

Approximate Bayesian Inference using Gibbs sampling

Mark Johnson

CG168 notes

Bayesian inference in complex models

- If the model is simple enough we can calculate the posterior exactly (conjugate priors)
- When the model is more complicated, we can only approximate the posterior
- *Variational Bayes* calculate the function closest to the posterior within a class of functions
- *Sampling algorithms* produce samples from the posterior distribution
 - ▶ *Markov chain Monte Carlo algorithms* (MCMC) use a Markov chain to produce samples
 - ▶ A *Gibbs sampler* is a particular MCMC algorithm
- *Particle filters* are a kind of *on-line* sampling algorithm (on-line algorithms only make one pass through the data)

Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

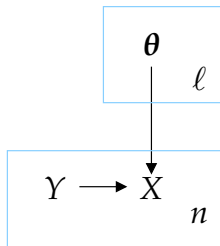
Topic modeling with Dirichlet multinomial mixtures

Summary

Mixture models

- Observations X_i are a *mixture* of ℓ source distributions $F(\boldsymbol{\theta}_k), k = 1, \dots, \ell$
- The value of Y_i specifies which source distribution is used to generate X_i (Y is like a switch)
- If $Y_i = k$, then $X_i \sim F(\boldsymbol{\theta}_k)$
- Here we assume the Y_i are not observed, i.e., *hidden*

$$X_i \mid Y_i, \boldsymbol{\theta} \sim F(\boldsymbol{\theta}_{Y_i}) \quad i = 1, \dots, n$$

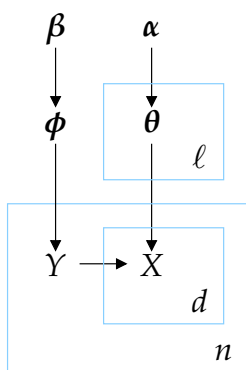


Applications of mixture models

- *Blind source separation*: data X_i come from ℓ different sources
 - ▶ Which X_i come from which source?
(Y_i specifies the source of X_i)
 - ▶ What are the sources?
(θ_k specifies properties of source k)
- X_i could be a document and Y_i the topic of X_i
- X_i could be an image and Y_i the object(s) in X_i
- X_i could be a person's actions and Y_i the "cause" of X_i
- These are *unsupervised learning problems*, which are kinds of *clustering problems*
- In a Bayesian setting, compute posterior $P(\mathbf{Y}, \boldsymbol{\theta} | \mathbf{X})$
But how can we compute this?

Dirichlet Multinomial mixtures

$$\begin{array}{l|l} \boldsymbol{\phi} & \boldsymbol{\beta} \sim \text{DIR}(\boldsymbol{\beta}) \\ Y_i & \boldsymbol{\phi} \sim \text{CATEGORICAL}(\boldsymbol{\phi}) \quad i = 1, \dots, n \\ \boldsymbol{\theta}_k & \boldsymbol{\alpha} \sim \text{DIR}(\boldsymbol{\alpha}) \quad k = 1, \dots, \ell \\ X_{i,j} & Y_i, \boldsymbol{\theta} \sim \text{CATEGORICAL}(\boldsymbol{\theta}_{Y_i}) \quad i = 1, \dots, n; j = 1, \dots, d_i \end{array}$$



- Y_i is generated from a multinomial $\boldsymbol{\phi}$
- Dirichlet priors on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}_k$
- Easy to modify this framework for other applications
- Why does each observation X_i consist of d_i elements?
- What effect do the priors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have?

Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

Why sample?

- Setup: Bayes net has variables \mathbf{X} , whose value \mathbf{x} we observe, and variables \mathbf{Y} , whose value we don't know
 - ▶ \mathbf{Y} includes any *parameters* we want to estimate, such as θ
- Goal: compute the *expected value* of some function f :

$$E[f|\mathbf{X} = \mathbf{x}] = \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$$

- ▶ E.g., $f(\mathbf{x}, \mathbf{y}) = 1$ if x_1 and x_2 are both generated from same hidden state, and 0 otherwise
- In what follows, everything is conditioned on $\mathbf{X} = \mathbf{x}$, so take $P(\mathbf{Y})$ to mean $P(\mathbf{Y}|\mathbf{X} = \mathbf{x})$
- Suppose we can produce n samples $\mathbf{y}^{(t)}$, where $\mathbf{Y}^{(t)} \sim P(\mathbf{Y})$. Then we can estimate:

$$E[f|\mathbf{X} = \mathbf{x}] = \frac{1}{n} \sum_{t=1}^n f(\mathbf{x}, \mathbf{y}^{(t)})$$

Markov chains

- A (first-order) *Markov chain* is a distribution over random variables $S^{(0)}, \dots, S^{(n)}$ all ranging over the same *state space* \mathcal{S} , where:

$$P(S^{(0)}, \dots, S^{(n)}) = P(S^{(0)}) \prod_{t=0}^{n-1} P(S^{(t+1)} | S^{(t)})$$

$S^{(t+1)}$ is *conditionally independent* of $S^{(0)}, \dots, S^{(t-1)}$ given $S^{(t)}$

