COMP 330: Mixture Models

Chris Jermaine and Kia Teymourian

Rice University

Mixture Model Intro

At highest level:

- Have a set of data
- And a set of random variables
- Don't know which one prodcued which point

This is a mixture model!

In one line:

- MM: "hierarchical," stochastic, latent variable model

Why Use Them?

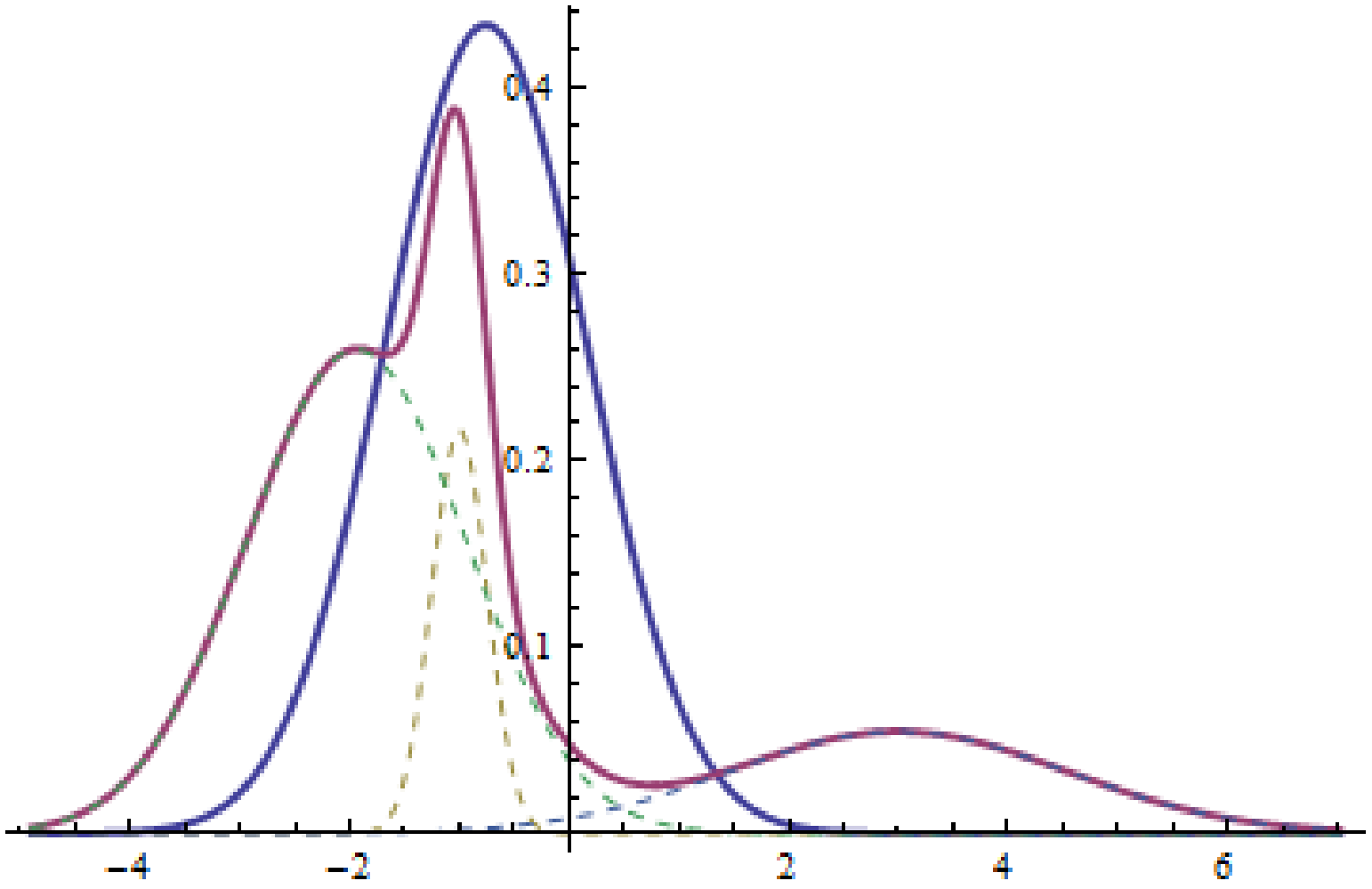
Sometimes, we want to segment the data

- Obsever a set of test scores
- Want 3 types of students: good, average, bad
- Associate each with a different Normal

