Mixture Models

David M. Blei

COS424
Princeton University

March 2, 2012

Unsupervised learning

- Unsupervised learning is about taking data and finding structure in it.
- It is about finding patterns without knowing what we are looking for.
- In the old days, classification and prediction were the most important problems in statistics and machine learning.
- They still are important, but unsupervised learning now plays an equal role. Why?
- Because of the “information explosion”, “data innondation problem,” whatever you want to call it.

Examples

- Attorneys used to get fileboxes of documents to read; Now they get gigabytes of emails. The trial is tomorrow. What is in these emails?
- We regularly search an indexed collection of documents. Queries might mean multiple things. Consider “Jaguar” as (a) an amimal (b) a car and (c) an operating system. Given search results, can we identify these groups?
- Biologists run microarray experiments, simultaneously testing many genes with many conditions. The idea is to uncover genes that behave similarly. How can they do it?
- Neuroscientists run fMRI studies resulting in thousands of high-resolution images of brain data, in a time series, while subjects are performing cognitive tasks. One goal is to find patterns in the brain.

More examples

- Historians collect reams of historical documents, scan them, and run them through OCR software. How can unsupervised learning help them with close reading and forming hypotheses?
- A reporter receives 5M emails from WikiLeaks. Where is the scoop?
- A physicist collects terrabytes of measurements from the universe. What happened that we didn't expect? What should we examine?
- Others?

Seeing observations in a different way

- One way we use unsupervised learning is to help us see observations in a different way. When observations happen at large scale, we need methods for summarizing them and visualizing them. In this sense, it's like a microscope or, more accurately, a camera filter/lense.
- Main idea: Posit the kind of structure that you think exists in your data set, and then to let the algorithm find the particular instantiation of that structure.
- For example:
  - Latent patterns of brain activity that correspond to cognitive variables
  - Groups of documents that are about a single subject
  - Collections of genes that have the same function
- We think such patterns exit, but don't know what they are.
- A lot of machine learning research is about taking a new type of data set, positing reasonable structures, and showing that the resulting unsupervised learning algorithm does something that "makes sense."

Dimension reduction

- Mathematically, we can think of unsupervised learning as “dimension reduction,” taking high dimensional data (consider, fMRI) and turning it into low dimensional data (the weight of one of 50 types of brain patterns coupled with what those pattens are).
- This has several effects.
- It smooths idiosyncracies in individual data points, e.g., “cat” and “feline” might mean the same thing in a reduced representation.
  (This also helps in prediction.)
- It summarizes the data in terms of patterns and how each data point expresses them.
- It compresses the data.

Difficulty with unsupervised learning

- Problem with unsupervised learning: It's hard to measure success.
- Its vague and fuzzy, but that doesn't mean it's not important.
- (Cf visualization, HCI)

Next few lectures

- K-means
- Mixture models
- Expectation-Maximization

(See other slides for K-means)

Hidden variable models

- **Hidden random variables** are imaginary variables that give structure to the distribution of the data.
- For example, in a probability model for clustering—which is called a **mixture model**—the hidden random variables code the cluster assignments of the data.
- When we fit a model with hidden variables, the parameters reveal how the hidden structure that we imagined is manifest in the data we observed.

Hidden variables are powerful

- Hidden variable models give us a lot of power and room for creativity in modeling data.
- We can imagine the structure that we think exists—e.g., groups of documents, a hierarchy of concepts, a binary vector of active “features”, communities of people that know each other in the same way.
- We can then derive an algorithm (next slide) to discover that structure from observed data.
- (In practice, there are many computational and conceptual problems.)

Expectation-Maximization

- Repeat: When we fit a model with hidden variables, the parameters reveal how the hidden structure is manifest in the obsverations.
- But, as we'll see, computing the MLE is hard. That's why we need the **expectation-maximization** algorithm
- The EM algorithm is a general algorithm for finding the MLE in a model with hidden random variables.
- It was "invented" in 1977 by Dempster et al. I have heard rumors that it was known before then, in the forties as secret WWII defense research.

