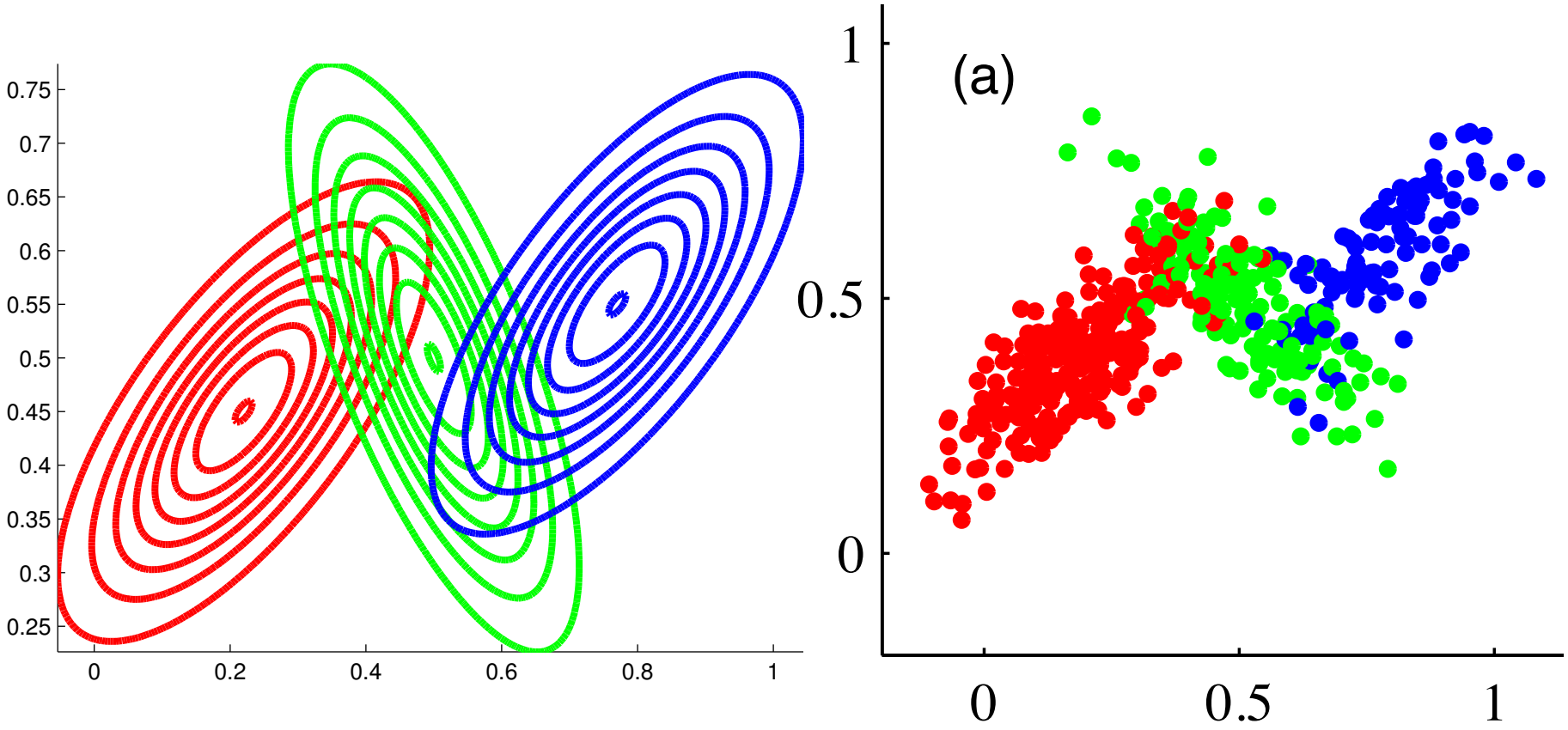
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 24:
Topic Models & Latent Dirichlet Allocation
Monte Carlo Methods

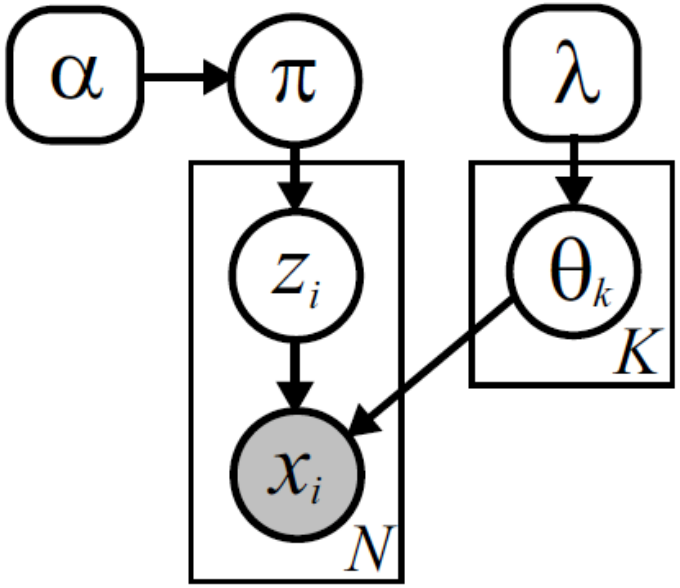
Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

