P(\theta_{j-1} \leq U_i \leq \theta_j)$$

where these cutpoints have the property that $-\infty = \theta_0 < \theta_1 < \cdots < \theta_C = \infty$.

- Defining the random quantity:

$$U_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \psi_i, \sigma^2)$$

where $\mathbf{X}_i$ are covariates associated with the i^{th} observation, $\boldsymbol{\beta}$ is the coefficient vector, and ψ_i gives a random effect to account for subject-specific deviation from the underlying model.

Models for Ordinal Outcomes

- The y_i are drawn from a mixture of distributions, $F(y_i|\psi_i, \boldsymbol{\beta}, \boldsymbol{\theta})$.
- The mixing over ψ is dictated by a distribution G, with Dirichlet prior.
- Putting this together gives a mixture model of the form:

$$\begin{aligned} y_i &\sim F(y_i|\psi_i, \boldsymbol{\beta}, \boldsymbol{\theta}) \\ \psi_i &\sim G \\ G &\sim \mathcal{DP}(m, G_{\mu,\tau^2}). \end{aligned}$$

- The $\mathcal{DP}$ defines a distribution over distributions, and given ψ_i, the y_i are independent of G and each other.
- Note that this the most basic form of the Dirichlet process mixtures model and that additional levels of the hierarchy can be specified by through priors on m and the parameters of G.

Setting Up the Estimation Process

- Since realizations of the $\mathcal{DP}$ select a discrete distribution with probability one (even though the generating mechanism is continuous), the model for ψ is a countably infinite mixture (key papers: Ferguson 1973, Berry & Christensen 1979, Lo 1984).
- Blackwell and MacQueen (1973) discovered/invented the following:
  - If G is a $\mathcal{DP}$, where $\psi_1, \ldots, \psi_n$ iid from G,
  - and G is marginalized over its prior distribution,
  - then $\psi_1, \ldots, \psi_n$ are equal in distribution to the first n steps of a Pólya process.
- Therefore reference can be made to a finite rather than infinite dimensions, and Dirichlet process posterior calculations involve a single parameter over this space.

Pólya Process

- The Pólya Process for sampling ψ is equivalent to the following permutation scheme:
  - a restaurant has many large circular tables positioned according to $f()$.
  - n diners enter one-at-a-time to be seated, where the first person sits at the first table.
  - For a given weight, k, the ith person sits at the unoccupied ith table with probability: $k/(k+i-1)$.
  - Otherwise this diner selects from an *occupied* table with uniform probability.
- Now the table locations of the seated diners, $\xi_1, \ldots, \xi_n$, is a dependent exchangeable sequence.
- $\xi^* = (\xi_1, \ldots, \xi_k)$ with $k \leq n$, the set of non-empty tables, is a *sample* from $f()$.
- Also, $N(n)/\log(n) \to k$ (almost surely) as $n \to \infty$, where $N(n)$ is the number of non-empty tables.

The Model for Ψ

- Following Neal (2000), we write the Dirichlet process as a latent class model:

$$\begin{aligned} \psi_{c_i} &\sim g(\psi_i) = \mathcal{N}(\mu_i, \tau_i^2) \\ c_i|\mathbf{q} &\sim \text{Discrete}(q_1, \ldots, q_k) \\ \mathbf{q} &\sim \text{Dirichlet}(m/k, \ldots, m/k), \end{aligned}$$

- Here the c_i are indicators from an unobserved process that serves only to group the ψ_i, and we get the restaurant process by letting $k \to \infty$.
- We get a resulting common value of $\psi_i = \psi_j$ if $c_i = c_j$ so the observations are from the same distribution.
- So on each iteration, the sampler updates: $\mathbf{c}$, k, ψ, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, τ, $\boldsymbol{\Theta}$, and calculates $\mathbf{U}$.
- It is easiest to treat the parameters of the Gibbs Sampler in two blocks: the Dirichlet parameters, and remaining parameters

Priors and Fixed Parameters

- To complete the Bayesian model we add the following priors:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_\beta^2)$$

$$\mu_i \sim \mathcal{N}(0, \tau_i^2)$$

$$\frac{1}{\tau_i^2} \sim \mathcal{GA}(a, b)$$

- The parameters in the priors on μ, and τ^2 are chosen to make the priors sufficiently diffuse to allow the random effect to do its work; σ^2 fixed for identifiability.
- The choice of prior mean zero for ψ does not lose generality, as the $\mathbf{X}_i\boldsymbol{\beta}$ term in $U_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \psi_i, \sigma^2)$ locates the distribution on the metric determined by fixing σ at 1 here.
- Treat m as an MCMC tuning parameter, for now.

