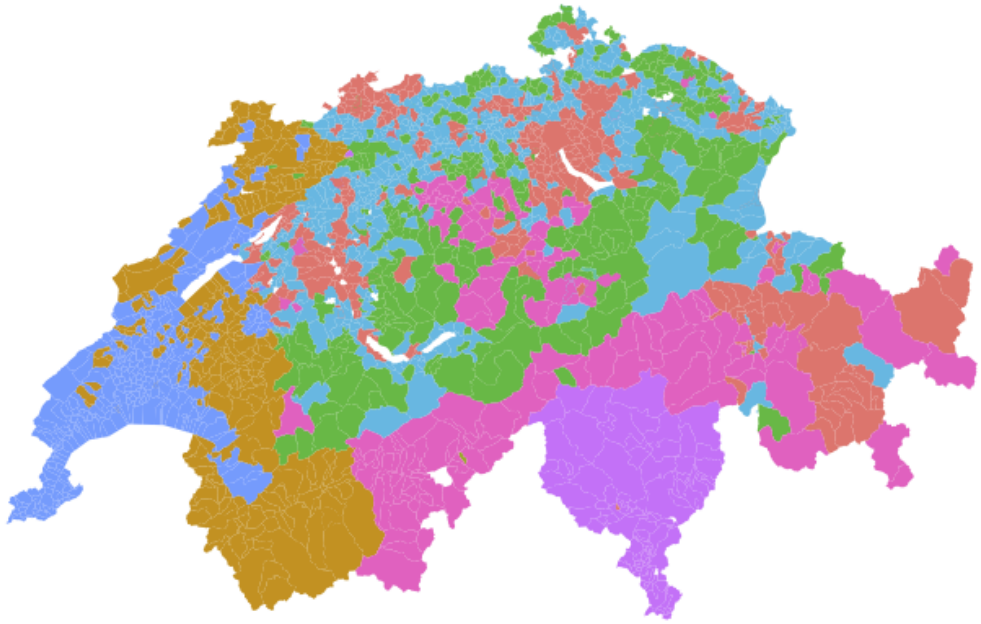
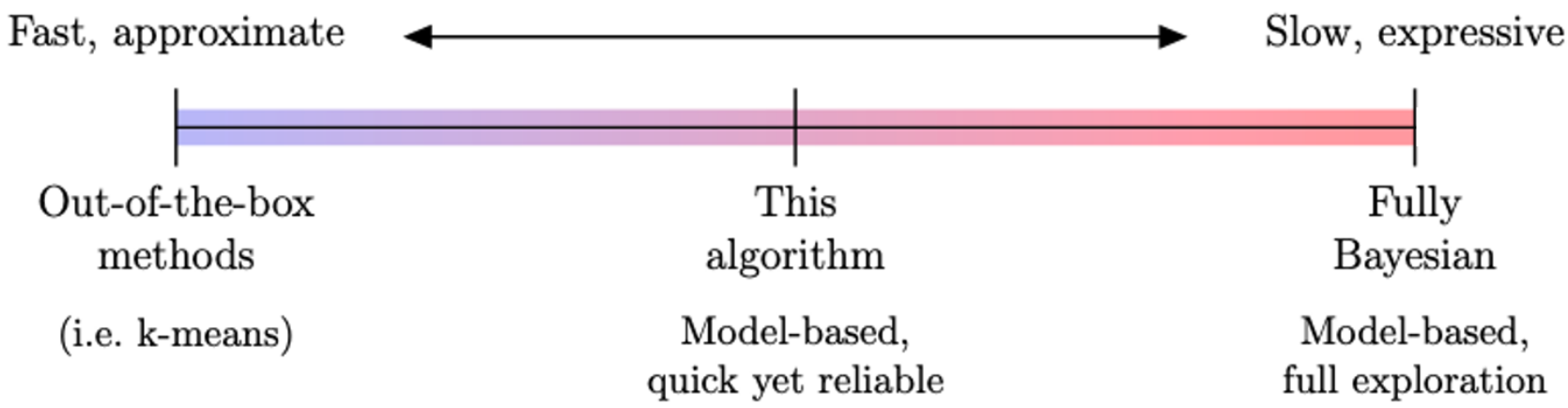
A rapid algorithm for inferring latent mixture structure in replicate social science data

Nathan Kellerman

Bowdoin College · Department of Mathematics




Fast, approximate
Slow, expressive
Out-of-the-box methods
(i.e. k-means)
This algorithm
Model-based, quick yet reliable
Fully Bayesian
Model-based, full exploration

Outline

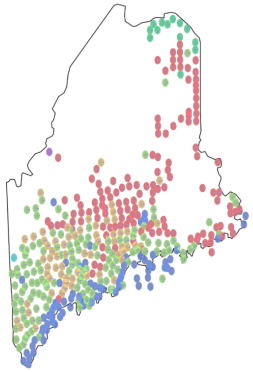
Theory

- Professor pedagogy toy example
  - How can we uncover hidden voting populations in data?
- Political science framing
  - Can we mathematically model voting ideology in referendum?

$$\log \mathcal{L}(\vec{\theta}; y)$$

Application

- Simulation studies
  - How rapid and accurate is the algorithm?
- Case studies
  - Maine referendum voting
  - Switzerland referendum voting (previously inaccessible)



Toy example: professor pedagogy referendum

Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Toy example: professor pedagogy referendum

Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Data: For each faculty, record binary support for the following:

Toy example: professor pedagogy referendum

Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Data: For each faculty, record binary support for the following:

1. I train my students for a professional workplace

Toy example: professor pedagogy referendum

Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Data: For each faculty, record binary support for the following:

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

Toy example: professor pedagogy referendum

Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Data: For each faculty, record 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: professor pedagogy referendum

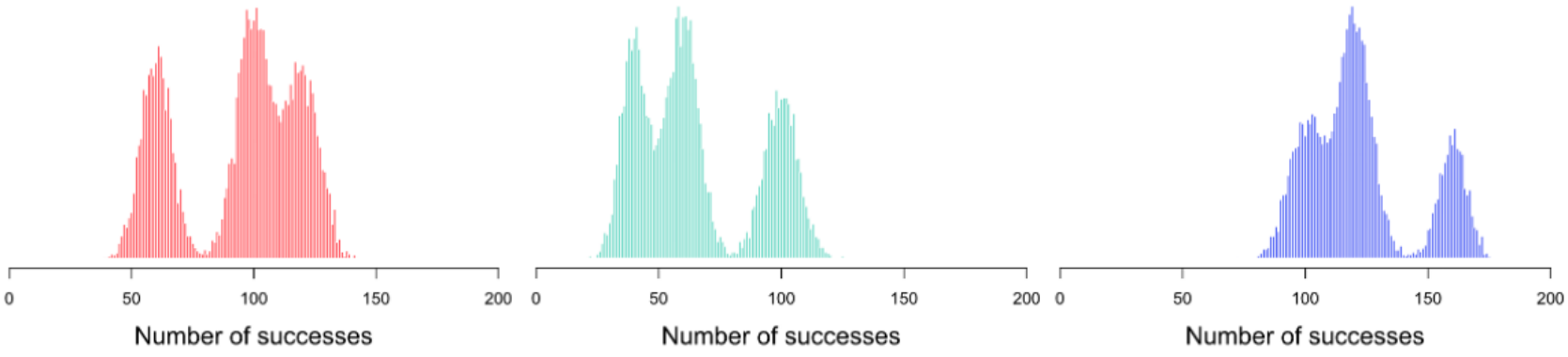
Assume: Bowdoin College has 200 faculty, each with unique membership to one of 33 academic departments

Data: For each faculty, record 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

	Q_1, Yes	Q_1, No	Q_2, Yes	Q_2, No	Q_3, Yes	Q_3, No
Department 1	$y_{1,1}$	$n_{1,1}$	$y_{1,2}$	$n_{1,2}$	$y_{1,3}$	$n_{1,3}$
Department 2	$y_{2,1}$	$n_{2,1}$	$y_{2,2}$	$n_{2,2}$	$y_{2,3}$	$n_{2,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Department 33	$y_{33,1}$	$n_{33,1}$	$y_{33,2}$	$n_{33,2}$	$y_{33,3}$	$n_{33,3}$

The shape of professor pedagogy



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

What cultural archetypes of professors exist at Bowdoin?

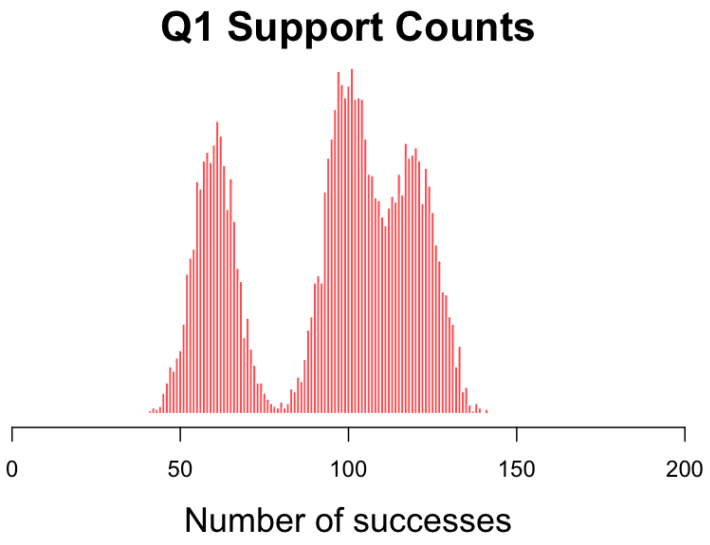
Three big questions

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

Some notation

- $i = 1, \ldots, M = 33$ departments
- $q = 1, \ldots, Q = 3$ referendum questions
- y_{iq} yes responses, per department, per question
- N_{iq} polled faculty per department, per question
- $k = 1, \ldots, K$ latent cultures
- $z_i \in \{1, \ldots, K\}$ latent assignment z per department

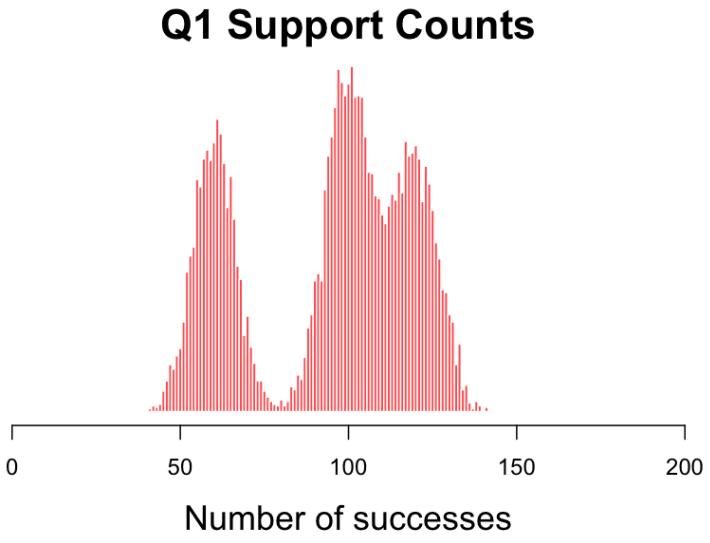
A simple mixture distribution



$$Y \sim \lambda_1 \cdot \pi_1 + \lambda_2 \cdot \pi_2 + \lambda_3 \cdot \pi_3$$

- $\lambda_1 + \lambda_2 + \lambda_3 = 1$
- each π_k arises from some well-behaved distributional family $\mathcal{T}(\theta)$

A simple mixture distribution



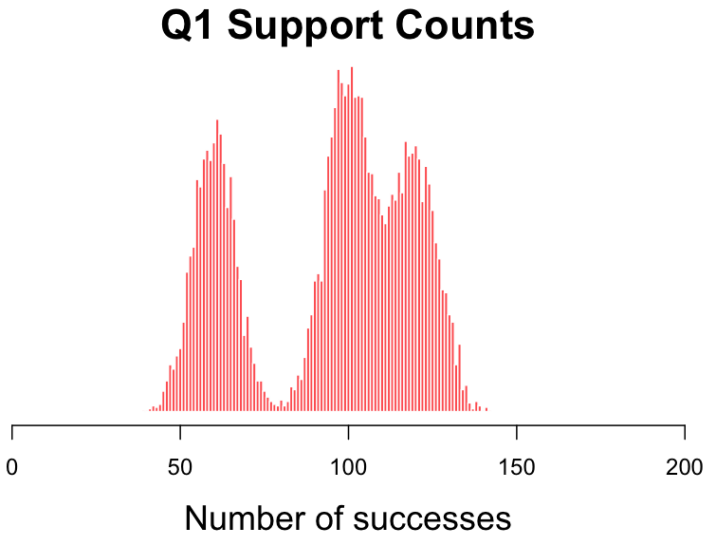
$$Y \sim \lambda_1 \cdot \pi_1 + \lambda_2 \cdot \pi_2 + \lambda_3 \cdot \pi_3$$

- $\lambda_1 + \lambda_2 + \lambda_3 = 1$
- each π_k arises from some well-behaved distributional family $\mathcal{T}(\theta)$

Each π_k is actually:

$$f_k(y_i) = \text{BetaBinomial}(y_i | N_i, \mu_k, \nu_k)$$

A simple mixture distribution



$$Y \sim \lambda_1 \cdot \pi_1 + \lambda_2 \cdot \pi_2 + \lambda_3 \cdot \pi_3$$

- $\lambda_1 + \lambda_2 + \lambda_3 = 1$
- each π_k arises from some well-behaved distributional family $\mathcal{T}(\theta)$

