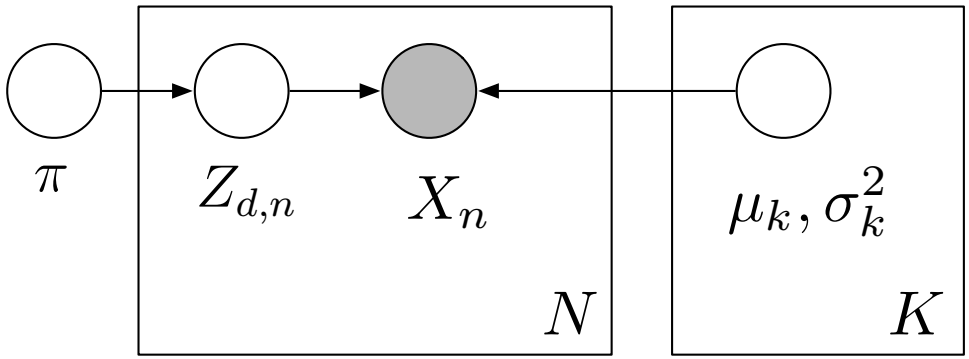
Mixture modeling examples

David M. Blei

COS424
Princeton University

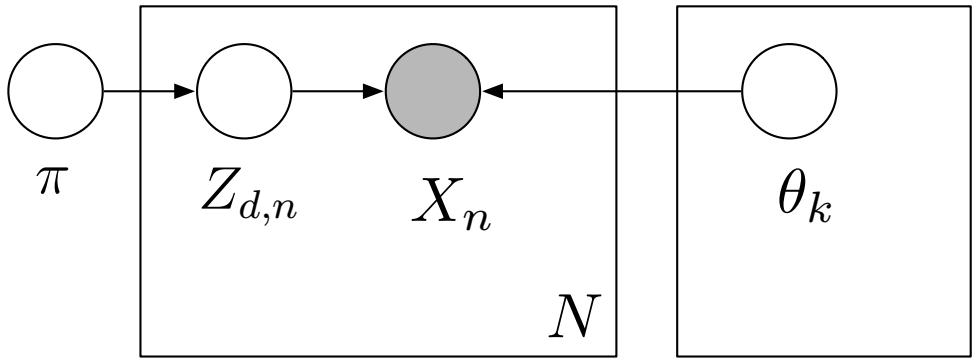
March 11, 2008

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)

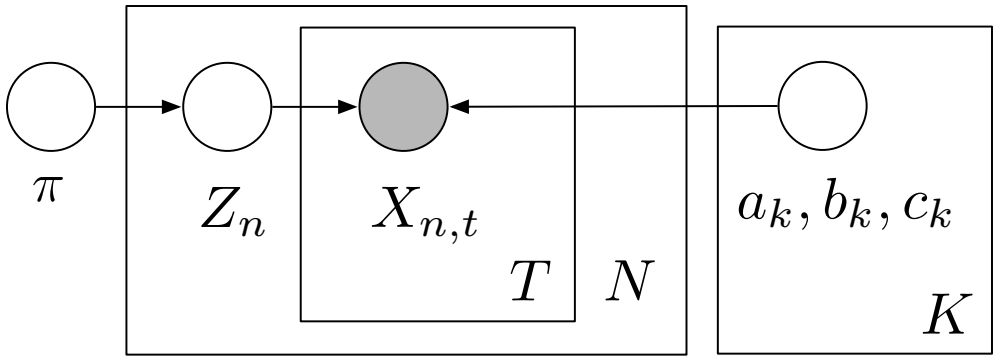


- Fit a mixture model to the survival times of animals. Data are $\mathcal{D} = \{x_n\}$, where x_n is the number of weeks for which a mouse survived.
- Assumes different populations in the data, independent of the experimental condition. For example, some animals are "long term survivors"; others are 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)