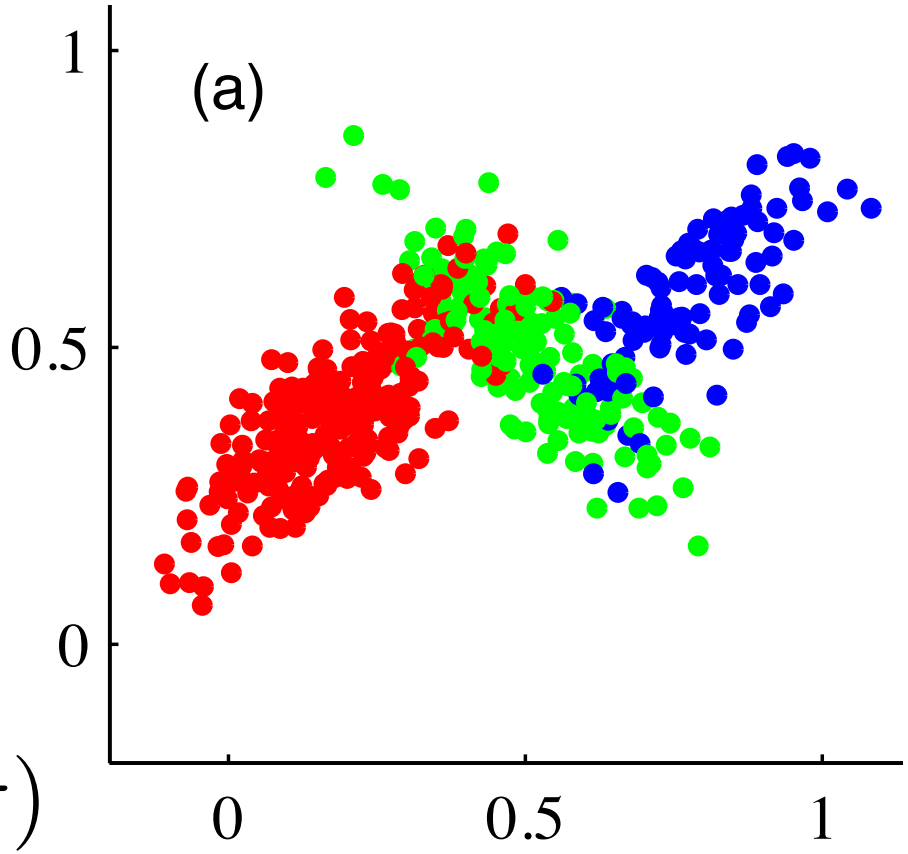
Gaussian Mixture Models



Mixture models describe a single, "flat" dataset.

Probabilistic Mixture Models





$$\theta_k = \{\mu_k, \Sigma_k\}$$

$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

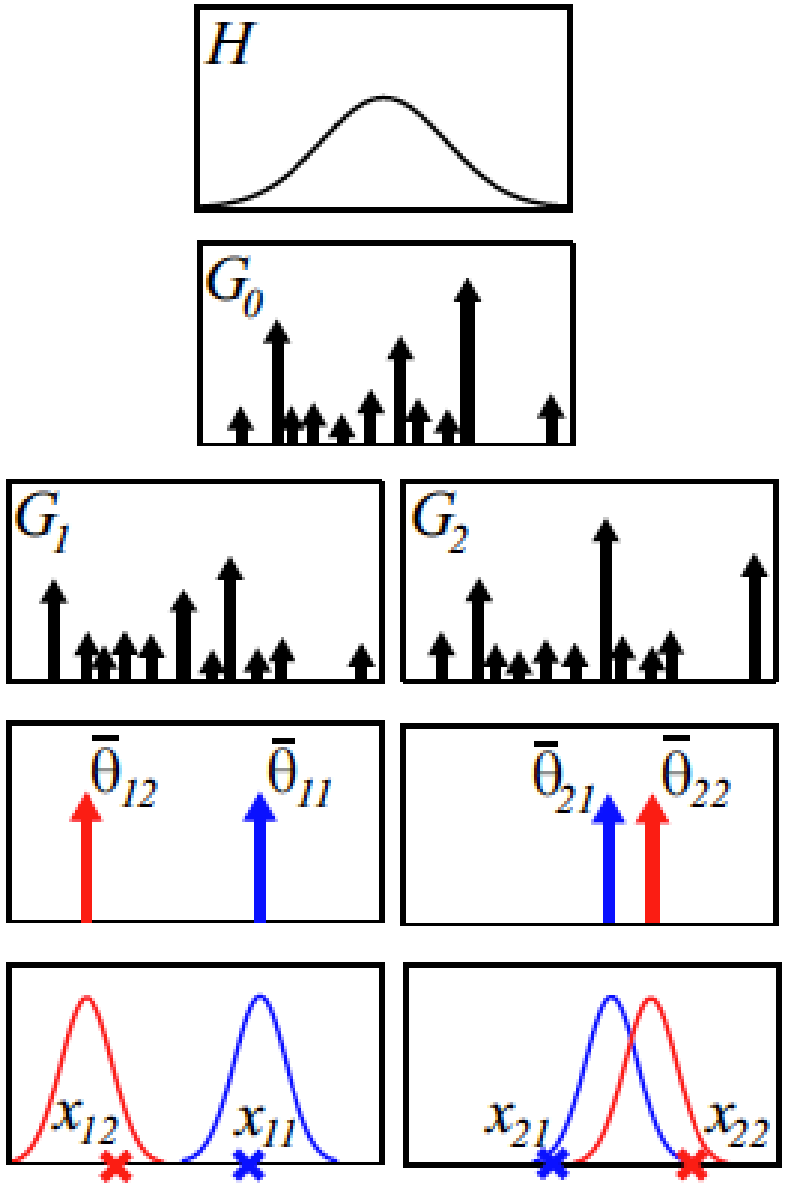
$$p(x_i \mid z_i, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Mixture models describe a single, "flat" dataset.

