A Bayesian Nonparametric Approach to Topic Consistent Latent Variables: Application in the U.S. Congress

Kevin McAlister

University of Michigan - Ann Arbor

July 19th, 2018

Bill Level Context in Roll Call Voting

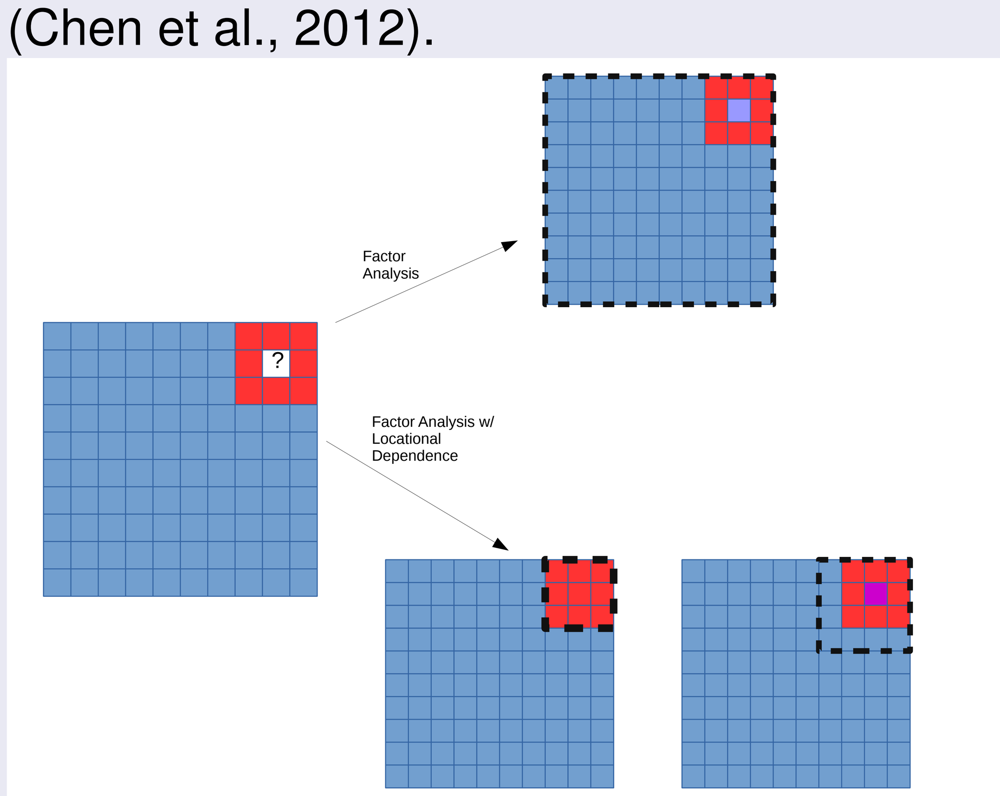
- The standard roll call scaling model:

$$y_{i,j} = \beta_j \omega_i - \alpha_j + \epsilon_{i,j}$$

- Estimation of bill level parameters treats information from **all** bills equally
- Context matters! Bill topic, committees, proposer, etc.
- Meaningful theories of Congress use ideal point scores to test concepts that are not taken into account during model estimation.
- Interpretation of uncovered dimensions is difficult.
- I seek to develop new (relatively assumption-free) estimation techniques that allow bill context to be included in estimation and explicit interpretation of dimensions in terms of **covariates**.

