Models for Probability/Discrete Vectors with Bayesian Non-parametric Methods

Wray Buntine
Monash University

Thanks to Lan Du and Mark Johnson of Macquarie Uni., Kar Wai Lim and Swapnil Mishra of The ANU, and Changyou Chen of Duke Uni. for many slides and ideas, and Lancelot James of HK UST for teaching me generalised IBP.

19th Jan 2015

MONASH University

Legend

Color coding:

blue phrases: important terms and phrases;

green phrases: new terms;

red phrases: important phrases with negative connotations.

Wikipedia is thin on the more esoteric parts of tutorial, but recommended content is marked so$^{\mathcal{W}}$.

Outline

What are Non-parametric Methods?

By classical statistics: statistical modelling without parameters,
e.g., nearest neighbour methods

By Bayesian statistics: statistical modelling with an infinite number of parameters.

More accurately: statistical modelling:

- with a finite but variable number of parameters;
- more parameters are "unfurled" as needed;
- the model selection problem is converted into a parameter fitting problem (using a parameter that induces model complexity).

Exchangeability

Definition of exchangeability

Exchangeable random variables$^{\mathcal{W}}$, for instance a sequence $(x_1, x_2, ..., x_N)$, are such that their probability is invariant w.r.t. the order of presentation of the data.

- exchangeable data are not independent, though independent data are exchangable
- de Finetti's Theorem$^{\mathcal{W}}$ says (roughly) that:

  exchangeable data are independent with an underlying but unknown common cause

i.e. marginalising out the unknown removes dependence but leaves exchangeability

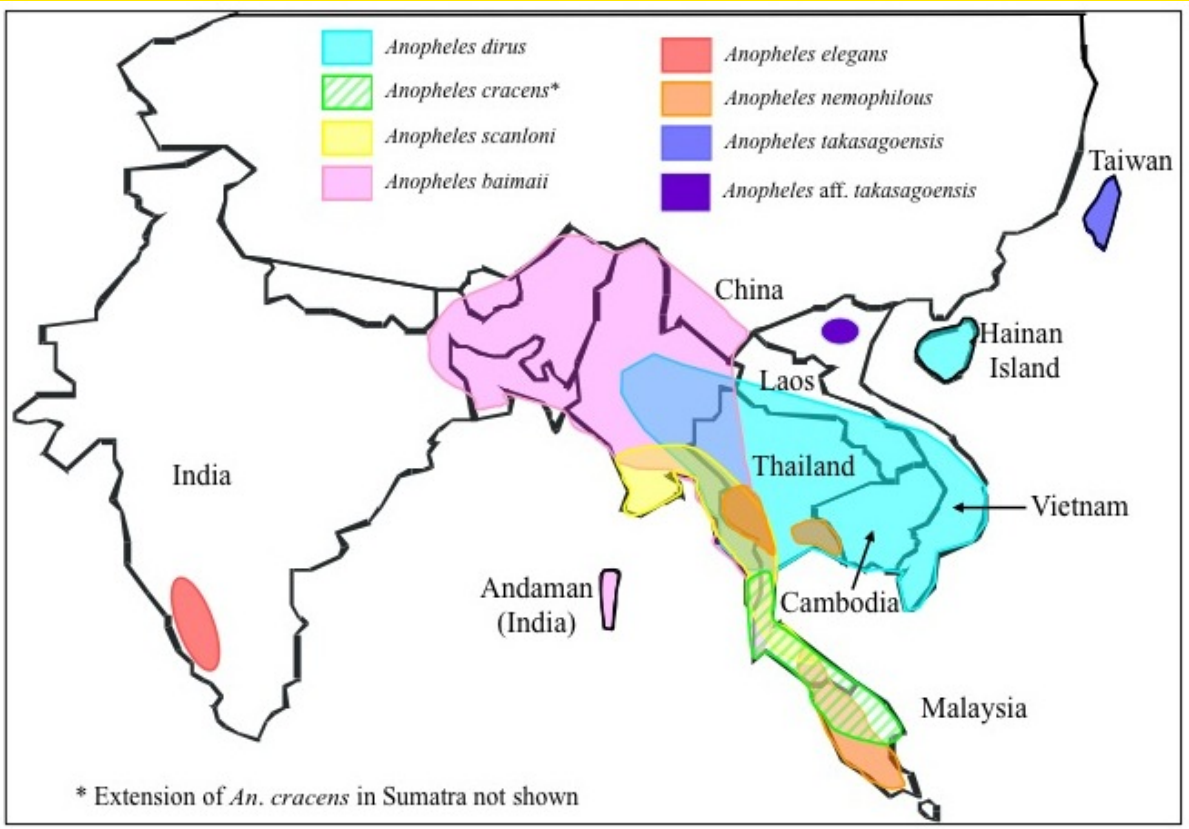
What is a Stochastic Process?

Definition of stochastic process

A stochastic process$^{\mathcal{W}}$ is a collection of random variables, representing the evolution of some system of random values over time (or some other "indexing").

- some definitions of the Dirichlet Process (DP) and the Pitman-Yor Process (PYP) define them via a stochastic process;
- in our use, they are exchangeable processes
  i.e., sample order is irrelevant;
- thus we use them as distributions;
- *i.e.*, ignore the word "process," its a historical artifact

How Many Species of Mosquitoes are There?



e.g. Given some measurement points about mosquitoes in Asia, how many species are there?

K=4? K=5? K=6 K=8?

How Many Words in the English Language are There?

> ... lastly, she pictured to herself how this same little sister of hers would, in the after-time, be herself a grown woman; and how she would keep, through all her riper years, the simple and loving heart of her childhood: and how she would gather about her other little children, and make their eyes bright and eager with many a strange tale, perhaps even with the dream of wonderland of long ago: ...

e.g. Given 10 gigabytes of English text, how many words are there in the English language?

K=1,235,791? K=1,719,765? K=2,983,548?

How Many are There?

How many species of mosquitoes are there?

- we expect there to be a finite number of species,
- we could use a Dirichlet of some fixed dimension K, and do model selection on K

$\longrightarrow$ Model with a finite mixture model of unknown dimension K.

How many words in the English language are there?

- This is a trick question.
- The *Complete Oxford English Dictionary* might attempt to define the language at some given point in time.
- The language keeps adding new words.
- The language is unbounded, it keeps growing.

$\longrightarrow$ Model with a countably infinite mixture model.

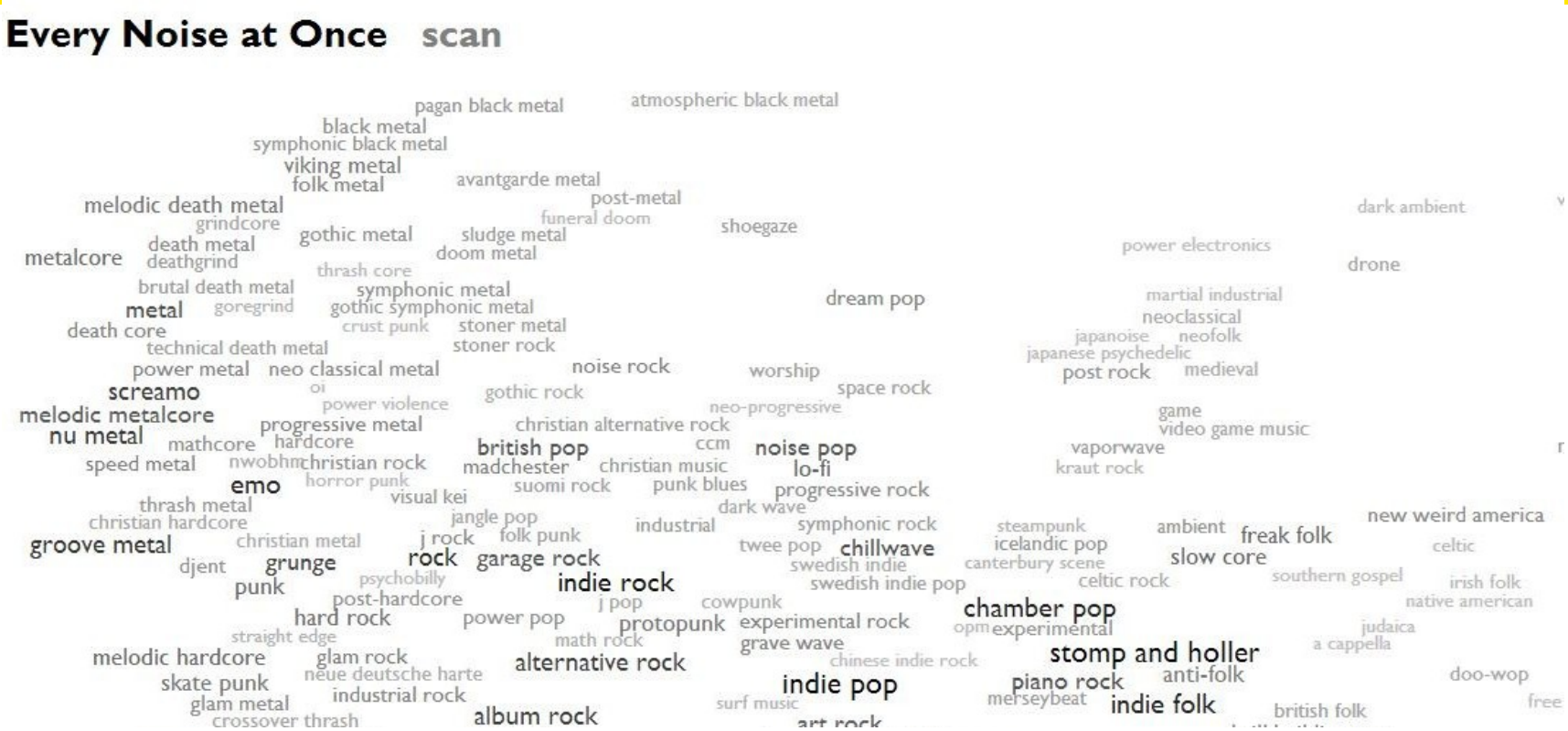
Probability Vectors

Problems in modern natural language processing and intelligent systems often have probability vectors for:

- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,
- hashtag in a tweet given the author.

We need to work with distributions over probability vectors to model these sorts of phenomena well.

Which Music Genre's do You Listen to?



(see http://everynoise.com)

Music genre's are constantly developing.

Which ones do you listen to?

What is the chance that a new genre is seen?

Which Ones are There and How Many?

Which music genres do you listen to?

- The available list is constantly expanding.
- You (may) have a small, fixed list you are aware of and actively listen to.

$\longrightarrow$ Model with a finite Boolean vector of unknown dimension K.

What Data is at Arnold Schwarzenegger's Freebase page?

- The list of entries is expanding: film performances, honorary degrees, profession, children, books, quotations, ...
- Which entries are there and how many of each?