- A Markov chain in *homogeneous* or *time-invariant* iff:

$$P(S^{(t+1)} = s' | S^{(t)} = s) = P_{s',s} \quad \text{for all } t, s, s'$$

The matrix P is called the *transition probability matrix* of the Markov chain

- If $P(S^{(t)} = s) = \pi_s^{(t)}$ (i.e., $\pi^{(t)}$ is a vector of state probabilities at time t) then:
 - ▶ $\pi^{(t+1)} = P \pi^{(t)}$
 - ▶ $\pi^{(t)} = P^t \pi^{(0)}$

Ergodicity

- A Markov chain with tpm P is *ergodic* iff there is a positive integer m s.t. all elements of P^m are positive (i.e., there is an m -step path between any two states)
- Informally, an ergodic Markov chain “forgets” its past states
- Theorem: For each homogeneous ergodic Markov chain with tpm P there is a *unique limiting distribution* D_P , i.e., as n approaches infinity, the distribution of S_n converges on D_P
- D_P is called the *stationary distribution* of the Markov chain
- Let π be the vector representation of D_P , i.e., $D_P(y) = \pi_y$.
Then:

$$\pi = P \pi, \quad \text{and}$$

$$\pi = \lim_{n \rightarrow \infty} P^n \pi^{(0)} \quad \text{for every initial distribution } \pi^{(0)}$$

Using a Markov chain for inference of $P(Y)$

- Set the state space \mathcal{S} of the Markov chain to the range of Y (\mathcal{S} may be *astronomically large*)
- Find a tpm P such that $P(Y) \sim D_P$
- “Run” the Markov chain, i.e.,
 - ▶ Pick $\mathbf{y}^{(0)}$ somehow
 - ▶ For $t = 0, \dots, n - 1$:
 - sample $\mathbf{y}^{(t+1)}$ from $P(\mathbf{Y}^{(t+1)} | \mathbf{Y}^{(t)} = \mathbf{y}^{(t)})$,
i.e., from $P_{\cdot, \mathbf{y}^{(t)}}$
 - ▶ After discarding the first *burn-in* samples, use remaining samples to calculate statistics
- **WARNING:** in general the samples $\mathbf{y}^{(t)}$ are *not independent*

Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

The Gibbs sampler

- The Gibbs sampler is useful when:
 - ▶ \mathbf{Y} is multivariate, i.e., $\mathbf{Y} = (Y_1, \dots, Y_m)$, and
 - ▶ easy to sample from $P(Y_j | \mathbf{Y}_{-j})$
- The *Gibbs sampler* for $P(\mathbf{Y})$ is the tpm $P = \prod_{j=1}^m P^{(j)}$, where:

$$P_{\mathbf{y}', \mathbf{y}}^{(j)} = \begin{cases} 0 & \text{if } \mathbf{y}'_{-j} \neq \mathbf{y}_{-j} \\ P(Y_j = y'_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j}) & \text{if } \mathbf{y}'_{-j} = \mathbf{y}_{-j} \end{cases}$$

- Informally, the Gibbs sampler cycles through each of the variables Y_j , replacing the current value y_j with a sample from $P(Y_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j})$
- There are *sequential scan* and *random scan* variants of Gibbs sampling

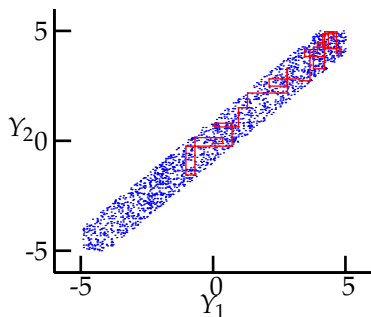
A simple example of Gibbs sampling

$$P(Y_1, Y_2) = \begin{cases} c & \text{if } |Y_1| < 5, |Y_2| < 5 \text{ and } |Y_1 - Y_2| < 1 \\ 0 & \text{otherwise} \end{cases}$$

- The Gibbs sampler for $P(Y_1, Y_2)$ samples repeatedly from:

$$P(Y_2|Y_1) = \text{UNIFORM}(\max(-5, Y_1 - 1), \min(5, Y_1 + 1))$$

$$P(Y_1|Y_2) = \text{UNIFORM}(\max(-5, Y_2 - 1), \min(5, Y_2 + 1))$$



Sample run

Y_1	Y_2
0	0
0	-0.119
0.363	-0.119
0.363	0.146
-0.681	0.146
-0.681	-1.551

A non-ergodic Gibbs sampler

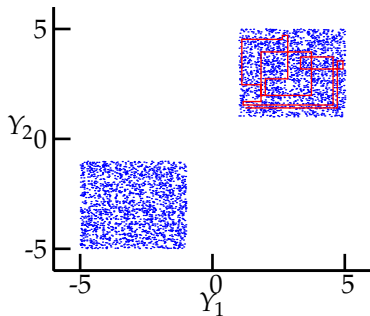
$$P(Y_1, Y_2) = \begin{cases} c & \text{if } 1 < Y_1, Y_2 \leq 5 \text{ or } -5 < Y_1, Y_2 < -1 \\ 0 & \text{otherwise} \end{cases}$$

- The Gibbs sampler for $P(Y_1, Y_2)$, initialized at (2,2), samples repeatedly from:

$$P(Y_2|Y_1) = \text{UNIFORM}(1,5)$$

$$P(Y_1|Y_2) = \text{UNIFORM}(1,5)$$

I.e., *never visits the negative values of Y_1, Y_2*



Sample run

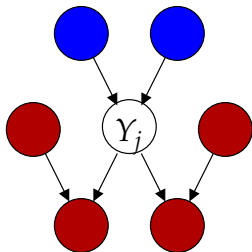
Y_1	Y_2
2	2
2	2.72
2.84	2.72
2.84	4.71
2.63	4.71
2.63	4.52
1.11	4.52

Why does the Gibbs sampler work?

