Bayesian Mixture Models and Admixtures

OVERVIEW

In the following

We will consider two variations:

- *Bayesian mixture models* (mixtures with priors).
- *Admixtures*, in which the generation of each observation (e.g. document) can be influenced by several components (e.g. topics).
- One particular admixture model, called *latent Dirichlet allocation*, is one of the most succesful machine learning models of the past ten years.

Inference: Sampling

These models are examples of models in which the exact posterior is intractable. Inference uses Markov chain Monte Carlo sampling, which will be our main topic for the last two lectures.

Bayesian Mixture Model

Recall: Finite mixture models

$$\pi(x) = \sum_{k=1}^{K} c_k p(x|\theta_k) = \int_{\mathcal{T}} p(x|\theta) m(\theta) d\theta \qquad \text{with} \qquad m := \sum_{k=1}^{K} c_k \delta_{\theta_k}$$

All parameters are summarized in the *mixing distribution m*.

Bayesian mixture model: Idea

In a Bayesian model, parameters are random variables. Here, that means a *random* mixing distribution:

$$M(\,.\,) = \sum_{k=1}^{K} C_k \delta_{\Theta_k}(\,.\,)$$

RANDOM MIXING DISTRIBUTION

How can we define a random distribution?

Since M is discrete with finitely many terms, we only have to generate the random variables C_k and Θ_k:

$$M(\,.\,) = \sum_{k=1}^{K} C_k \delta_{\Theta_k}(\,.\,)$$

More precisely

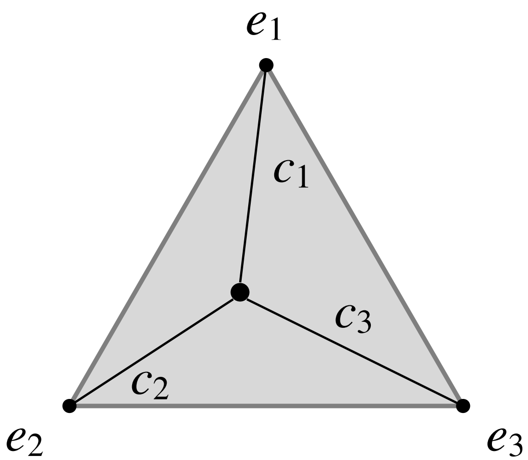
Specifically, the term BMM implies that all priors are natural conjugate priors. That is:

- The mixture components $p(x|\theta)$ are an exponential family model.
- The prior on each Θ_k is a natural conjugate prior of p.
- The prior of the vector $(C_1, \dots, C_K)$ is a Dirichlet distribution.

Explanation: Dirichlet distribution

- When we sample from a finite mixture, we choose a component k from a multinomial distribution with parameter vector $(c_1, \dots, c_k)$.
- The conjugate prior of the multinomial is the Dirichlet distribution.

THE DIRICHLET DISTRIBUTION



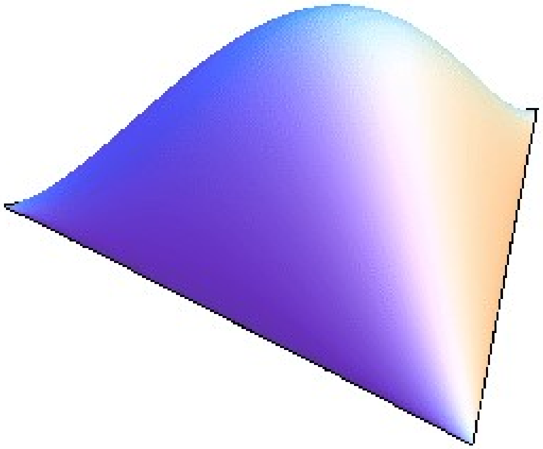
Recall: Probability simplex

The set of all probability distributions on K events is the *simplex*
$\triangle_K := \{(c_1,\ldots,c_k) \in \mathbb{R}^K \,|\, c_k \geq 0 \text{ and } \sum_k c_K = 1\}$.