Collections of Mixture Models

- Many applications involve multiple “groups” of data
  - Multiple documents in a text corpus
  - Multiple images in a photo repository
  - Multiple users with their own spam filtering decisions
  - Multiple hospitals in a clinical trial
  - Multiple companies in a financial market
- How can we jointly model this data?
- Lumping into single large dataset ignores group differences
- Modeling groups independently can be ineffective, especially when limited data about any one group
- Hierarchical Bayesian models share between groups

Multiple Gaussian Mixtures



J=2 Groups of Data

Use data to learn set of shared mixture identities, and their frequencies across groups

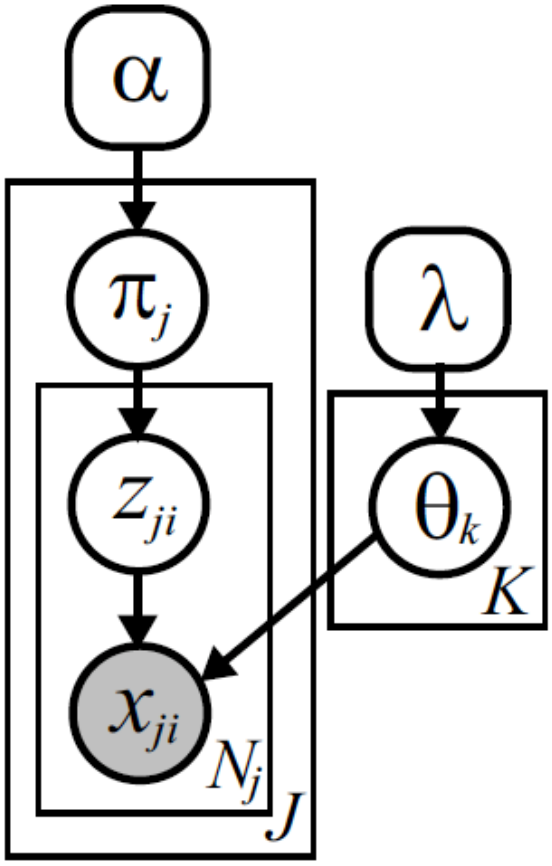
Each group has its own weight on shared mixture parameters:

$$\pi_j \sim \mathrm{Dir}(\alpha)$$

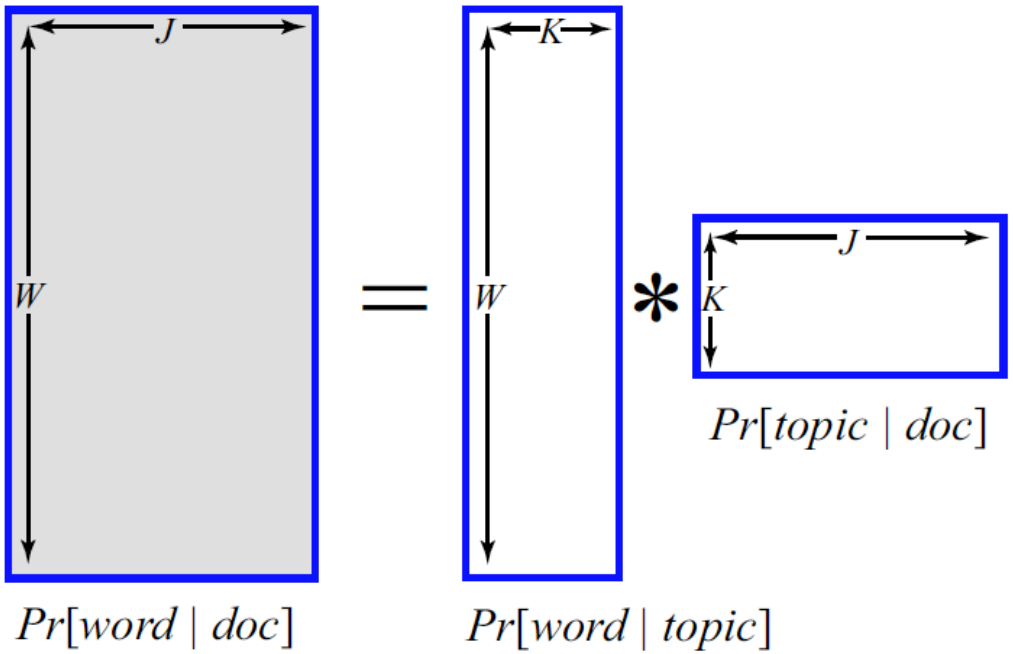
Each observation comes from some mixture component:

$$p(x_{ji} \mid \pi_j, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_{jk} f(x_{ji} \mid \theta_k)$$