- The Gibbs sampler tpm is $P = \prod_{j=1}^m P^{(j)}$, where $P^{(j)}$ replaces y_j with a sample from $P(Y_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j})$ to produce y'_j
 - But if \mathbf{y} is a sample from $P(\mathbf{Y})$, then so is \mathbf{y}' , since \mathbf{y}' differs from \mathbf{y} only by replacing y_j with a sample from $P(Y_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j})$
 - Since $P^{(j)}$ maps samples from $P(\mathbf{Y})$ to samples from $P(\mathbf{Y})$, so does P
- ⇒ $P(\mathbf{Y})$ is a stationary distribution for P
- If P is ergodic, then $P(\mathbf{Y})$ is the unique stationary distribution for P , i.e., the sampler converges to $P(\mathbf{Y})$

Gibbs sampling with Bayes nets

- Gibbs sampler: update y_j with sample from $P(Y_j | \mathbf{Y}_{-j}) \propto P(Y_j, \mathbf{Y}_{-j})$
- Only need to evaluate terms that depend on Y_j in Bayes net factorization
 - ▶ Y_j appears once in a term $P(Y_j | \mathbf{Y}_{Pa_j})$
 - ▶ Y_j can appear multiple times in terms $P(Y_k | \dots, Y_j, \dots)$
- In graphical terms, need to know value of:
 - ▶ Y_j 's parents
 - ▶ Y_j 's children, and their other parents



Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

Review of Dirichlet distributions

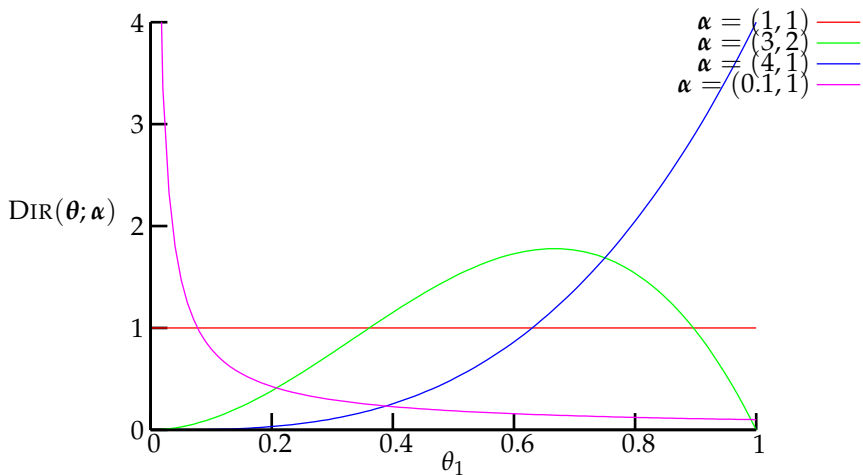
- *Dirichlet distributions* are probability distributions over multinomial parameter vectors
 - ▶ called *Beta distributions* when $m = 2$
- Parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ where $\alpha_j > 0$ that determines the shape of the distribution

$$\text{DIR}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^m \theta_j^{\alpha_j - 1}$$

$$C(\boldsymbol{\alpha}) = \int \prod_{j=1}^m \theta_j^{\alpha_j - 1} d\boldsymbol{\theta} = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^m \alpha_j)}$$

- Γ is a generalization of the factorial function
- $\Gamma(k) = (k - 1)!$ for positive integer k
- $\Gamma(x) = (x - 1)\Gamma(x - 1)$ for all x

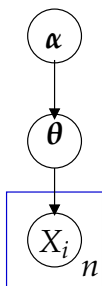
Plots of the Dirichlet distribution



$$\text{DIR}(\theta; \alpha) = \frac{\Gamma(\sum_{j=1}^m \alpha_j)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m \theta_j^{\alpha_j - 1}$$

Dirichlet-multinomial distributions

$$\begin{array}{l|l} \boldsymbol{\theta} & \boldsymbol{\alpha} \sim \text{DIR}(\boldsymbol{\alpha}) \\ X_i & \boldsymbol{\theta} \sim \text{CATEGORICAL}(\boldsymbol{\theta}), \quad i = 1, \dots, n \end{array}$$



$$\begin{aligned} P(\mathbf{X}, \boldsymbol{\theta} | \boldsymbol{\alpha}) &= P(\boldsymbol{\theta} | \boldsymbol{\alpha}) P(\mathbf{X} | \boldsymbol{\theta}) \\ &= \left(\frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^m \theta_j^{\alpha_j - 1} \right) \left(\prod_{j=1}^m \theta_j^{N_j(\mathbf{X})} \right) \\ &= \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^m \theta_j^{N_j(\mathbf{X}) + \alpha_j - 1}, \text{ so} \end{aligned}$$

$$\begin{aligned} P(\boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}) &\propto P(\mathbf{X}, \boldsymbol{\theta} | \boldsymbol{\alpha}) \\ &\propto \prod_{j=1}^m \theta_j^{N_j(\mathbf{X}) + \alpha_j - 1}, \text{ so} \end{aligned}$$

$$\boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha} \sim \text{DIR}(\mathbf{N} + \boldsymbol{\alpha})$$