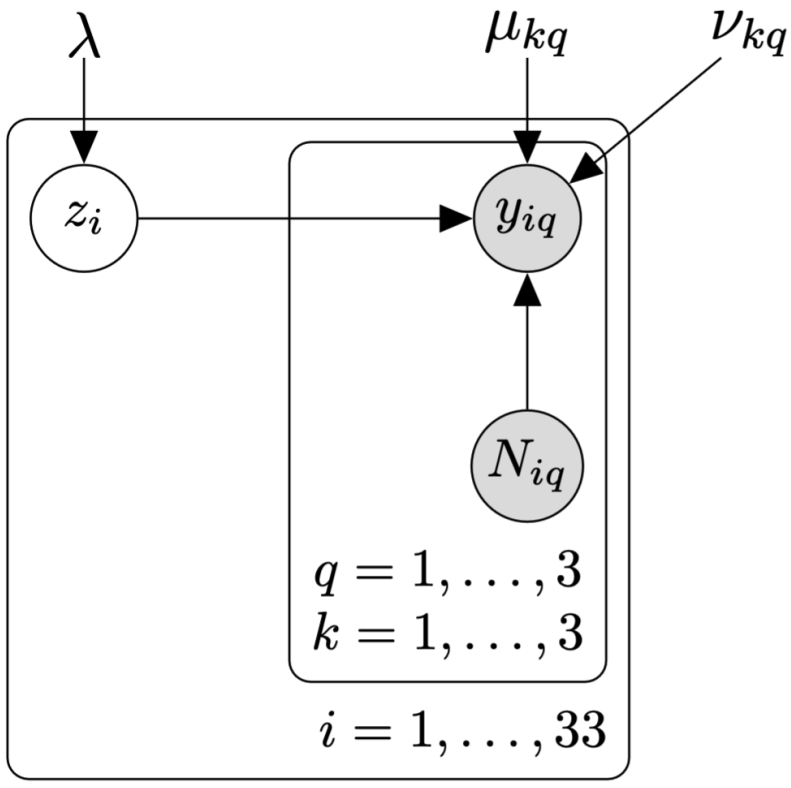
Each π_k is actually:

$$f_k(y_i) = \text{BetaBinomial}(y_i | N_i, \mu_k, \nu_k)$$

Across all $Q = 3$ questions:

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

Data generating process



Data likelihood $\mathcal{L}(\vec{\theta}, y)$ across K cultures:

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

Bayesian vs Frequentist?

Bayesian approach

- Examine $\mathbb{P}(\theta|\mathcal{D})$
- Find distributions of parameters θ by conditioning upon the data
- Explore all parameters

Frequentist approach

- Examine $\mathbb{P}(\mathcal{D}|\theta)$
- Construct data likelihood $\mathcal{L}(\vec{\theta}, y)$
- Converge on single set of parameters θ

Bayesian vs Frequentist?

Bayesian approach

- Examine $\mathbb{P}(\theta|\mathcal{D})$
- Find distributions of parameters θ by conditioning upon the data
- Explore all parameters

Frequentist approach

- Examine $\mathbb{P}(\mathcal{D}|\theta)$
- Construct data likelihood $\mathcal{L}(\vec{\theta}, y)$
- Converge on single set of parameters θ

Case Study	# of Locales	# of Questions	Polling Year Range
Pedagogy Toy Example	33	3	NA
Maine Referendum Data	423	70	$2008-2024$
Switzerland Referendum Data	2048	390	$1981-2026$

Rapid algorithm $\equiv$ maximum likelihood estimation procedure

Approach: Try to take derivatives

$$\frac{\partial}{\partial \mu_{kq}} \left[\sum_{i=1}^{M} \log \left\{ \sum_{k=1}^{K} \lambda_k \left(\prod_{q=1}^{Q} \text{BetaBinomial}(\mathbf{y}_{iq} | N_{iq}, \mu_{kq}, \nu_{kq}) \right) \right\} \right]$$

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

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

Approach: Try to take derivatives

$$\frac{\partial}{\partial \mu_{kq}} \left[\sum_{i=1}^{M} \log \left\{ \sum_{k=1}^{K} \lambda_k \left(\prod_{q=1}^{Q} \text{BetaBinomial}(\mathbf{y}_{iq} | N_{iq}, \mu_{kq}, \nu_{kq}) \right) \right\} \right]$$

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

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

---

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

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

Approach: Try to take derivatives

$$\frac{\partial}{\partial \mu_{kq}} \left[\sum_{i=1}^{M} \log \left\{ \sum_{k=1}^{K} \lambda_k \left(\prod_{q=1}^{Q} \text{BetaBinomial}(\mathbf{y}_{iq} | N_{iq}, \mu_{kq}, \nu_{kq}) \right) \right\} \right]$$

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

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

---

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

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

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

Expectation Maximization iteration scheme

Useful definition: The probability that department i is in cluster k is

$$\gamma_{ik} = \mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1}) = \frac{\lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{kq}, \nu_{kq})}{\sum_{k=1}^{K} \lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{jq}, \nu_{jq})}$$

Expectation Maximization iteration scheme

Useful definition: The probability that department i is in cluster k is

$$\gamma_{ik} = \mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1}) = \frac{\lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{kq}, \nu_{kq})}{\sum_{k=1}^{K} \lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{jq}, \nu_{jq})}$$

Amazing fact: $\mathcal{L}(\theta_t; X) \geq \mathcal{Q}(\theta_t, \theta_{t+1})$, for

$$\mathcal{Q}(\theta_t, \theta_{t+1}) = \sum_i \sum_k \mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1}) \log \frac{\mathbb{P}(X_{iq} = x_{iq}, Z_i = k | \theta_t)}{\mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1})}$$

Expectation Maximization iteration scheme

Useful definition: The probability that department i is in cluster k is

$$\gamma_{ik} = \mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1}) = \frac{\lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{kq}, \nu_{kq})}{\sum_{k=1}^{K} \lambda_k \prod_{q=1}^{Q} \mathrm{BB}(x_{iq} | N_{iq}, \mu_{jq}, \nu_{jq})}$$

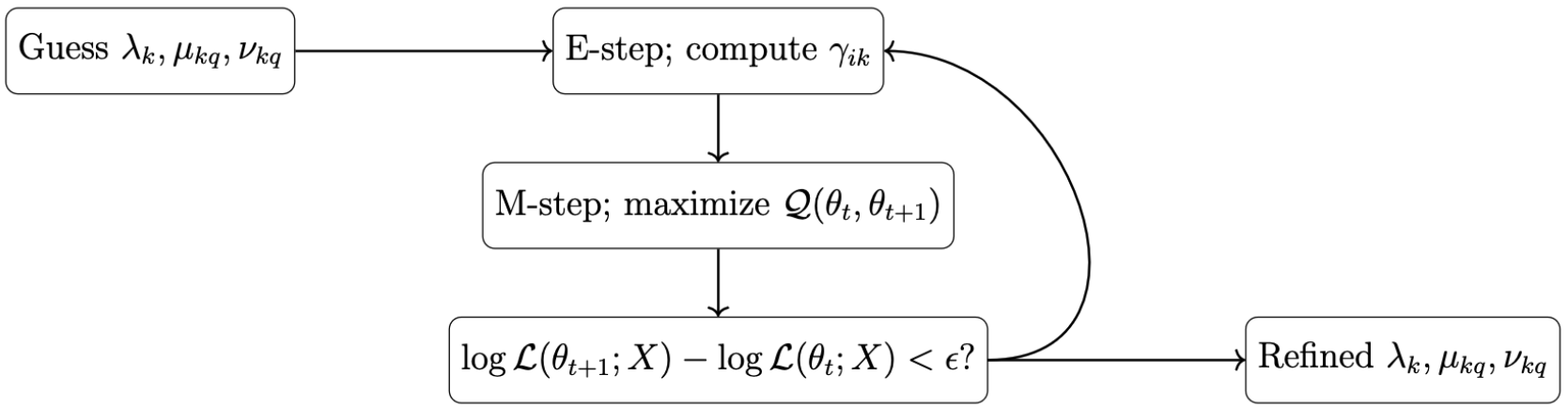
Amazing fact: $\mathcal{L}(\theta_t; X) \geq \mathcal{Q}(\theta_t, \theta_{t+1})$, for

$$\mathcal{Q}(\theta_t, \theta_{t+1}) = \sum_i \sum_k \mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1}) \log \frac{\mathbb{P}(X_{iq} = x_{iq}, Z_i = k | \theta_t)}{\mathbb{P}(Z_i = k | x_{iq}, \theta_{t+1})}$$

One EM iteration:

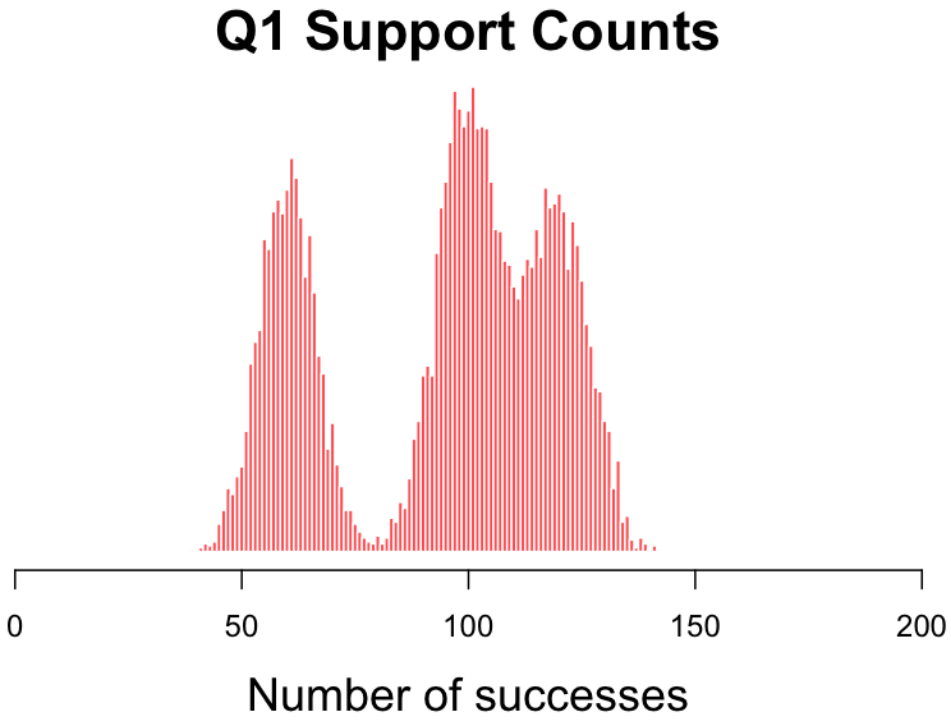
1. E-Step: Compute responsibilities γ_{ik} for each department
2. M-Step: Find the updated parameters which maximize $\mathcal{Q}(\theta_t, \theta_{t+1})$ and thus improve the likelihood $\mathcal{L}(\theta; X)$.

