Technical Appendix for “Estimating Preferred Outcomes from Votes and Text”

July 29, 2016

Abstract

We include here several technical appendices. First, we formally derive the expressions for $c_i^{vote}$ and $b_p^{vote}$ that appear in Section 2 of the paper. The next contains some details on the choice theoretic model underlying the standard topic model. The remainder contain the details for implementing the SFA model.

A The Legislator and Proposal Intercepts from the Voting Model

Let’s start with the legislator’s utility from the *Aye*:

$$U_i^{vote}(\text{Aye}; \{x_{ld}\}_{d=1}^D, \{z_{pd}^{aye}\}_{d=1}^D) = -\frac{1}{2} \sum_{d=1}^D a_d (z_{pd}^{aye} - x_{ld})^2 + \tilde{\xi}_{lp}^{aye} \quad (1)$$

and “Nay” alternatives:

$$U_i^{vote}(\text{Nay}; \{x_{ld}\}_{d=1}^D, \{z_{pd}^{nay}\}_{d=1}^D) = -\frac{1}{2} \sum_{d=1}^D a_d (z_{pd}^{nay} - x_{ld})^2 + \tilde{\xi}_{lp}^{nay} \quad (2)$$

Next, let’s calculate the difference between these expressions to get the legislator’s preference intensity for the *Aye* outcome. Substituting from expressions $^{(1)}$ and $^{(2)}$ we have:
\[ V_{l^*} = U_{l}^{\text{vote}} \left( \text{Aye}; \{ x_{ld} \}_{d=1}^{D}, \{ z_{pd}^{\text{aye}} \}_{d=1}^{D} \right) - U_{l}^{\text{vote}} \left( \text{Nay}; \{ x_{ld} \}_{d=1}^{D}, \{ z_{pd}^{\text{nay}} \}_{d=1}^{D} \right) \]

\[ = -\frac{1}{2} \sum_{d=1}^{D} a_d (z_{pd}^{\text{aye}} - x_{ld})^2 + \tilde{z}_{lp}^{\text{aye}} - \left( -\frac{1}{2} \sum_{d=1}^{D} a_d (z_{pd}^{\text{nay}} - x_{ld})^2 + \tilde{z}_{lp}^{\text{nay}} \right) \]

\[ = \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + \sum_{d=1}^{D} \left( \frac{a_d}{2} \cdot 2x_{ld} (z_{pd}^{\text{aye}} - z_{pd}^{\text{nay}}) \right) + \tilde{z}_{lp}^{\text{aye}} - \tilde{z}_{lp}^{\text{nay}} \]

\[ = \left( \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + E \{ \tilde{z}_{lp}^{\text{aye}} \} - E \{ \tilde{z}_{lp}^{\text{nay}} \} \right) + \sum_{d=1}^{D} a_d x_{ld} (z_{pd}^{\text{aye}} - z_{pd}^{\text{nay}}) \]

\[ = \left( \pi_l^{\text{aye}} - \pi_l^{\text{nay}} \right) + \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + \varphi_l^{\text{aye}} - \varphi_l^{\text{nay}} \]

Now let:

\[ E \{ \tilde{z}_{lp}^{\text{aye}} \} = \pi_l^{\text{aye}} + \varphi_l^{\text{aye}} \] and \[ E \{ \tilde{z}_{lp}^{\text{nay}} \} = \pi_l^{\text{nay}} + \varphi_l^{\text{nay}} \]

substituting this into the first part of the equation (3),

\[ \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + E \{ \tilde{z}_{lp}^{\text{aye}} \} - E \{ \tilde{z}_{lp}^{\text{nay}} \} = \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + \left( \pi_l^{\text{aye}} + \varphi_l^{\text{aye}} \right) - \left( \pi_l^{\text{nay}} + \varphi_l^{\text{nay}} \right) \]

\[ = \pi_l^{\text{aye}} - \pi_l^{\text{nay}} + \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + \varphi_l^{\text{aye}} - \varphi_l^{\text{nay}} \]

\[ = c_l^{\text{vote}} + b_l^{\text{vote}} \]