$\longrightarrow$ Model with a finite count vector of unknown dimension K.

Intelligent Systems Inference

Problems in modern natural language processing and intelligent systems require:

- inheritance, sharing and fusion of information;
- arbitrary numbers of relations and features;
- mixture models and symbols of arbitrary size;
- boot-strapping from relational (non-probabilistic) resources;
- inference and learning.

And all this needs to be done with probability vectors and with discrete feature vectors/structures.

Bayesian Inference

Bayesian inference$^{\mathcal{W}}$ is particularly suited for intelligent systems in the context of the previous requirments:

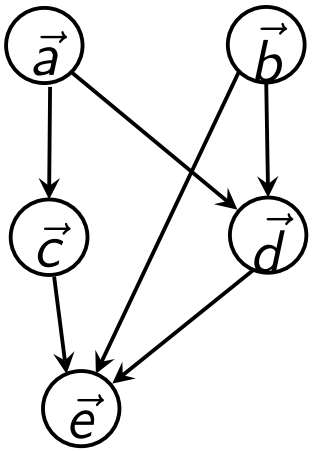
- Bayesian model combination$^{\mathcal{W}}$ and Bayes factors$^{\mathcal{W}}$ for model selection$^{\mathcal{W}}$ can be used;
- marginal likelihood$^{\mathcal{W}}$, a.k.a. the evidence for efficient estimation;
- collapsed Gibbs samplers$^{\mathcal{W}}$, a.k.a. Rao-Blackwellised samplers, for Monte-Carlo Markov chain$^{\mathcal{W}}$ (MCMC) estimation;
- also blocked Gibbs samplers$^{\mathcal{W}}$.

NB. Wikipedia coverage of Bayesian non-parametrics is poor.

But can the non-parametric inference be made practical?

Networks/Hierarchies of Probability Vectors

- Early inference on Bayesian networks$^{\mathcal{W}}$ had categorical or Gaussian variables only.
- Subsequent research gradually extends the range of distributions.



- A large class of problems in NLP and machine learning require inference and learning on networks of probability vectors.
- For this we use MCMC methods and discrete non-parametric models called species sampling models:
  - Dirichlet Process, Pitman-Yor process, normalised Generalised Gamma distribution.
- Discrete non-parametric methods do far more than just estimate "how many"!

ASIDE: Aspects, Ratings and Sentiments

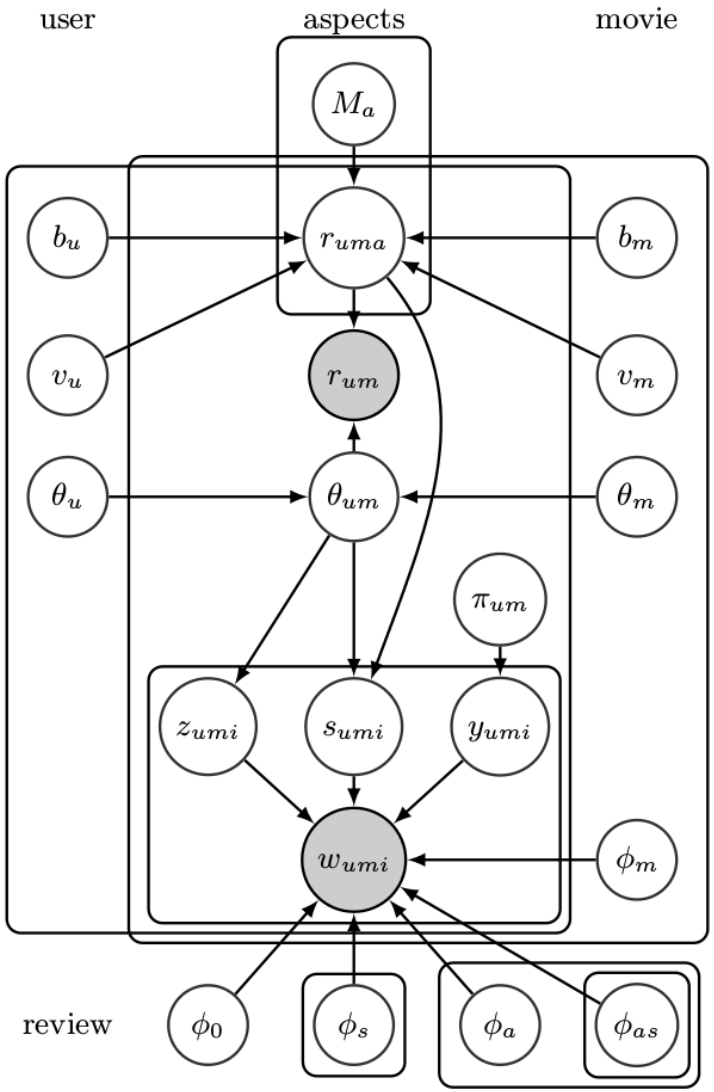


Figure 1: Factorized rating and review model.

"Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS)," Diao, Qiu, Wu, Smola, Jiang and Wang, KDD 2014.

State of the art sentiment model.

Typical methods currently lack probability vector hierarchies.

Chinese Restaurants and Breaking Sticks

- Standard machine learning methods for dealing with probability vectors are based on:
  - stick-breaking definitions for infinite probability vectors, and
  - Chinese restaurant processes (CRP) for distributions on probability vectors.
- They tend to hide aspects of the standard Bayesian framework: the actual posterior, the underlying model.
- The CRP requires considerable dynamic memory for use on larger problems.
- The stick-breaking model seems to interact poorly with variational algorithms.

We will consider more recent techniques that improve on these.

Goals of the Tutorial

- We'll see how to address the problems:
  - distributions on probability vectors,
  - countably infinite mixture models,
  - infinite discrete feature vectors.
- We'll use the Dirichlet Process (DP), the Pitman-Yor Process (PYP), and a generalisation of Indian Buffet Process (IBP)
- We'll see how to develop complex models and samplers using these.
- The methods are ideal for tasks like sharing, inheritance and arbitrarily complex models, all of which are well suited for Bayesian methods.
- The analysis will be done in the context of the standard Bayesian practice.

Outline

Dirichlet distributions

Definition of Dirichlet distribution

The Dirichlet distribution$^{\mathcal{W}}$ is used to sample finite probability vectors.

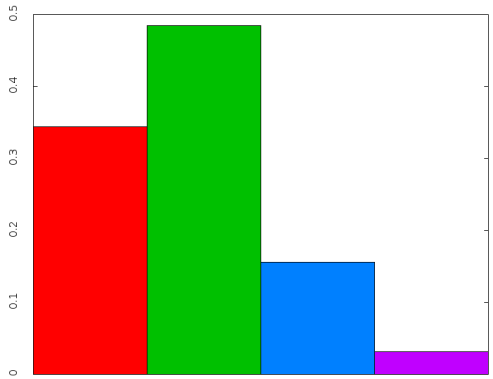
$$\vec{p} \sim \text{Dirichlet}_K(\alpha_0, \vec{\mu}) \quad \text{or} \quad \text{Dirichlet}_K(\vec{\alpha})$$

where $\alpha_0 > 0$ and $\vec{\mu}$ is a positive K-dimensional probability vector; alternatively $\vec{\alpha}$ is a positive K-dimensional vector.

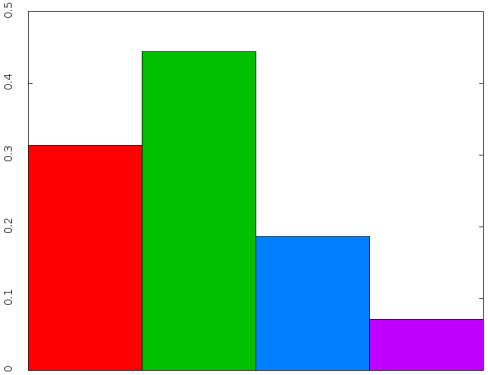
- first form comparable to the circular multivariate Gaussian $\vec{x} \sim \text{Gaussian}_K\left(\sigma^2, \vec{\mu}\right)$ (mean, concentration, *etc.*),
- second form more common,
- said to be a conjugate prior$^{\mathcal{W}}$ for the multinomial distribution, *i.e.*, makes math easy.

4-D Dirichlet samples

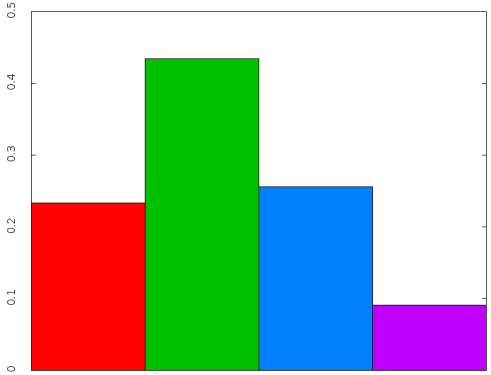
$\vec{p}_0$



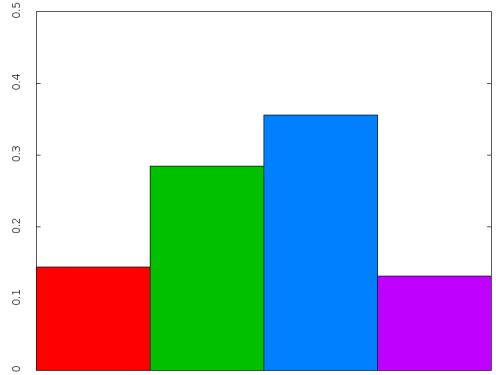
$\vec{p}_1 \sim \text{Dirichlet}_4(500, \vec{p}_0)$



$\vec{p}_2 \sim \text{Dirichlet}_4(5, \vec{p}_0)$

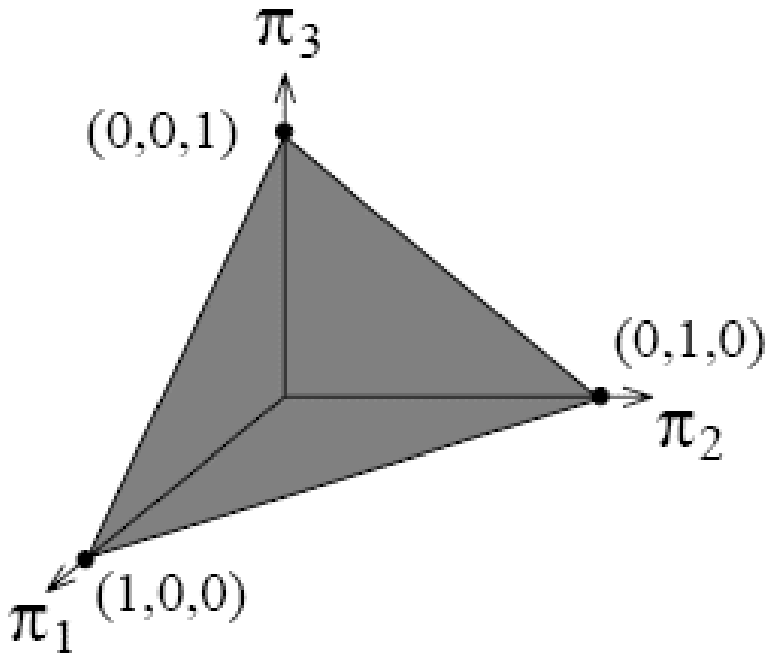


$\vec{p}_3 \sim \text{Dirichlet}_4(0.5, \vec{p}_0)$



Two Forms for 3-D Dirichlet

Consider $\vec{p} = (p_1, p_2, p_3)$ where $\sum_k p_k = 1$.