Gibbs Sampler Strategy, Updating $\mathbf{c}$

- Define for each possible table, $\ell \in [1 : L],\ L \leq n$:

$$\text{table assignment vector excluding } i: \qquad \mathbf{c}_{-i} = (c_1, c_2, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$$

$$\text{number assigned to the } \ell\text{th table:} \qquad n_{-i,\ell} = \#(c_j = \ell), \quad j \neq i$$

$$\text{probability } i \text{ selects category } j \text{ given current } \psi_i: \qquad f(y_i|\psi_i) = p_j = P(\theta_{j-1} \leq U_i \leq \theta_j),$$

- then, for $i = 1, \ldots, n$, the Pólya process gives us the probability that the ith case selects the ℓth table:

$$P(c_i = \ell|\mathbf{c}_{-i}) \propto \begin{cases} \frac{n_{-i,\ell}}{n-1+m} f(y_i|\psi_i) \text{ if } n_{-i,\ell} > 0 \\ \\ \frac{m}{n-1+m} H_i \text{ if } n_{-i,\ell} = 0 \end{cases} \qquad \text{where:} \qquad H_i = \int_{-\infty}^{\infty} \int_{\theta_{j-1}}^{\theta_j} f(u|\psi) g(\psi)\, du\, d\psi.$$

Gibbs Sampler Strategy

- When the set $\mathbf{c}$ has been updated we can then update $(\psi_1, \ldots, \psi_n)$, which is done with a common value for all $c_i = c$:

$$\psi_C \sim \mathcal{N}\left(\frac{n_C\tau^2 T}{\sigma^2+n_C\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+n_C\tau^2}\right)$$

where $n_c = \#(c_i = c)$ and $T = (1/n_c)\sum_{i\in I_c}(U_i - X_i\boldsymbol{\beta})$.

Gibbs Sampler Strategy

Once the Dirichlet parameters have been generated, the remainder of the sampler is straightforward.

- The conditional distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathcal{N}\left(A^{-1}b, A^{-1}\right)$$

where: $A = \frac{1}{\sigma^2}X'X + \frac{1}{\sigma_\beta^2}I, \quad b = \frac{1}{\sigma^2}X'(\mathbf{U} - \psi) + \frac{1}{\sigma_\beta^2}\beta_0.$

- The distribution of $\boldsymbol{\theta}$ is updated from $\prod_{i=1}^{n}\prod_{j=1}^{C} I[\theta_{j-1} \leq U_i \leq \theta_j]^{y_{ij}}$ and is given by

$$\boldsymbol{\theta}_j \sim \mathcal{U}\left(\max_{i:y_{ij}=1}(U_i, \boldsymbol{\theta}_{j-1}), \min_{i:y_{i,j+1}=1}(U_i, \boldsymbol{\theta}_{j+1})\right)$$