Now let’s return to the last line of expression (3) and substitute:

\[ V_{l^*} = U_l^1(\{x_{ld}\}_{d=1}^{D}, \{p_d\}_{d=1}^{D}) - U_l^1(\{x_{ld}\}_{d=1}^{D}, \{q_d\}_{d=1}^{D}) \]

\[ = \left( \sum_{d=1}^{D} \frac{a_d}{2} (z_{pd}^{\text{nay}} - z_{pd}^{\text{aye}}) + E \{ \tilde{z}_{lp}^{\text{aye}} \} - E \{ \tilde{z}_{lp}^{\text{nay}} \} \right) - \left( E \{ \tilde{z}_{lp}^{\text{nay}} \} - E \{ \tilde{z}_{lp}^{\text{aye}} \} \right) \]

\[ = c_l^{\text{vote}} + b_l^{\text{vote}} + \sum_{d=1}^{D} a_d x_{ld} g_{pd}^{\text{vote}} - \epsilon_{lp}^{\text{vote}} \]

Equation (4) matches equation (2) in the text.
B Choice Theoretic Underpinnings of Topic Models

While we opt for SFA, it is useful to consider the behavior that would lead one to adopt a topic model for legislative speech. One way to do this is to suppose that a legislator’s speech is generated by the random arrival of opportunities to speak. At each opportunity the legislator must choose one word from a lexicon, which we represent by a $W \times 1$ vector $\omega$, with each entry corresponding to a different word. Each word has a spatial location, which for the moment we place on a single dimension. Legislator $j \in \{1 \ldots V\}$ would derive utility $u(\tilde{w}_l|x_j) + \eta_{j,t}$ from uttering word $j \in \{1 \ldots W\}$ at time $t$. Should the opportunity to speak at time $t$ actually arise, the legislator utters the word offering the greatest utility. To keep things simple we assume that $\eta_{j,t}$ and $\eta_{r,s}$ are independent if either $j \neq r$ or $t \neq s$.

Paralleling the development in Maddala (1983), we operationalize our model with a distributional assumption for $\eta_{j,t} \in \mathbb{R}$, which we take to follow a type I extreme value distribution, with probability density:

$$f(\eta) = e^{-(\eta + e^{-\eta})}$$

and by concretizing the utility function $u(\tilde{w}_l|x_j)$:

$$u(\tilde{w}_l|x_j) = -\frac{1}{2}(\tilde{w}_l - x_j)^2$$

(5)

where $x_j$ is the preferred ideological signal that legislator $j$ would like to convey, and $\tilde{w}_l$ is the ideological connotation of word $i$.

Again following Maddala (1983) we see that the probability that at a randomly chosen time $t$ legislator $j$ prefers word $i$ to all other elements of the lexicon is:
\[ q_{lj} = \frac{e^{u(\tilde{w}_l|x_j)}}{\sum_{k=1}^{W} e^{u(\tilde{w}_k|x_j)}} \]

Let word 1 correspond to a “stop word”. We can rewrite the probability \( j \) uses word \( i \) if she has the opportunity to speak at \( t \) as:

\[ q_{lj} = \frac{e^{u(\tilde{w}_l|x_j)} - u(\tilde{w}_1|x_j)}{\sum_{k=1}^{W} e^{u(\tilde{w}_k|x_j)} - u(\tilde{w}_1|x_j)} \]

substituting from equation (5) into our expression for \( q_{lj} \) we have:

\[ q_{lj}(x, g, b) = \frac{e^{x_j g_l + b_l}}{1 + \sum_{k=2}^{W} e^{x_j g_k + b_k}} \quad (6) \]

where \( g_k = \tilde{w}_k - \tilde{w}_1 \) and \( b_k = -\frac{\tilde{w}_k + \tilde{w}_1}{2} \) for \( k \in \{2 \ldots W\} \).

The probability of an observed \( W \times 1 \) vector \( c \) of word counts is:

\[ \prod_{w=1}^{W} q_{lj}(x, g, b)^{c_w} \quad (7) \]

With the right choice of Dirichlet priors this turns into the latent Dirichlet model of Blei, Ng, and Jordan (2003) if we set \( x_j = g_l = 0 \) for all \( i \) and \( j \). In the ideal point setting, though, \( x_j \) and \( g_i \) correspond with precisely the preferred outcomes and term ideologies with which we are most interested.