Dirichlet distribution

The **Dirichlet distribution** is the distribution on $\triangle_K$ with density

$$q_{\text{Dirichlet}}(c_{1:K}|\alpha, g_{1:K}) := \frac{1}{K(\alpha, g_{1:K})} \exp\Big(\sum_{k=1}^{K} (\alpha g_k - 1)\log(c_k)\Big)$$



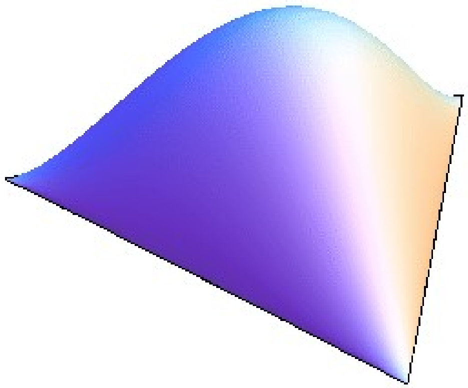
Parameters:

- $g_{1:K} \in \triangle_K$: Mean parameter, i.e. $\mathbb{E}[c_{1:K}] = g_{1:K}$.
- $\alpha \in \mathbb{R}_+$: Concentration.
  Larger $\alpha \to$ sharper concentration around $g_{1:K}$.

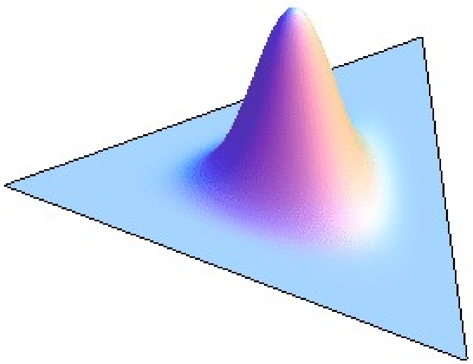
THE DIRICHLET DISTRIBUTION

In all plots, $g_{1:K} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Light colors = large density values.

Density plots

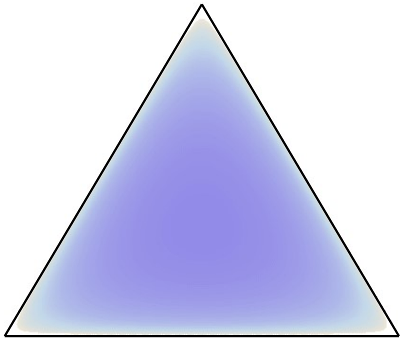


$\alpha = 1.8$

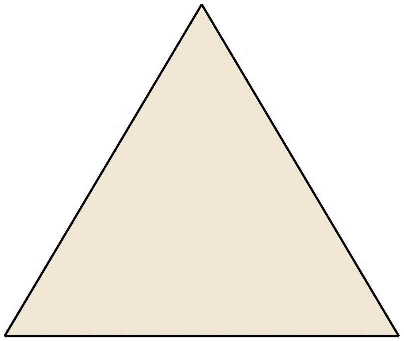


$\alpha = 10$

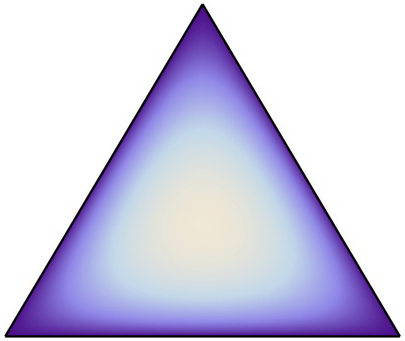
As heat maps



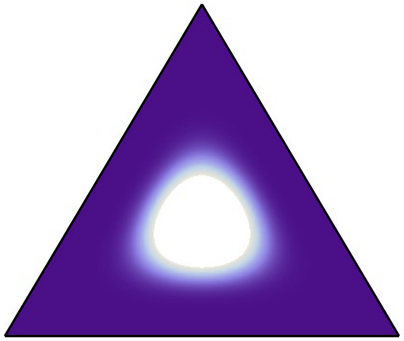
$\alpha = 0.8$
Large density values
at extreme points