Gibbs Sampler Strategy

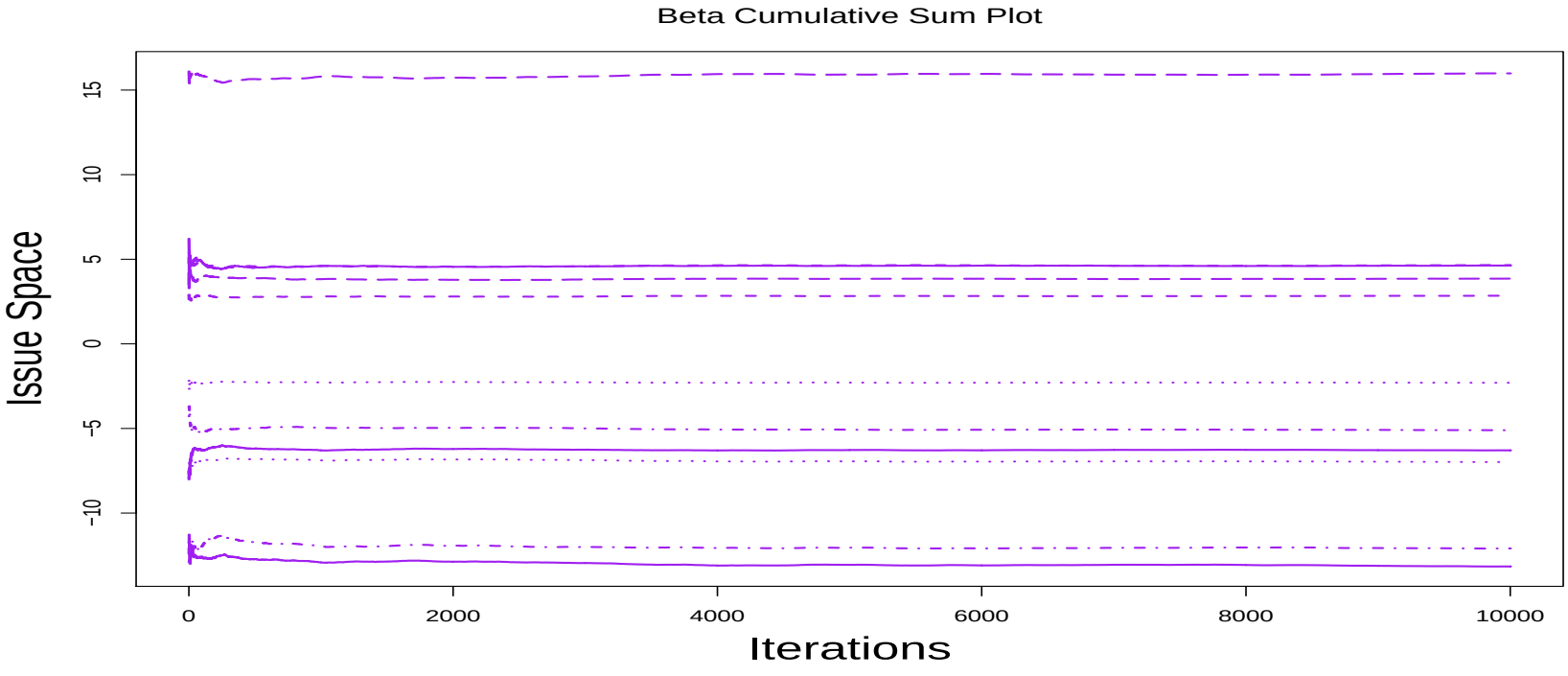
- When updating μ and τ^2 we *only use the distinct values* of $(\psi_1, \ldots, \psi_n)$. Denote these distinct values by $(\psi_{c_1}, \ldots, \psi_{c_r})$, $r \leq n$. Then

$$\mu \sim \mathcal{N}\left(\frac{d\tau^2}{rd\tau^2+1}\sum_{i=1}^{r}\psi_{c_i}, \frac{d\tau^2}{rd\tau^2+1}\right)$$