Estimation for these models are not straightforward, requiring a Metropolis algorithm or variational approximations. We favor SFA on theoretical grounds, as it allows legislators to select words as a function of their preferred outcomes. We also favor it because it offers a tractable Gibbs sampling scheme for most of the parameters, which we address in the next section.
C  Estimation of SFA

We now shift to a more condensed notation. Hereafter, we reindex the vote and term outcomes using a common index, $j$, which falls into two sets: $J_{term}$ and $J_{vote}$ for whether the observed outcome (now a common $Y_{lj}$) is a term outcome or vote outcome, and $J = |J_{term}| + |J_{vote}|$. We will denote the systematic components of the vote and term selection as

$$\theta_{lp}^{vote} = c_{l}^{vote} + b_{p}^{vote} + \sum_{d=1}^{D} a_{d} x_{ld} g_{pd}^{vote} \quad (8)$$

$$\theta_{lw}^{term} = c_{l}^{term} + b_{w}^{term} + \sum_{d=1}^{D} a_{d} x_{ld} z_{wd}^{term} \quad (9)$$

We will also suppress the superscript for the $\theta_{lw}^{term}$ and $\theta_{lp}^{vote}$ while changing to the joint subscript $j$. The likelihood is given by:

$$L(\theta_{\cdot}^{vote}, \theta_{\cdot}^{term}, \tau_{\cdot}, \bar{T}_{\cdot}, \bar{V}_{\cdot}) = \prod_{l=1}^{L} \left( \prod_{p=1}^{P} Pr\{ V_{lp} = \bar{V}_{lp} | \cdot \} \right)^{w_{lp}^{P}} \prod_{w=1}^{W} Pr\{ T_{lw} = \bar{T}_{lw} | \cdot \}^{w_{lw}^{T}} \cdot (10)$$

where:

$$Pr\{ T_{lw} = \bar{T}_{lw} | \cdot \} = \begin{cases} \Phi (\theta_{lw}^{term} - \tau_{0}) & T_{lw} = 0 \\ \Phi \left( \theta_{lw}^{term} - \tau_{\bar{T}_{lw}} \right) - \Phi \left( \theta_{lw}^{term} - \tau_{\bar{T}_{lw} - 1} \right) & 0 < T_{wl} \end{cases} \quad (11)$$

$$Pr\{ V_{lp} = \bar{V}_{lp} | \cdot \} = \Phi((2\bar{V}_{lp} - 1)\theta_{lp}^{vote}) \quad (12)$$

and the prior structure is given by:
\[ c_l, b_w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1) \]
\[ \mu \sim \mathcal{N}(0, 1) \]
\[ g^\text{wd} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4) \]

\[ \log(\beta_1), \log(\beta_2) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \]
\[ \Pr(a_d) = \frac{1}{2\lambda} e^{-\lambda|a_d|} \]
\[ \Pr(\lambda) = 1.78 e^{-1.78\lambda} \]

Combining the likelihood and prior gives us the posterior:

\[
\Pr(\theta_{lj}, \tau, \beta_1, \beta_2 | Y.) = \prod_{1 \leq l \leq L} \left\{ \Phi (\theta_{lj}) Y_{lj} (1 - \Phi (\theta_{lj}))^{1 - Y_{lj}} \right\} \mathbf{1}(j \in \{ \text{votes} \})
\times \left\{ \Phi (\tau Y_{lj} - \theta_{lj}) - \Phi (\theta_{lj} - \tau Y_{lj} - 1) \right\} \mathbf{1}(j \in \{ \text{terms} \})
\times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \mu^2} \times \prod_{1 \leq l \leq L} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (c_l - \mu)^2} \times \prod_{1 \leq j \leq J} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (b_j - \mu)^2}
\times \prod_{1 \leq d \leq D} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x_{ld})^2} \times \prod_{1 \leq d \leq D} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (g_{jd})^2} \times \prod_{1 \leq d \leq D} \frac{1}{\sqrt{2\pi}} e^{-\lambda|a_d|}
\times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (\log \beta_1)^2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (\log \beta_2)^2} \times e^{-1.78\lambda}
\]

We implement two forms of data augmentation. In the first, for each observation we introduce a normal random variable \( Z^*_{lj} \) as is standard in latent probit models [Albert and Chib, 1993]. This transforms the likelihood into a least squares problem, as:

\[
\Pr(Y_{lj} = k | Z^*_{lj}, \theta_{lj}, \tau, \beta_1, \beta_2) = \prod_{1 \leq l \leq L} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (Z^*_{lj} - \theta_{lj})^2}
\]