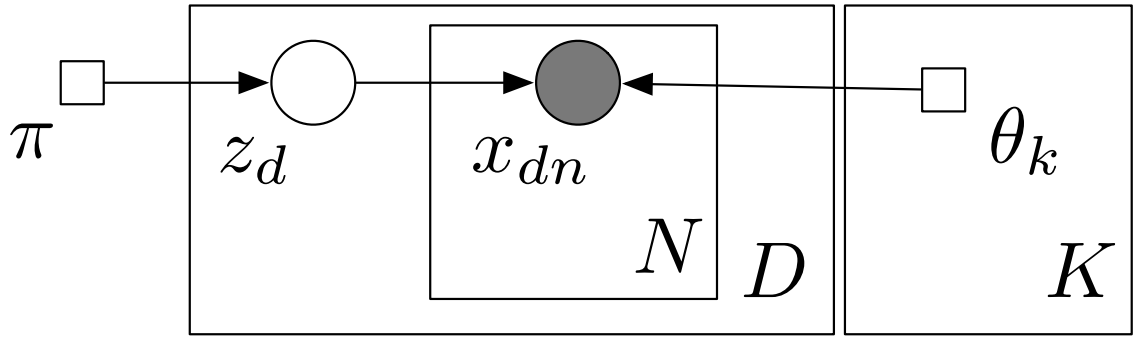
Mixture models

- The first hidden variable model we will consider is the mixture of multinomials.
- Recall how model-based classification let us plug in simpler IID models into a classification algorithm.
- Mixture models mirror this idea, but for clustering problem.
- The difference is that the labels (now called cluster assignments) are *never* observed. There is no training data.

Clustering documents

- Problem: You have 10M emails. You want to group them into 25 groups so you can better see what they contain.
  - You want to describe each group
  - You want to identify each document with a group
- The model will assume that the words of each document were drawn independently from one of 25 multinomial distributions.
- How is this similar to and different from k-means?
- Some reasons—
  - The clustering problem is similar.
  - k-means is defined for continuous data. Here, data are discrete and there is no explicit distance function.
  - k-means does not reference a probability model
- How does this compare to the text classification model?

Mixture of multinomials



- The data are D documents

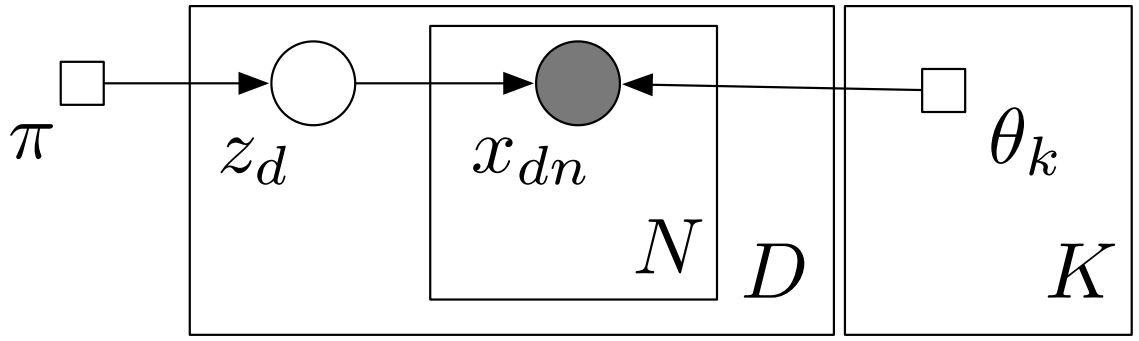
$$\mathscr{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_D\}, \tag{1}$$

where each document is a sequence of N words

$$\mathbf{x}_d = \{x_{d,1}, \ldots, x_{d,N}\}. \tag{2}$$

- Recall we represent discrete data with an indicator vector.
- Note: This is the **only** variable observed in the model.