Flow diagram

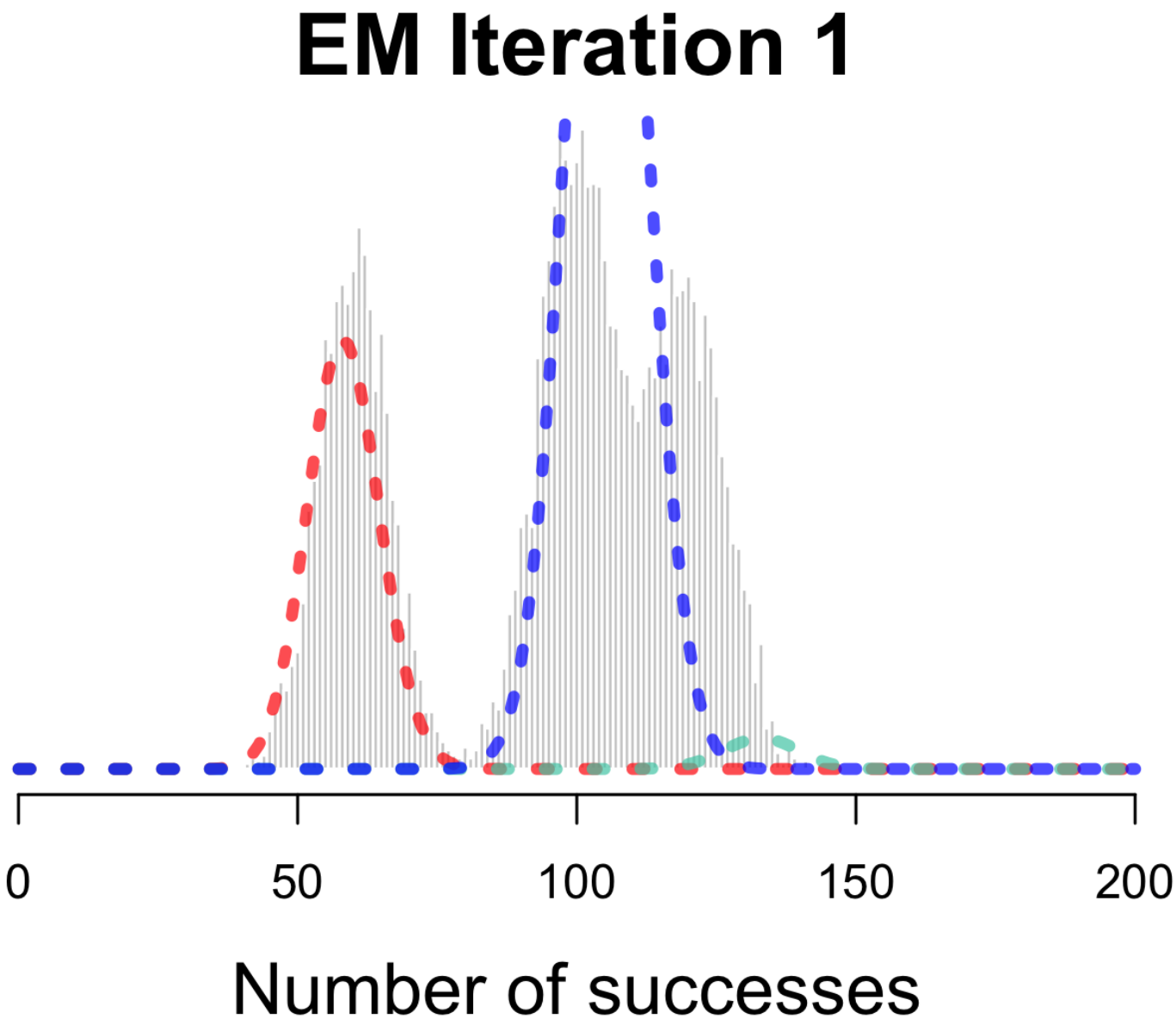


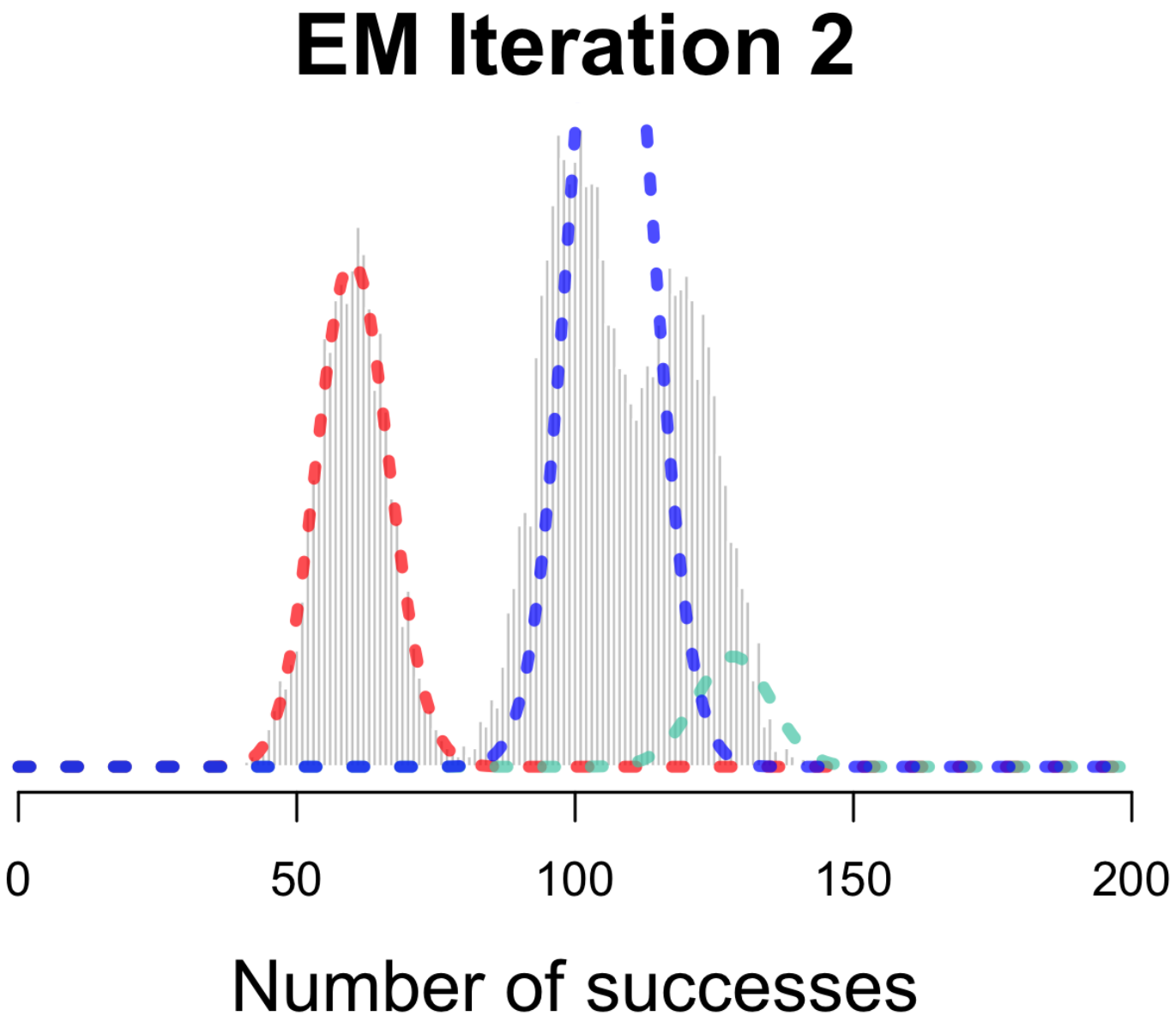
EM over a single question

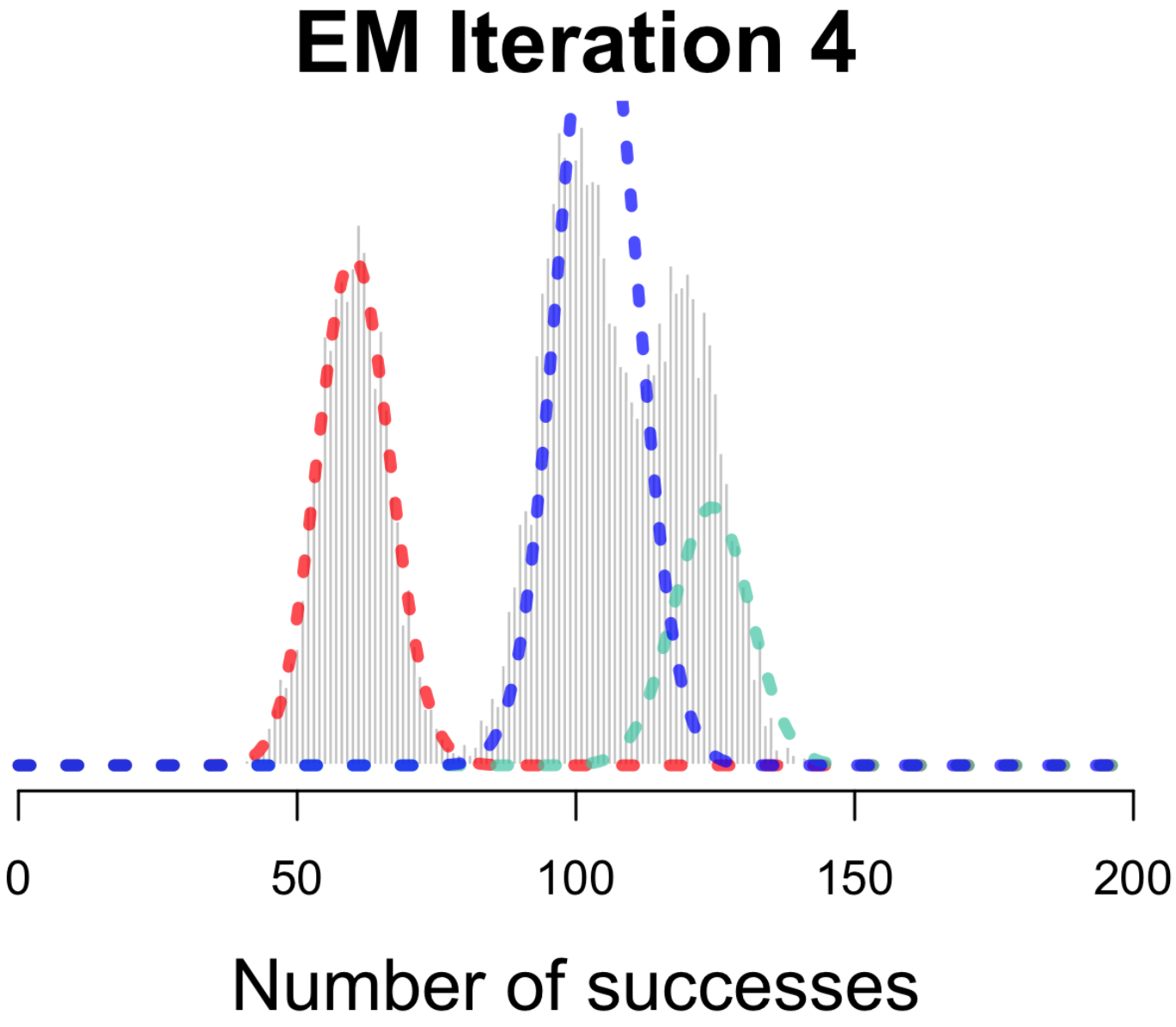
Question 1: I train my students for a professional workplace