Sampling from a Dirichlet

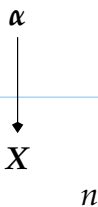
$$\boldsymbol{\theta} | \boldsymbol{\alpha} \sim \text{DIR}(\boldsymbol{\alpha}) \quad \text{iff} \quad P(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^m \theta_j^{\alpha_j - 1}, \quad \text{where:}$$

$$C(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^m \alpha_j)}$$

- There are several algorithms for producing samples from $\text{DIR}(\boldsymbol{\alpha})$. A simple one relies on the following result:
- If $V_k \sim \text{GAMMA}(\alpha_k)$ and $\theta_k = V_k / (\sum_{k'=1}^m V_{k'})$, then $\boldsymbol{\theta} \sim \text{DIR}(\boldsymbol{\alpha})$
- This leads to the following algorithm for producing a sample $\boldsymbol{\theta}$ from $\text{DIR}(\boldsymbol{\alpha})$
 - ▶ Sample v_k from $\text{GAMMA}(\alpha_k)$ for $k = 1, \dots, m$
 - ▶ Set $\theta_k = v_k / (\sum_{k'=1}^m v_{k'})$

“Collapsed” model, i.e., integrate out θ

$$\begin{array}{l|l} \theta & \alpha \sim \text{DIR}(\alpha) \\ X_i & \theta \sim \text{CATEGORICAL}(\theta), \quad i = 1, \dots, n \end{array}$$



$$\begin{aligned} P(\mathbf{X}|\alpha) &= \int P(\mathbf{X}, \theta|\alpha) d\theta = \int P(\mathbf{X}|\theta) P(\theta|\alpha) d\theta \\ &= \int \left(\prod_{j=1}^m \theta_j^{N_j} \right) \left(\frac{1}{C(\alpha)} \prod_{j=1}^m \theta_j^{\alpha_j - 1} \right) d\theta \\ &= \frac{1}{C(\alpha)} \int \prod_{j=1}^m \theta_j^{N_j + \alpha_j - 1} d\theta \\ &= \frac{C(N + \alpha)}{C(\alpha)}, \quad \text{where } C(\alpha) = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^m \alpha_j)} \end{aligned}$$

The Dirichlet-multinomial posterior

$$P(\mathbf{X}|\boldsymbol{\alpha}) = \frac{C(\mathbf{N} + \boldsymbol{\alpha})}{C(\boldsymbol{\alpha})} \quad \text{where} \quad C(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\alpha_{\bullet})} \quad \text{and} \quad \alpha_{\bullet} = \sum_{j=1}^m \alpha_j$$

$$\begin{aligned} P(\mathbf{X}|\boldsymbol{\alpha}) &= \left(\frac{\prod_{j=1}^m \Gamma(N_j + \alpha_j)}{\Gamma(n + \alpha_{\bullet})} \right) \left(\frac{\Gamma(\alpha_{\bullet})}{\prod_{j=1}^m \Gamma(\alpha_j)} \right) \\ &= \left(\prod_{j=1}^m \frac{\Gamma(N_j + \alpha_j)}{\Gamma(\alpha_j)} \right) \left(\frac{\Gamma(\alpha_{\bullet})}{\Gamma(n + \alpha_{\bullet})} \right) \\ &= \frac{\alpha_1}{\alpha_{\bullet}} \times \frac{\alpha_1 + 1}{\alpha_{\bullet} + 1} \times \dots \times \frac{\alpha_1 + N_1 - 1}{\alpha_{\bullet} + N_1 - 1} \\ &\quad \times \frac{\alpha_2}{\alpha_{\bullet} + N_1} \times \frac{\alpha_2 + 1}{\alpha_{\bullet} + N_1 + 1} \times \dots \times \frac{\alpha_2 + N_2 - 1}{\alpha_{\bullet} + N_1 + N_2 - 1} \\ &\quad \times \dots \\ &\quad \times \frac{\alpha_m}{\alpha_{\bullet} + n - N_m - 1} \times \frac{\alpha_m + 1}{\alpha_{\bullet} + n - N_m} \times \dots \times \frac{\alpha_m + N_m - 1}{\alpha_{\bullet} + n - 1} \end{aligned}$$

Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

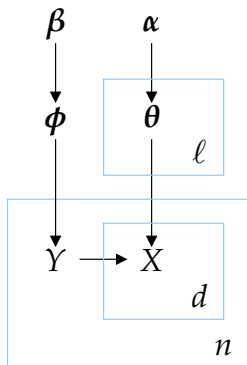
Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

Dirichlet-Multinomial mixtures

ϕ		β	\sim	DIR(β)	
Y_i		ϕ	\sim	CATEGORICAL(ϕ)	$i = 1, \dots, n$
θ_k		α	\sim	DIR(α)	$k = 1, \dots, \ell$
$X_{i,j}$		Y_i, θ	\sim	CATEGORICAL(θ_{Y_i})	$i = 1, \dots, n; j = 1, \dots, d_i$