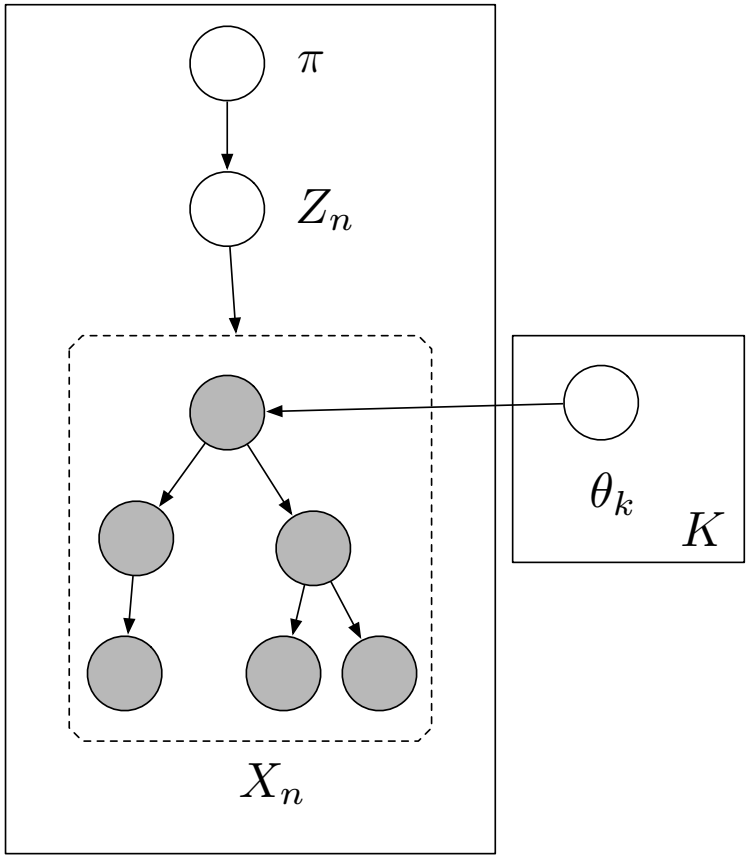
"For the toxicological experiment discussed by Pierce et al. (1979), mixture models postulating a subpopulation of long-term survivors are appealing from both the biological and statistical viewpoints. The use of such models should be restricted, however, to problems in which there is strong scientific 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)

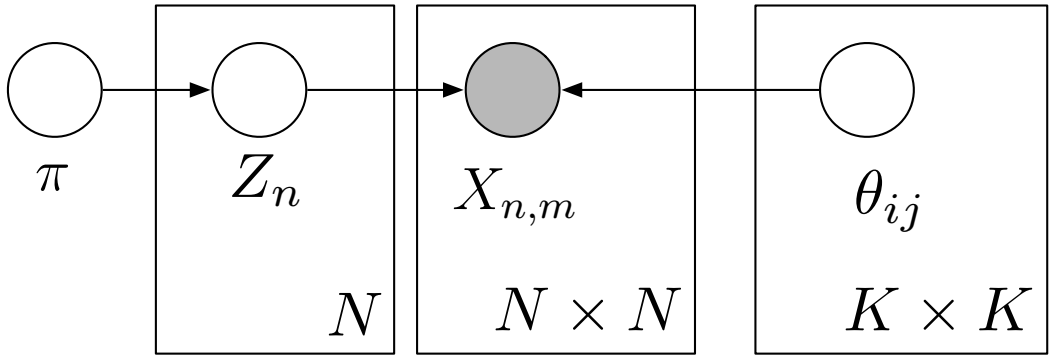


- Considered a mixture model over the rate of mutation of different places on the genome.
- Data are $\{x_{i,n}\}$, where $x_{i,n}$ is the DNA value at site n in the position i on the evolutionary tree.

Mixtures of genetic data (Pagel and Meader, 2004)

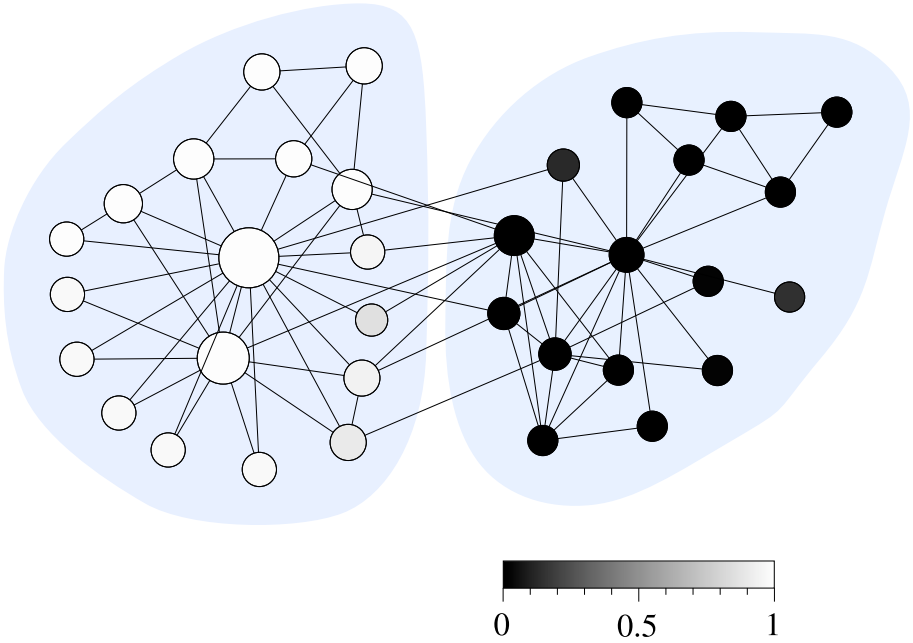
- The authors observe that the mixture model fits the data better. They use a number of statistics to determine this, as well as empirical evaluations.
- Conclusion: "The results we have reported for the pattern-heterogeneity mixture model send the encouraging message that phylogenetically structured 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 $\mathcal{D} = \{x_{nm}\}$, 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

Mixture models are a natural way to build a clustering model out of an existing probabilistic model.