$\vec{p} \sim \text{Dirichlet}_3(\vec{\alpha})$ means that $p(\vec{p}\,|\,\vec{\alpha})$ is

$$\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1+\alpha_2+\alpha_3)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} \ ,$$

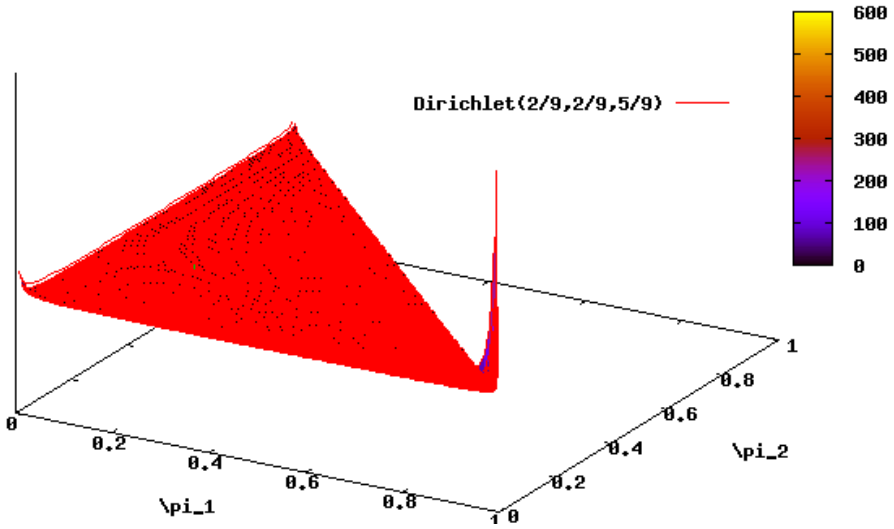
where $\Gamma(\cdot)$ is the Gamma function.

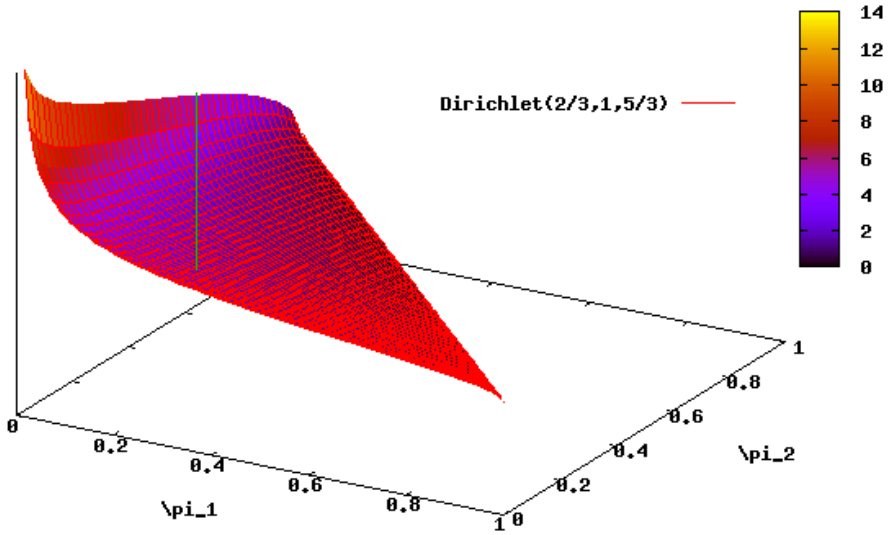
Alternatively $\vec{p} \sim \text{Dirichlet}_3(\alpha_0, \vec{\mu})$ where

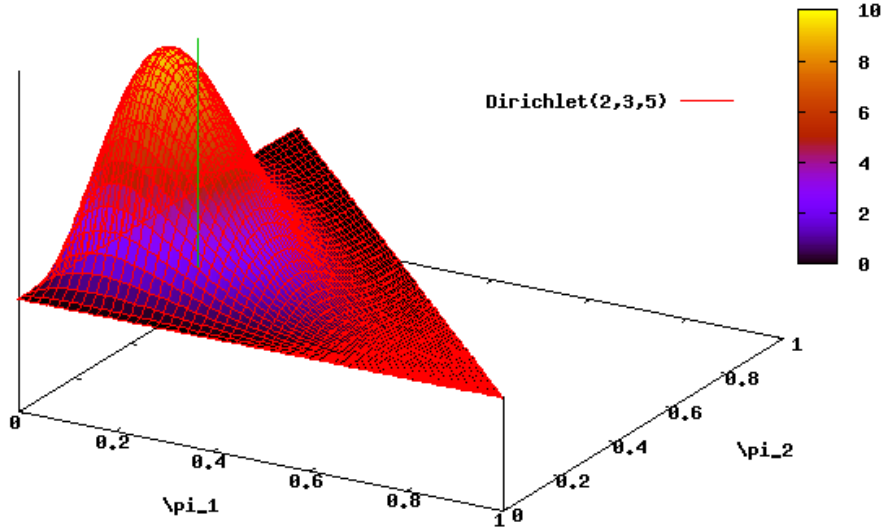
$$\begin{aligned} \alpha_0 &= \alpha_1 + \alpha_2 + \alpha_3 \\ \vec{\mu} = (\mu_1, \mu_2, \mu_3) &= \left(\frac{\alpha_1}{\alpha_0}, \frac{\alpha_2}{\alpha_0}, \frac{\alpha_3}{\alpha_0} \right) \ , \end{aligned}$$

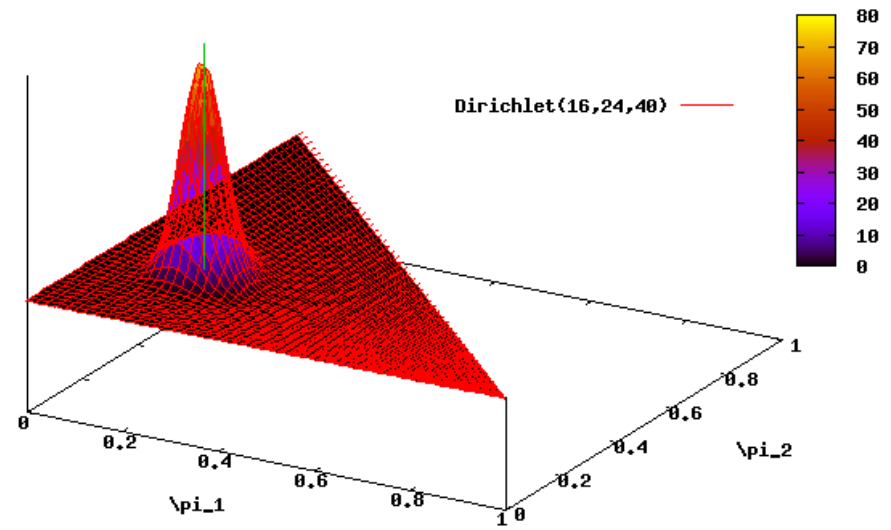
and the mean of $\vec{p}$ is $\vec{\mu}$.

The Dirichlet Distribution









Dirichlet Details

- $\alpha_0 = \alpha_1 + \alpha_2 + \alpha_3$ is called the concentration parameter[W]. Its significance is shown by the variance

$$Var_{\vec{\alpha}}[\vec{p}_i] = \frac{1}{\alpha_0 + 1} \mathbb{E}_{\vec{\alpha}}[\vec{p}_i](1 - \mathbb{E}_{\vec{\alpha}}[\vec{p}_i]) \ .$$

- Let $\vec{n} \sim$ multinomial $(N, \vec{p})$, then

$$p(\vec{n} \,|\, \vec{p}) = C_{\vec{n}}^{N} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

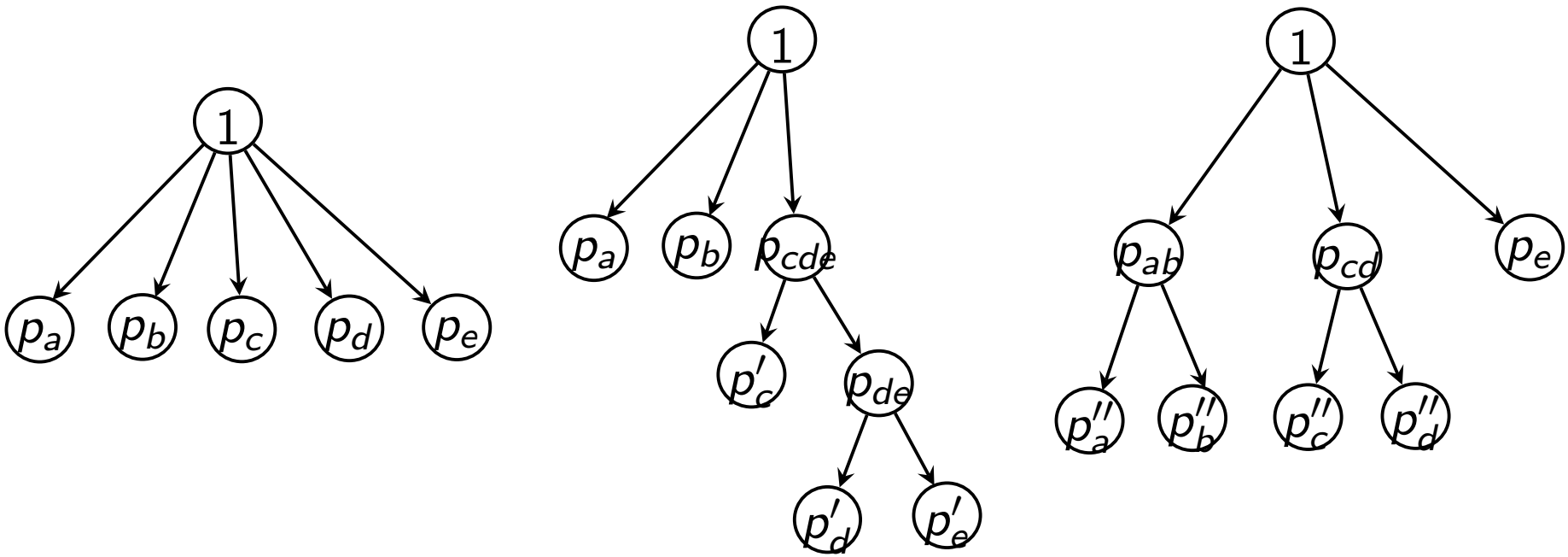
  So data sampled from a multinomial using $\vec{p}$ has the same functional form as the Dirichlet distribution on $\vec{p}$, *i.e.*, conjugacy.