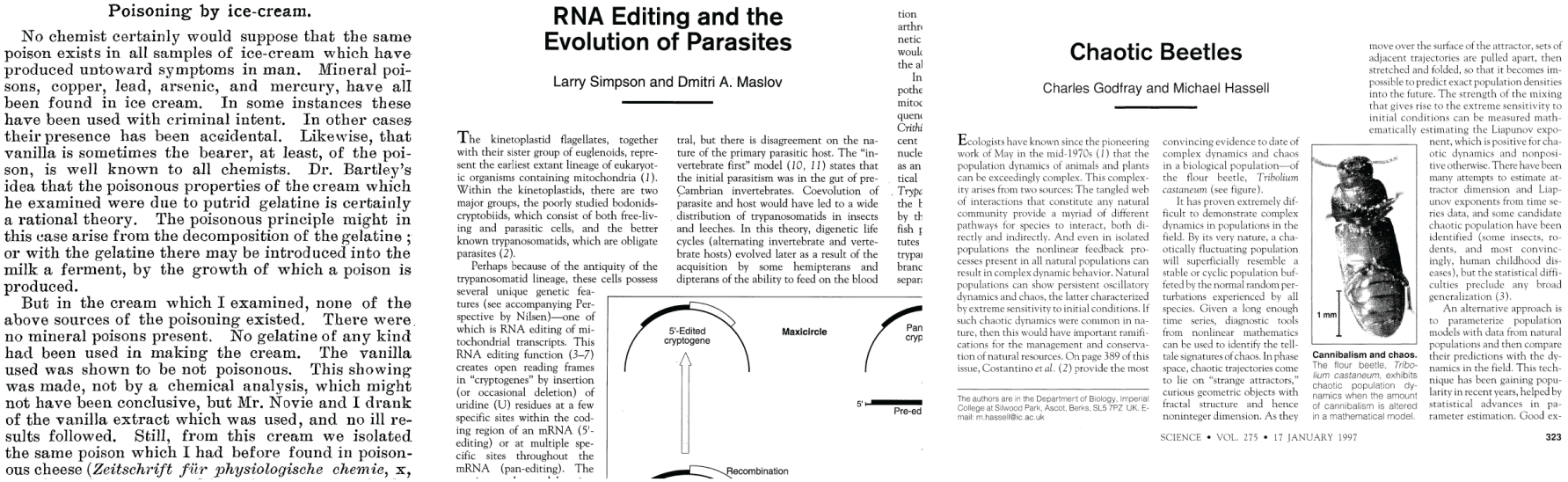
Admixture or Topic Models: Multiple Multinomial Mixtures



Latent Dirichlet Allocation (LDA)



Example Data for a Topic Model



Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Novie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (*Zeitschrift für physiologische chemie*, x,

RNA Editing and the Evolution of Parasites

Larry Simpson and Dmitri A. Maslov

The kinetoplastid flagellates, together with their sister group of euglenoids, represent the earliest extant lineage of eukaryotic organisms containing mitochondria (*1*). Within the kinetoplastids, there are two major groups, the poorly studied bodonids-cryptobiids, which consist of both free-living and parasitic cells, and the better known trypanosomatids, which are obligate parasites (*2*).

Perhaps because of the antiquity of the trypanosomatid lineage, these cells possess several unique genetic features (see accompanying Perspective by Nilsen)—one of which is RNA editing of mitochondrial transcripts. This RNA editing function (*3*–*7*) creates open reading frames in "cryptogenes" by insertion (or occasional deletion) of uridine (U) residues at a few specific sites within the coding region of an mRNA (5′-editing) or at multiple specific sites throughout the mRNA (pan-editing). The tral, but there is disagreement on the nature of the primary parasitic host. The "invertebrate first" model (*10*, *11*) states that the initial parasitism was in the gut of pre-Cambrian invertebrates. Coevolution of parasite and host would have led to a wide distribution of trypanosomatids in insects and leeches. In this theory, digenetic life cycles (alternating invertebrate and vertebrate hosts) evolved later as a result of the acquisition by some hemipterans and dipterans of the ability to feed on the blood

5′-Edited cryptogene

Maxicircle

Recombination

5′

Pre-ed

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

1 mm

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ, UK. E-mail: m.hassell@ic.ac.uk

- Our data are the pages *Science* from 1880-2002 (from JSTOR)
- No reliable punctuation, meta-data, or references.
- Note: this is just a subset of JSTOR's archive.

D. Blei, 2008

Example Output: 4 Topics

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Columns sorted by probability of word given topic.