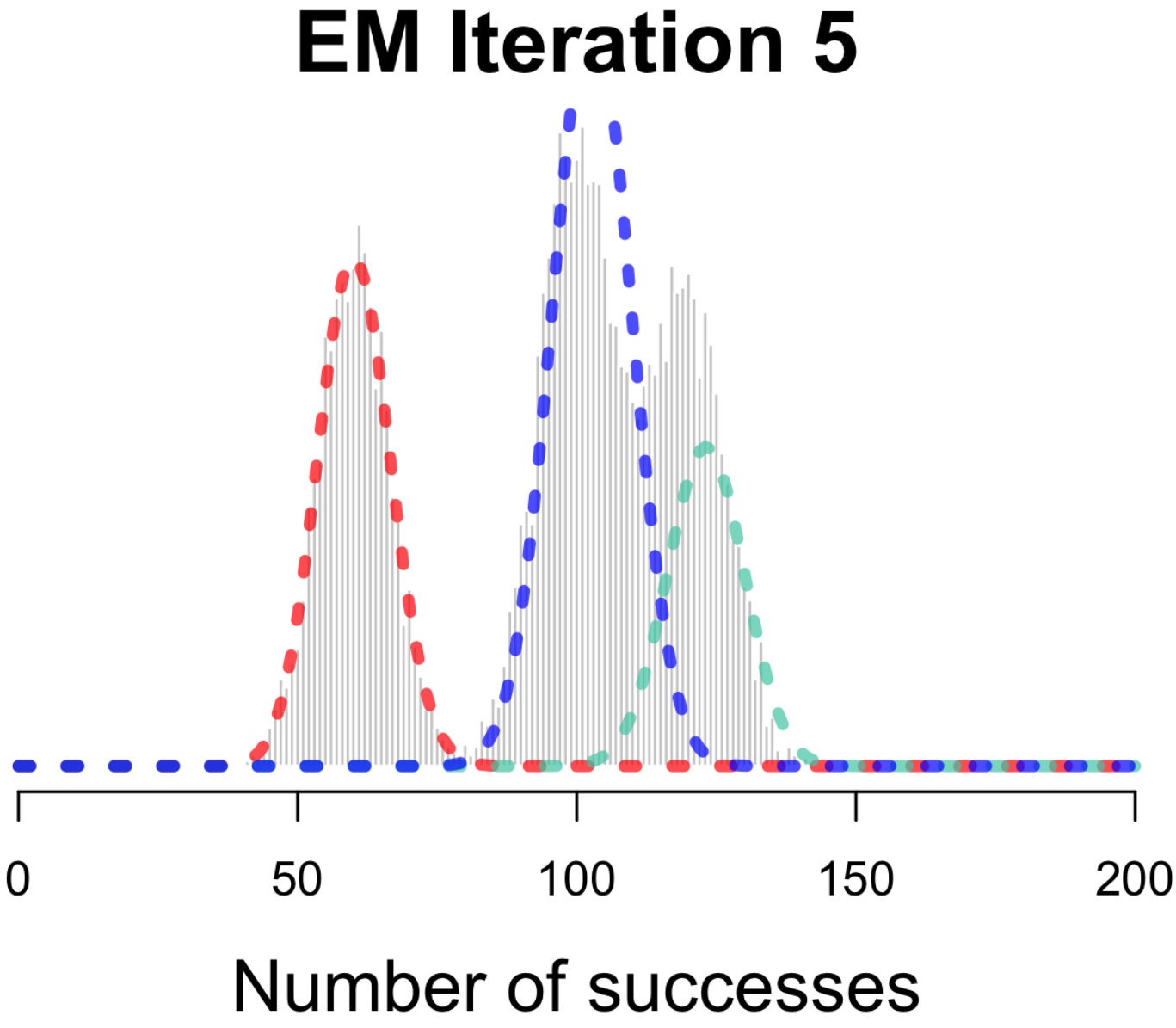


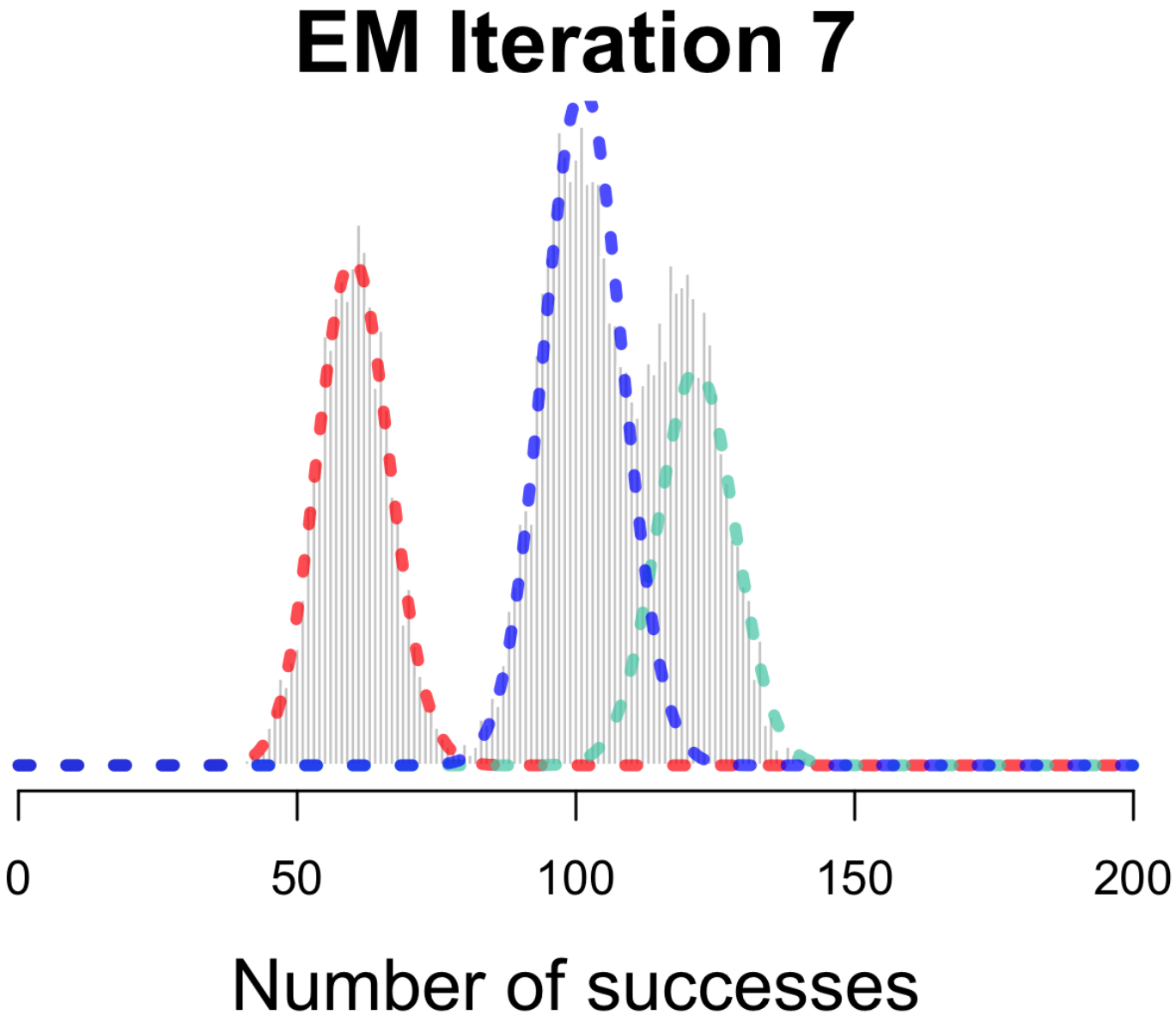
Choose $K = 3$. Guess parameters λ_k, μ_k, and ν_k.

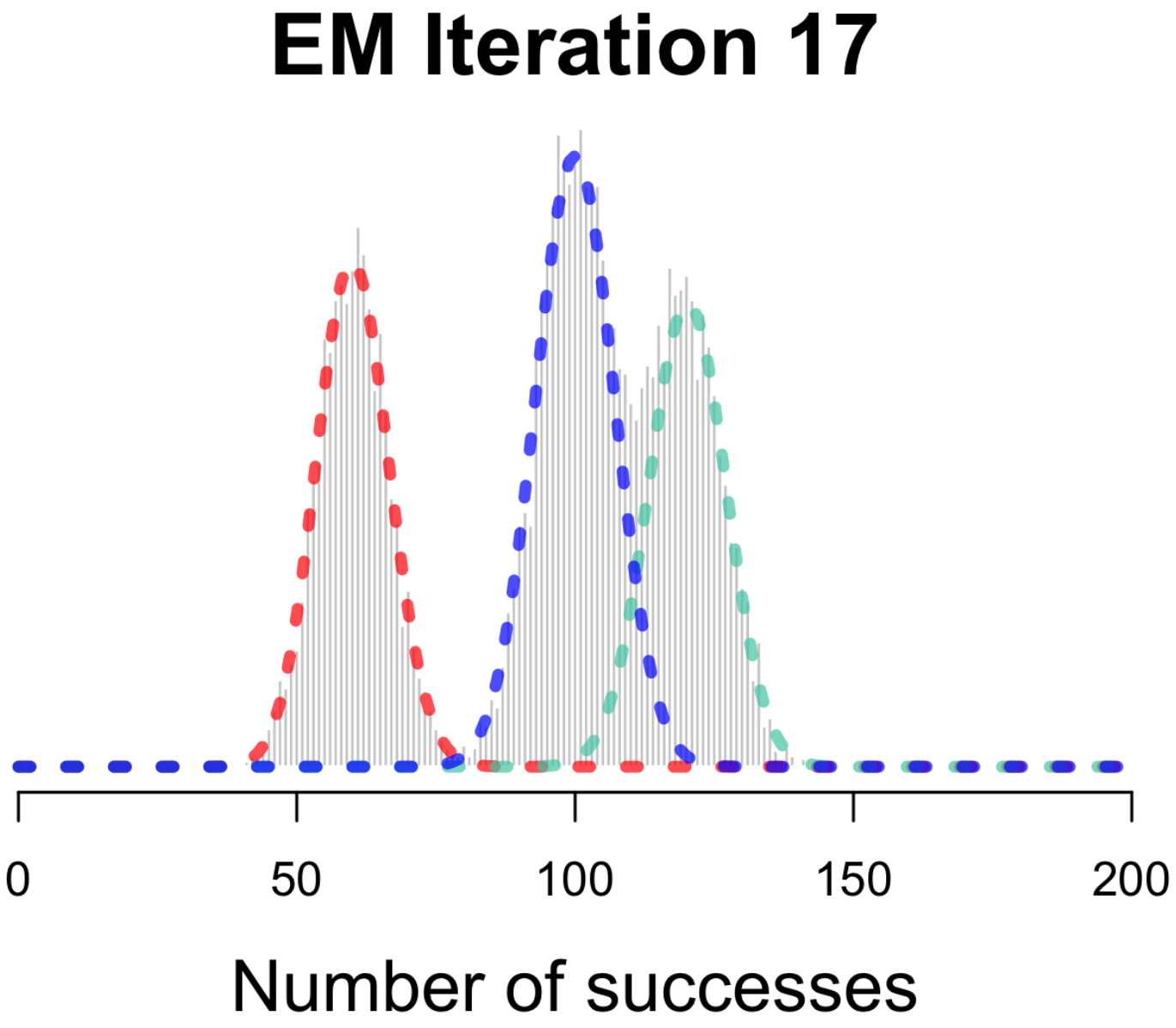

EM Iteration 1
0
50
100
150
200
Number of successes

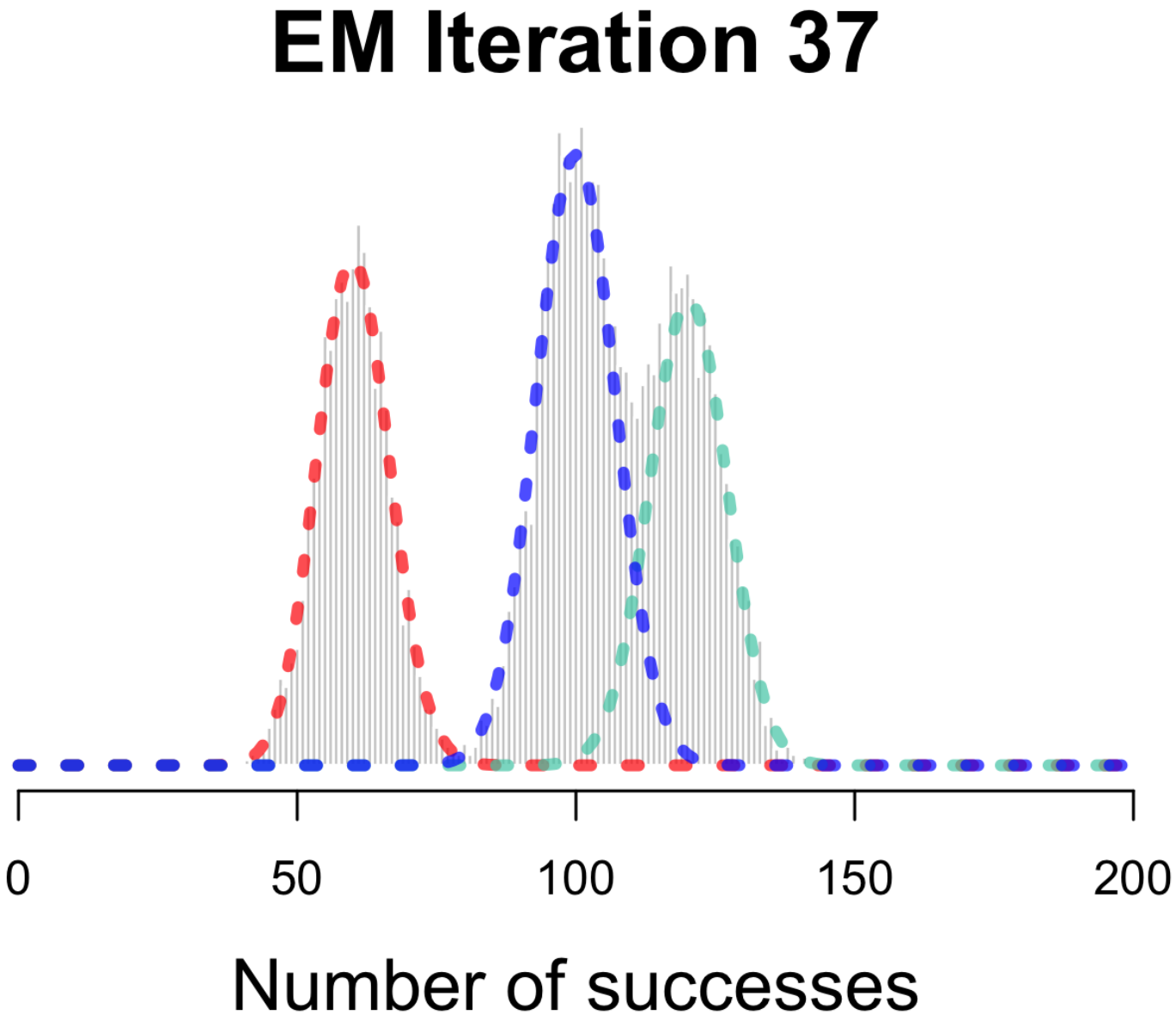

EM Iteration 2
0
50
100
150
200
Number of successes


EM Iteration 4
0
50
100
150
200
Number of successes


EM Iteration 5
0
50
100
150
200
Number of successes


EM Iteration 7
0
50
100
150
200
Number of successes

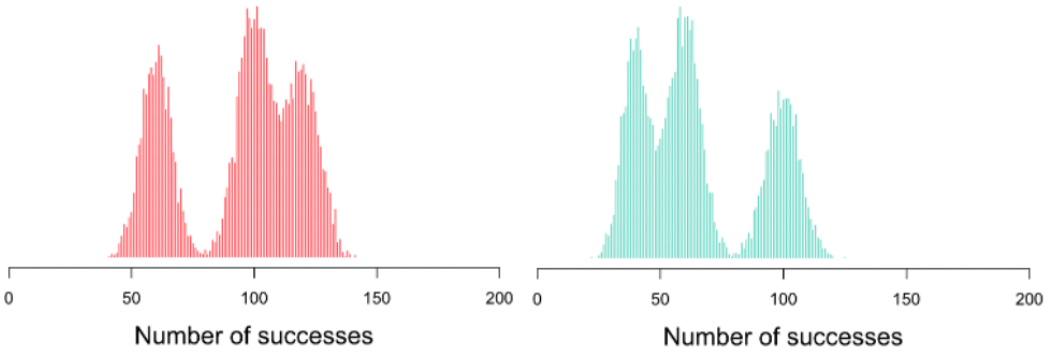

EM Iteration 17
0
50
100
150
200
Number of successes

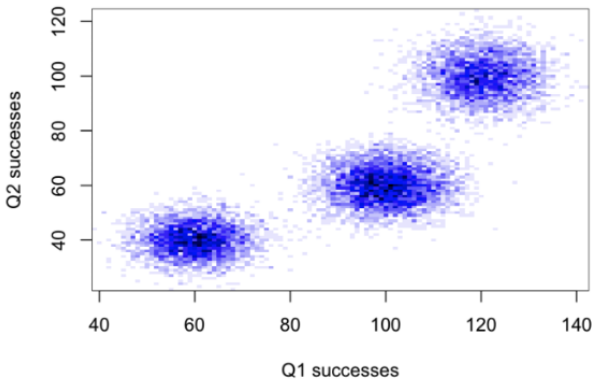

EM Iteration 37
0
50
100
150
200
Number of successes

EM over two questions (novel approach)