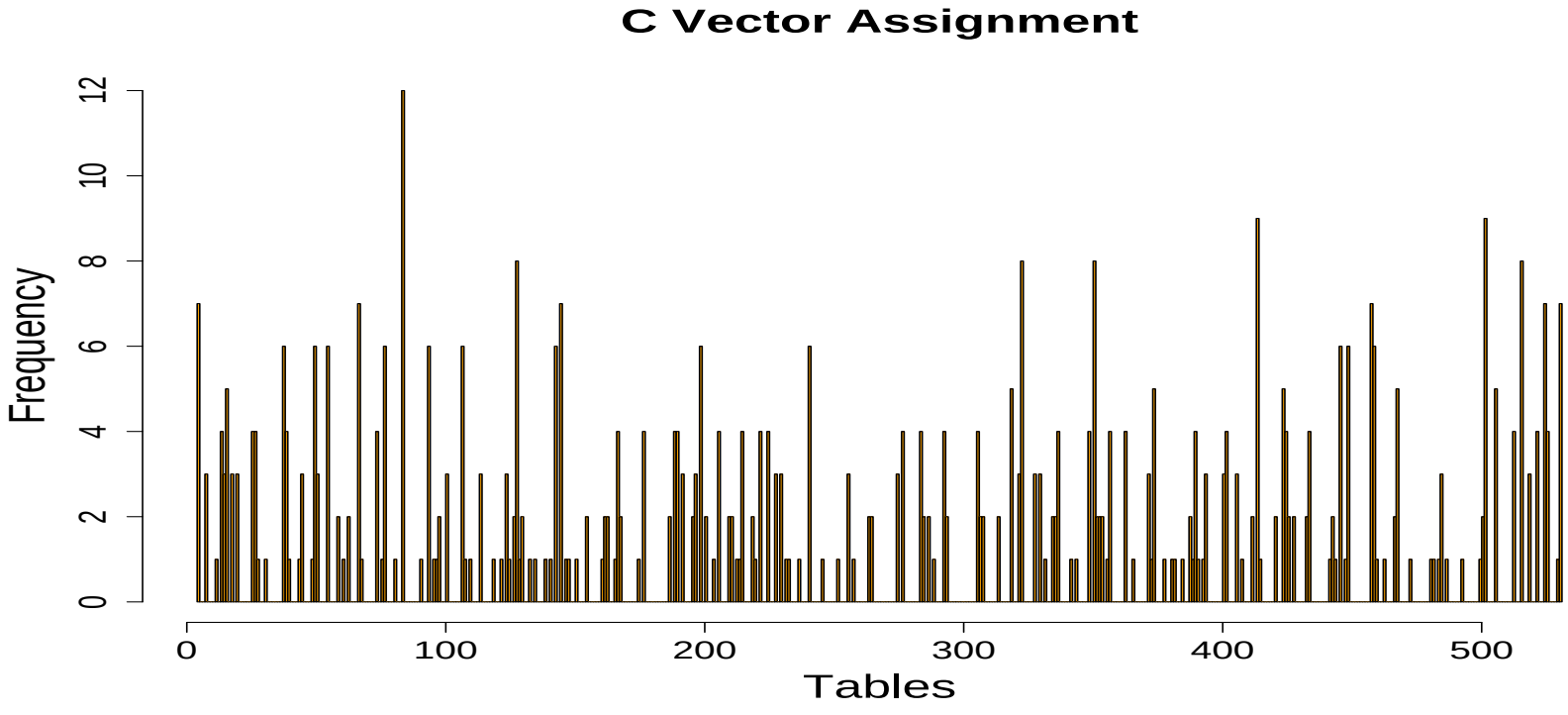
$$\frac{1}{\tau^2} \sim \mathcal{GA}\left(\frac{r+1}{2}+a, \frac{1}{2}\sum_{i=1}^{r}(\psi_{c_i}-\mu)^2+\frac{1}{b}\right)$$

- U_i is updated from $U_i \sim \mathcal{N}\left(X_i\boldsymbol{\beta}+\psi_{c_i}, \sigma^2\right)$ truncated to (θ_{j-1}, θ_j).

Gibbs Sampler Output



Gibbs Sampler Output



Posterior	Mean	95% HD Interval
Government Experience	0.185	[-0.023 : 0.394]
Republican	0.152	[0.026 : 0.278]
Committee Relationship	-0.309	[-0.467 : -0.150]
Career.Exec-Compet	-0.276	[-0.480 : -0.071]
Career.Exec-Liaison/Bur	0.129	[0.046 : 0.340]
Career.Exec-Liaison/Cong	-0.049	[-0.154 : 0.055]
Career.Exec-Day2day	-0.310	[-0.516 : -0.104]
Career.Exec-Diff	0.246	[0.055 : 0.436]
Confirmation Preparation	-0.434	[-0.724 : -0.145]
Hours/Week	0.757	[0.465 : 1.048]
President Orientation	-0.610	[-0.976 : -0.244]
Cutpoints: **(None) (Little)**	0.324	[-0.859 : 0.410]
(Little) (Some)	0.883	[-0.344 : 1.011]
(Some) (Significant)	1.614	[0.266 : 1.864]
(Significant) (Extreme)	3.638	[1.477 : 4.701]

Posterior Mean Comparison, Intra-Bureaucracy Explanatory Variables

	Uninformed Prior	Dirichlet Process Prior
Career.Exec-Compet	-0.175	-0.276
Career.Exec-Liaison/Bur	0.105	0.129
Career.Exec-Liaison/Cong	-0.029	-0.049
Career.Exec-Day2day	-0.153	-0.310
Career.Exec-Diff	0.114	0.246