- Normalising constant for Dirichlet is called a Beta function[W]:

$$\text{Beta}_3(\alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}$$

Outline

Multinomial Properties



The multinomial distribution can be arbitrarily rearranged as a tree and probabilities appropriately renormalised:

$$p_{cde} = p_c + p_d + p_e,\ p'_c = p_c/p_{cde},\ p_{ab} = p_a + p_b,\ p''_a = p_a/p_{ab},\ \textit{etc.}$$

Dirichlet Properties (second form)

If $(x_1, ..., x_K) \sim \text{Dir}_K(\alpha_1, ..., \alpha_K)$ then the following hold:

Symmetry: for any order σ,

$$(x_{\sigma(1)}, ..., x_{\sigma(K)}) \sim \text{Dir}_K(\alpha_{\sigma(1)}, ..., \alpha_{\sigma(K)}) \ .$$

Aggregation: if cells $1, ..., C < K$ are merged giving combination $y_C = x_1 + ... + x_C$, then the resultant distribution is

$$(y_C, x_{C+1}, ..., x_K) \sim \text{Dir}_{K-C+1}(\alpha_{1..C}, \alpha_{C+1}, ..., \alpha_K) \ ,$$

where $\alpha_{1..C} = \sum_{i=1}^{C} \alpha_i$ is sum for the cells $1, ..., C$.

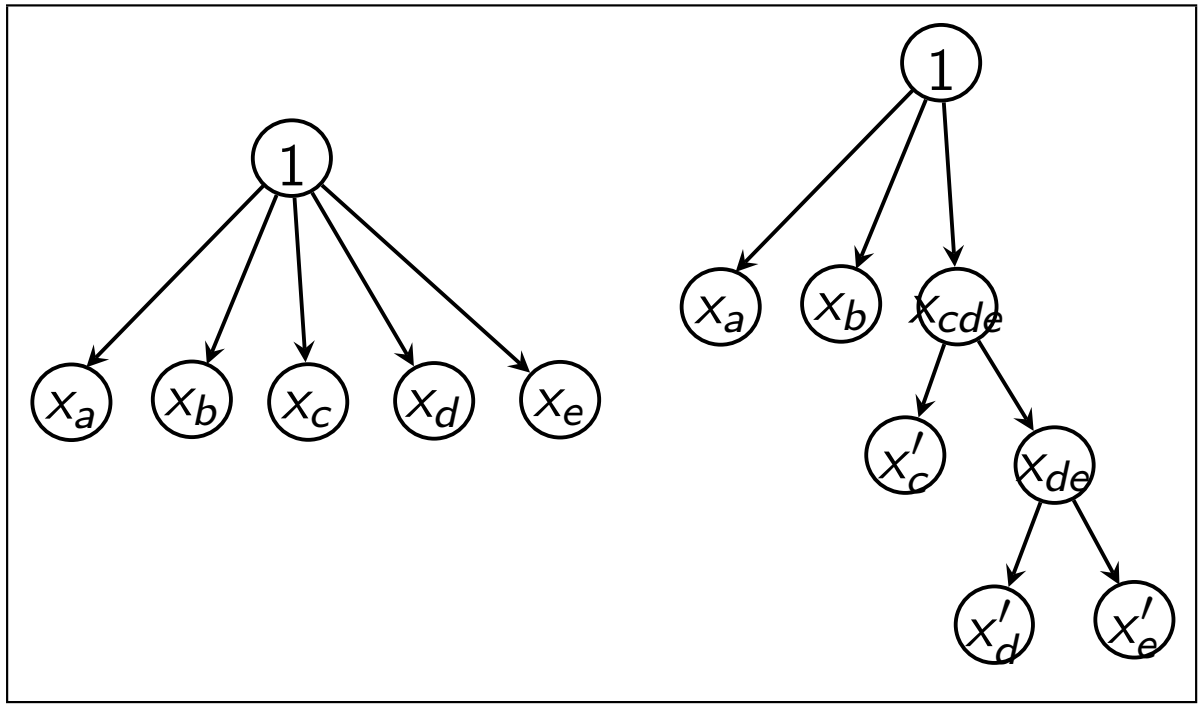
Restriction: if cells $1, ..., C < K$ are considered in isolation, then the distribution on the subset is

$$\frac{(x_1, ..., x_C)}{\sum_{c=1}^{C} x_c} \sim \text{Dir}_C(\alpha_1, ...\alpha_C) \ .$$

Aggregation Properties (second form), example

Assume

$(x_a, x_b, x_c, x_d, x_e) \sim$
$\mathrm{Dir}_5(\alpha_a, \alpha_b, \alpha_c, \alpha_d, \alpha_e))$



Then:

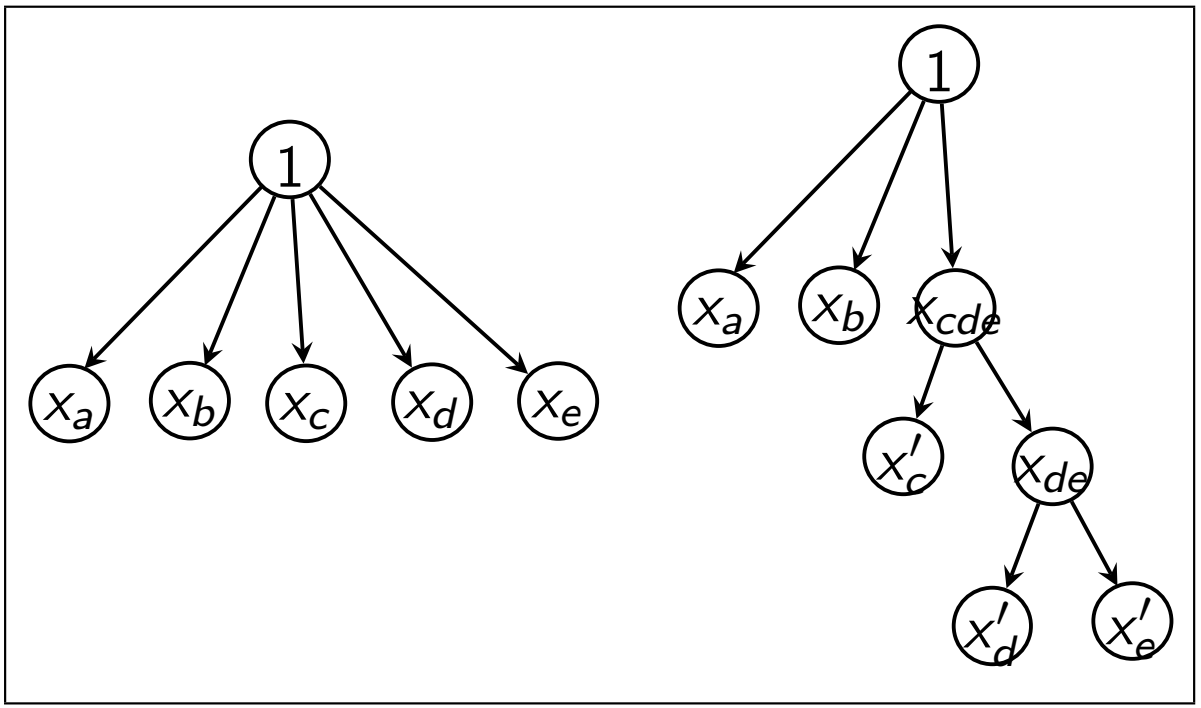
$$
\begin{aligned}
(x_a, x_b, x_c + x_d + x_e) &\sim \mathrm{Dir}_3(\alpha_a, \alpha_b, \alpha_c + \alpha_d + \alpha_e) \ , \\
\left(\frac{x_c}{x_c + x_d + x_e}, \frac{x_d + x_e}{x_c + x_d + x_e}\right) &\sim \mathrm{Dir}_2(\alpha_c, \alpha_d + \alpha_e) \ , \\
\left(\frac{x_d}{x_d + x_e}, \frac{x_e}{x_d + x_e}\right) &\sim \mathrm{Dir}_2(\alpha_d, \alpha_e) \ .
\end{aligned}
$$

Aggregation Properties (second form), example cont.

$$\begin{aligned}
\text{Assume } (x_a, x_b, x_c + x_d + x_e) &\sim \text{Dir}_3(\alpha_a, \alpha_b, \alpha_{cde}) \ , \\
\left(\frac{x_c}{x_c + x_d + x_e}, \frac{x_d + x_e}{x_c + x_d + x_e}\right) &\sim \text{Dir}_2(\alpha_c, \alpha_{de}) \ , \\
\left(\frac{x_d}{x_d + x_e}, \frac{x_e}{x_d + x_e}\right) &\sim \text{Dir}_2(\alpha_d, \alpha_e) \ .
\end{aligned}$$

To collapse back to $\text{Dir}_5(\alpha_a, \alpha_b, \alpha_c, \alpha_d, \alpha_e)$ the concentrations must sum appropriately:

$$\begin{aligned}
\alpha_{de} &= \alpha_d + \alpha_e \ , \\
\alpha_{cde} &= \alpha_c + \alpha_{de} \ .
\end{aligned}$$



ASIDE: Dirichlet Properties (first form)

If $(x_1, ..., x_K) \sim \text{Dir}_K(\alpha, (p_1, ..., p_K))$ then the following hold:

Symmetry: for any order σ,

$$(x_{\sigma(1)}, ..., x_{\sigma(K)}) \sim \text{Dir}_K(\alpha, (p_{\sigma(1)}, ..., p_{\sigma(K)})) \ .$$

Aggregation: if cells $1, ..., C < K$ are merged giving combination $y_C = x_1 + ... + x_C$, then the resultant distribution is

$$(y_C, x_{C+1}, ..., x_K) \sim \text{Dir}_{K-C+1}(\alpha, (q_C, p_{C+1}, ..., p_K)) \ ,$$

where $q_C = \sum_{i=1}^{C} p_i$ is the normaliser for the cells $1, ..., C$.