Joint distribution



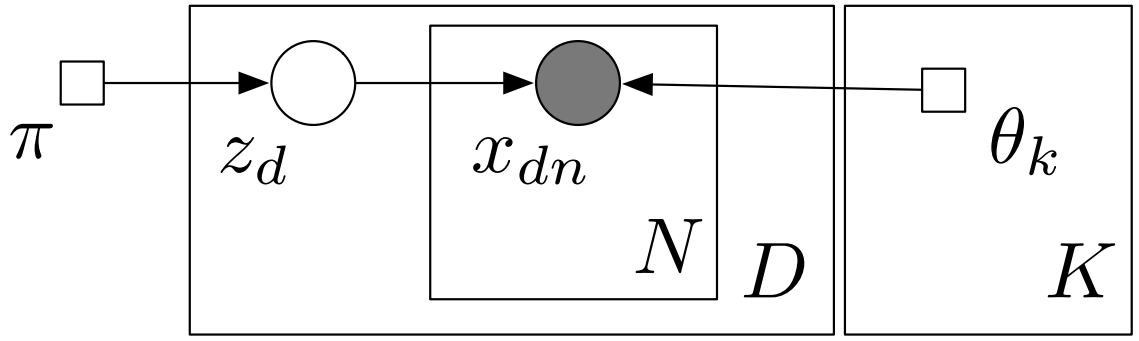
- The joint distribution of a single document and cluster assignment is

$$p(c, \mathbf{x}) = p(c|\pi) \prod_{n=1}^{N} p(x_n|\theta_{1:K}, c). \tag{3}$$

- Recall the "selection mechanism" for getting the right cluster.
- Recall that words drawn independently from the same distribution amount to a product over the vocabulary terms exponentiated by the counts

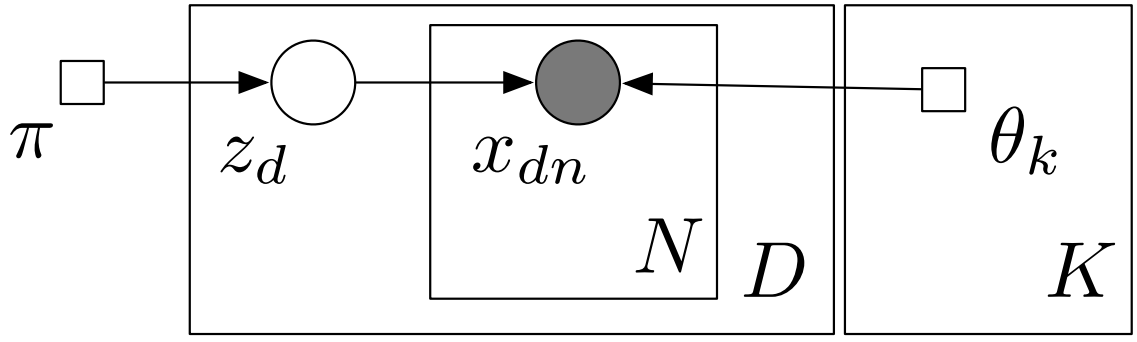
$$\prod_{n=1}^{N} p(x_n|\theta_c) = \prod_{v=1}^{V} \theta_{c,v}^{n_v(\mathbf{x})} \tag{4}$$

Generative process



- It is convenient to think of models (all models) by their **generative process**, the imaginary process by which the data are assumed to have been drawn.
- For the *d*th document:
  - Draw $z_d \sim \pi$
  - For the *n*th word in the *d*th document:
    - Draw $w_{d,n} \sim \theta_{z_d}$