The second form of augmentation involves representing the double exponential prior for \( a_d \) to maintain conjugacy. Following [Park and Casella, 2008], we introduce latent variables \( \tilde{\tau}_l \), such that:
\[
d.d. | \tilde{\tau}^2 \sim \mathcal{N}(0_D, \tilde{D}_\tau) \tag{16}
\]
\[
\tilde{D}_\tau = \text{diag}(\tilde{\tau}_1^2, \tilde{\tau}_2^2, \ldots, \tilde{\tau}_D^2) \tag{17}
\]
\[
\tilde{\tau}_1^2, \tilde{\tau}_2^2, \ldots, \tilde{\tau}_D^2 \sim \prod_{1 \leq d \leq D} \frac{\lambda^2}{2} e^{-\lambda^2 \tilde{\tau}_d^2 / 2 d \tilde{\tau}_d^2} \tag{18}
\]
\]

where, after integrating out \( \tilde{\tau}_l^2 \), we are left with the LASSO prior. The proposed method differs from the presentation in Park and Casella (2008) in that we know \( \sigma^2 = 1 \), by assumption.

C.1 The Gibbs Sampler

Next, we outline the Gibbs sampler. All conditional posterior densities are conjugate normals except \( \lambda, \tilde{\tau}, \beta_1, \) and \( \beta_2 \). For a derivation of the posterior densities of \( \lambda \) and \( \tilde{\tau} \), see Park and Casella (2008). We fit \( \beta_1 \) and \( \beta_2 \), which determine \( \tau \), using a Hamiltonian Monte Carlo algorithm, but first we describe the Gibbs updates.

The Gibbs updates occur in two steps. First, we place all data on the latent \( z \) scale. Second, we update all of the remaining parameters. For the first step, we sample as:

\[
Z_{ij}^* \sim \begin{cases}
\mathcal{T} \mathcal{N}(\theta_{ij}, 1, 0, \infty); & Y_{ij} = 1, \ j \in j_{votes} \\
\mathcal{T} \mathcal{N}(\theta_{ij}, 1, -\infty, 0); & Y_{ij} = 0, \ j \in j_{votes} \\
\mathcal{T} \mathcal{N}(\theta_{ij}, 1, \tau_{k-1}, \tau_k); & Y_{ij} = k, \ j \in J_{terms} \\
\mathcal{N}(\theta_{ij}, 1); & Y_{ij} \text{ missing}
\end{cases} \tag{19}
\]

Note that we have ignored missing values up to this point. In the Bayesian framework used here, imputing is straightforward: the truncated normal is replaced with a standard normal, whether term or vote data.

Next, we update all of \( \theta_{ij} \) except for \( \tau \) using a Gibbs sampler, as:
\[ \mu | \sim \mathcal{N} \left( \frac{\sum_{l=1}^{L} \sum_{j=1}^{J} Z_{l,j}^*}{LJ + 1}, \frac{1}{L^2J^2 + 1} \right) \]  
\[ c_l | \sim \mathcal{N} \left( \frac{\sum_{j=1}^{J} Z_{l,j}^*}{J + 1}, \frac{1}{J^2 + 1} \right) \]  
\[ b_j | \sim \mathcal{N} \left( \frac{\sum_{l=1}^{L} Z_{l,j}^*}{L + 1}, \frac{1}{L^2 + 1} \right) \]  
\[ Z_{l,j}^{**} = Z_{l,j}^* - c_l - b_j + \mu \]  
\[ a | \sim \mathcal{N} \left( A^{-1} \tilde{X}^\top \text{vec}(Z^{**}), A^{-1} \right) \]  
where
\[ \tilde{X} = \left[ \text{vec} \left( x_{1}g_{1}^\top \right) : \text{vec} \left( x_{2}g_{2}^\top \right) : \ldots : \text{vec} \left( x_{L}g_{L}^\top \right) \right] \text{ and } \]  
\[ A = \tilde{X}^\top \tilde{X} + T^{-1} \text{ with } T = \text{diag}(\tau_l^2) \]  
\[ x_{l,d} | \sim \mathcal{N} \left( \frac{\sum_{j=1}^{J} Z_{l,j,-d}^{**}a_d g_j}{\sqrt{\sum_{j=1}^{J} (a_d^2 g_j^2 + \frac{1}{4J})}}, \frac{1}{\sqrt{\sum_{j=1}^{J} (a_d^2 g_j^2 + \frac{1}{4J})}} \right) \]  
\[ g_{jd} | \sim \mathcal{N} \left( \frac{\sum_{l=1}^{L} Z_{l,j,-d}^{**}a_d x_{l,d}}{\sqrt{\sum_{l=1}^{L} (a_d^2 x_{l,d}^2 + \frac{1}{4L})}}, \frac{1}{\sqrt{\sum_{l=1}^{L} (a_d^2 x_{l,d}^2 + \frac{1}{4L})}} \right) \]  
where
\[ Z_{l,j,-d}^{**} = Z_{l,j}^{**} - \sum_{d \neq d} x_{l,d} g_{jd} a_d \]  
\[ \tau_l^2 | \sim \text{InvGauss} \left( \sqrt{\frac{\lambda^2}{a_d^2}}, \lambda^2 \right) \]  
\[ \lambda^2 | \sim \text{Gamma} \left( L + 1, L \sum_{l=1}^{L} \frac{\tau_l^2}{2} + 1.78 \right) \]  