Question 1: I train my students for a professional workplace

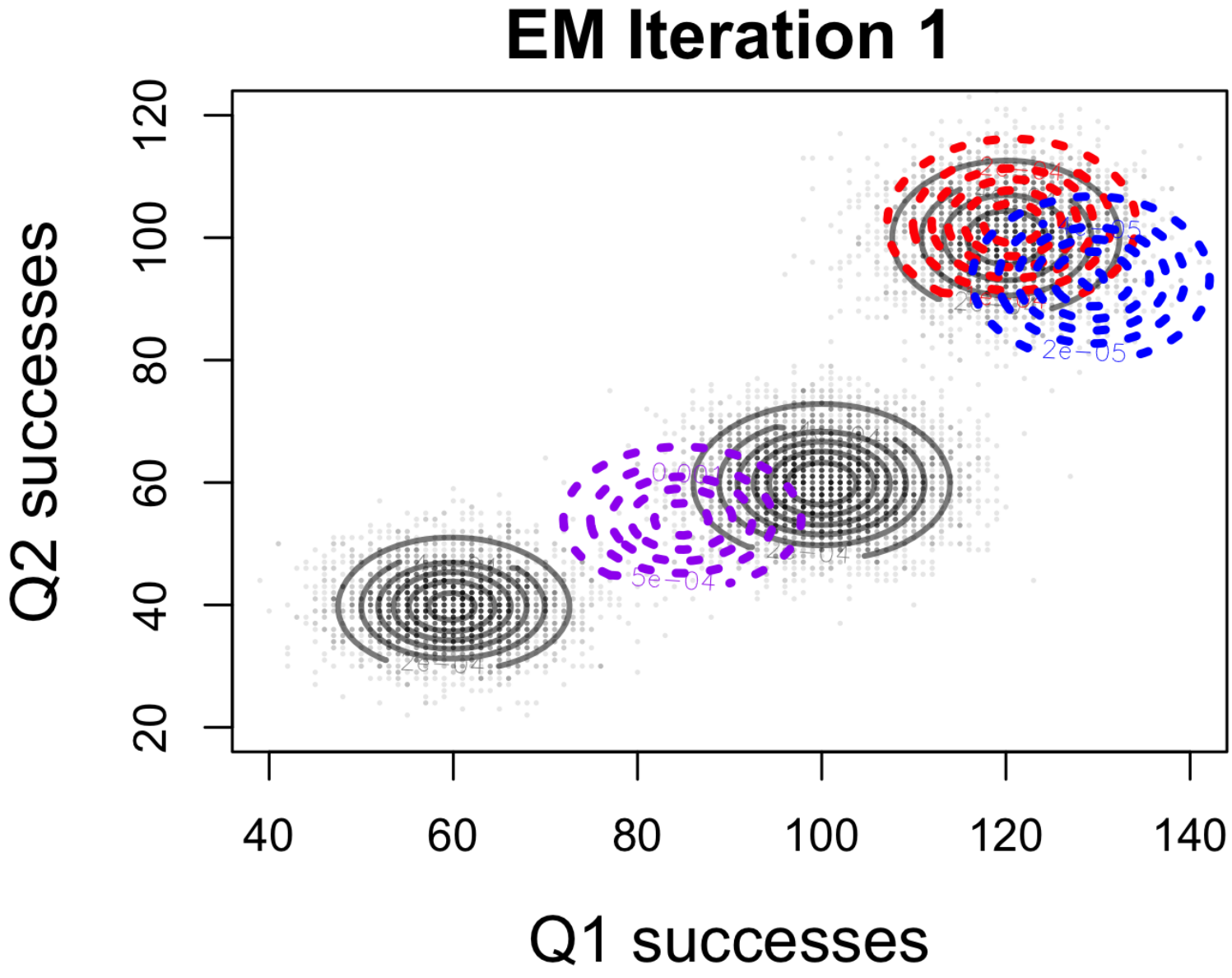
Question 2: Learning should always be difficult



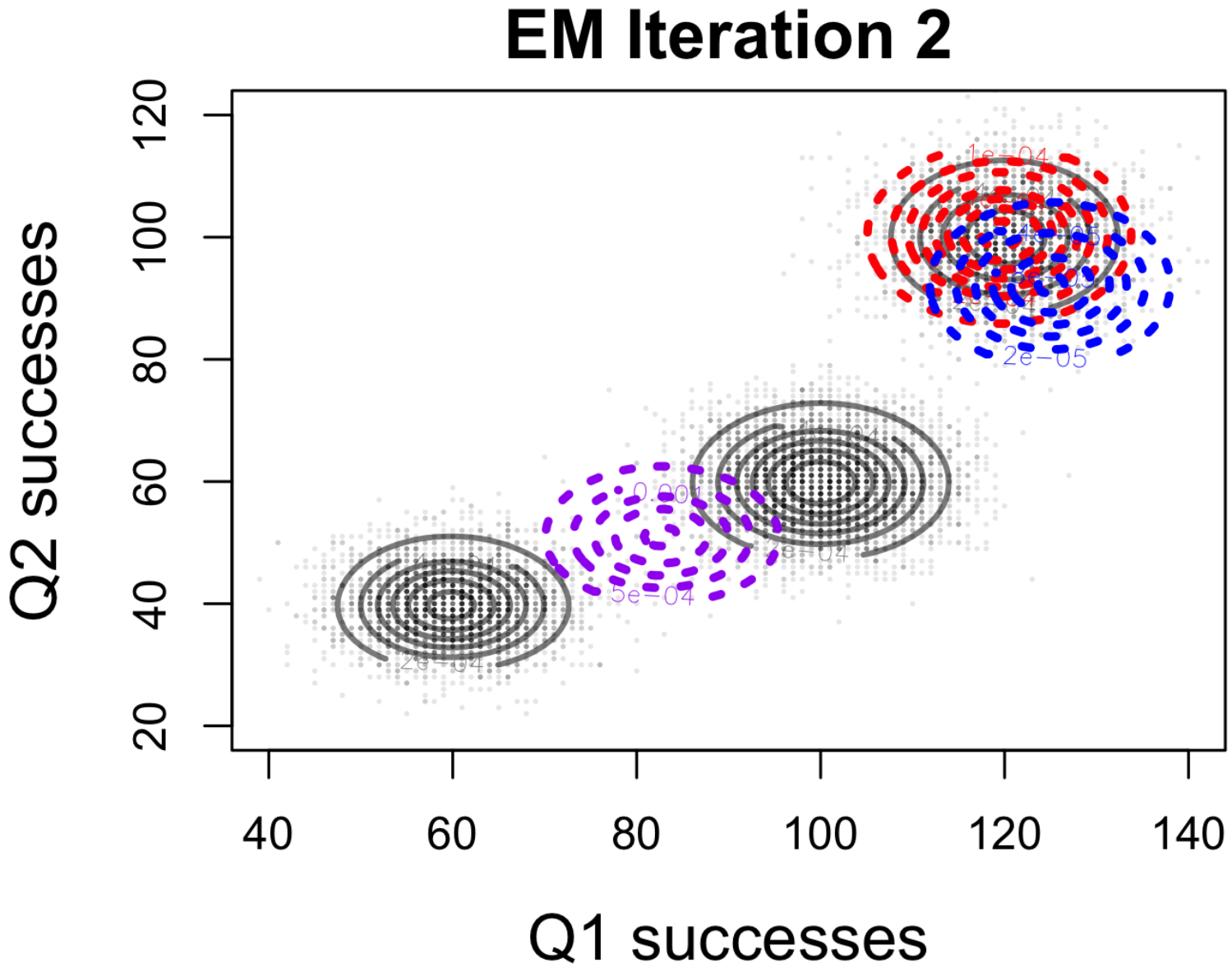


Choose $K = 3$. Guess parameters λ_k, μ_{kq}, and ν_{kq}.

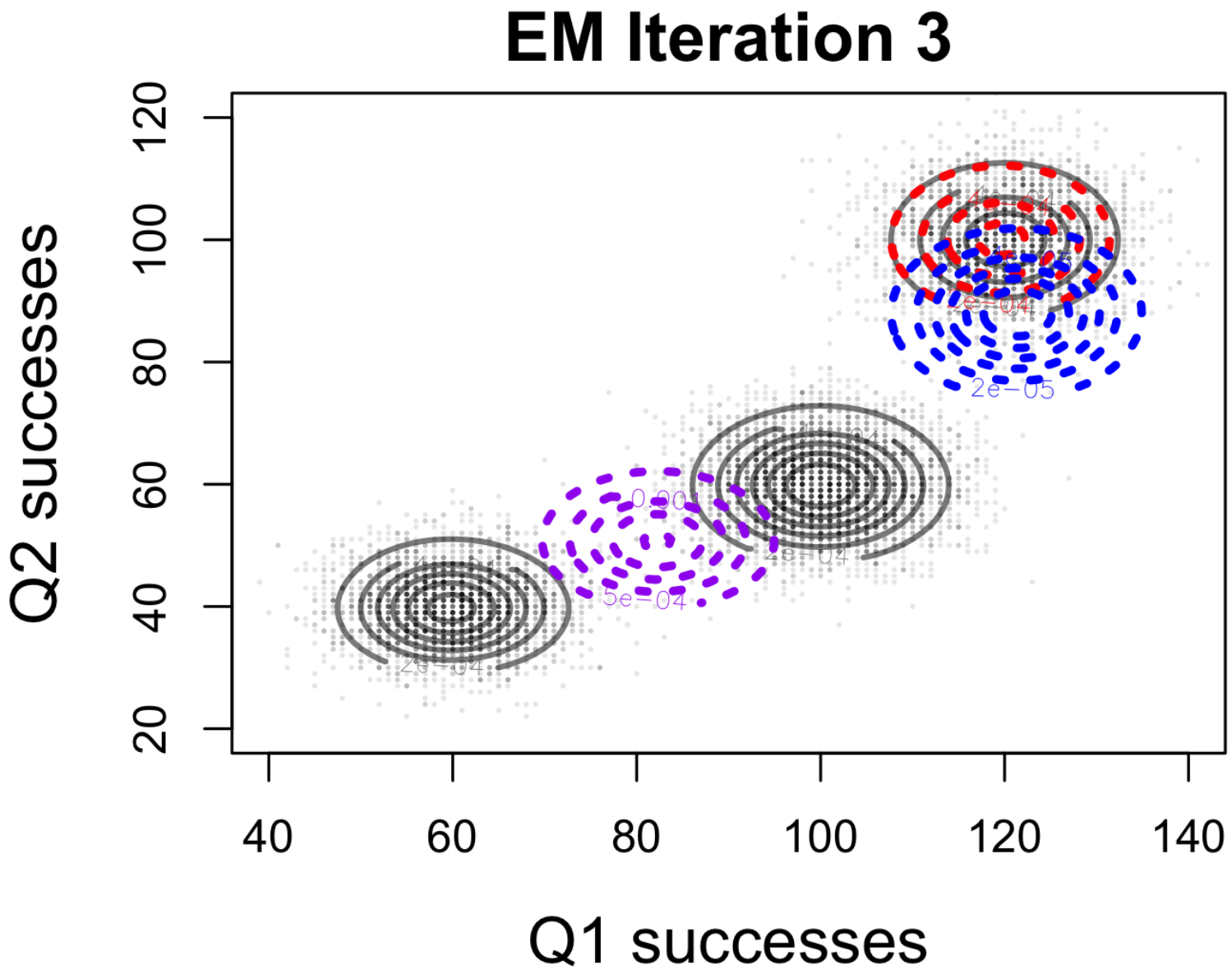
EM over two questions (novel approach)



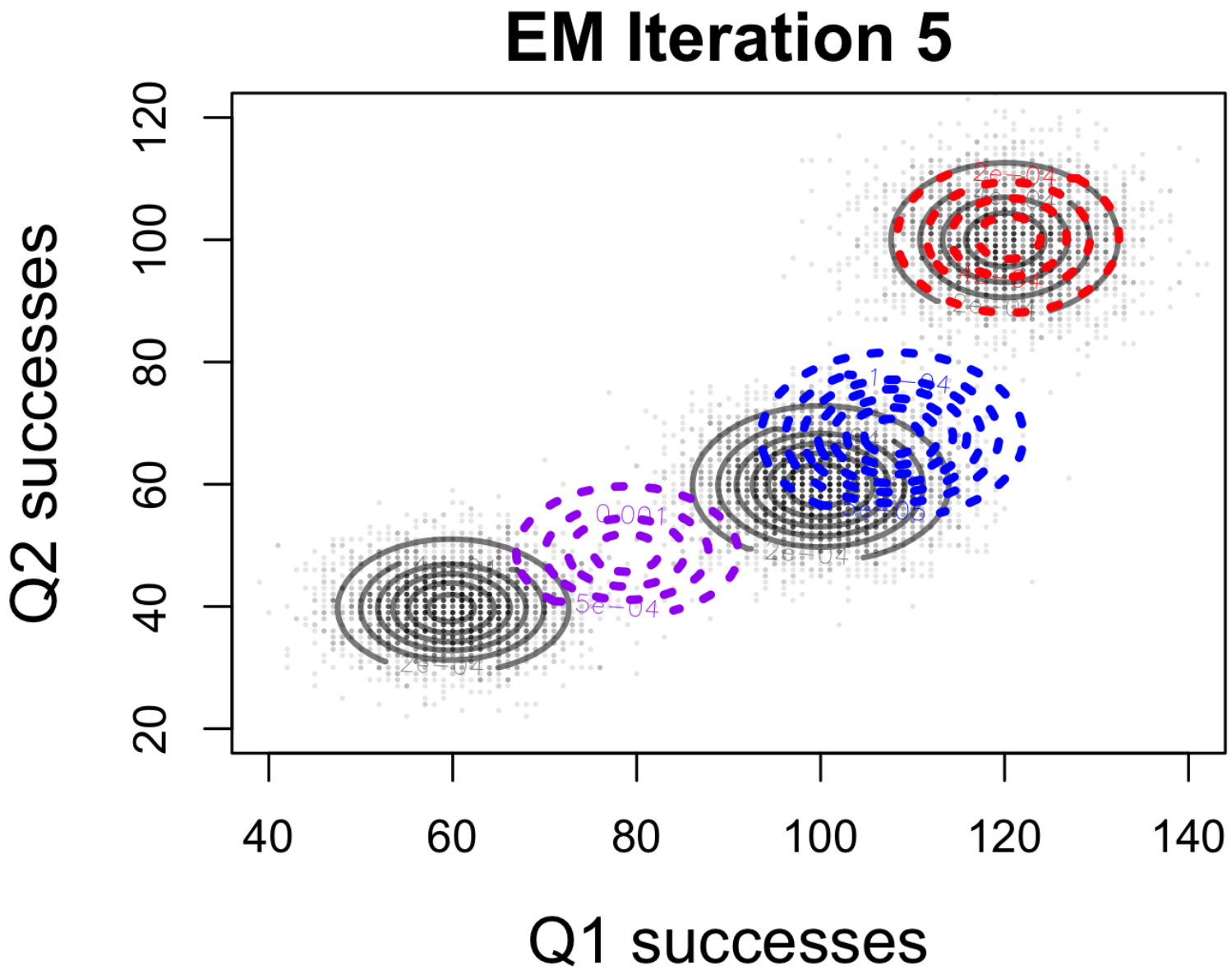
EM over two questions (novel approach)