Sometimes, we just want a very flexible model

- A mixture can give a complicated, muti-modal distribution

GMM Example



Mixture Model Intro

Used to produce a set of data $x_1, x_2, ..., x_n$

- And a set of hidden (latent) indictors $c_1, c_2, ..., c_n$

MM begins with a distribution fuction f

- Common f: Gaussian, Multinomial, Gamma, etc.
- We have k sets of parameters for f: $\theta_1, \theta_2, ..., \theta_k$
- And a probability vector π that tells us how important each component is

Pseudo-code to produce n observations is:

for i = 1 to n **do:**
  $c_i \sim \text{Categorical}(\pi)$
  $x_i \sim f(\theta_{c_i})$

Mixture Model PDF

In general, PDF is:

$$P(x_1, x_2, ..., x_n) = \prod_i \left(\sum_j \pi_j f(x_i|\theta_j) \right)$$

Why?

Learning

Would be easy if we knew $c_1, c_2, ..., c_n$

But we don't!

Standard methods:

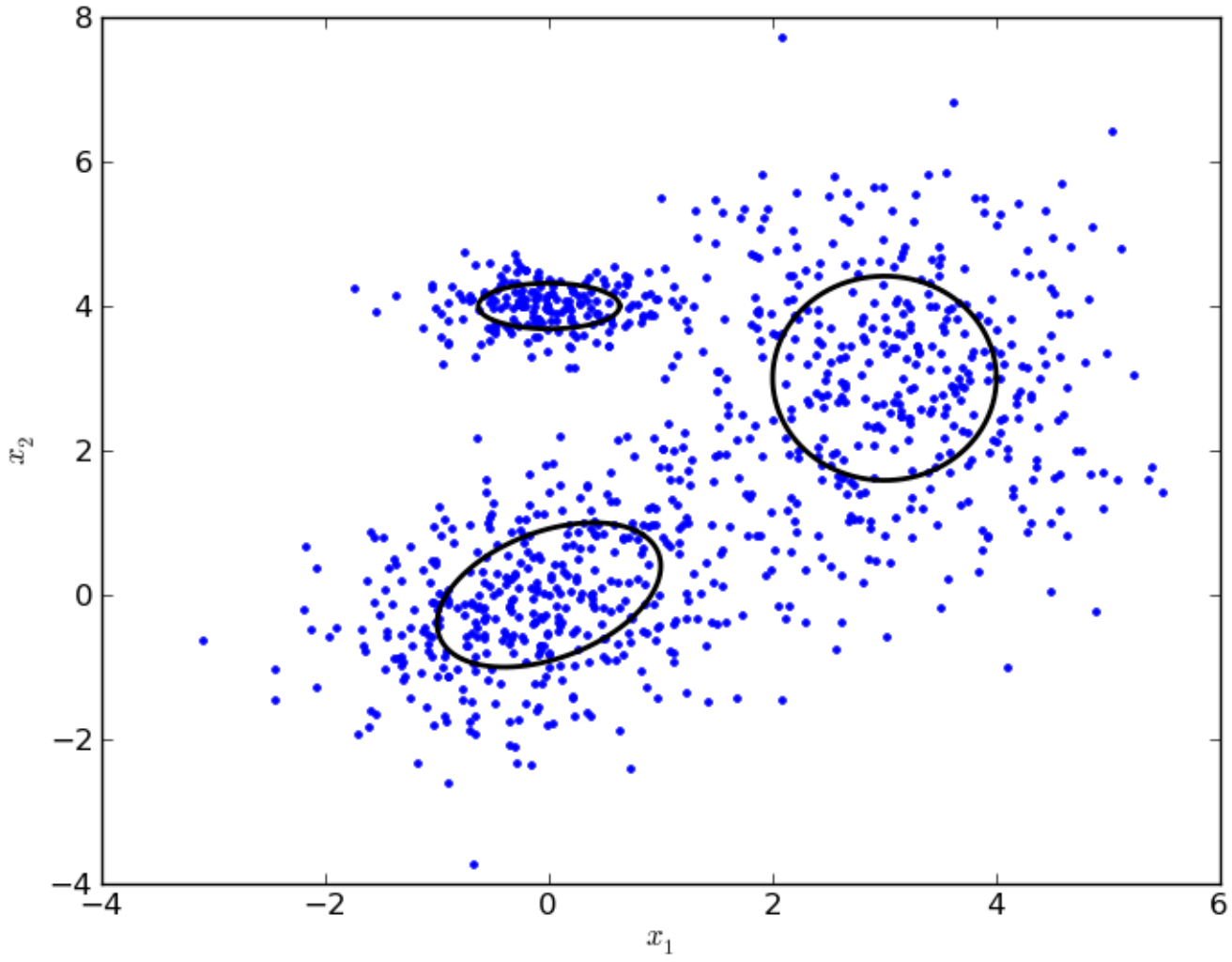
- MCMC (sample $c_1, c_2, ..., c_n$)
- EM