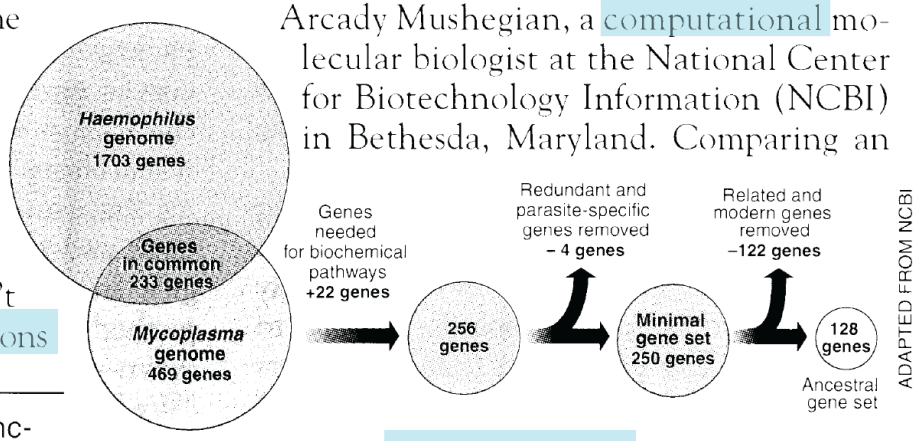
D. Blei, 2008

LDA: Intuition

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

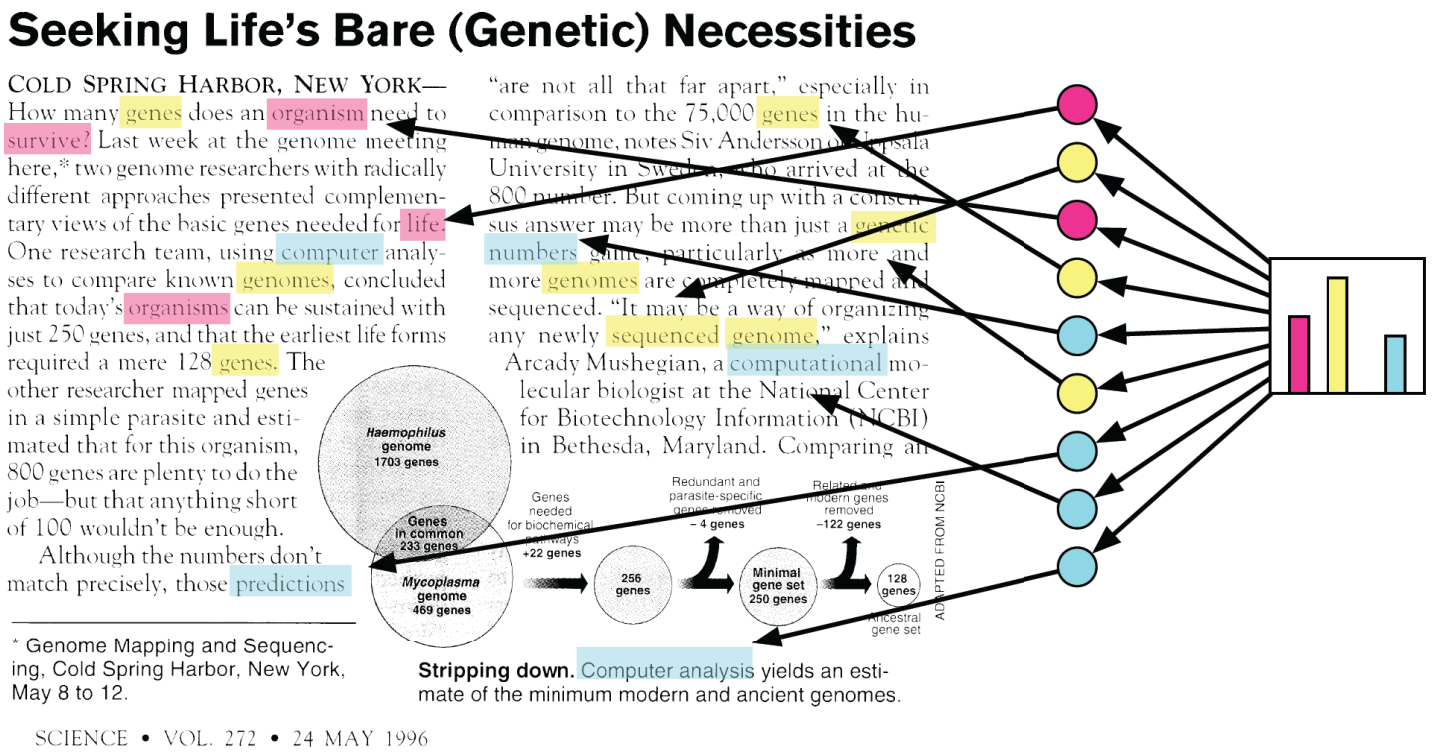
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Every document discusses a mixture of multiple topics.