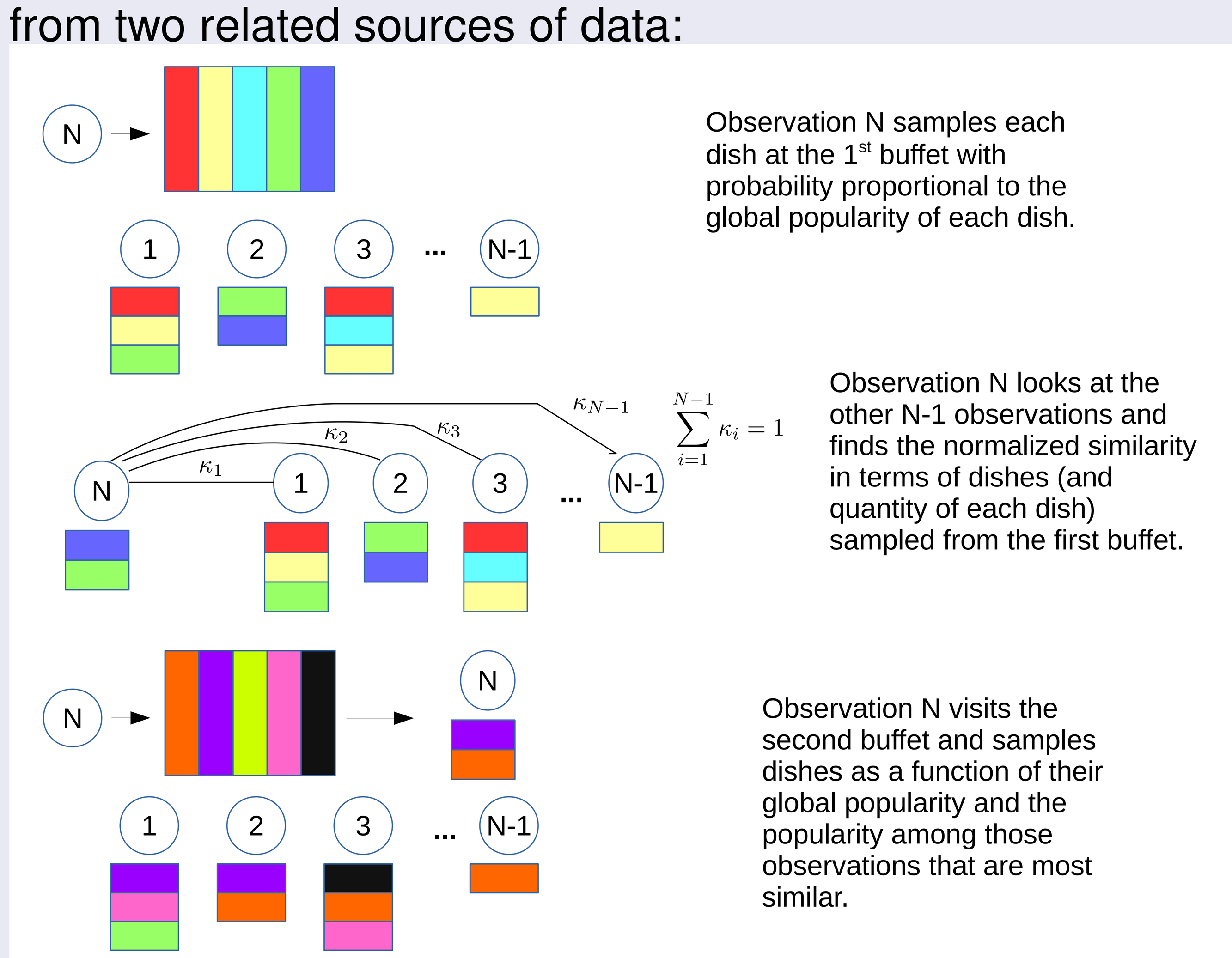
Location Dependent Unsupervised Learning for Images

- Pixel location matters for unsupervised learning with images (Chen et al., 2012).