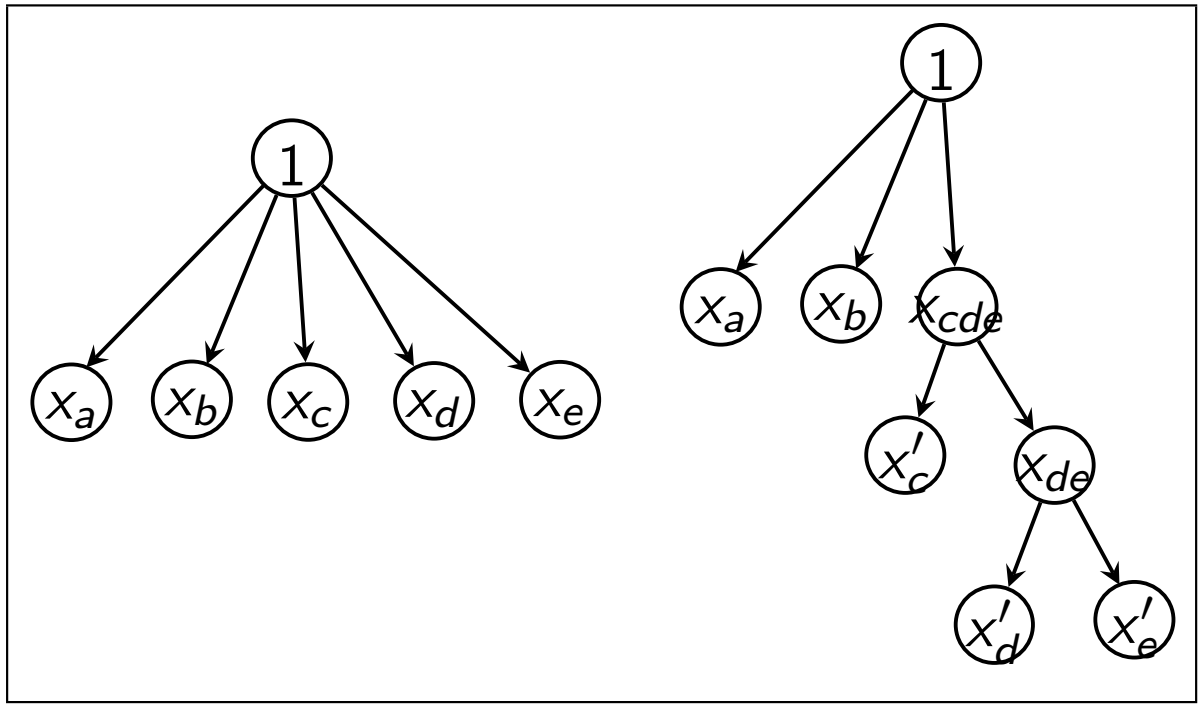
Restriction: if cells $1, ..., C < K$ are considered in isolation, then the distribution on the subset is

$$\frac{(x_1, ..., x_C)}{\sum_{c=1}^{C} x_c} \sim \text{Dir}_C\left(\alpha q_C, \left(\frac{p_1}{q_C}, ... \frac{p_C}{q_C}\right)\right) \ .$$

ASIDE: Aggregation Properties (first form), example

Assume

$(x_a, x_b, x_c, x_d, x_e) \sim$
$\mathrm{Dir}_5(\alpha, (p_a, p_b, p_c, p_d, p_e))$



Then:

$$(x_a, x_b, x_c + x_d + x_e) \sim \mathrm{Dir}_3(\alpha, (p_a, p_b, p_{cde})) ,$$

$$\left(\frac{x_c}{x_c + x_d + x_e}, \frac{x_d + x_e}{x_c + x_d + x_e}\right) \sim \mathrm{Dir}_2\left(\alpha p_{cde}, \left(\frac{p_c}{p_{cde}}, \frac{p_d + p_e}{p_{cde}}\right)\right) ,$$

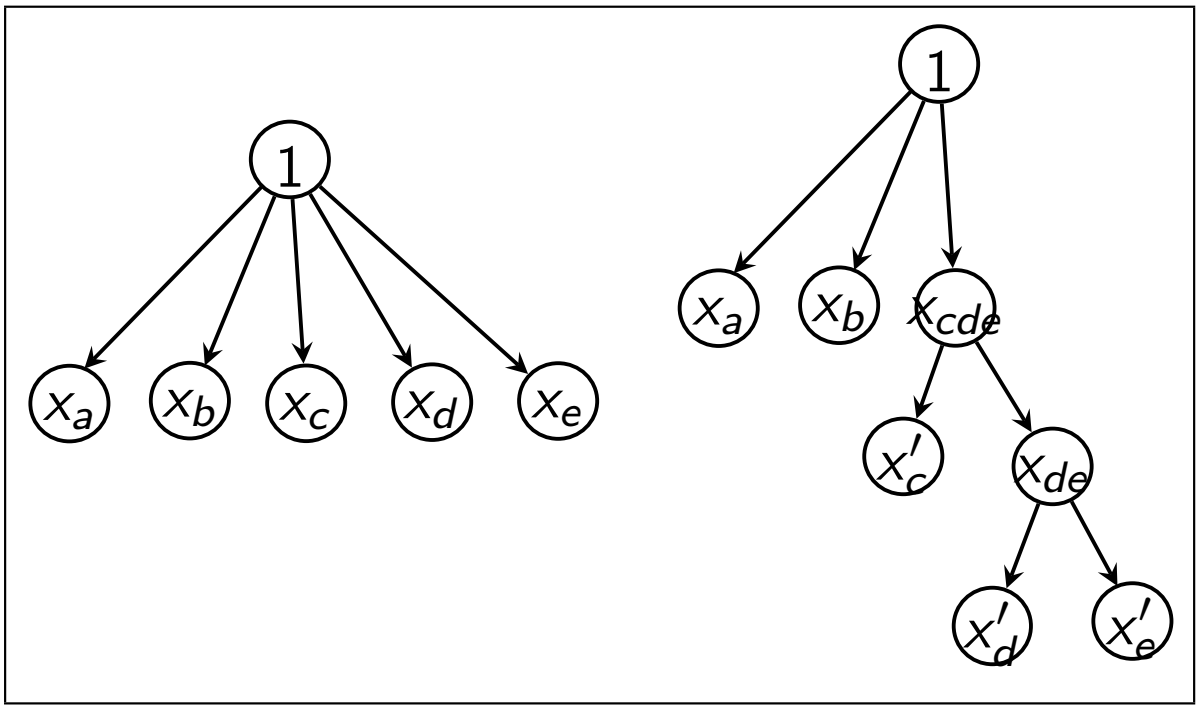
$$\left(\frac{x_d}{x_d + x_e}, \frac{x_e}{x_d + x_e}\right) \sim \mathrm{Dir}_2\left(\alpha(p_d + p_e), \left(\frac{p_d}{p_d + p_e}, \frac{p_e}{p_d + p_e}\right)\right)$$

ASIDE: Aggregation Properties (first form), eg. cont.

$$
\begin{aligned}
\text{Assume } (x_a, x_b, x_c + x_d + x_e) &\sim \text{Dir}_3\left(\alpha_0, (p_a, p_b, p_{cde})\right) , \\
\left(\frac{x_c}{x_c + x_d + x_e}, \frac{x_d + x_e}{x_c + x_d + x_e}\right) &\sim \text{Dir}_2\left(\alpha_1, \left(\frac{p_c}{p_{cde}}, \frac{p_d + p_e}{p_{cde}}\right)\right) , \\
\left(\frac{x_d}{x_d + x_e}, \frac{x_e}{x_d + x_e}\right) &\sim \text{Dir}_2\left(\alpha_2, \left(\frac{p_d}{p_d + p_e}, \frac{p_e}{p_d + p_e}\right)\right) .
\end{aligned}
$$

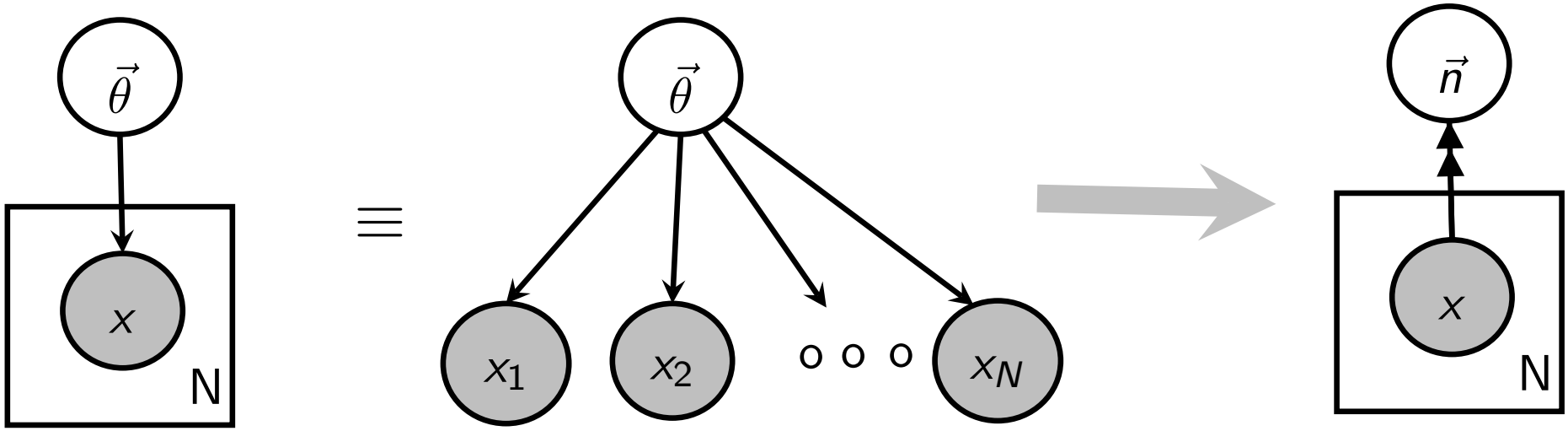
To collapse back to $\text{Dir}_5\left(\alpha_0, (p_a, p_b, p_c, p_d, p_e)\right)$ the concentrations must scale appropriately:

$$
\begin{aligned}
\alpha_1 &= \alpha_0 p_{cde} , \\
\alpha_2 &= \alpha_0 p_{de} .
\end{aligned}
$$



Outline

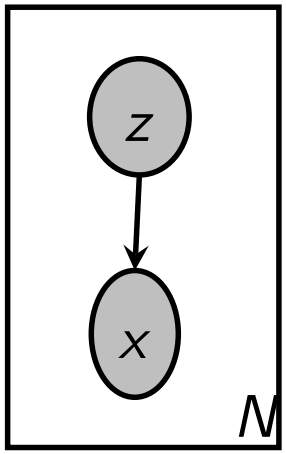
Reading a Graphical Model



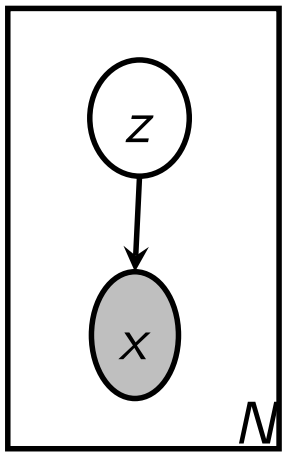
arcs	=	"depends on"
double headed arcs	=	"deterministically computed from"
shaded nodes	=	"supplied variable/data"
unshaded nodes	=	"unknown variable/data"
boxes	=	"replication"