Some Example Mixture Models...

The Patriarch: The GMM

$$P(x_1, x_2, ..., x_n) = \prod_i \left(\sum_j \pi_j \text{Normal}(x_i | \mu_j, \sigma_j^2) \right)$$

GMM



Miture of Dirichlets

Imagine that I have a large corpus of text documents

And I want to understand the various types of documents present

View TF vector as being produced as a sample from Dirichlet distribution

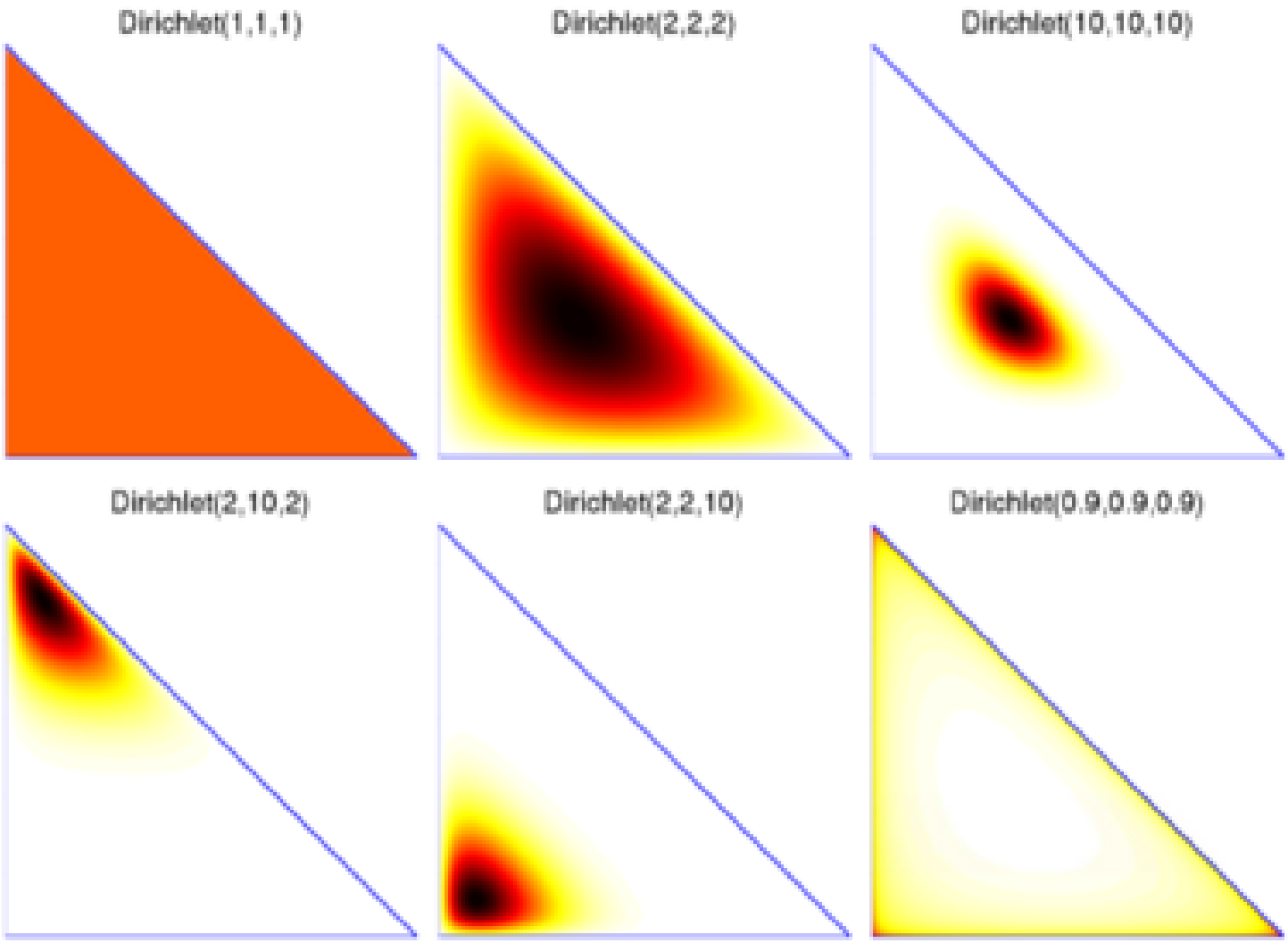
What's a Dirichlet?

