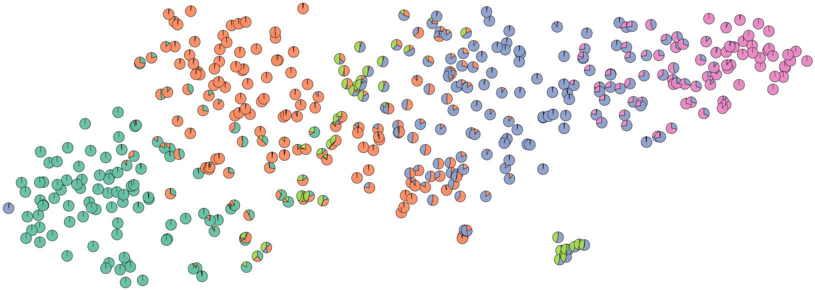
Mixture modeling political cultures (and more?)

Nathan Kellerman

Bowdoin College · Department of Mathematics



Toy example: educational ideology referendum

Assume: Bowdoin has 33 academic departments (unique membership)

Toy example: educational ideology referendum

Assume: Bowdoin has 33 academic departments (unique membership)

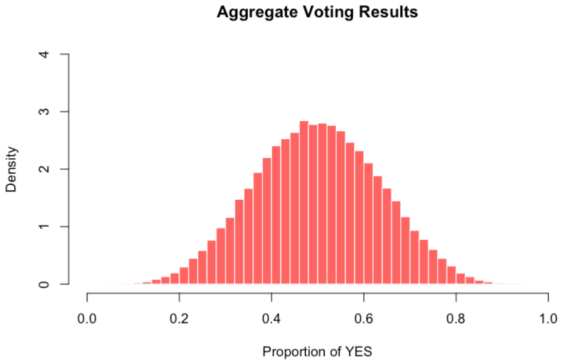
Within each department, determine binary support for the following:

Toy example: educational ideology referendum

Assume: Bowdoin has 33 academic departments (unique membership)

Within each department, determine binary support for the following:

1. I train my students for a professional workplace
2.
3.

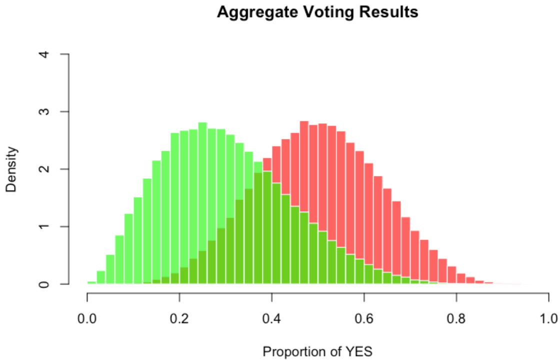


Toy example: educational ideology referendum

Assume: Bowdoin has 33 academic departments (unique membership)

Within each department, determine binary support for the following:

1. I train my students for a professional workplace
2. Learning should always be difficult
3.

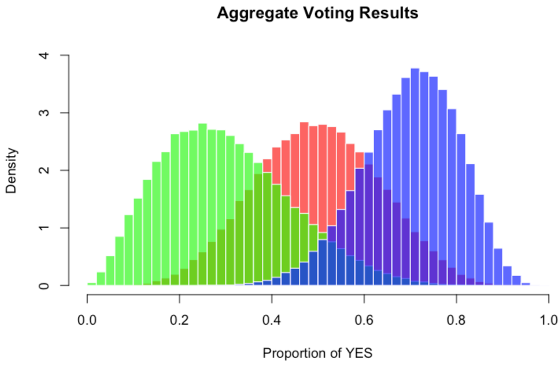


Toy example: educational ideology referendum

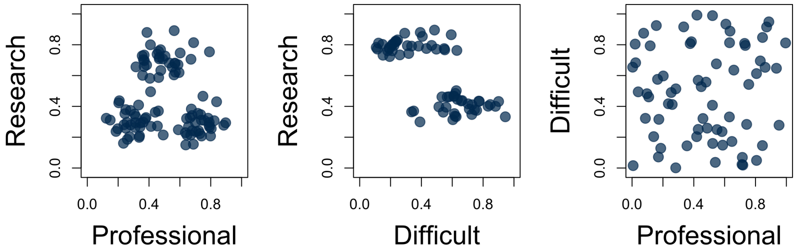
Assume: Bowdoin has 33 academic departments (unique membership)