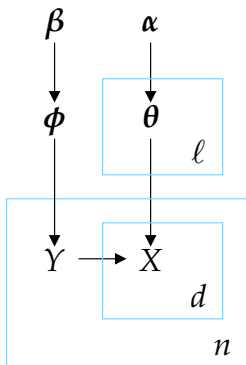


$$\begin{aligned}
 & P(\phi, Y, \theta, X | \alpha, \beta) \\
 &= \frac{1}{C(\beta)} \prod_{k=1}^{\ell} \left(\phi_k^{\beta_k - 1 + N_k(Y)} \right. \\
 & \quad \left. \frac{1}{C(\alpha)} \prod_{j=1}^m \theta_{k,j}^{\alpha_j - 1 + \sum_{i: Y_i=k} N_j(X_i)} \right)
 \end{aligned}$$

where
$$C(\alpha) = \frac{\prod_{j=1}^m \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^m \alpha_j)}$$

Gibbs sampling for D-M mixtures

$\boldsymbol{\phi}$		$\boldsymbol{\beta}$	\sim	DIR($\boldsymbol{\beta}$)	
Y_i		$\boldsymbol{\phi}$	\sim	CATEGORICAL($\boldsymbol{\phi}$)	$i = 1, \dots, n$
$\boldsymbol{\theta}_k$		$\boldsymbol{\alpha}$	\sim	DIR($\boldsymbol{\alpha}$)	$k = 1, \dots, \ell$
$X_{i,j}$		$Y_i, \boldsymbol{\theta}$	\sim	CATEGORICAL($\boldsymbol{\theta}_{Y_i}$)	$i = 1, \dots, n; j = 1, \dots, d_i$

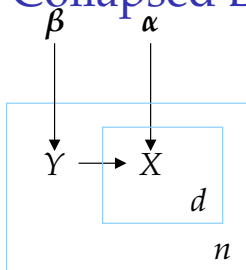


$$P(\boldsymbol{\phi} | \mathbf{Y}, \boldsymbol{\beta}) = \text{DIR}(\boldsymbol{\phi}; \boldsymbol{\beta} + \mathbf{N}(\mathbf{Y}))$$

$$P(Y_i = k | \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{X}_i) \propto \phi_k \prod_{j=1}^m \theta_{k,j}^{N_j(\mathbf{X}_i)}$$

$$P(\boldsymbol{\theta}_k | \boldsymbol{\alpha}, \mathbf{X}, \mathbf{Y}) = \text{DIR}(\boldsymbol{\theta}_k; \boldsymbol{\alpha} + \sum_{i: Y_i=k} \mathbf{N}(\mathbf{X}_i))$$

Collapsed Dirichlet Multinomial mixtures



$$P(\mathbf{Y}|\boldsymbol{\beta}) = \frac{C(\mathbf{N}(\mathbf{Y}) + \boldsymbol{\beta})}{C(\boldsymbol{\beta})}$$

$$P(\mathbf{X}|\boldsymbol{\alpha}, \mathbf{Y}) = \prod_{k=1}^{\ell} \frac{C(\boldsymbol{\alpha} + \sum_{i:Y_i=k} \mathbf{N}(\mathbf{X}_i))}{C(\boldsymbol{\alpha})}, \text{ so}$$

$$P(Y_i = k | \mathbf{Y}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{N_k(\mathbf{Y}_{-i}) + \beta_k}{n - 1 + \beta_{\bullet}} \frac{C(\boldsymbol{\alpha} + \sum_{i' \neq i: Y_{i'}=k} \mathbf{N}(\mathbf{X}_{i'}) + \mathbf{N}(\mathbf{X}_i))}{C(\boldsymbol{\alpha} + \sum_{i' \neq i: Y_{i'}=k} \mathbf{N}(\mathbf{X}_{i'}))}$$

- $P(Y_i = k | \mathbf{Y}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is proportional to the prob. of generating:
 - ▶ $Y_i = k$, given the other \mathbf{Y}_{-i} , and
 - ▶ \mathbf{X}_i in cluster k , given \mathbf{X}_{-i} and \mathbf{Y}_{-i}

Gibbs sampling for Dirichlet multinomial mixtures

- Each \mathbf{X}_i could be generated from one of several Dirichlet multinomials
- The variable Y_i indicates the source for \mathbf{X}_i
- The *uncollapsed sampler* samples \mathbf{Y} , $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$
- The *collapsed sampler* integrates out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and just samples \mathbf{Y}
- Collapsed samplers often (but not always) converge faster than uncollapsed samplers
- Collapsed samplers are usually easier to implement

Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

Topic modeling of child-directed speech

- Data: Adam, Eve and Sarah's mothers' child-directed utterances

I like it .

why don't you read Shadow yourself ?

that's a terribly small horse for you to ride .

why don't you look at some of the toys in the basket .

want to ?

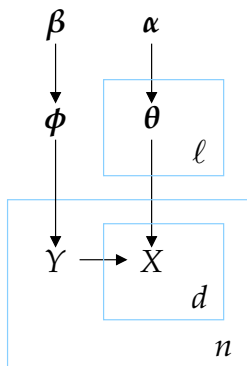
do you want to see what I have ?

what is that ?

not in your mouth .