Dirichlet Distribution

Generalization of Beta to many dimensions

- Takes params $v_1, v_2, \ldots, v_m$
- Produces m values that sum to 1
- To generate, sample from m Gammas, each with shape v_j
- ith value is ith Gamma's fracitonal contribution to total
- Expected value of ith is $v_j / \sum_{j'} v_{j'}$

Dirichlet Distribution



Miture of Dirichlets

If param vector for jth Dirichlet is α_j, then PDF is:

$$P(x_1, x_2, ..., x_n) = \prod_i \left(\sum_j \pi_j \text{Dirichlet}(x_i | \alpha_j) \right)$$

Example:

- Take all the papers I've written, all Luay has written
- Dictionary has words: $\langle$ biology, databases, phylogenetics, statistics $\rangle$
- Might get $\alpha_{\text{Chris}} = \langle .5, 12, 1.2, 8 \rangle$
- Might get $\alpha_{\text{Luay}} = \langle 11, 2, 18, 3.2 \rangle$

Mixture of Experts

Used for regresion/classification

- In "classical" GLMs
- Use dot product $x \cdot r = \sum_j x_j \times r_j$ to get "natural" parameter for error dist
- Ex: Normal (least squares) regression:

$$f(y|x) = \text{Normal}(y|x \cdot r, \sigma^2)$$

Mixture of Experts

Used for regresion/classification

- In "classical" GLMs
- Use dot product $x \cdot r = \sum_j x_j \times r_j$ to get "natural" parameter for error dist
- Ex: Normal (least squares) regression:

$$f(y|x) = \text{Normal}(y|x \cdot r, \sigma^2)$$

Now, let's have a mixture of these

- Instead of r, we have $r_1, r_2, \ldots, r_k$
- Instead of σ^2, we have $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$

And then we use a "softmax gating network" to get π... that is:

$$\pi_j = \frac{\exp(x \cdot \eta_j)}{\sum_{j'} \exp(x \cdot \eta_{j'})}$$

Mixture of Experts

So PDF is:

$$P(x_1, x_2, ..., x_n) = \prod_i \left(\sum_j \frac{\exp(x_i \cdot \eta_j)}{\sum'_j \exp(x_i \cdot \eta_{j'})} \text{Normal}(x_i | x_i \cdot r_j, \sigma_j^2) \right)$$

General Question: How to Choose Mixture Size

Guess!

Information theoretic methods

Bayesian methods

Questions?