Generative process

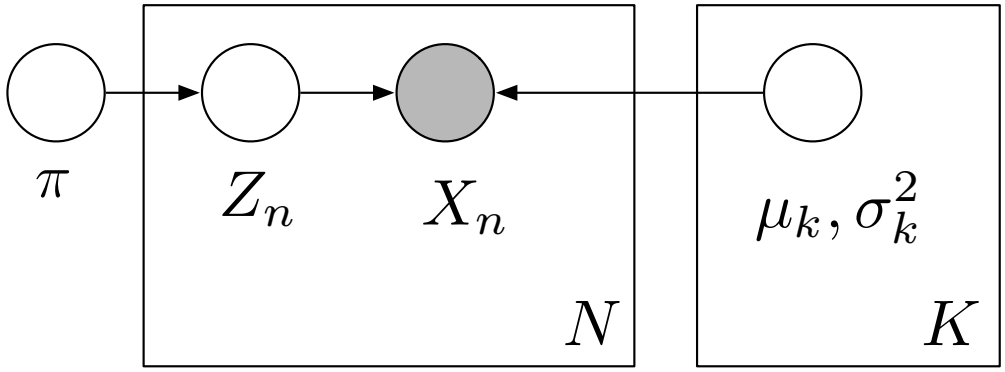


- This is a third way—in addition to the joint and the graphical model—to articulate some of the dependencies that we are assuming.
- Parameter estimation and inference can be thought of as "reversing" the generative process. What is the hidden structure, distributions and assignments, that best explain the observed documents?
- The generative process also provides a program for generating data from the model—this is useful for many reasons.

Mixtures of *

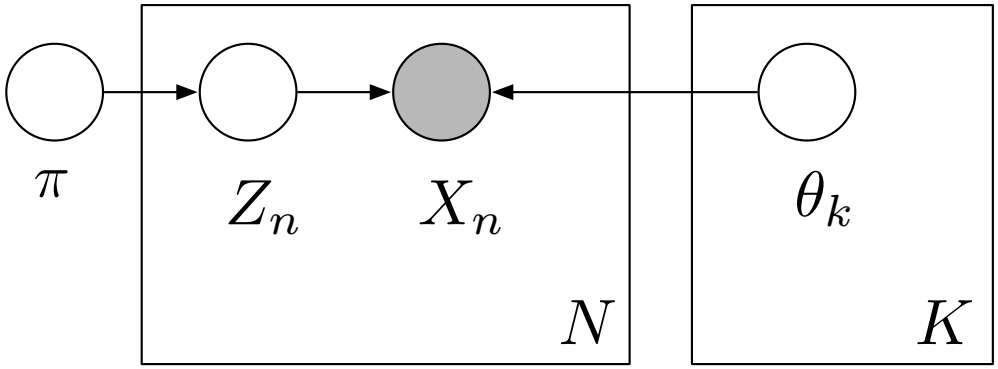
- Before discussing how to fit, I want to show examples of other mixtures.
- By replacing the simple independent words model with another model (of other data), we obtain different kinds of mixture models.
- For example, a mixture of Gaussians replaces the multinomial observation with continuous observation.
- We can even define mixtures of complicated hidden variable models—but we won't do that yet.
- Mixture models are a natural way to build a clustering model out of an existing probabilistic model.

Mixtures of educational data (Schnipke and Scrams 1997)



- Data are $\{x_n\}$, the time to respond to a GRE question
- Model this data with a mixture of two log-normal distributions
- There are different kinds of behaviors; speedy behavior and careful behavior. They fit a mixture model and find this to be true.
- High level point: Interesting behavior can be drawn from exploratory analysis of the shape of distributions, rather than summary statistics.

Mixtures of survival data (Farewell, 1982)

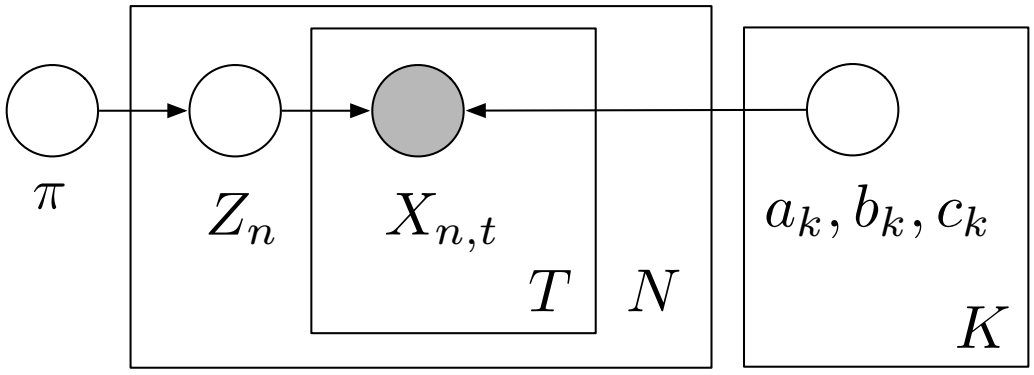


- Fit a mixture model to the survival times of animals. Data are $\mathscr{D} = \{x_n\}$, where x_n is the number of weeks for which a mouse survived.
- Assumes different populations in the data, independent of the experimental condition. For example, some animals are “long term survivors”; others are affected by “experimental stresses”
- Idea: the previously developed simple parametric model is not appropriate, and can skew the inferences. Populations are more heterogenous.