Dependent Hierarchical Beta Processes with Unknown Locations

- Bill similarities are unknown, *a priori*.
- Chen et al. (2011 & 2012) develop hierarchical Beta process.
- Requires known locations.
- Develop new model that simultaneously learns latent features from two related sources of data:



A Topic Consistent Roll Call Scaling Model

- Let **Y** be a matrix of P roll call votes for N members of Congress. For each of these P bills, let **M** be a matrix of P associated texts that have been tokenized into D distinct terms.
- Two data equations:

$$y_{i,j} = (z_j \odot \beta_j)(w_i \odot \omega_i) - \alpha_j + \epsilon_{i,j}$$
$$m_{j,d} = b_d(r_j \odot a_j) + e_{j,d}$$

where Z, W, and R are infinite dimension binary matrices that only have a sparse set of active dimensions (K, K, and L, respectively).
- The vote model (top-level) is over-parameterized to induce sparsity.
- The models link through Z and A, which create dependence in feature selection in Z through similarities in A:

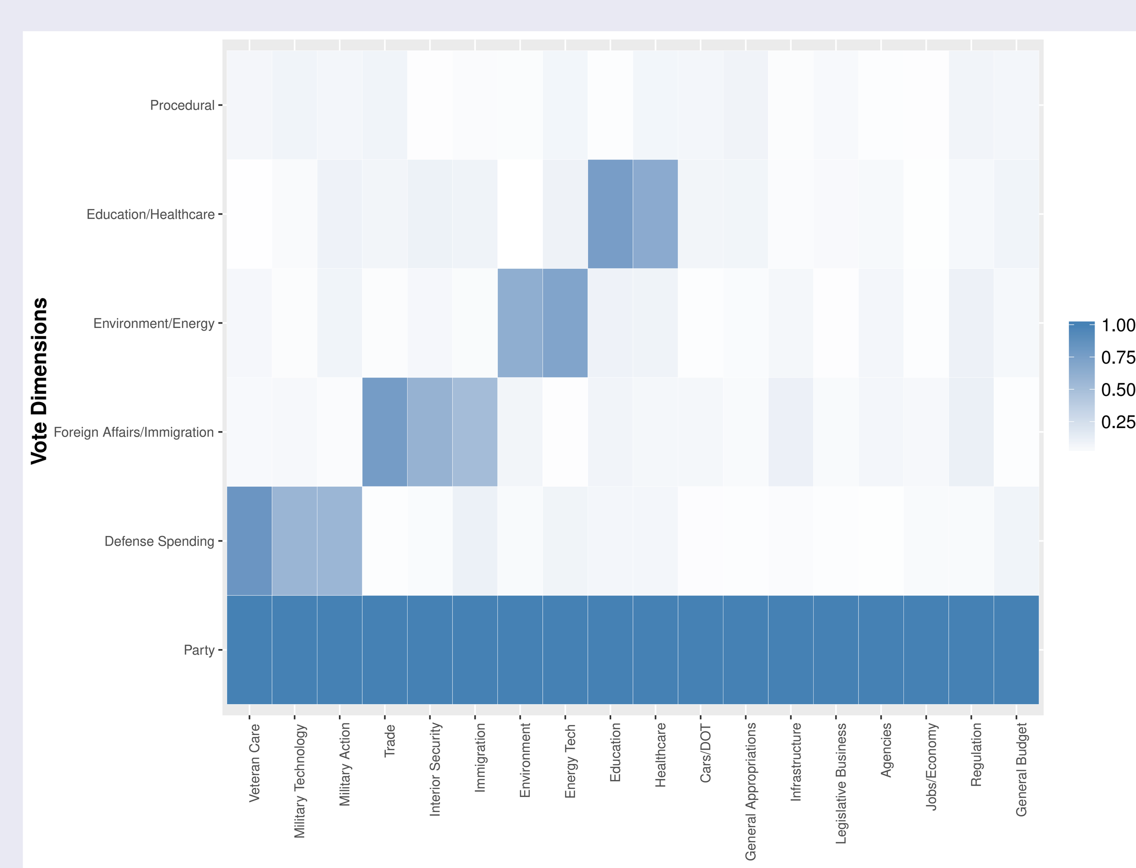
$$a_j \sim \mathcal{N}_L(a_j; 0, \mathcal{I}_L) \qquad \pi^*_{j,k} \sim \text{Beta}(\pi^*_{j,k}; \eta_k, 1-\eta_k)$$
$$\kappa_{j,h} = \frac{\|(r_j \odot a_j) - (r_j \odot a_h)\|}{\sum_{q=1}^{P} \|(r_j \odot a_j) - (r_j \odot a_q)\|} \qquad \pi_{j,k} = \kappa_k^T \pi_k^*$$
$$z_{j,k} \sim \text{Bern}(\pi_{j,k})$$