$\alpha = 1$
Uniform distribution
on $\triangle_K$



$\alpha = 1.8$
Density peaks
around its mean



$\alpha = 10$
Peak sharpens
with increasing α

Multinomial-Dirichlet Model

Model

The Dirichlet is the natural conjugate prior on the multinomial parameters. If we observe h_k counts in category k, the posterior is

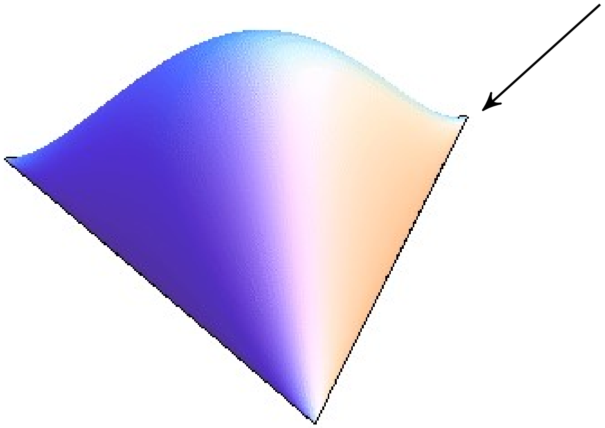
$$\Pi(c_{1:K}|h_1,\ldots,h_k) = q_{\text{Dirichlet}}(c_{1:K}|\alpha+n,(\alpha g_1+h_1,\ldots,\alpha g_K+h_K))$$

where $n = \sum_k h_k$ is the total number of observations.

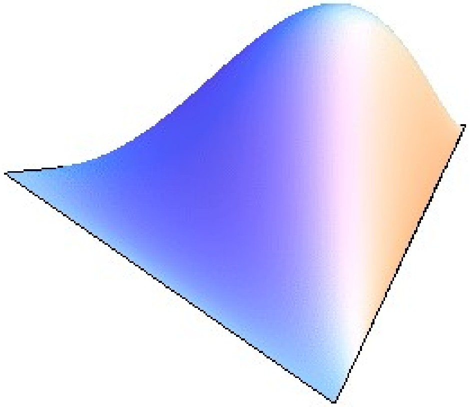
Illustration: One observation

Suppose $K = 3$ and we obtain a single observation in category 3.

This extreme point correponds to $k = 3$.



Prior: Mean at the center.



Posterior: Shifted mean, increased concentration.

Bayesian Mixture Models

Definition

A model of the form

$$\pi(x) = \sum_{k=1}^{K} c_k p(x|\theta_k) = \int_{\mathcal{T}} p(x|\theta) M(\theta) d\theta$$

is called a **Bayesian mixture model** if $p(x|\theta)$ is an exponential family model and M a random mixing distribution, where:

- $\Theta_1, \dots, \Theta_k \sim_{\text{iid}} q(\,.\,|\lambda, y)$, where q is a natural conjugate prior for p.
- $(C_1, \dots, C_K)$ is sampled from a K-dimensional Dirichlet distribution.

Bayesian Mixture: Inference

Posterior distribution

The posterior of a BMM under observations $x_1, \ldots, x_n$ is (up to normalization):

$$\Pi(c_{1:K}, \theta_{1:K} | x_{1:n}) \propto \prod_{i=1}^{n} \Big(\sum_{k=1}^{K} c_k p(x_i | \theta_k) \Big) \Big(\prod_{k=1}^{K} q(\theta_k | \lambda, y) \Big) q_{\text{Dirichlet}}(c_{1:K})$$

The posterior is analytically intractable

- Thanks to conjugacy, we *can* evaluate each term of the posterior.
- However: Due to the $\prod_{k=1}^{K} \big(\sum_{i=1}^{n} \ldots \big)$ bit, the posterior has K^n terms!
- Even for 10 clusters and 100 observations, that is impossible to compute.