Models in Graphical Form

Supervised learning or Prediction model

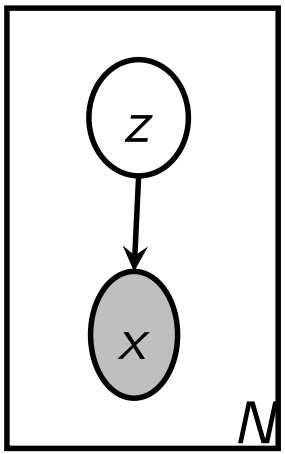


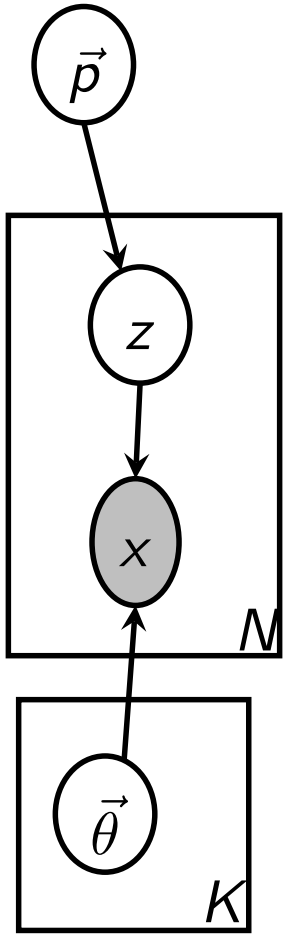
Clustering or Mixture model

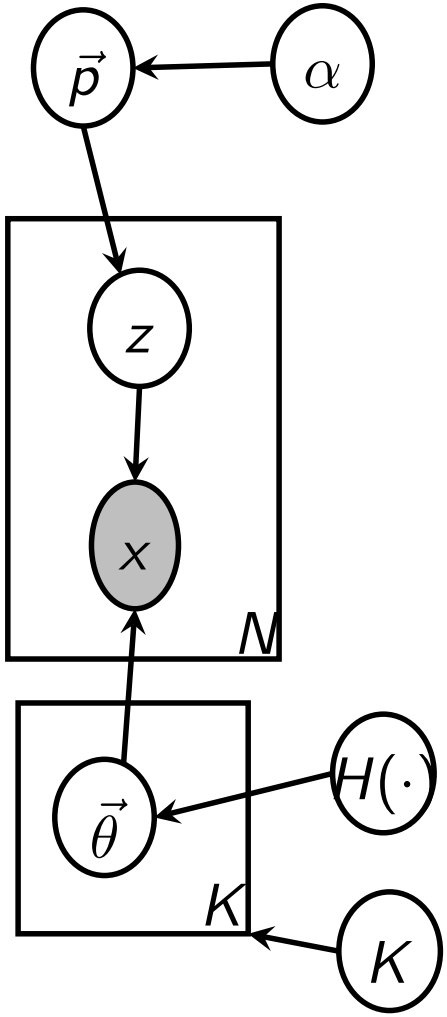


Mixture Models in Graphical Form

Building up the parts:



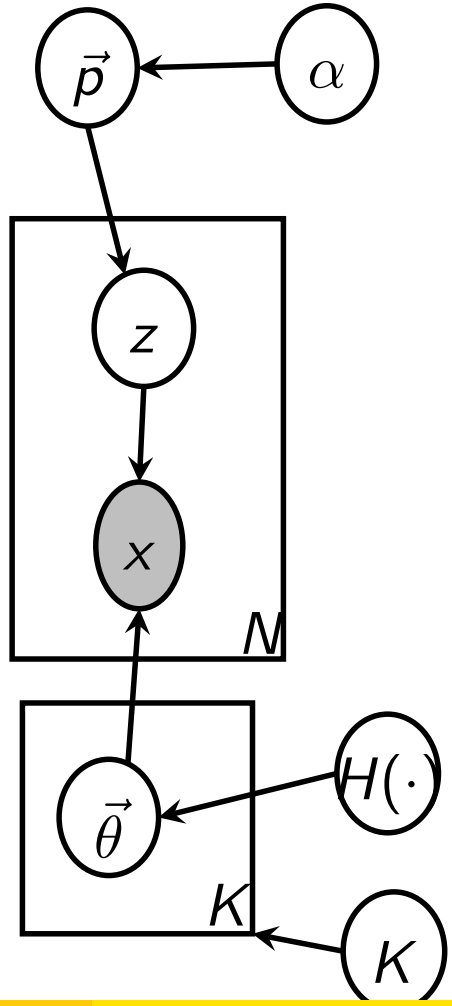




The Classic Mixture Model

Data is a mixture of unknown dimension K and base distribution $H(\cdot)$ generating mixture entries with proportions $\vec{p}$.

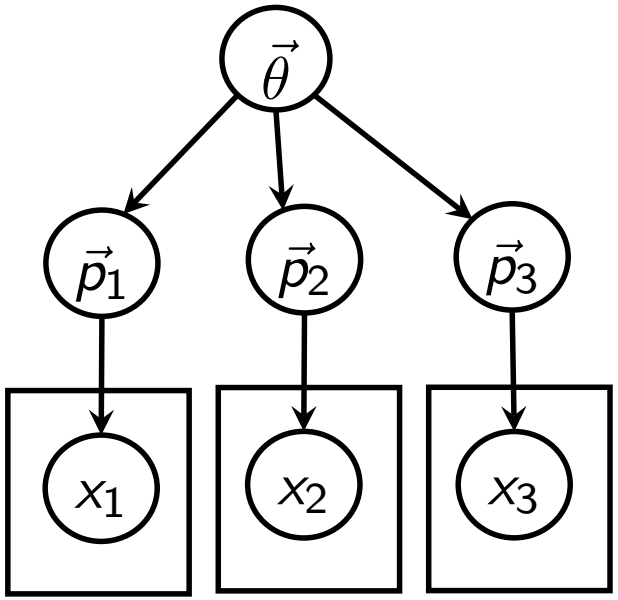
$$
\begin{aligned}
K &\sim p(K) \\
\vec{p} &\sim \text{Dirichlet}_K\left(\frac{\alpha}{K}\vec{1}\right) \\
\vec{\theta}_k &\sim H(\cdot) && \forall_{k=1,\ldots,K} \\
z_n &\sim \vec{p} && \forall_{n=1,\ldots,N} \\
x_n &\sim \vec{\theta}_{z_n} && \forall_{n=1,\ldots,N}
\end{aligned}
$$

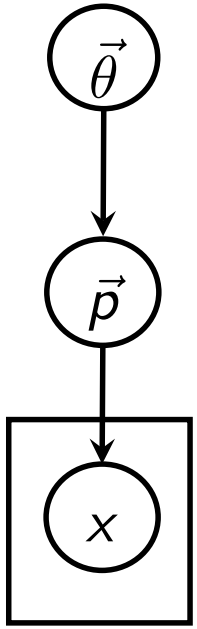


Outline

Dirichlet-Multinomial Motivation

A useful inference component is the Dirichlet-Multinomial. Begin by adding multinomials off the samples from a Dirichlet.





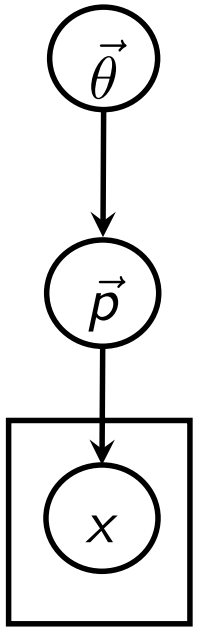
$$\vec{p}_l \sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right) \quad \forall_l$$
$$x_{l,n} \sim \text{Discrete}(\vec{p}_l) \quad \forall_{l,n}$$

$$\vec{p} \sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right)$$
$$x_n \sim \text{Discrete}(\vec{p}) \quad \forall_n$$

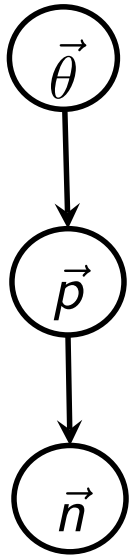
We will analyse the simplest case on the right.

The Dirichlet-Multinomial

First convert the categorical data into a set of counts.

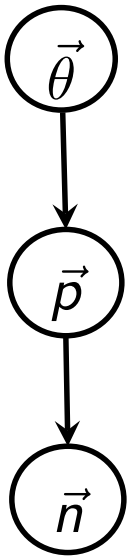


$$\begin{aligned} \vec{p} &\sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right) \\ x_n &\sim \text{Discrete}(\vec{p}) \quad \forall_n \end{aligned}$$



$$\begin{aligned} \vec{p} &\sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right) \\ \vec{n} &\sim \text{Multinomial}\left(\vec{p}, N\right) \end{aligned}$$

The Dirichlet-Multinomial, cont



$$\vec{p} \sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right) \quad \forall_k$$

$$\vec{n} \sim \text{Multinomial}\left(\vec{p}, N\right)$$

$$p\left(\vec{p}, \vec{n} \middle| \vec{\theta}, \ldots\right) = \frac{1}{\text{Beta}(\alpha\vec{\theta})} \prod_k p_k^{\alpha\theta_k - 1} \binom{N}{\vec{n}} \prod_k p_k^{n_k}$$

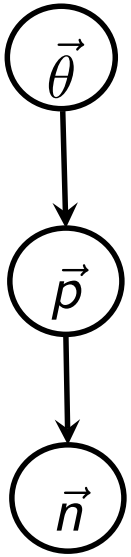
Integrate out (or eliminate) $\vec{p}$:

$$p\left(\vec{n} \middle| \vec{\theta}, \ldots\right) = \frac{1}{\text{Beta}(\alpha\vec{\theta})} \binom{N}{\vec{n}} \int_{\text{simplex}} \prod_k p_k^{n_k + \alpha\theta_k - 1} \mathrm{d}\vec{p}$$

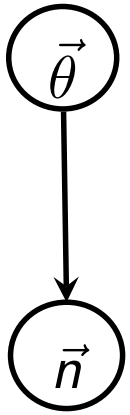
$$= \binom{N}{\vec{n}} \frac{\text{Beta}(\vec{n} + \alpha\vec{\theta})}{\text{Beta}(\alpha\vec{\theta})}$$

The Dirichlet-Multinomial, cont

The distribution with $\vec{p}$ marginalised out is given on right:



$$\vec{p} \sim \text{Dirichlet}\left(\alpha, \vec{\theta}\right)$$
$$\vec{n} \sim \text{Multinomial}\left(\vec{p}, N\right)$$



$$\vec{n} \sim \text{MultDir}\left(\alpha, \vec{\theta}, N\right)$$
$$\text{where } \sum_k n_k = N$$
$$p\left(\vec{n} \middle| N, \text{MultDir}, \alpha, \vec{\theta}\right) = \binom{N}{\vec{n}} \frac{\text{Beta}(\vec{n} + \alpha\vec{\theta})}{\text{Beta}(\alpha\vec{\theta})}$$

The Dirichlet-Multinomial, cont

Definition of Dirichlet-Multinomial

Given a concentration parameter α, a probability vector $\vec{\theta}$ of dimension K, and a count N, the Dirichlet-multinomial distribution creates count vector samples $\vec{n}$ of dimension K. Now $\vec{n} \sim \text{MultDir}\left(\alpha, \vec{\theta}, N\right)$ denotes

$$p\left(\vec{n}\middle| N, \text{MultDir}, \alpha, \vec{\theta}\right) = \binom{N}{\vec{n}} \frac{\text{Beta}\left(\vec{n} + \alpha\vec{\theta}\right)}{\text{Beta}\left(\alpha\vec{\theta}\right)}$$

where $\sum_{k=1}^{K} n_k = N$ and $\text{Beta}(\cdot)$ is the normalising function for the Dirichlet distribution.