Within each department, determine binary support for the following:

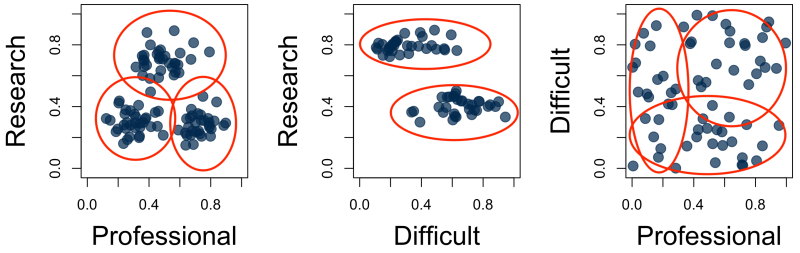
1. I train my students for a professional workplace
2. Learning should always be difficult
3. All students should pursue undergraduate research



Toy example: educational ideology referendum



Toy example: educational ideology referendum



Toy example:

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

Political data:

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

Political data:

1. How many political cultures exist in a given state?

Toy example $\mapsto$ political data

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

Political data:

1. How many political cultures exist in a given state?
2. How are municipalities mixed by culture?

Toy example $\mapsto$ political data

Toy example:

1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

Political data:

1. How many political cultures exist in a given state?
2. How are municipalities mixed by culture?
3. What parameters define these cultures?

Toy example:

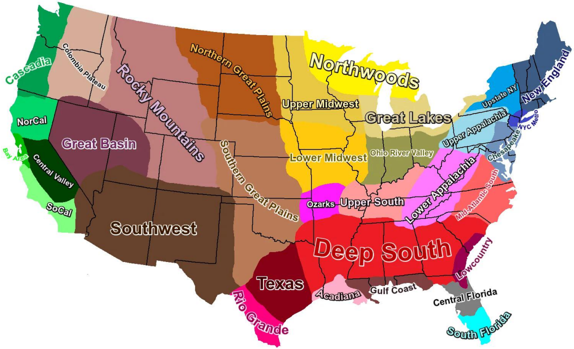
1. How many cultures of educational ideology exist among Bowdoin faculty?
2. How are departments mixed by culture?
3. What parameters define these cultures?

~~**Political data:**~~ Three Big Qs

1. How many political cultures exist in a given state?
2. How are municipalities mixed by culture?
3. What parameters define these cultures?

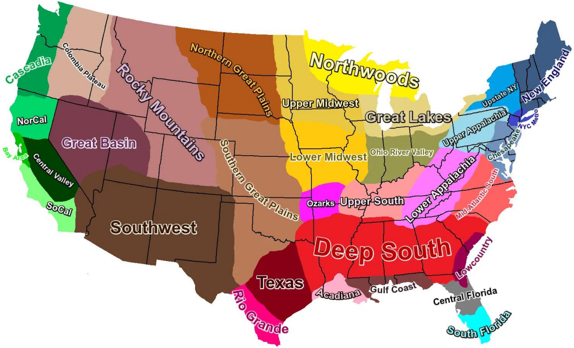
Assumptions:

1. Political cultures manifest as a consistent set of beliefs and principles
2. Voting populations are finite mixtures of political cultures



Assumptions:

1. Political cultures manifest as a consistent set of beliefs and principles
2. Voting populations are finite mixtures of political cultures



Can we quickly detect hidden populations in high-dimensional data?

Organizing our data

1. $q = 1, \dots, Q$ referendum questions
2. $i = 1, \dots, M$ municipalities
3. $k = 1, \dots, K$ political cultures
4. $N_{iq} = y_{iq} + n_{iq}$ voters per municipality, question
5. $p_{iq} = y_{iq}/N_{iq}$ observed proportion of support per municipality, question

	Q_1, Yes	Q_1, No	$\cdots$	Q_Q, Yes	Q_Q, No
Municipality 1	y_{11}	n_{12}	$\cdots$	y_{1Q}	n_{1Q}
Municipality 2	y_{21}	n_{22}	$\cdots$	y_{2Q}	n_{2Q}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Municipality M	y_{M1}	n_{M1}	$\cdots$	y_{MQ}	n_{MQ}