D. Blei, 2008

LDA: Generative Model



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

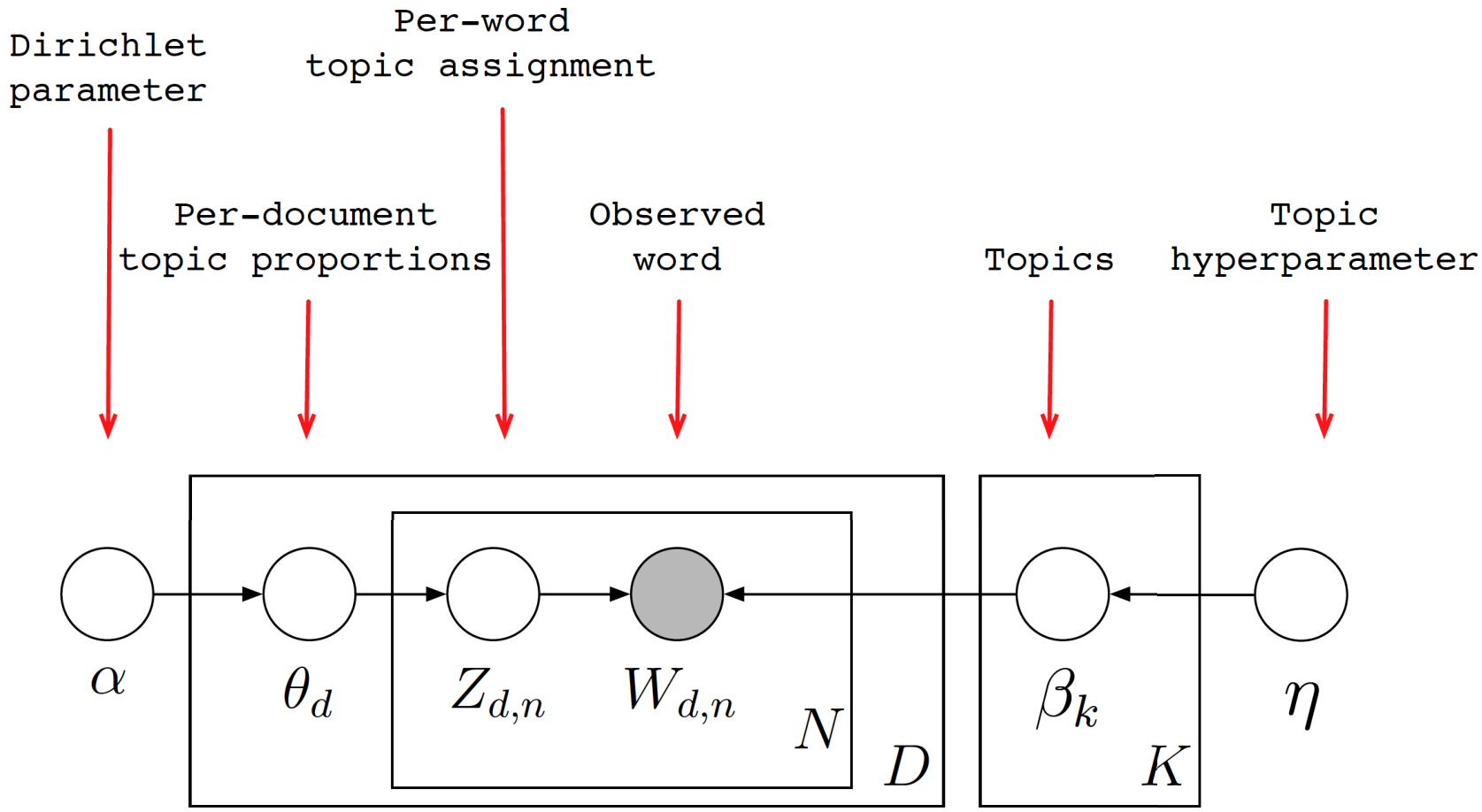
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

- Cast these intuitions into a generative probabilistic process
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