Solution

The posterior can be sampled with a very simple MCMC sampler (which looks strikingly similar to an EM algorithm). We will discuss sampling algorithms in the next lecture.

Text Models

Recall: Multinomial text clustering

We assume the corpus is generated by a multinomial mixture model of the form

$$\pi(\mathbf{H}) = \sum_{k=1}^{K} c_k P(\mathbf{H}|\theta_k) \ ,$$

where $P(\mathbf{H}|\theta_k)$ is multionmial.

- A document is represented by a histogram $\mathbf{H}$.
- Topics $\theta_1, \dots, \theta_K$.
- $\theta_{kj} = \Pr\{ \text{ word } j \text{ in topic } k\}$.

Problem

Each document is generated by a single topic; that is a very restrictive assumption.

Sampling Documents

Parameters

Suppose we consider a corpus with K topics and a vocubulary of d words.

- $\phi \in \triangle_K$ topic proportions ($\phi_k = \Pr\{$ topic $k\}$).
- $\theta_1, \dots, \theta_K \in \triangle_d$ topic parameter vectors ($\theta_{kj} = \Pr\{$ word j in topic $k\}$).

Note: For random generation of documents, we assume that ϕ and the topic parameters θ_k are given (they properties of the corpus). To train the model, they have to be learned from data.

Model 1: Multinomial mixture

To sample a document containing M words:

1. Sample topic $k \sim \text{Multinomial}(\phi)$.
2. For $i = 1, \dots, M$: Sample word$_i \sim \text{Multinomial}(\theta_k)$.

The entire document is sample from topic k.

Latent Dirichlet Allocation

Mixtures of topics

Whether we sample words or entire documents makes a big difference.

- When we sample from the multinomial mixture, we choose a topic at random, then sample the *entire* document from that topic.
- For several topics to be represented in the document, we have to sample each word individually (i.e. choose a new topic for each word).
- Problem: If we do that in the mixture above, every document has the same topic proportions.

Model 2: Admixture model

Each document explained as a *mixture* of topics, with mixture weights $c_{1:K}$.

1. Sample topic proportions $c_{1:K} \sim \text{Dirichlet}(\phi)$.
2. For $i = 1, \ldots, M$:
   2.1 Sample topic for word i as $k_i \sim \text{Multinomial}(c_{1:K})$.
   2.2 Sample $\text{word}_i \sim \text{Multinomial}(\theta_k)$.

This model is known as **Latent Dirichlet Allocation** (LDA).

Comparison: LDA and BMM

Observation

LDA is *almost* a Bayesian mixture model: Both use multinomial components and a Dirichlet prior on the mixture weights. However, they are not identical.

Comparison

Bayesian MM	Admixture (LDA)
Sample $c_{1:K} \sim \text{Dirichlet}(\phi)$. Sample topic $k \sim \text{Multinomial}(c_{1:K})$. For $i = 1, \dots, M$: Sample word$_i \sim \text{Multinomial}(\theta_k)$.	Sample $c_{1:K} \sim \text{Dirichlet}(\phi)$. For $i = 1, \dots, M$: Sample topic $k_i \sim \text{Multinomial}(c_{1:K})$. Sample word$_i \sim \text{Multinomial}(\theta_{k_i})$.

In admixtures:

- $c_{1:K}$ is generated at random, *once for each document*.
- Each word is sampled from its own topic.

What do we learn in LDA?

LDA explains each document by a separate parameter $c_{1:K} \in \triangle_K$. That is, LDA models documents as *topic proportions*.

Example: Mixture of Topics

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

From Blei, Ng, Jordan, "Latent Dirichlet Allocation", 2003