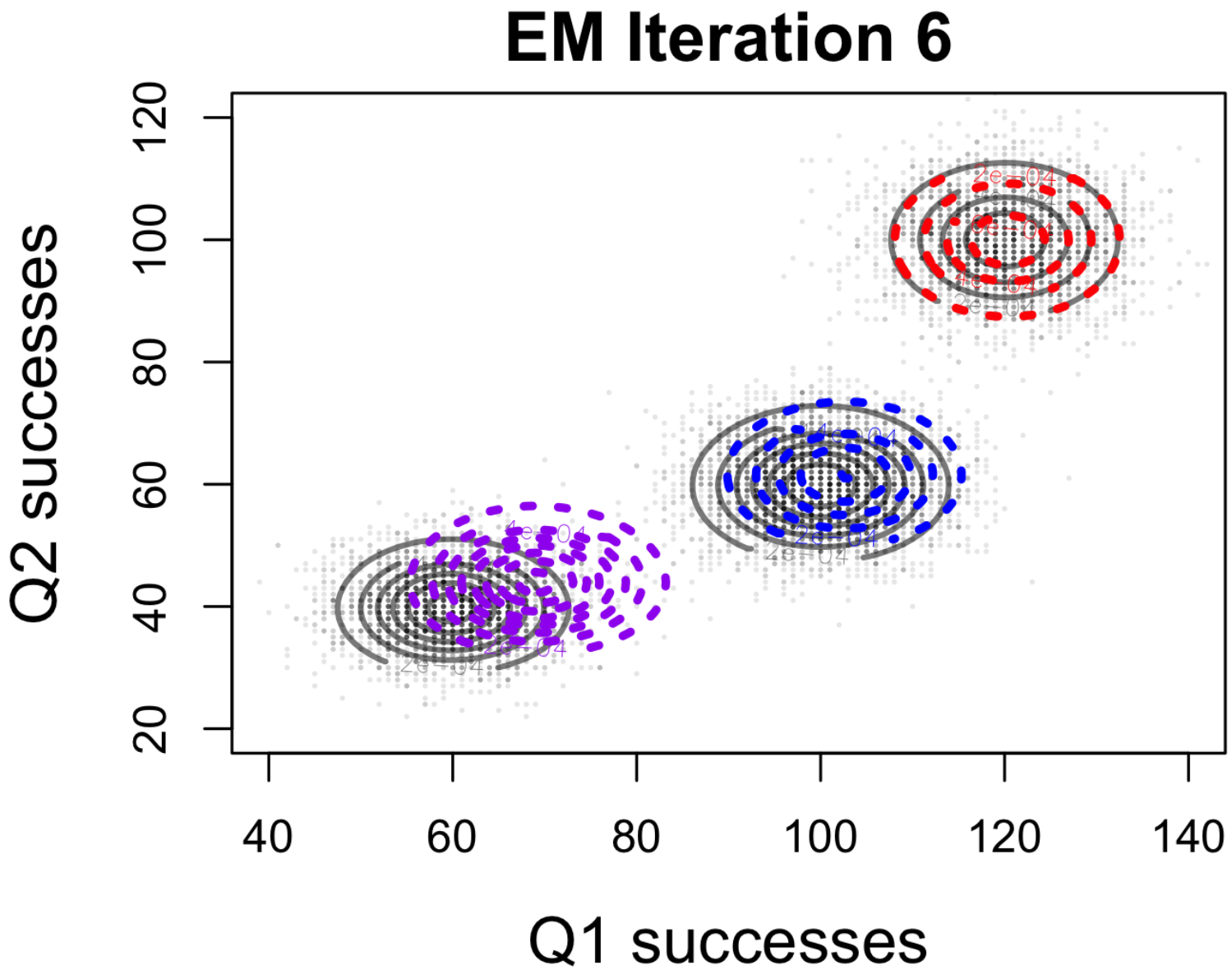
EM over two questions (novel approach)



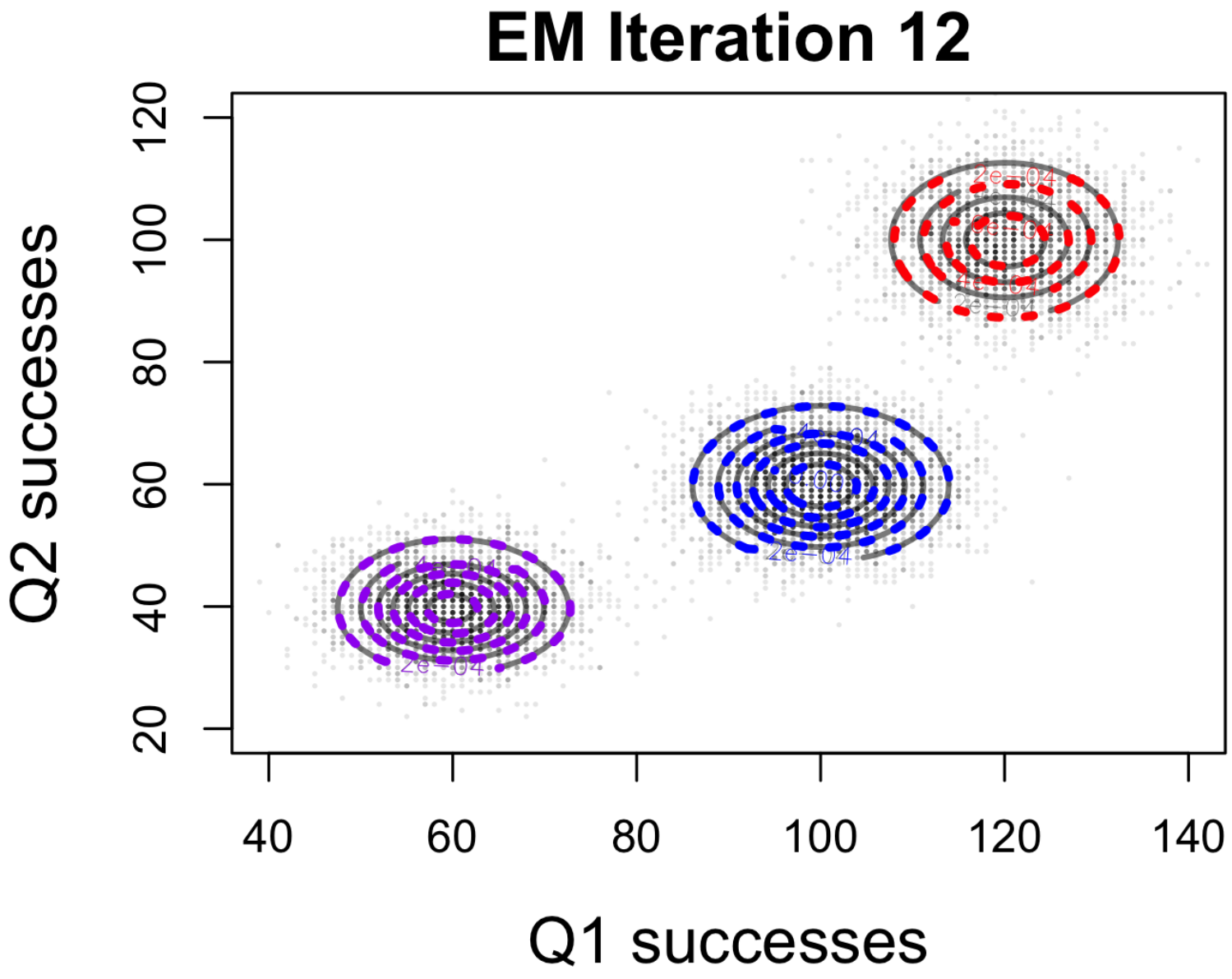
EM over two questions (novel approach)



EM over two questions (novel approach)



EM over two questions (novel approach)



Three big questions, revisited

When fed Q questions, the EM converges to local maxima of $\lambda_k, \mu_{kq}, \nu_{kq}$

1. How many educational ideologies exist among Bowdoin faculty?
    1.1 K, to be obtained through a model selection heuristic
2. How are departments mixed by ideology?
    2.1 Examine responsibilities γ_{ik}, or $\widehat{z_i} = \arg\max \gamma_{ik}$
3. What parameters define these ideologies?
    3.1 Examine learned μ_{kq}, ν_{kq}

Three big questions, revisited (again)

When fed Q questions, the EM converges to local maxima of $\lambda_k, \mu_{kq}, \nu_{kq}$

1. How many ~~educational~~ political ideologies exist among ~~Bowdoin faculty~~ a given electorate?
   - 1.1 K, to be obtained through a model selection heuristic
2. How are ~~departments~~ locales mixed by ideology?
   - 2.1 Examine responsibilities γ_{ik}, or $\widehat{z_i} = \arg\max \gamma_{ik}$
3. What parameters define these ideologies?
   - 3.1 Examine learned μ_{kq}, ν_{kq}

Assumptions:

1. Ideologies are latent, consistent sets of principles and beliefs that guide political engagement
   1.1 Is the basis for ideology demographic, linguistic, religious, regional, economic?
2. Voting populations are finite mixtures of political cultures, which are informed by ideologies

Assumptions:

1. Ideologies are latent, consistent sets of principles and beliefs that guide political engagement
   1.1 Is the basis for ideology demographic, linguistic, religious, regional, economic?
2. Voting populations are finite mixtures of political cultures, which are informed by ideologies

Referendum are very direct, not just voting for a political candidate

Referendum ballot measure, Maine 2016
Legalize marijuana for personal use.
3% tax on household income over $\$200,000$.
Background checks for gun sales and transfers.
Increase minimum wage to \$12 per hour by 2020.
Establish ranked-choice voting.
\$100 million in bonds for transportation projects.

Sanity check: Maine case study

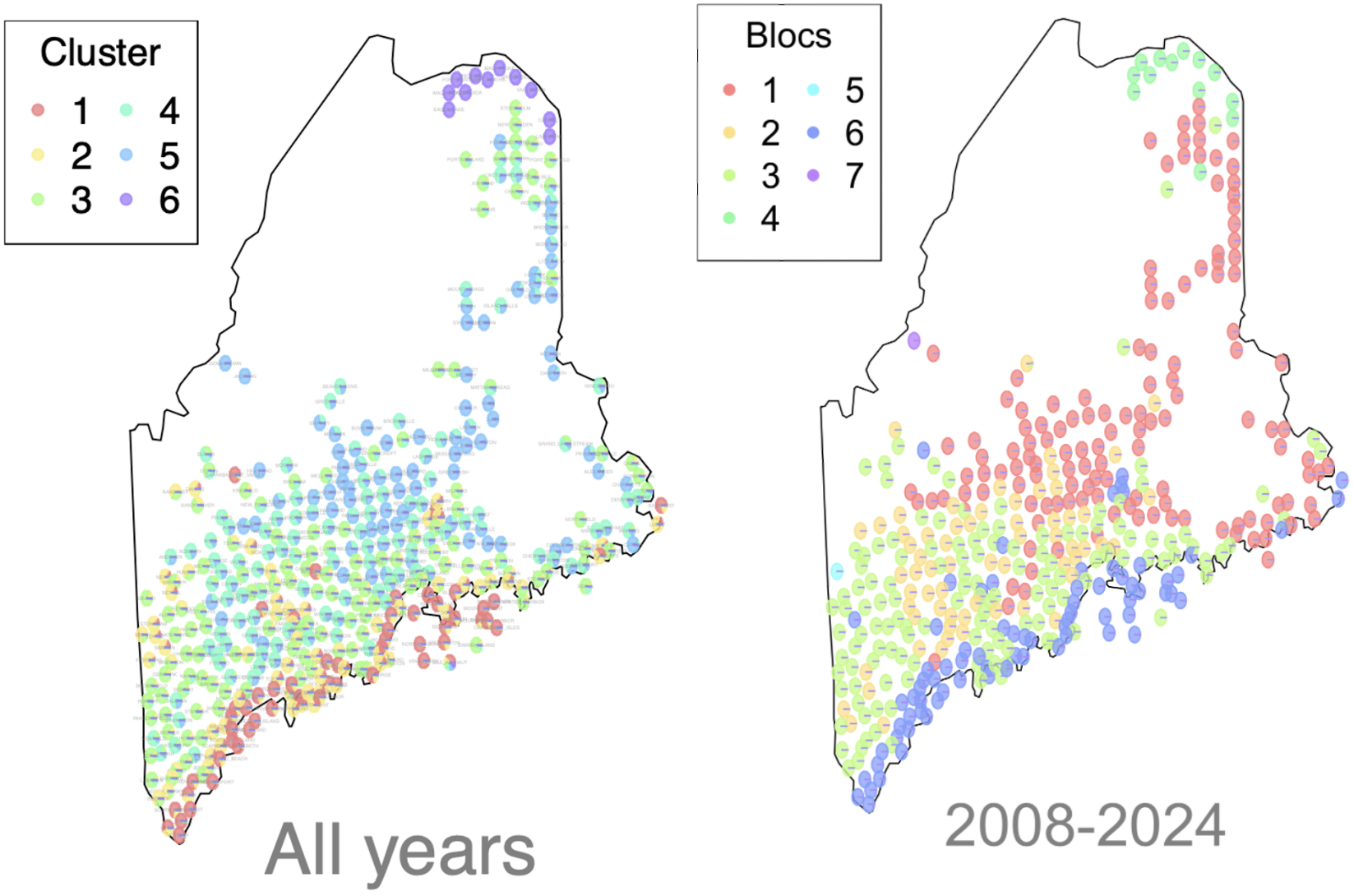


Figure 1: Responsibilities per Maine municipality. Left: Bayesian, 6 weeks of computation. Right: EM, 16.12 seconds of computation.