- 59,959 utterances, composed of 337,751 words

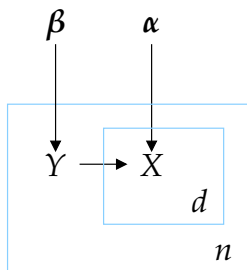
Uncollapsed Gibbs sampler for topic model



- Data consists of “documents” \mathbf{X}_i
- Each \mathbf{X}_i is a sequence of “words” $X_{i,j}$
- Initialize by *randomly* assign each document \mathbf{X}_i to a topic Y_i
- Repeat the following:
 - ▶ Replace ϕ with a sample from a Dirichlet with parameters $\beta + \mathbf{N}(\mathbf{Y})$
 - ▶ For each topic k , replace θ_k with a sample from a Dirichlet with parameters $\alpha + \sum_{i:Y_i=k} \mathbf{N}(\mathbf{X}_i)$
 - ▶ For each document i , replace Y_i with a sample from

$$P(Y_i = k | \phi, \theta, \mathbf{X}_i) \propto \phi_k \prod_{j=1}^m \theta_{k,j}^{N_j(\mathbf{X}_i)}$$

Collapsed Gibbs sampler for topic model



- Initialize by *randomly* assign each document X_i to a topic Y_i
- Repeat the following:
 - ▶ For each document i in $1, \dots, n$ (in random order):
 - Replace Y_i with a random sample from $P(Y_i | Y_{-i}, \alpha, \beta)$

$$P(Y_i = k | Y_{-i}, \alpha, \beta) \propto \frac{N_k(Y_{-i}) + \beta_k}{n - 1 + \beta_{\bullet}} \frac{C(\alpha + \sum_{i' \neq i: Y_{i'} = k} N(X_{i'}) + N(X_i))}{C(\alpha + \sum_{i' \neq i: Y_{i'} = k} N(X_{i'}))}$$

Topics assigned after 100 iterations

- 1 big drum ?
- 3 horse .
- 8 who is that ?
- 9 those are checkers .
- 3 two checkers # yes .
- 1 play checkers ?
- 1 big horn ?
- 2 get over # Mommy .
- 1 shadow ?
- 9 I like it .
- 1 why don't you read Shadow yourself ?
- 9 that's a terribly small horse for you to ride .
- 2 why don't you look at some of the toys in the basket .
- 1 want to ?
- 1 do you want to see what I have ?
- 8 what is that ?
- 2 not in your mouth .
- 2 let me put them together .
- 2 no # put floor .
- 3 no # that's his pencil .
- 3 that's not Daddy # that's Colin

Most probable words in each cluster

P(Y=4) = 0.4334		P(Y=9) = 0.3111		P(Y=7) = 0.2555		P(Y=3) = 5.003e	
X	P(X Y)	X	P(X Y)	X	P(X Y)	X	P(X Y)
.	0.12526	?	0.19147	.	0.2258	quack	0.85
#	0.045402	you	0.062577	#	0.0695	.	0.15
you	0.040475	what	0.061256	that's	0.034538		
the	0.030259	that	0.022295	a	0.034066		
it	0.024154	the	0.022126	no	0.02649		
I	0.021848	#	0.021809	oh	0.023558		
to	0.018473	is	0.021683	yeah	0.020332		
don't	0.015473	do	0.016127	the	0.014907		
a	0.013662	it	0.015927	xxx	0.014288		
?	0.013459	a	0.015092	not	0.013864		
in	0.011708	to	0.013783	it's	0.013343		
on	0.011064	did	0.012631	?	0.013033		
your	0.010145	are	0.011427	yes	0.011795		
and	0.009578	what's	0.011195	right	0.0094166		
that	0.0093303	your	0.0098961	alright	0.0088953		
have	0.0088019	huh	0.0082591	is	0.0087975		
no	0.0082514	want	0.0076782	you're	0.0076571		
put	0.0067486	where	0.0072346	one	0.006647		
know	0.0064239	why	0.0070656	!	0.0057673		

Remarks on cluster results

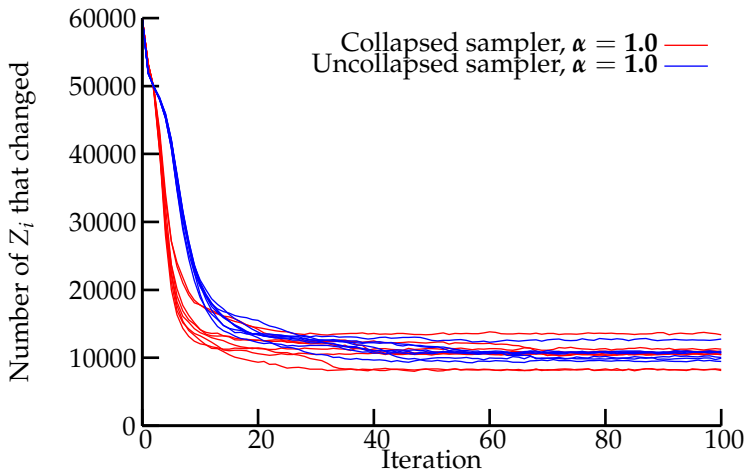
- The samplers cluster words by clustering the documents they appear in, and cluster documents by clustering the words that appear in them
 - Even though there were $\ell = 10$ clusters and $\alpha = \mathbf{1}, \beta = \mathbf{1}$, typically only 4 clusters were occupied after convergence
 - Words x with high marginal probability $P(X = x)$ are typically so frequent that they occur in all clusters
- ⇒ Listing the most probable words in each cluster may not be a good way of characterizing the clusters
- Instead, we can Bayes invert and find *the words that are most strongly associated with each class*

$$P(Y = k | X = x) = \frac{N_{k,x}(\mathbf{Y}, \mathbf{X}) + \epsilon}{N_x(\mathbf{X}) + \epsilon \ell}$$