Organizing our data

1. $q = 1, \ldots, Q$ referendum questions
2. $i = 1, \ldots, M$ municipalities
3. $k = 1, \ldots, K$ political cultures
4. $N_{iq} = y_{iq} + n_{iq}$ voters per municipality, question
5. $p_{iq} = y_{iq}/N_{iq}$ observed proportion of support per municipality, question

	Q_1, Yes	Q_1, No	$\cdots$	Q_Q, Yes	Q_Q, No
Municipality 1	y_{11}	n_{12}	$\cdots$	y_{1Q}	n_{1Q}
Municipality 2	y_{21}	n_{22}	$\cdots$	y_{2Q}	n_{2Q}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Municipality M	y_{M1}	n_{M1}	$\cdots$	y_{MQ}	n_{MQ}

Our observed votes are aggregated within each town!

Idea: a finite mixture model

Traditional political science inference pipelines:

- generalized regression; ecological regression; latent class analysis

[1] Holmes et al. 2012
[2] Hellenthal et al. 2014
[3] Blei et al. 2003

Idea: a finite mixture model

Traditional political science inference pipelines:

- generalized regression; ecological regression; latent class analysis

Question: How do researchers in other empirical fields answer questions about aggregated mixtures?

1. How can we probabilistically model taxa frequencies in microbial metagenomics data?[1]
2. To what extent can the DNA of admixed populations explain the rise and fall of human populations?[2]
3. How can we uncover and understand an underlying set of topics from a series of text documents?[3]

[1] Holmes et al. 2012

[2] Hellenthal et al. 2014

[3] Blei et al. 2003

Idea: a finite mixture model

Traditional political science inference pipelines:

- generalized regression; ecological regression; latent class analysis

Question: How do researchers in other empirical fields answer questions about aggregated mixtures?

1. How can we probabilistically model taxa frequencies in microbial metagenomics data?[1]
2. To what extent can the DNA of admixed populations explain the rise and fall of human populations?[2]
3. How can we uncover and understand an underlying set of topics from a series of text documents?[3]

Idea: Initialize a finite mixture model to explore our Three Big Qs

[1] Holmes et al. 2012

[2] Hellenthal et al. 2014

[3] Blei et al. 2003

Modeling our data as a mixture distribution

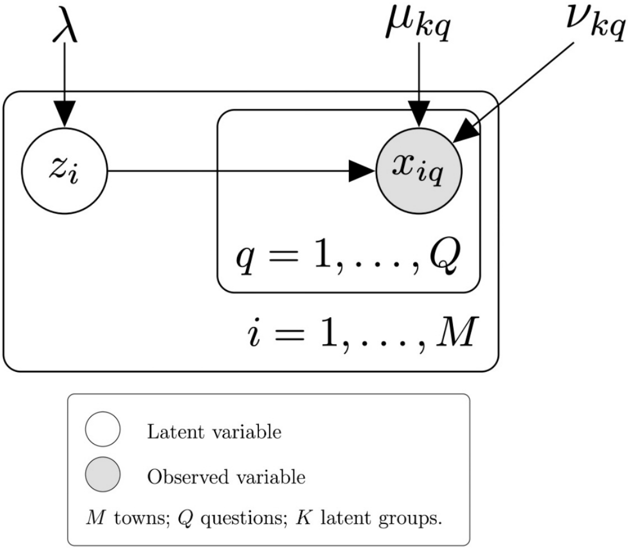
- Each observation $\mathbf{c}_{iq} = (y_{iq}, n_{iq})$ is generated by a finite mixture
- Each response $y_{iq}|z_i = k \sim \text{BetaBinomial}(N_{iq}, \mu_{kq}, \nu_{kq})$
- Thus, $f_{kq}(y_{iq}) = \text{BetaBinomial}(y_{iq}|N_{iq}, \mu_{kq}, \nu_{kq})$
- $\vec{\theta}_k = (\mu_k, \nu_k)$
- λ_k is the mixture proportion of being in cluster k

$$\pi(\vec{\theta}_k) = \prod_{q=1}^{Q} f_{kq}(y_{iq}) \tag{1}$$

$$(y_{iq}, n_{iq}) \sim \sum_{k=1}^{K} \lambda_k \pi(\vec{\theta}_k) \tag{2}$$

for mixture weights $\lambda_k \geq 0$ and $\sum_{j=1}^{K} \lambda_j = 1$

The model's substructure



Explicitly defining a Beta-binomial mixture model

Crossroads: Bayesian approach or maximum likelihood approach?