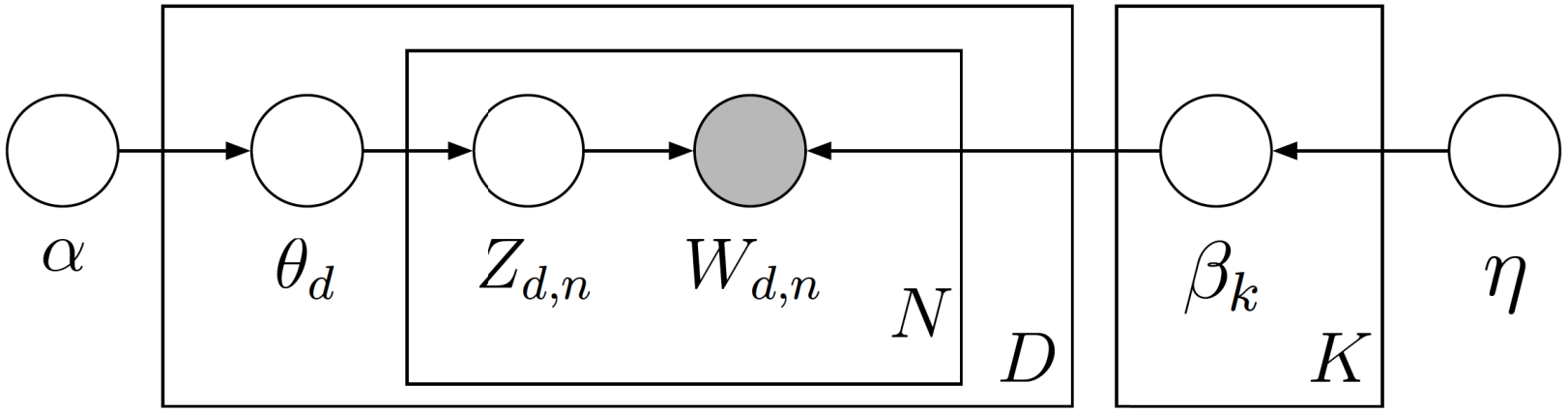
D. Blei, 2008

LDA: Graphical Model



D. Blei, 2008

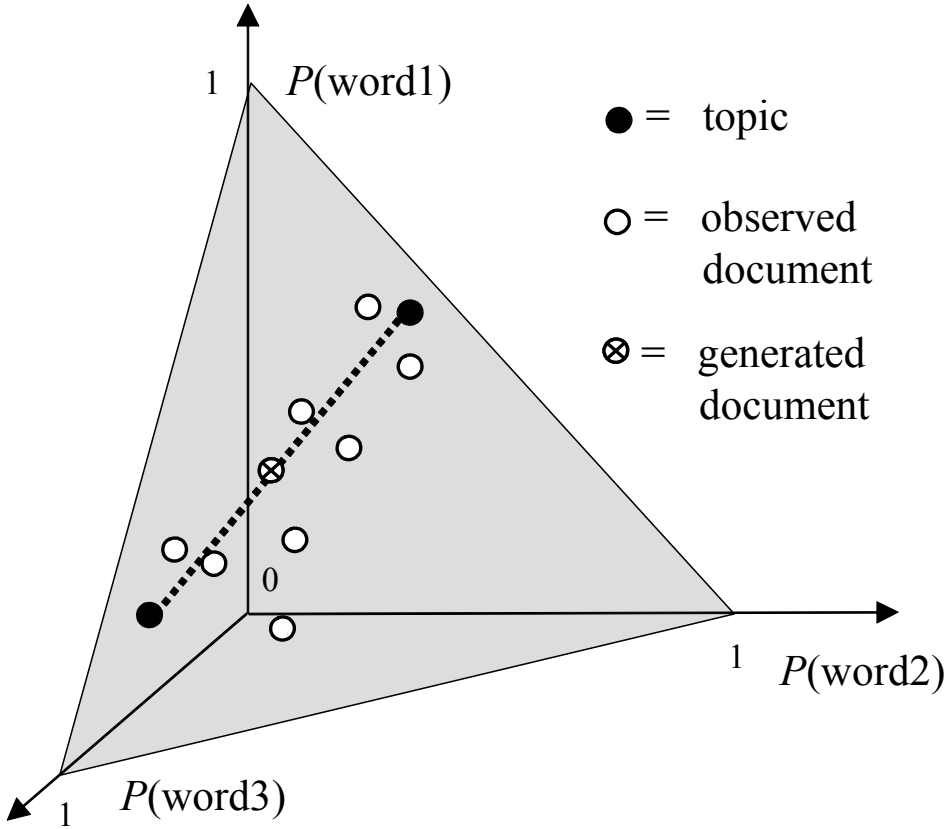
LDA: Graphical Model



1. Draw each topic $\beta_i \sim \mathrm{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$.

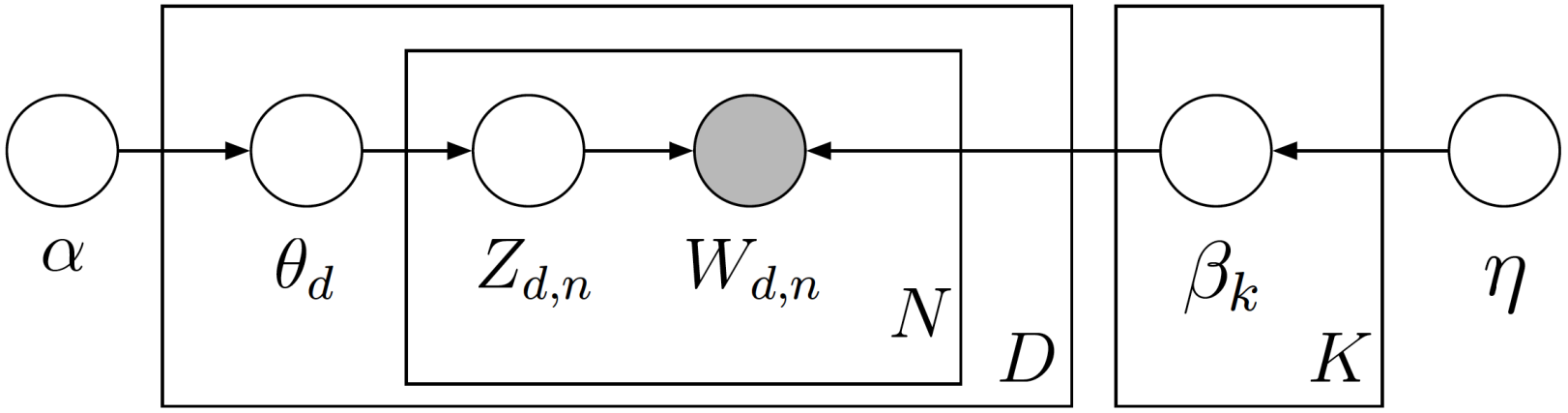
D. Blei, 2008

Geometry of Topic Models



- Documents are multinomial distributions over some predefined vocabulary of (tens of thousands) of words
- Topics are multinomial distributions on same vocabulary
- Generative model: Each document is (nearly) a convex combination of the topic distributions

LDA: Learning via EM Algorithm?