Important exploratory questions:

1. What is the empirical time complexity of this algorithm?
2. What is the probabilistic accuracy of this inference procedure?

Important exploratory questions:

1. What is the empirical time complexity of this algorithm?
2. What is the probabilistic accuracy of this inference procedure?

Creating our own data:

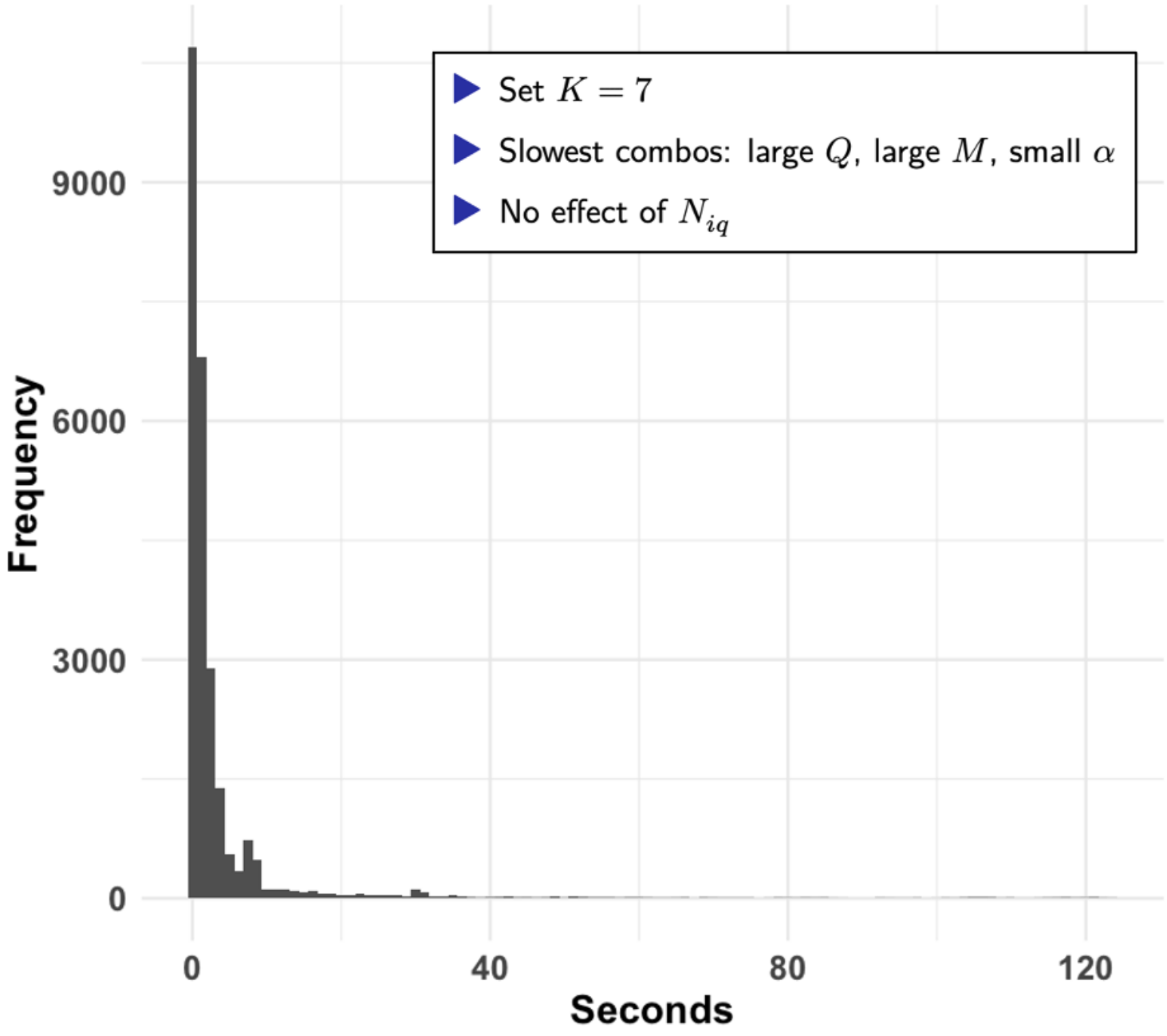
1. Draw a single global mixture weight over the K ideologies
2. Draw a support pattern per latent ideology per question.
3. For each locale, draw a latent ideology label
4. Combine support pattern per ideology with each question, per drawn locale labels

Spanned parameter sets: [a]

$$M \in \{10, 50, 200, 500, 2000\}$$
$$Q \in \{1, 5, 20, 70\}$$
$$N_{iq} \in \{100, 1000, 5000, 30000\}$$
$$\nu_{kq} = 100$$
$$\alpha \in \{0.1, 0.3, 0.5, 0.8\}$$

[a] 320 unique parameter combos

Computational time simulations

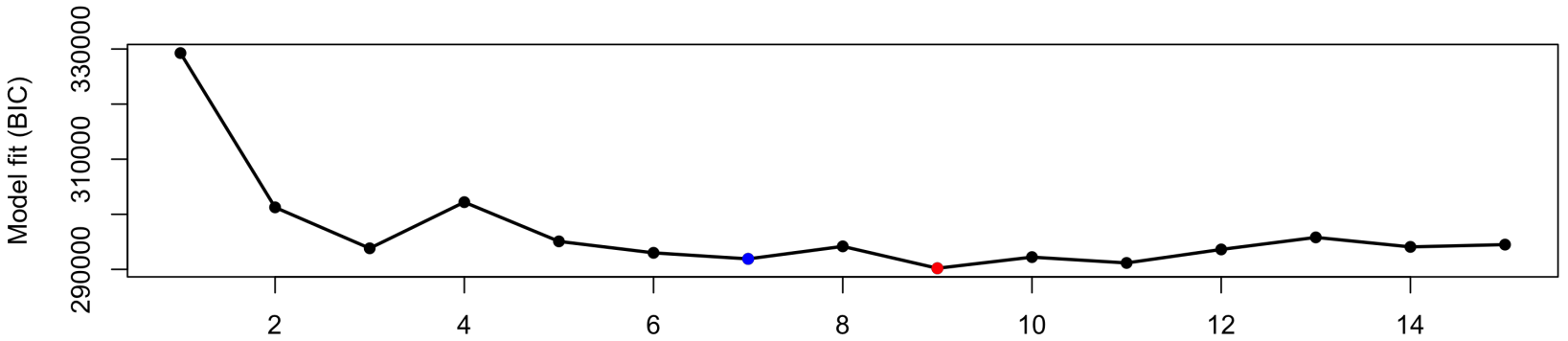


Recovering K simulations

1. Fit model over a range of K
2. Compute model selection heuristic BIC
3. Choose K such that $BIC(K) - BIC(K-1) < \epsilon$

$$M \in \{10, 50, 200, 500, 2000\}$$
$$Q \in \{1, 5, 20, 70\}$$
$$N_{iq} \in \{100, 1000, 5000, 30000\}$$
$$\nu_{kq} = 100$$
$$\alpha \in \{0.1, 0.3, 0.5, 0.8\}$$
$$K_{true} \in \{3, \ldots, 9\}$$
$$K_{search} = 1, \ldots, 15$$

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

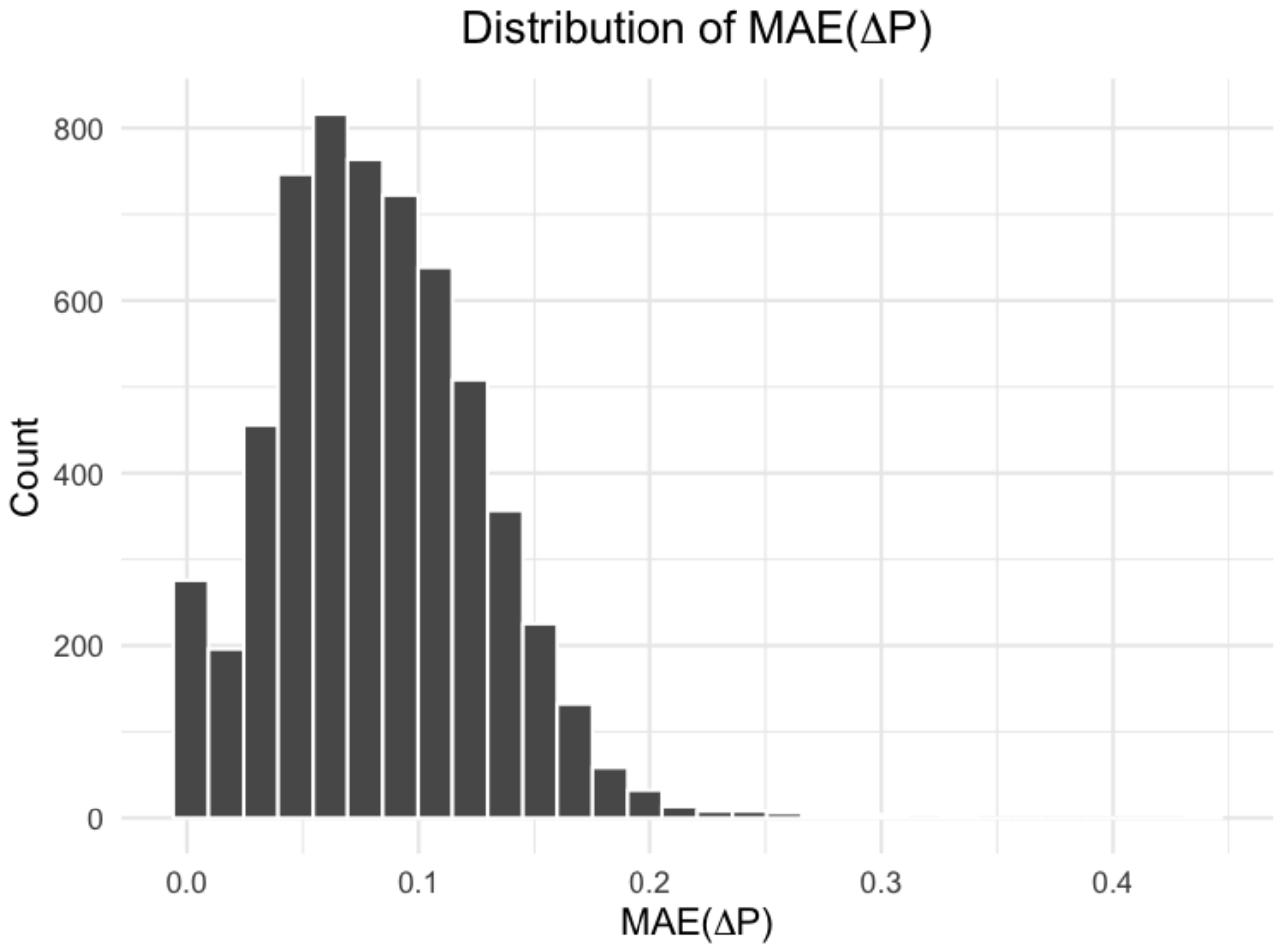


Recovering K simulations

- Slowest recovery of 77.5 minutes
  - $(M, N, Q, \alpha, K) = (2000, 100, 70, 0.5, 3)$
- Larger Q improves model ability to find K
- Model over-predicts K in most cases
  - I.e. Jackman on Maine map
- Note: $1,680,000$ EM runs for this study!