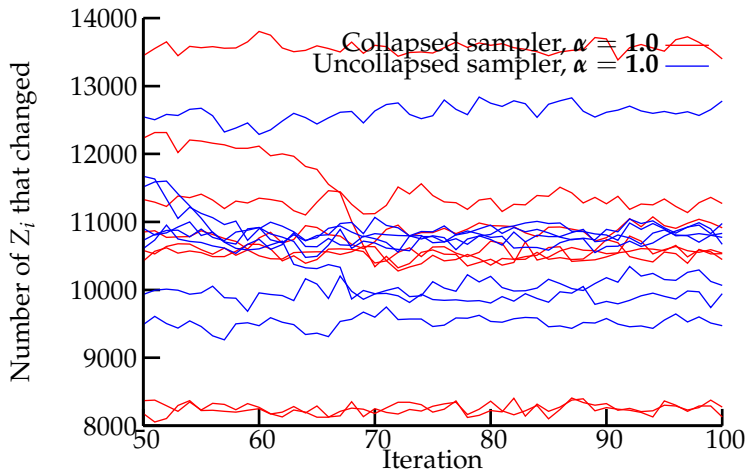
Purest words of each cluster

P(Y=4) = 0.4334		P(Y=9) = 0.3111		P(Y=7) = 0.2555		P(Y=3) = 5.0	
X	P(Y X)	X	P(Y X)	X	P(Y X)	X	P(Y X)
I'll	0.97168	d(o)	0.97138	0	0.94715	quack	0.64
we'll	0.96486	what's	0.95242	mmhm	0.944	.	0.00
c(o)me	0.95319	what're	0.94348	www	0.90244		
you'll	0.95238	happened	0.93722	m:hm	0.83019		
may	0.94845	hmm	0.93343	uhhuh	0.81667		
let's	0.947	whose	0.92437	uh(uh)	0.78571		
thought	0.94382	what	0.9227	uhuh	0.77551		
won't	0.93645	where's	0.92241	that's	0.7755		
come	0.93588	doing	0.90196	yep	0.76531		
let	0.93255	where'd	0.9009	um	0.76282		
I	0.93192	don't]	0.89157	oh+boy	0.73529		
(h)ere	0.93082	whyn't	0.89157	d@l	0.72603		
stay	0.92073	who	0.88527	goodness	0.7234		
later	0.91964	how's	0.875	s@l	0.72		
thank	0.91667	who's	0.85068	sorry	0.70588		
them	0.9124	[:	0.85047	thank+you	0.6875		
can't	0.90762	?	0.84783	o:h	0.68		
never	0.9058	matter	0.82963	nope	0.67857		
em	0.89922	what'd	0.8125	hi	0.67213		

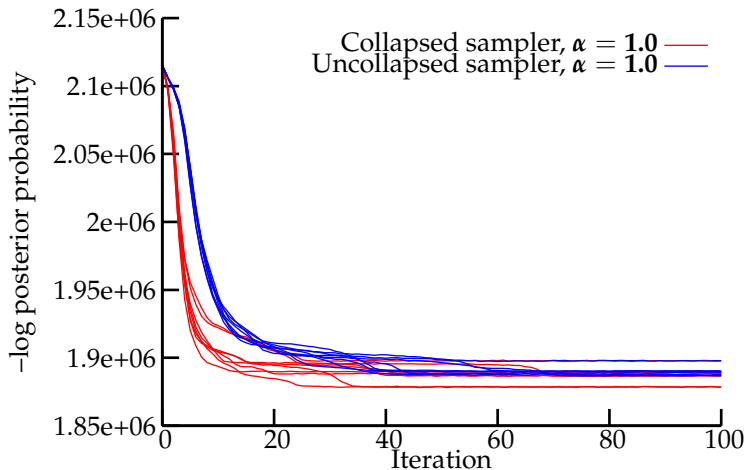
Number of Y_i that changed each iteration with $\ell = 10, \alpha = 1$



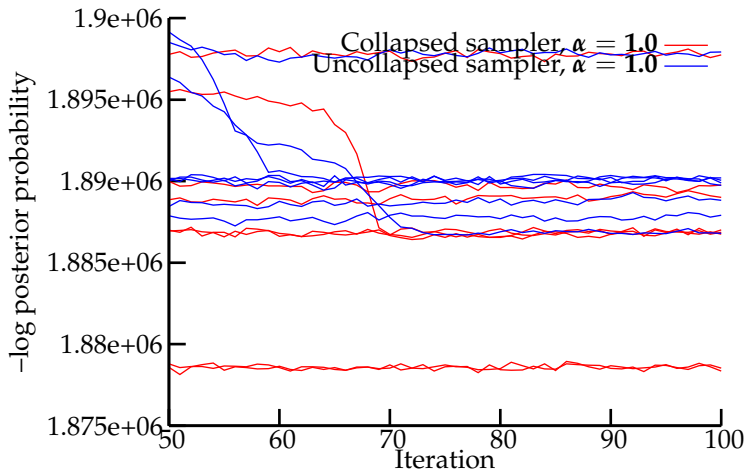
Number of Y_i that changed each iteration with $\ell = 10, \alpha = 1$