C.2 The Hamiltonian Monte Carlo Sampler

We have no closed form estimates for the conditional posterior densities of \( \beta_1 \) and \( \beta_2 \). To estimate these, we implement a Hamiltonian Monte Carlo scheme adapted directly from [Neal (2011)].
adapt the algorithm in one important manner: rather than taking a negative gradient step, we calculate the numerical Hessian and take a fraction ($\alpha$) of a Newton-Raphson step at each. We select $\alpha$ so that the acceptance ratio of proposed ($\beta_1, \beta_2$) is about .4.

Specifically, let $\hat{\text{dev}}(\beta_1, \beta_2)$ denote the estimate deviance at the point $(\beta_1, \beta_2)$. Define the numerical gradients, $\hat{\nabla}_1\text{dev}(\beta_1, \beta_2)$ and $\hat{\nabla}_2\text{dev}(\beta_1, \beta_2)$ as the estimated gradient at $(\beta_1, \beta_2)$ and $\hat{\nabla}_{11}\text{dev}(\beta_1, \beta_2)$, $\hat{\nabla}_{22}\text{dev}(\beta_1, \beta_2)$, and $\hat{\nabla}_{12}\text{dev}(\beta_1, \beta_2)$ as the cross derivative. Next, define the empirical Hessian as:

$$\hat{H}(\beta_1, \beta_2) = \begin{pmatrix} \hat{\nabla}_{11}\text{dev}(\beta_1, \beta_2) & \hat{\nabla}_{12}\text{dev}(\beta_1, \beta_2) \\ \hat{\nabla}_{12}\text{dev}(\beta_1, \beta_2) & \hat{\nabla}_{22}\text{dev}(\beta_1, \beta_2) \end{pmatrix}$$  \hspace{1cm} (30)

We implement the algorithm in Neal (2011) exactly, except instead taking updates of the form:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}^+ := \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \alpha \begin{pmatrix} \hat{\nabla}_1(\beta_1, \beta_2) \\ \hat{\nabla}_2(\beta_1, \beta_2) \end{pmatrix}$$  \hspace{1cm} (31)

we instead do updates of the form:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} := \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \alpha \times \left\{ \hat{H}(\beta_1, \beta_2) \right\}^{-1} \begin{pmatrix} \hat{\nabla}_1(\beta_1, \beta_2) \\ \hat{\nabla}_2(\beta_1, \beta_2) \end{pmatrix}$$  \hspace{1cm} (32)

where the Hessian and gradients are updated every third update of the parameters. The step length parameter $\alpha$ is adjust every 50 iterations to by a factor of $4/5$ if the acceptance rate is below 10%, $5/4$ if the acceptance rate is above 90%, and left the same otherwise. After the burn-in period, the acceptance rate levels off around 45%. We implement twenty steps in order to produce a proposal.
C.3 Numerical Approximation of the Deviance

Calculating the gradient and Hessian terms, and assessing the proposal, in the Hamiltonian Monte Carlo scheme requires evaluating functions of the form $l(a, b) = \log(\Phi(a) - \Phi(b))$. Unfortunately, for values of $a$ and $b$ much larger in magnitude than 5.3 produces returns values of 1 or 0, leaving it impossible to evaluate the logarithm.

Extrapolating from the observed values yields the linear approximation:

$$l(a, b) = \begin{pmatrix} 1 \\ a \\ b \\ a^2 \\ b^2 \\ \log(|a - b|) \\ \{\log(|a - b|)\}^2 \\ ab \end{pmatrix}^\top \gamma$$  \hspace{1cm} (33)

where
We derived the values for $\gamma$ from fitting a model over the range $4 \leq b < a \leq 8$. We get a mean absolute error of 0.0165, or 0.08% error as a fraction of the value returned by R. We use this approximation in order to extrapolate to values where R returns values of NA or Inf for $f(a,b)$.
References