Recovering μ_{kq} in simulation

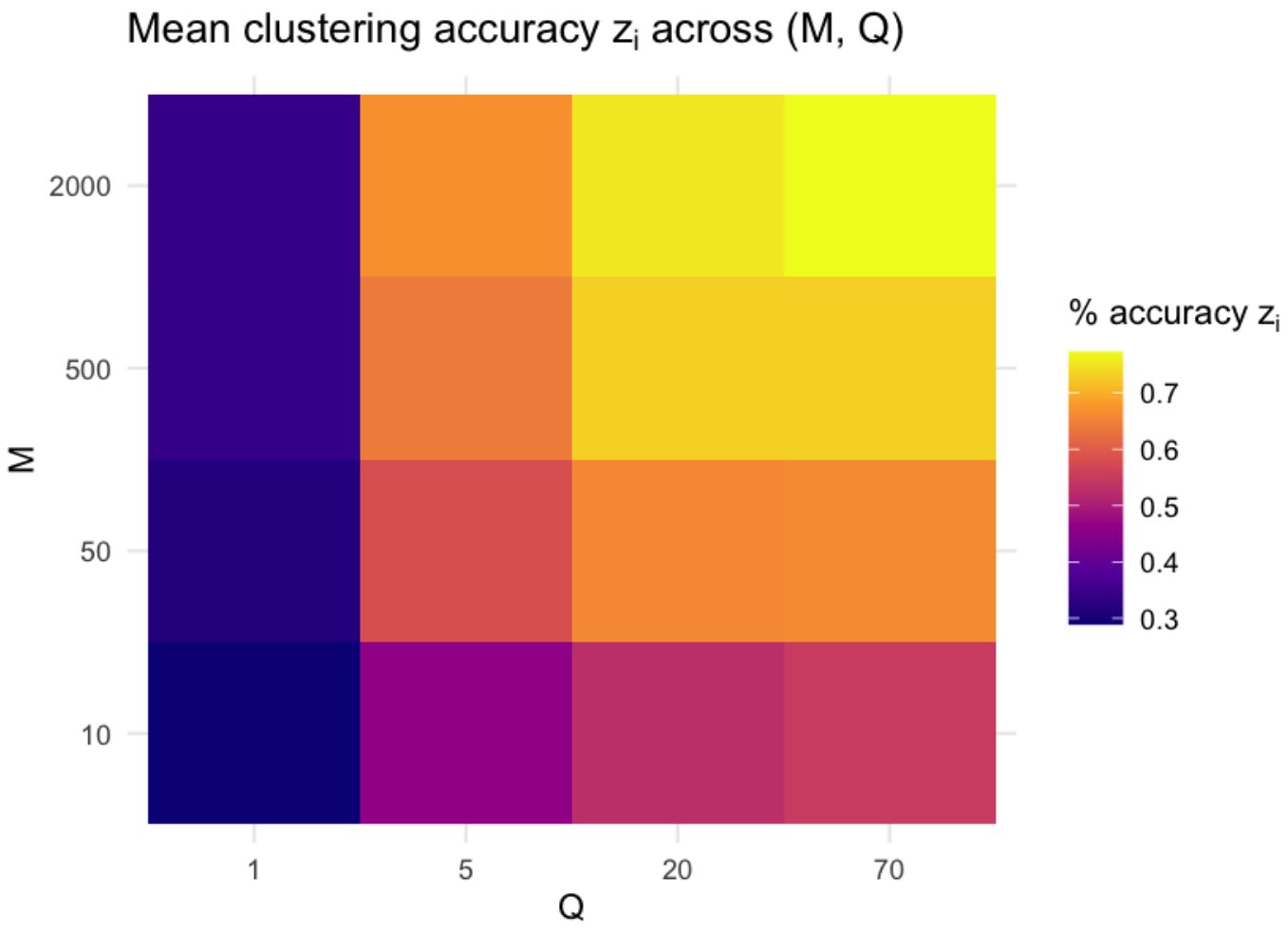
$P_{K\times Q}$ vs $\widehat{P}_{K\times Q}$, across usual parameters with $K \in \{4, \dots, 10\}$



- On average 8% wrong
- Q, M have weak effect

Recovering mixture weights in simulation

z_i vs $\widehat{z_i} = \arg\max_k \hat{\gamma}_{ik}$, across usual parameters with $K \in \{4, \dots, 10\}$

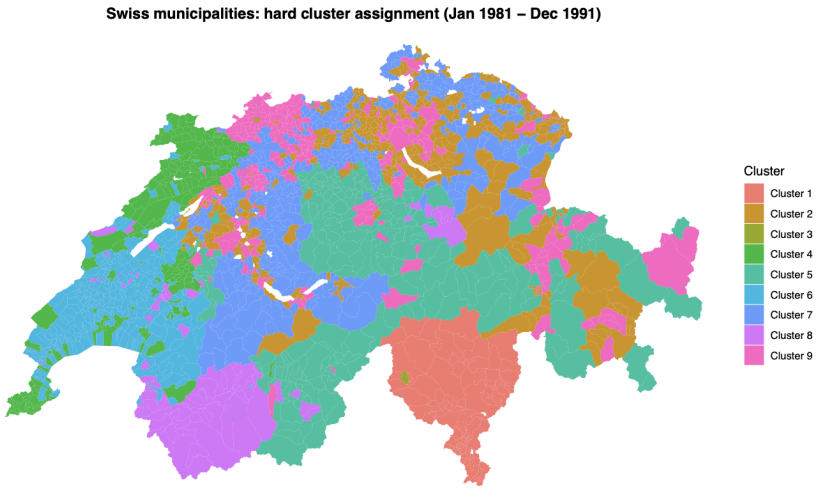


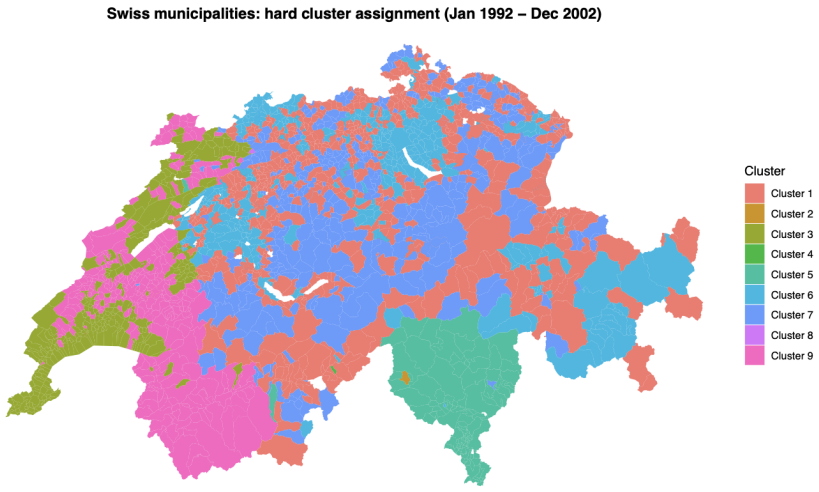
- Feature of the algorithm: sparse responsibility vectors

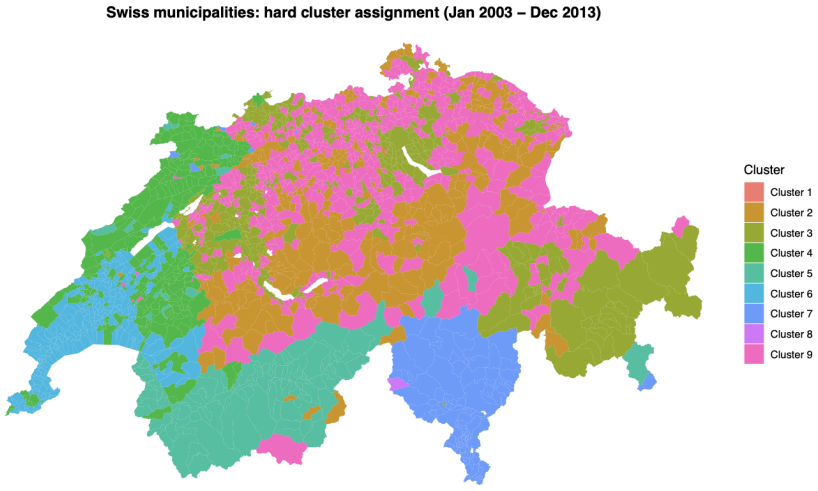
Switzerland case study

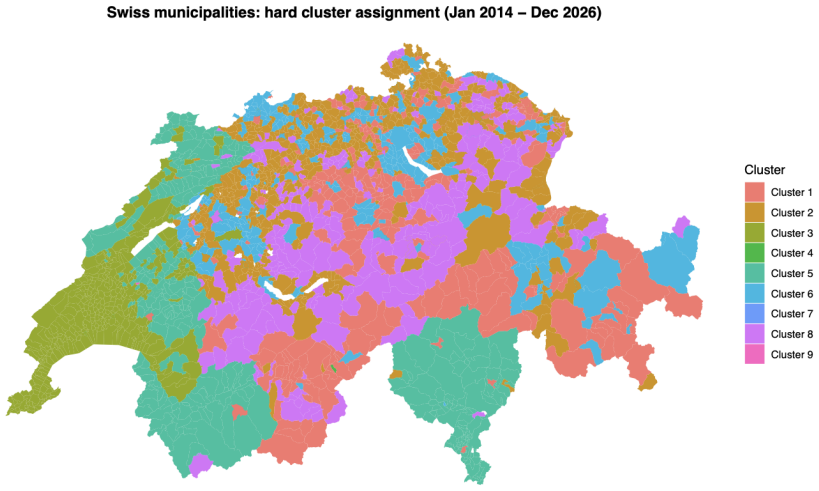
- Switzerland system of direct democracy
- Rich longitudinal referendum data
  - Since 1981, 390 questions across 2048 municipalities (26 cantons)
  - Bayesian method limits out at ~ 500 municipalities
- 7.25 minutes to infer model and below hard assignments

Consistency across 10-year periods

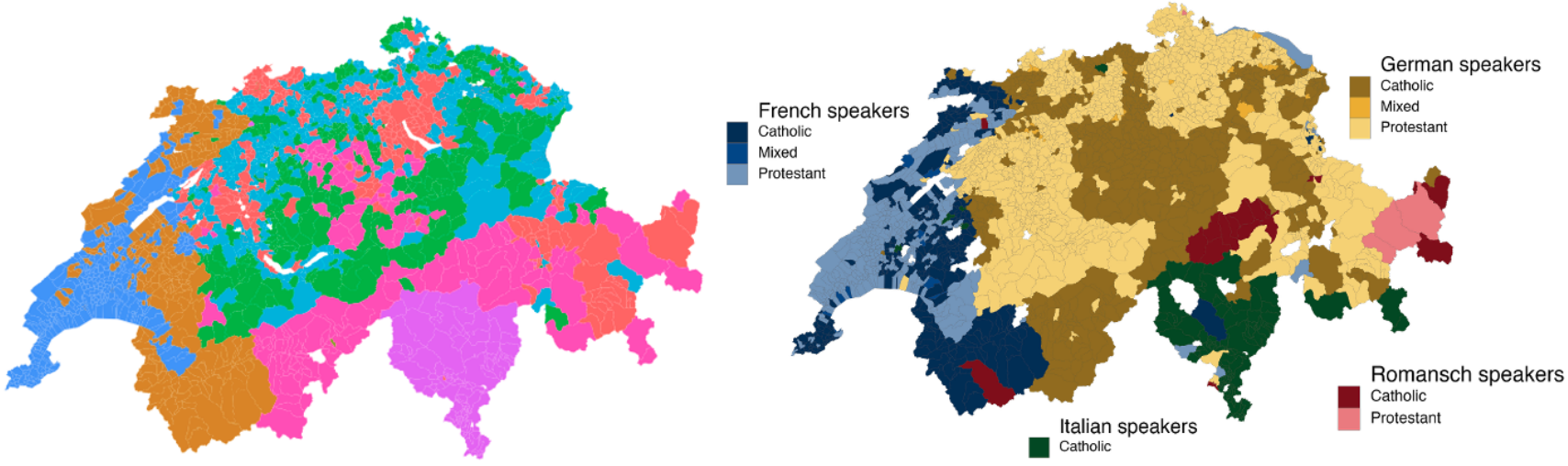








Referendum ideology $\iff$ religious and linguistic patterns



Map (right) procured from 2000 Census data

Back to ideology

- Religion and language are bases for political ideologies!
  - Derived from settlement patterns?

Back to ideology

- Religion and language are bases for political ideologies!
  - Derived from settlement patterns?

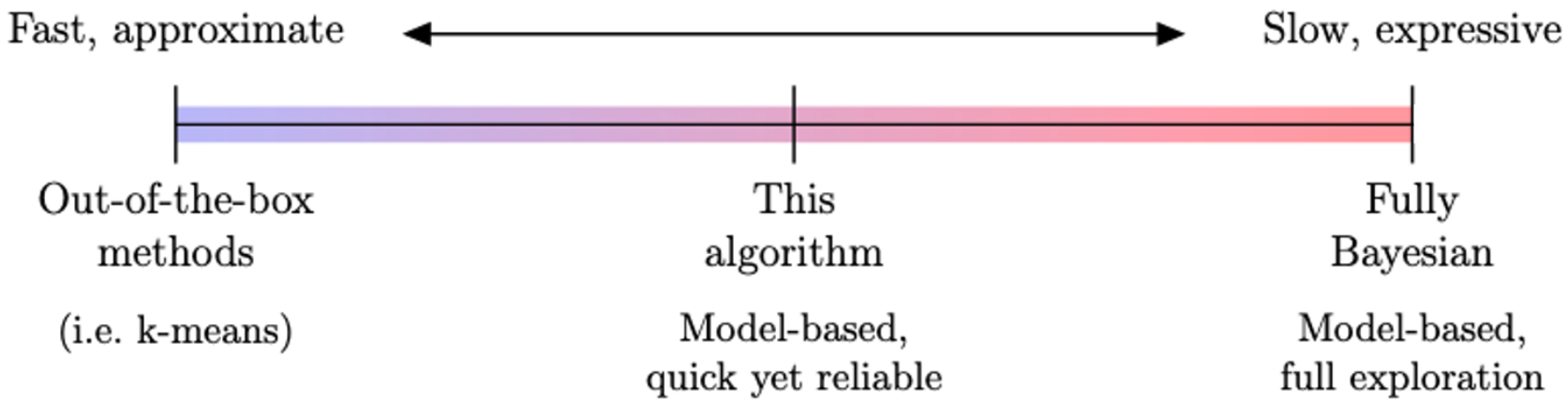
Can we mathematically model voting ideology in referendum? Yes!

Back to ideology

- Religion and language are bases for political ideologies!
  - Derived from settlement patterns?

Can we mathematically model voting ideology in referendum? Yes!

Can we pinpoint exactly where ideology comes from? Not quite...



Acknowledgements

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

Questions, please!

Extra: Future considerations

- How well does the model handle `NA` values in data?
  - Miami referendum data
- Invariance of method to spatial boundaries?
  - Canton vs Bezirke vs municipality
- Methods for more informative exploration of mixture
  - Laplacian approximation near MLE?