Mixtures of survival data (Farewell, 1982)

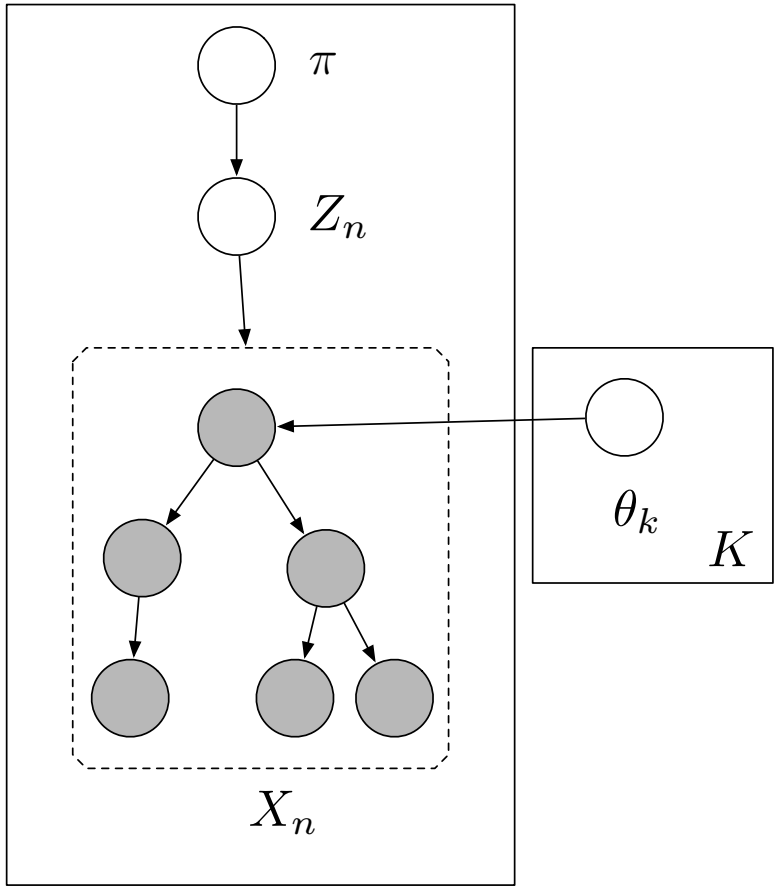
"For the toxicological experiment discussed by Pierce et al. (1979), mixture models postulating a subpopulation of long-term survivors are appealing from both the biological and statistical viewpoints. The use of such models should be restricted, however, to problems in which there is strong scientific evidence for the possibility that the individuals come from two or more separate populations. Otherwise, the modelling assumptions are too strong for widespread use. If two populations are assumed, then inferences will be made about the two populations whether they exist or not."

Mixtures of financial data (Liesenfeld, 1998)



- Two-component mixture model of stock price closing
- Use a complicated known model that's good for modeling stocks, and turn it into a mixture.
- Tests indicate that the mixture model is better in some respects, but not better in others.
- Good example of taking well-worn parametric models and using them in a mixture.

Mixtures of genetic data (Pagel and Meader, 2004)