- Our goal is computational speed!

Crossroads: Bayesian approach or maximum likelihood approach?

- Our goal is computational speed!

To answer our Three Big Qs, we want to find the most probable:

- K, number of political cultures
- λ_k, global mixture weights (which sum to 1)
- μ_{kq}, mean proportion of support per political culture, question
- ν_{kq}, dispersion of support per political culture, question
- z_i, latent assignments for each town

Idea: Expectation-Maximization (EM).

Consider $\mathcal{D} = (X, \mathcal{Z}) \equiv$ (observed X, unobserved latent assignments $\mathcal{Z}$)

Idea: Expectation-Maximization (EM).

Consider $\mathcal{D} = (X, \mathcal{Z}) \equiv$ (observed X, unobserved latent assignments $\mathcal{Z}$)

Definition: the complete data log-likelihood is $\log \mathcal{L}(\vec{\theta}; X, \mathcal{Z})$.

Idea: Expectation-Maximization (EM).

Consider $\mathcal{D} = (X, \mathcal{Z}) \equiv$ (observed X, unobserved latent assignments $\mathcal{Z}$)

Definition: the complete data log-likelihood is $\log \mathcal{L}(\vec{\theta}; X, \mathcal{Z})$.

Fact: Finding $\arg\max_{\vec{\theta}} \log \mathcal{L}(\vec{\theta}; X)$ is hard.

Claim: Finding $\arg\max_{\vec{\theta}} \log \mathcal{L}(\vec{\theta}; X, \mathcal{Z})$ may be easier.

Let's "pretend" to know $\mathcal{Z}$ and let our data guide us in the right direction

Expectation-Maximization iteration scheme

E-Step: Computes expected value of $\log \mathcal{L}(\vec{\theta}; X, \mathcal{Z})$ given X and $\vec{\theta}_{old}$.

$$Q(\theta; \theta_{old}) = \mathbb{E}[\log \mathcal{L}(\vec{\theta}; X, \mathcal{Z}) | X, \theta_{old}] = \sum_i \sum_k \gamma_{ik} \log \mathbb{P}(X_i = x, Z_i = k | \theta)$$

$$\gamma_{ik} = \mathbb{P}(z_i = k | x_i, \theta_{old}) = \frac{\lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | n_{iq}, \mu_{kq}, \nu_{kq})}{\sum_{j=1}^{K} \lambda_j \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | n_{iq}, \mu_{jq}, \nu_{jq})}$$

M-Step: Maximizes[4] the expectation we just computed over $\vec{\theta}$.

Procedure:

1. $\theta_{new} := \max_{\theta} Q(\theta; \theta_{old})$
2. Set $\theta_{old} = \theta_{new}$

BB derivations:

- $\widehat{\lambda_k} = \frac{1}{M} \sum_{i=1}^{M} \gamma_{ik}$
- $\widehat{\mu_{kq}} \equiv$ approximated by Newton-Raphson method

[4] Initialize Newton steps at Binomial MLE, almost guarantees convergence

Performing E-M iterations

With these estimates for updated paramaters $\hat{\lambda}_k$ and $\hat{\mu}_{kq}$, we perform:

1. Initialize λ_k and μ_{kq}, find log-likelihood with these parameters.
2. Find γ_{ik} with current parameters.
3. Estimate $\hat{\lambda}_k$ and $\hat{\mu}_{kq}$ with found γ_{ik}.
4. Evaluate new log-likelihood with these new parameters.
5. Repeat (1) through (4) until the log-likelihood converges.

We are left with our most probable parameters!

- 70 referendum questions from 2008-2024, i.e.

Back to Maine

- 70 referendum questions from 2008-2024, i.e.
  - 2012: Would legalize same-sex marriage

- 70 referendum questions from 2008-2024, i.e.
  - 2012: Would legalize same-sex marriage
  - 2017: Casino or slot machines allowed in York county

Back to Maine

- 70 referendum questions from 2008-2024, i.e.
  - 2012: Would legalize same-sex marriage
  - 2017: Casino or slot machines allowed in York county
  - 2023: Prohibit election spending by foreign governments

- 70 referendum questions from 2008-2024, i.e.
  - 2012: Would legalize same-sex marriage
  - 2017: Casino or slot machines allowed in York county
  - 2023: Prohibit election spending by foreign governments
  - 2024: Replace Maine state flag

- 70 referendum questions from 2008-2024, i.e.
  - 2012: Would legalize same-sex marriage
  - 2017: Casino or slot machines allowed in York county
  - 2023: Prohibit election spending by foreign governments
  - 2024: Replace Maine state flag
- 423 municipalities (Acton, ..., York)
  - $N_M \in [3, 37829], \mathrm{mean}(N_M) \approx 1220$

Back to Maine

Three Big Qs, revisited Maine-style:

1. How many different cultures of Mainers are there?
2. How does each culture of Mainer vote in referendum?
3. What cultures of Mainers compose each municipality?