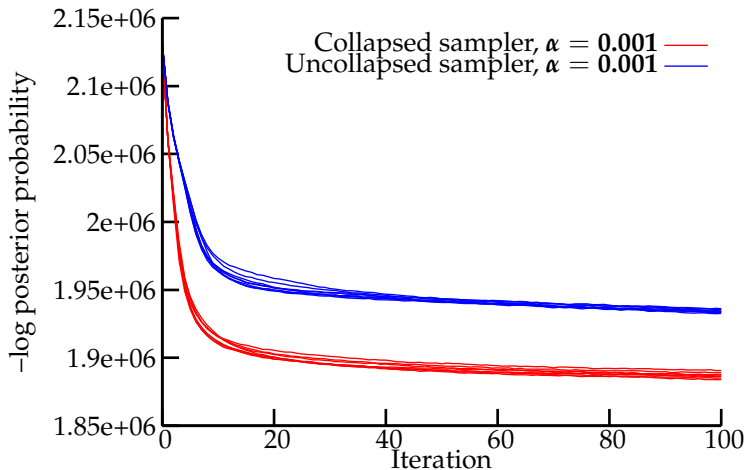
Posterior probability $P(Y, X|\alpha, \beta)$ with $\ell = 10, \alpha = 1$



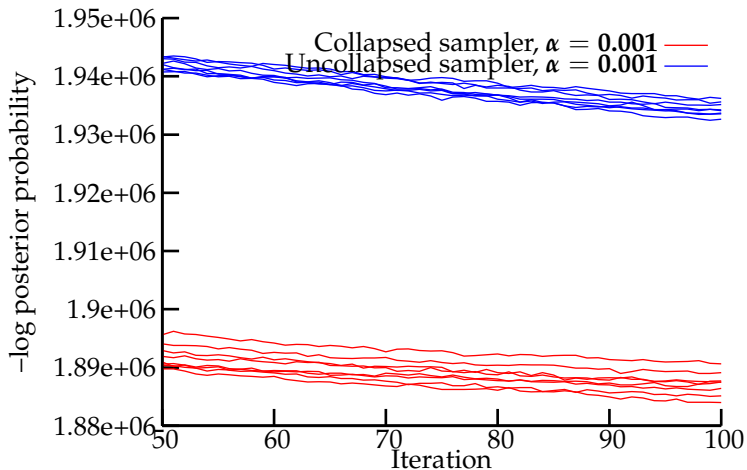
Posterior probability $P(Y, X|\alpha, \beta)$ with $\ell = 10, \alpha = 1$



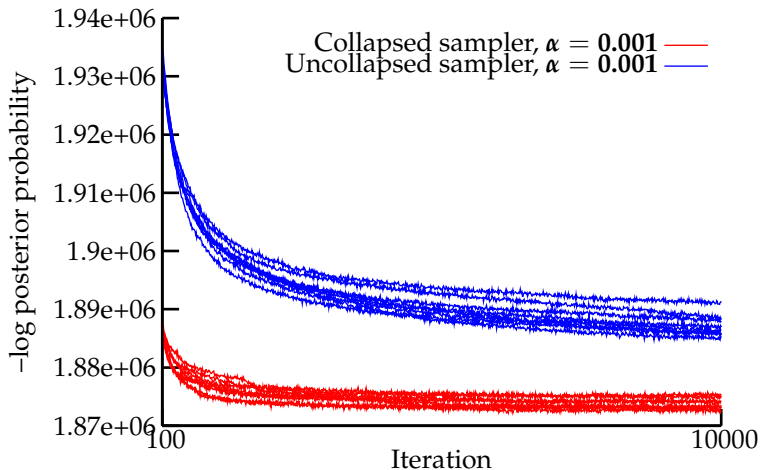
Posterior probability $P(Y, X|\alpha, \beta)$ with $\ell = 10, \alpha = 0.001$



Posterior probability $P(Y, X|\alpha, \beta)$ with $\ell = 10, \alpha = 0.001$



Posterior probability $P(Y, X|\alpha, \beta)$ with $\ell = 10, \alpha = 0.001$



Outline

Mixture models

Sampling with Markov Chains

The Gibbs sampler

Review of Dirichlet-Multinomial distributions

Gibbs sampling for Dirichlet-Multinomial mixtures

Topic modeling with Dirichlet multinomial mixtures

Summary

Summary

- Complex models often don't have analytic solutions
- Approximate inference can be used on many such models
- Monte Carlo Markov chain methods produce samples from (an approximation to) the posterior distribution
- Gibbs sampling is an MCMC procedure that resamples each variable conditioned on the values of the other variables
- If you can sample from the conditional distribution of each hidden variable in a Bayes net, you can use Gibbs sampling to sample from the joint posterior distribution
- We applied Gibbs sampling to Dirichlet-multinomial mixtures to cluster sentences