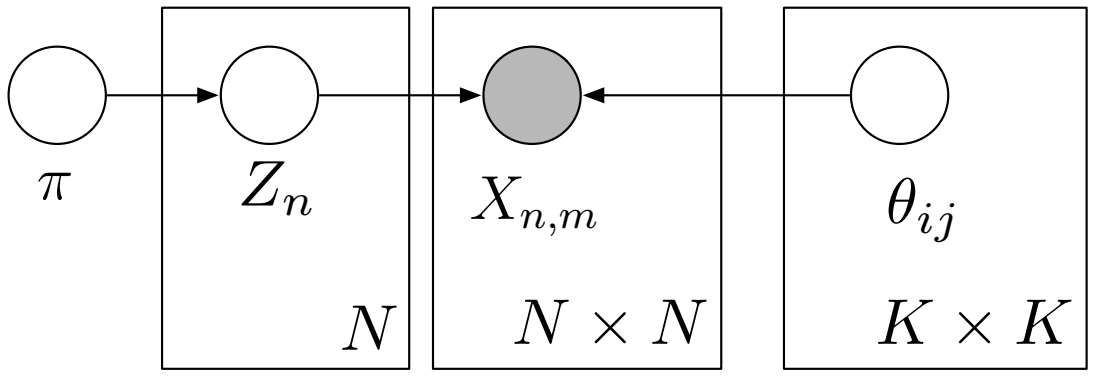


- Considered a mixture model over the rate of mutation of different places on the genome.
- Data are $\{x_{i,n}\}$, where $x_{i,n}$ is the DNA value at site n in the position i on the evolutionary tree.

Mixtures of genetic data (Pagel and Meader, 2004)

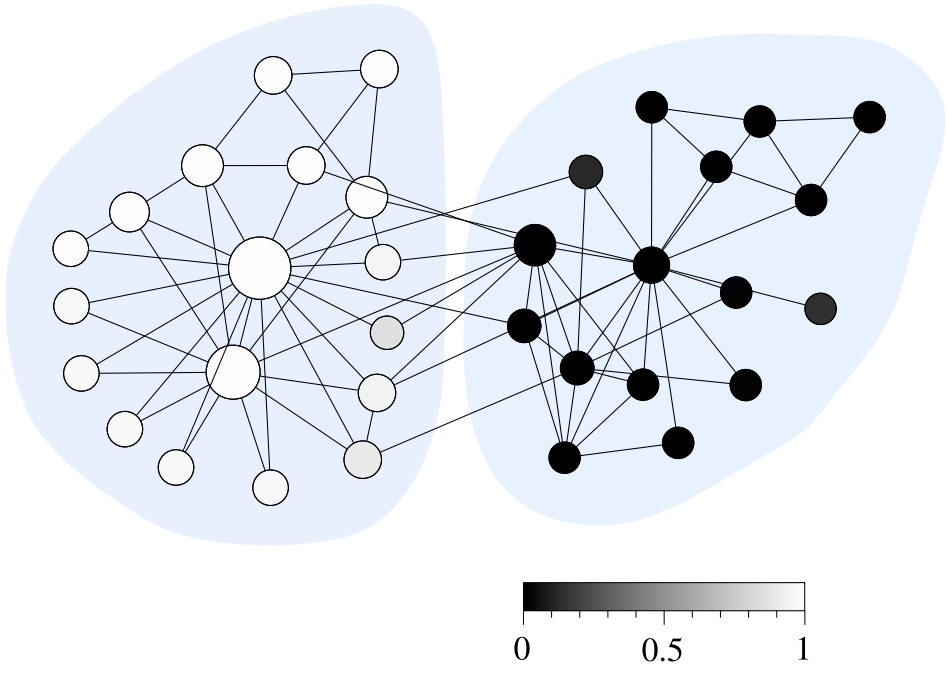
- The authors observe that the mixture model fits the data better. They use a number of statistics to determine this, as well as empirical evaluations.
- Conclusion: “The results we have reported for the pattern-heterogeneity mixture model send the encouraging message that phylogenetically structured data harbor complex signals of the history of evolution, and that it is possible to design general models to detect those signals.”

Mixtures of social networks (Newman, 2007)



- Data are $\mathscr{D} = \{x_{nm}\}$, where $x_{nm} = 1$ if there is a connection between actor n and actor m.
- Parameters are θ, a $K \times K$ matrix of probabilities. The element θ_{ij} is the probability that a pair in group i and j are connected.

Example inference



- Friendships in a Karate school
- Two groups split. This is “ground truth” (shaded regions)
- Mixture of two components; these are node colorings

Fitting a mixture model

(Now let's go to the blackboard.)