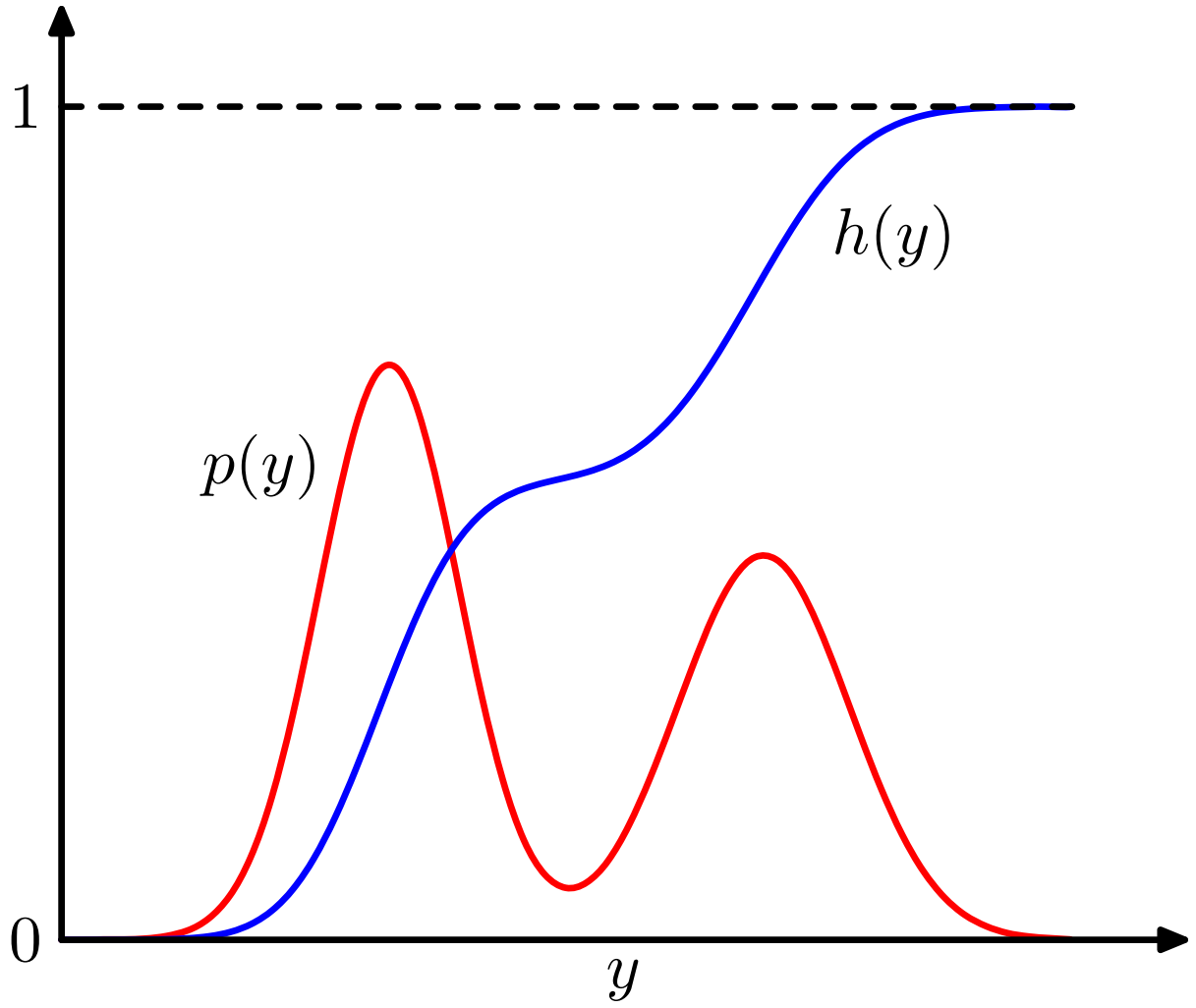
- **M-Step**: Maximize likelihood bound with respect to global topic usage distribution α and topic-word distributions β_k
- **E-Step:** Find posterior distribution of document-specific topic frequencies θ_d and word token assignments z_{dn}
- **Problem:** Posterior is intractable for large N
- **Variational methods:** Create a looser but more tractable log-likelihood bound by constraining form of posterior approx.
- Alternative: **Markov Chain Monte Carlo (MCMC)**

Uses of Monte Carlo Methods

$$z^{(\ell)} \sim p(z) \qquad \mathbb{E}[f] = \int f(z)p(z)\,dz$$

- Basic goals: Sampling or estimation of expectations
- Instead of learning by *optimization*, do *simulation*
- Given estimated parameters for some statistical model, quantitatively or qualitatively assess accuracy of fit
- Parameter estimation when closed forms unavailable
- Parameter estimation for models with hidden "nuisance" variables (alternative to the EM algorithm)
- General approach to applying computational resources to solve statistical learning problems…

Random Number Generation



Chaotic dynamical systems generate sequences of pseudo-random numbers approximately distributed uniformly on [0,1]

Monte Carlo Estimators

$$\mathbb{E}_p[f(x)] = \int_{\mathcal{X}} f(x)p(x)\,dx$$

$$\approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) = \mathbb{E}_{\tilde{p}}[f(x)]$$

$\{x^{(\ell)}\}_{\ell=1}^{L}$ *independent samples*

$$\tilde{p}(x) = \frac{1}{L}\sum_{\ell=1}^{L} \delta(x, x^{(\ell)})$$

Good properties if *L* sufficiently large:

- Unbiased for any sample size
- Variance inversely proportional to sample size (and independent of dimension of space)
- Weak law of large numbers
- Strong law of large numbers
- **Problem:** Drawing samples from complex distributions…

Alternatives for hard problems:

- Importance sampling
- Markov chain Monte Carlo (MCMC)