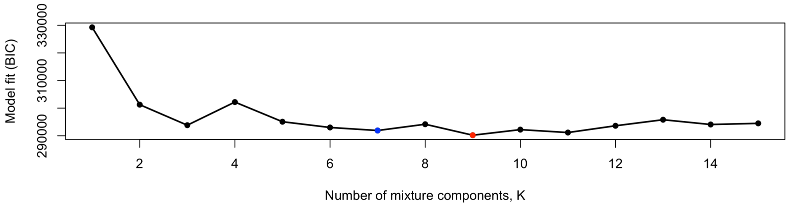
Big question 1: BIC analysis

Bayesian information criterion:

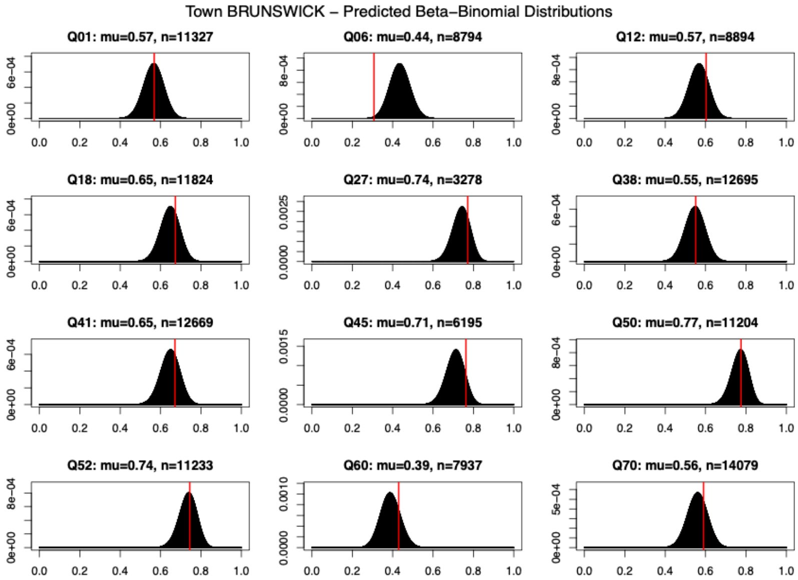
$$BIC = \underbrace{\mathcal{K} \ln(M)}_{\text{complexity}} - 2 \underbrace{\ln \mathcal{L}}_{\text{fit}},$$

where

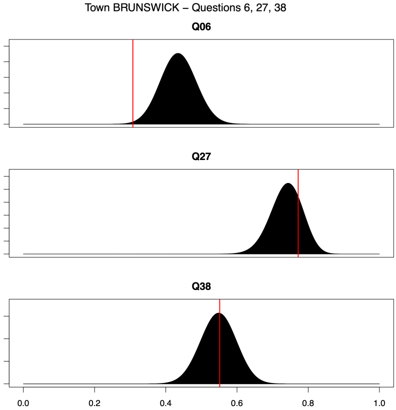
$$\mathcal{K} = \underbrace{(K-1)}_{\text{mixture weights}} + \underbrace{(K \cdot Q)}_{\mu \text{ parameters}} + \underbrace{1}_{\text{fixed } \nu \text{ parameter}}$$