- Estimation mostly utilizes Gibbs sampling. One tough step:

$$P(a_{j,l}|-) \propto \mathcal{N}(a_{l,j}; \varrho_{j,l}, \rho_{j,l}) \prod_{k=1}^{K^+} \prod_{j=1}^{P} Bern\left(z_{j,k}; \sum_{v=1}^{P} \kappa_{j,v} \pi^*_{v,k}\right)$$

114th U.S. House

- 1117 votes and 441 Voters
- Bill text for each vote scraped from ProPublica API as close to time of vote as possible.
- Votes not associated with a specific bill use the Library of Congress question field.
- L = 18 meaningful text dimensions. Tuned LDA returned 16. Rank-tuned NMF returned 18.
- K = 6 meaningful vote dimensions. NOMINATE returns 1 (using Scree plot). IBP-FA returns 4.



- Within party disagreement on specific policy areas
- Method fails to pick up procedural questions (approval of journal, motion to adjourn, etc.) that is present in the vote dimensions.

Where is the status quo?

- Context is key in roll call scaling - an ideal point must be meaningful in the context of a vote.
- Given α_j and β_j, the alternative (ξ_j) and the status quo (ϕ_j) in the vote space can be identified up to a multiplicative constant:

$$\alpha_j = (\xi_j'\xi_j - \phi_j'\phi_j)/\sigma_j^2 \; ; \; \beta_j = 2(\xi_j - \phi_j)/\sigma_j^2$$

- σ_j^2 cannot be directly estimated in the model.
- Distance interpretation of the model:

$$P(y_{i,j} = 1) = P(\|\omega_i - \xi_j\| < \|\omega_i - \phi_j\|)$$

Rules for Estimation

- Goal: project L-dimension topic space to K-dimension vote space
- Only use votes on passage of a bill
- Votes occur temporally - ξ_j vs. $f(\phi_{j-1}, \{0, \xi_{j-1}\})$
- A policy only changes the status quo in a subset of $\mathbb{R}^K$ in accordance with z_j.
- The topic space representation of bills should be maintained in vote space - the projection of **A** to Ω should be affine.
- Given the posterior distributions for Ω and **A**, define a projection of the topic space into the vote space for bill j (the t^{th} passage vote) as:

$$\omega^*_{jt} = (\tilde{z}_j \odot \mathbb{P})(r_j \odot a_j)$$

where $\mathbb{P}$ is a $K \times L$ matrix.
- Determine the probability of the alternative location and choose $\mathbb{P}$ to maximize the probability of ω^*:

$$P(\omega^*_j) = \prod_{i=1}^{N} P(\|\omega_i - \omega^*_{jt}\| > \|\omega_i - \phi_{jt}\|)^{y_{i,j}} P(\|\omega_i - \omega^*_{jt}\| < \|\omega_i - \phi_{jt}\|)^{1-y_{i,j}}$$

$$\mathbb{P} = \underset{\mathbb{P}}{\text{argmax}} \prod_{j=1}^{P} P(\omega^*_j)$$

Tracking Alternatives and the Status Quo

- HR 5055. A vote on passage - failed on the floor.
- Related to spending for the DOE and Army Corp.
- Written to increase spending to nuclear energy and, particularly, power plants. Split Republican vote.
- Algorithm finds that proposal was too far left to beat status quo on Party dimension.