Big question 2: per-question support

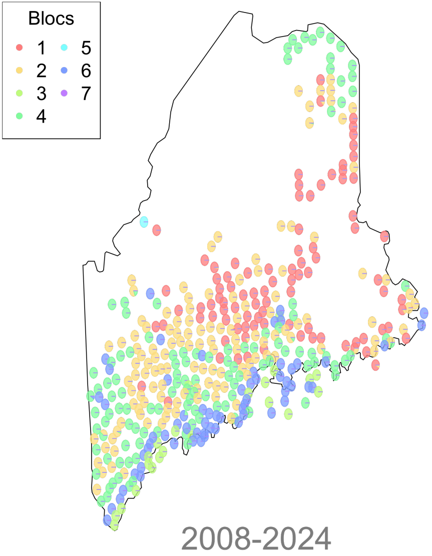


Big question 2: per-question support

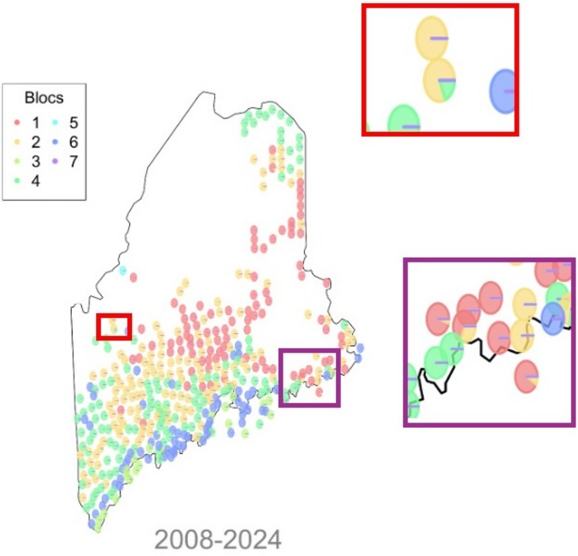


1. Repeal law mandating school district restructuring
2. $15,500,000 bond to upgrade buildings in Maine CC system
3. Legalize marijuana for personal use

Big question 3: by-municipality mixtures



Big question 3: by-municipality mixtures



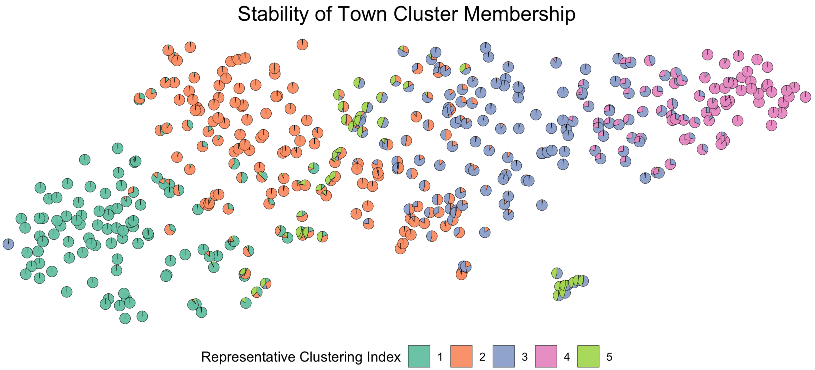
Multiple EM runs

Idea: Noise our count data, run EM many times on our "new" data

$$\underbrace{\widetilde{\mu_{iq}}}_{\text{jittered BB mean}} \sim \text{Normal}\Big(\underbrace{p_{iq}}_{\text{observed}}, \frac{\sqrt{\epsilon}}{N_{iq}} \Big)$$

Average town-level mixtures across multiple EM runs

Representative clustering



- Frame-by-frame t-SNE embedding animation

- Research-grade simulation study (use of Bowdoin HPC)
- Application to larger political science datasets
  - Voting data in Switzerland, Los Angeles (unique structures of ideology)
- Generalizing method for social science count data
  - Writing `R` package with general algorithm
- Longitudinal analysis of shifts in ideology across time and space
- Methods for aggregating multiple EM runs, to achieve more intense municipality-level mixtures

Acknowledgements

A special thanks to Prof. Jack O'Brien and the Mathematics Department

Questions, please!

NB: Writing our model's likelihood

The log-likelihood $\log \mathcal{L}(\vec{\theta}; X)$ of our M observations is

$$\sum_{i=1}^{M} \log \left\{ \sum_{k=1}^{K} \lambda_k \left(\prod_{q=1}^{Q} \mathrm{BetaBinomial}(X_{iq} | N_{iq}, \mu_{kq}, \nu_{kq}) \right) \right\}$$

If we pretend to know the latent assignments, our complete data log-likelihood $\log \mathcal{L}(\vec{\theta}; X, \mathcal{Z})$ is

$$\sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \left[\log \lambda_k + \sum_{q=1}^{Q} \log \mathrm{BetaBinomial}\big(X_{iq} \mid N_{iq}, \mu_{kq}, \nu_{kq}\big) \right]$$