This probability is also the evidence (probability of data with parameters marginalised out) for a Dirichlet distribution.

A Hierarchical Dirichlet-Multinomial Component?

Consider the functional form of the MultDir.

$$
\begin{aligned}
p\left(\vec{n}\middle| N, \text{MultDir}, \alpha, \vec{\theta}\right) &= \binom{N}{\vec{n}} \frac{\text{Beta}\left(\alpha\vec{\theta} + \vec{n}\right)}{\text{Beta}\left(\alpha\vec{\theta}\right)} \\
&= \binom{N}{\vec{n}} \frac{1}{(\alpha)_N} \prod_{k=1}^{K} (\alpha\theta_k)(\alpha\theta_k + 1)\cdots(\alpha\theta_k + n_k - 1) \\
&= \binom{N}{\vec{n}} \frac{1}{(\alpha)_N} \prod_{k=1}^{K} (\alpha\theta_k)_{n_k}
\end{aligned}
$$

where $(x)_n = x(x+1)...(x+n-1)$ is the rising factorial.

This is a complex polynomial we cannot deal with in a hierarchical model.

Outline

Topic Models of Text

Original news article:

Despite their separation, Charles and Diana stayed close to their boys William and Harry. Here, they accompany the boys for 13-year-old William's first day school at Eton College on Sept. 6, 1995, with housemaster Dr. Andrew Gayley looking on.

Bag of words:

13 1995 accompany and(2) andrew at boys(2) charles close college day despite diana dr eton first for gayley harry here housemaster looking old on on school separation sept stayed the their(2) they to william(2) with year

We'll approximate the bag with a linear mixture of text topics as probability vectors.

Components:

Words (probabilities not shown)	Human label
Prince, Queen, Elizabeth, title, son, ...	Royalty
school, student, college, education, year, ...	School
John, David, Michael, Scott, Paul, ...	Names
and, or, to , from, with, in, out, ...	Function

Matrix Approximation View

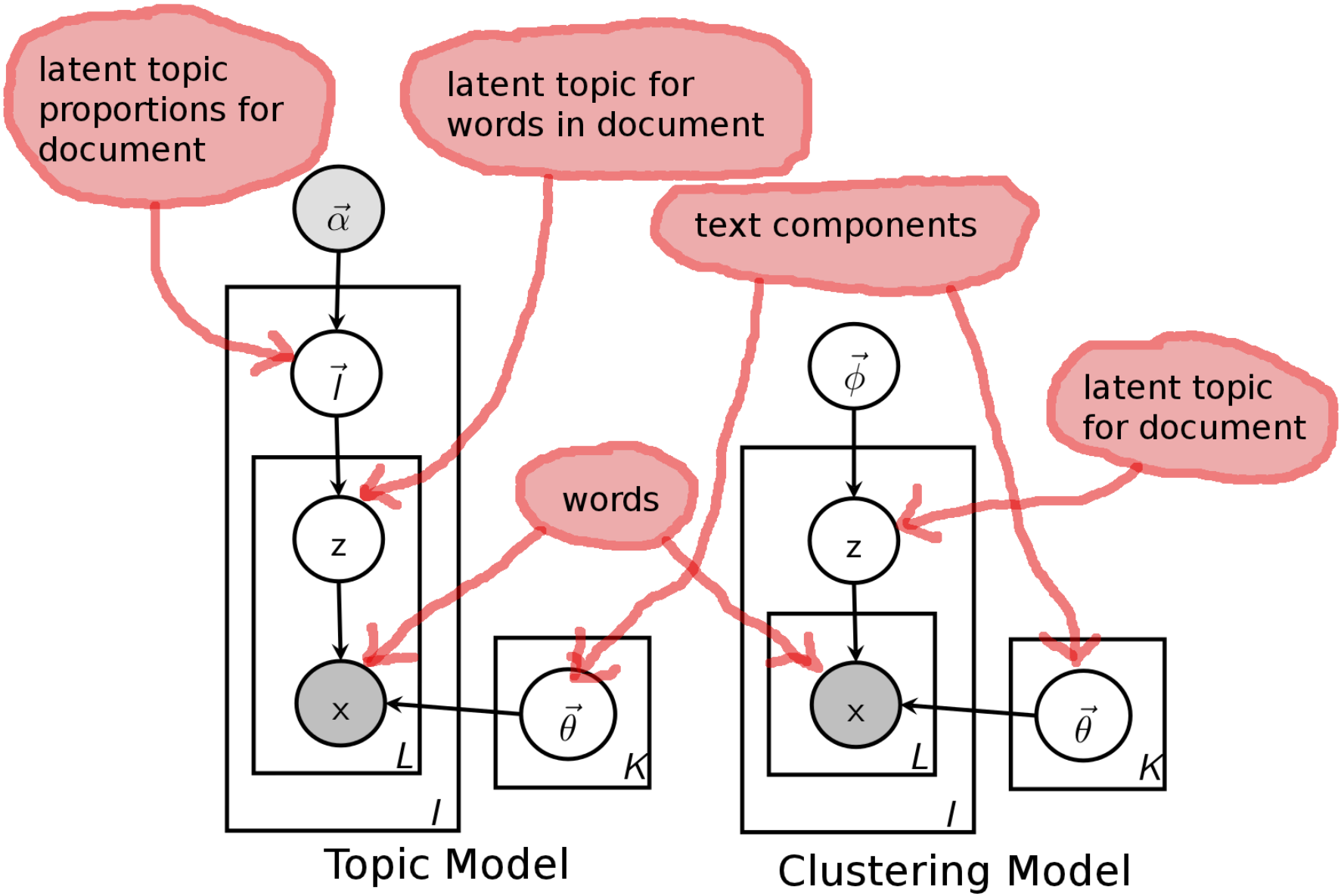
$$\boldsymbol{W} \simeq \boldsymbol{L} * \boldsymbol{\Theta}^T$$

$$\underbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}_{J \text{ words}} \Big\} \text{I documents} \simeq \underbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}_{K \text{ components}} * \underbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}_{J \text{ words}} \Big\} \text{K components}$$

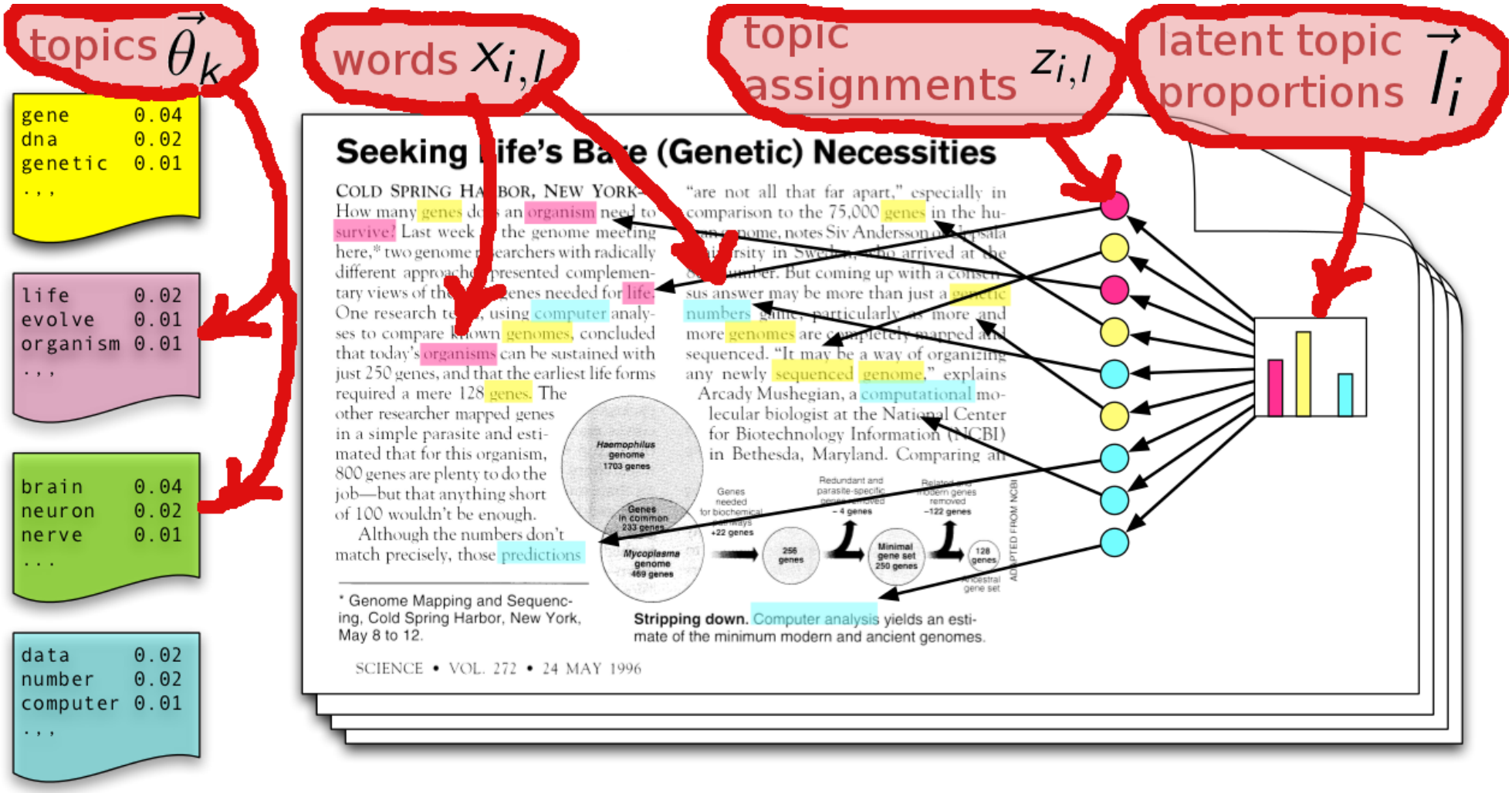
Different variants:

Data $\boldsymbol{W}$	Components $\boldsymbol{L}$	Error	Models
real valued	unconstrained	least squares	PCA and LSA
non-negative	non-negative	least squares	codebooks, NMF
non-neg integer	non-negative	cross-entropy	topic modelling, NMF
real valued	independent	small	ICA

LDA Topic Models versus Clustering

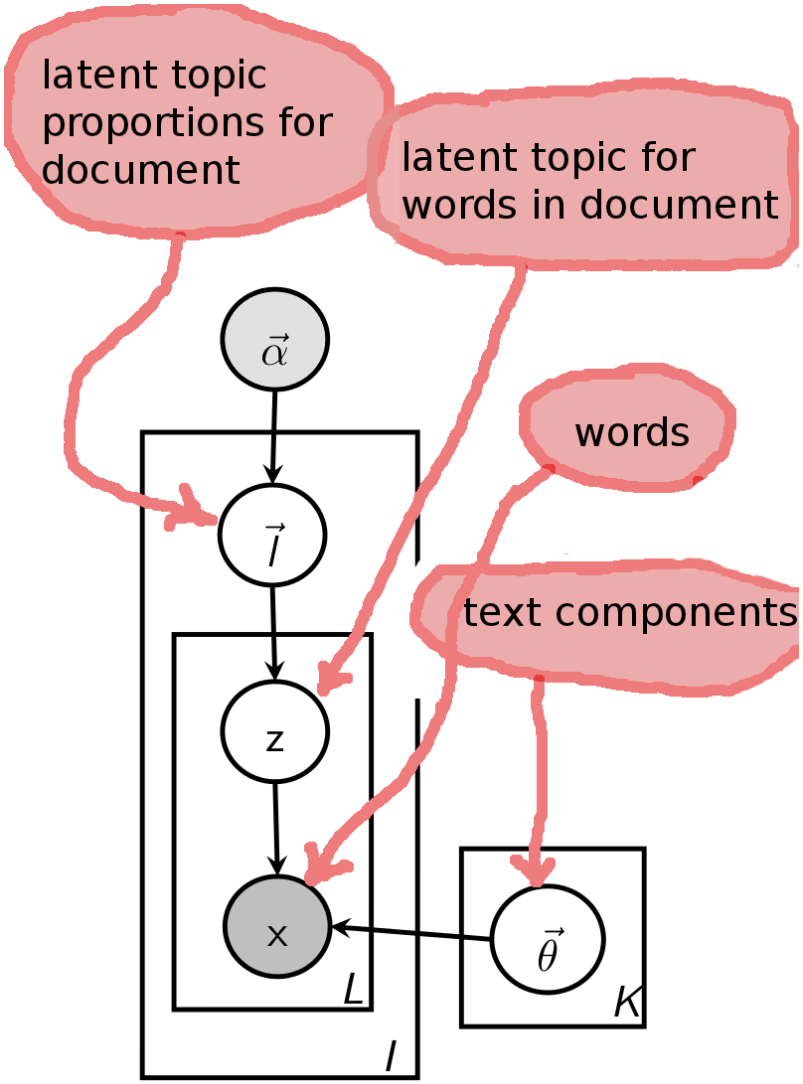


Clustering Words in Documents View



Blei's MLSS 2009 talk, with annotation by Wray.

LDA Topic Model



$$\vec{\theta}_k \sim \text{Dirichlet}_V(\vec{\gamma}) \quad \forall_{k=1}^{K} ,$$
$$l_i \sim \text{Dirichlet}_K(\vec{\alpha}) \quad \forall_{i=1}^{I} ,$$
$$z_{i,l} \sim \text{Discrete}\left(\vec{l}_i\right) \quad \forall_{i=1}^{I} \forall_{l=1}^{L_i} ,$$
$$x_{i,l} \sim \text{Discrete}\left(\vec{\theta}_{z_{i,l}}\right) \quad \forall_{i=1}^{I} \forall_{l=1}^{L_i} .$$

where

$$K := \#\text{ topics},$$
$$V := \#\text{ words},$$
$$I := \#\text{ documents},$$
$$L_i := \#\text{ words in doc } i$$

Collapsed LDA Inference

$$
\begin{aligned}
\vec{\theta}_k &\sim \text{Dirichlet}_V(\vec{\gamma}) && \forall_{k=1}^{K} \\
\vec{l}_i &\sim \text{Dirichlet}_K(\vec{\alpha}) && \forall_{i=1}^{I} \\
z_{i,l} &\sim \text{Discrete}\left(\vec{l}_i\right) && \forall_{i=1}^{I}\forall_{l=1}^{L_i} \\
x_{i,l} &\sim \text{Discrete}\left(\vec{\theta}_{z_{i,l}}\right) && \forall_{i=1}^{I}\forall_{l=1}^{L_i}
\end{aligned}
$$

where

$$
\begin{aligned}
K &:= \#\text{ topics}, \\
V &:= \#\text{ words}, \\
I &:= \#\text{ documents}, \\
L_i &:= \#\text{ words in doc } i
\end{aligned}
$$

The LDA posterior is collapsed by marginalising out $\forall_i \vec{l}_i$ and $\forall_i \vec{\theta}_i$:

$$
\prod_{i=1}^{I} \frac{\text{Beta}_K(\vec{\alpha} + \vec{m}_i)}{\text{Beta}(\vec{\alpha})} \prod_{k=1}^{K} \frac{\text{Beta}_J(\vec{\gamma} + \vec{n}_k)}{\text{Beta}(\vec{\gamma})}
$$

where

$$
\begin{aligned}
\vec{m}_i &:= \dim(K) \text{ data counts of topics for doc } i, \\
\vec{n}_k &:= \dim(V) \text{ data counts of words for topic } k.
\end{aligned}
$$

See the Dirichlet-multinomials!

So people have trouble making LDA hierarchical!

Summary: What You Need to Know

Dirichlet distribution: basic statistical unit for discrete data

Dirichlet properties: aggregration, symmetry, conjugacy

Graphical models: convey the structure of a probability model

Dirichlet-Multinomial: the evidence for a Dirichlet, a distribution itself

LDA topic model: a basic unsupervised component model

Outline

Conjugate Discrete Families

Conjugate Family	$p(x\|\lambda)$	$p(\lambda) \propto$
Bernoulli-Beta	$\lambda^x(1-\lambda)^{1-x}$	$\lambda^{\alpha-1}(1-\lambda)^{\beta-1}\delta_{0<\lambda<1}$
Poisson-Gamma	$\frac{1}{x!}\lambda^x e^{-\lambda}$	$\lambda^{\alpha-1}e^{-\beta\lambda}$
negtve-Binomial-Gamma	$\frac{1}{x!}(\lambda)_x\,\rho^x(1-\rho)^\lambda$	$\lambda^{\alpha-1}e^{-\beta\lambda}$

parameters $\alpha, \beta > 0$
$(\lambda)_x$ is rising factorial $\lambda(\lambda+1)...(\lambda+x-1)$

- multinomial-Dirichlet used in LDA
- Poisson-Gamma in some versions of NMF)
- Bernoulli-Beta is the basis of IBP
- negative-Binomial-Gamma is not quite conjugate; the negative-Binomial is a "robust" variant of a Poisson

ASIDE: Divisability of Discrete Families

Data can be split and merged across dimensions in some cases:

Poisson-Gamma: $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$ then $x_1 + x_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Bernoulli-Beta; $x_1 \sim \text{Bernoulli}(\lambda_1)$ and $x_2 \sim \text{Bernoulli}(\lambda_2)$ and x_1, x_2 mutually exclusive then $x_1 + x_2 \sim \text{Bernoulli}(\lambda_1 + \lambda_2)$.

negative-Binomial-Gamma: $x_1 \sim \text{NB}(\lambda_1, p)$ and $x_2 \sim \text{NB}(\lambda_2, p)$ then $x_1 + x_2 \sim \text{NB}(\lambda_1 + \lambda_2, p)$.

Infinite Vectors

- Let $\omega_k \in \Omega$ for $k = 1, ..., \infty$ be index points for an infinite vector.
- Have infinite parameter vector $\vec{\lambda} = \sum_{k=1}^{\infty} \lambda_k \delta_{\omega_k}$ for $\lambda_k \in (0, \infty)$,
- Generate I finite discrete feature vectors $\vec{x}_i = \sum_{k=1}^{\infty} x_{i,k} \delta_{\omega_k}$ pointwise using discrete distribution $p(x_{i,k}|\lambda_k)$.
- Intent is that only finite number of $x_{i,k} \neq 0$ for given i.
- This means we need $\sum_{k=1}^{\infty} \lambda_k < \infty$,
  - assuming lower λ_k makes $x_{i,k}$ more likely to be zero.
- Can arbitrarily rearrange dimensions k since all objects have term $\sum_{k=1}^{\infty}(\cdot)$.
- Don't know which dimensions k will have non-zero data so rearrange as needed.

See 2014 ArXiv paper by Lancelot James (`http://arxiv.org/abs/1411.2936`).

ASIDE: Improper Uniform Prior for Gaussian Data

- Suppose we want a "uniform" prior on location μ on the real line. A constant prior on μ must be over a finite domain:

$$p(\mu) \sim \frac{1}{2C} \qquad \text{for } \mu \in [-C, C]$$

- Consider data x_1 with Gaussian likelihood $p(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{(x-\mu)^2}$.
  - for C large enough, this is possible
- As $C \to \infty$, the corresponding posterior is

$$p(\mu|x_1) = \frac{1}{\sqrt{2\pi}} e^{(x_1-\mu)^2} .$$

- The limit of the constant prior as $C \to \infty$ is called an improper prior because:
  - it is the limit of a sequence of proper distributions,
  - by itself it is not a probability distribution, and
  - its posterior from any one data point is a proper distribution.

Generating Infinite Vectors

- Almost all λ_k must be infinitesimally small, so the prior for λ is not proper.
  - It is formally modelled using Poisson processes;
  - have Poisson process on domain $(0, \infty) \times \Omega$ with rate $p(\lambda|\omega)\mathrm{d}\lambda G(\mathrm{d}\omega)$ where $p(\lambda|\omega)$ may not normalise;
  - need $\int_0^\infty \int_\Omega p(\lambda|\omega)\mathrm{d}\lambda G(\mathrm{d}\omega) = \infty$ to have infinite number of points;
  - need $\int_0^\infty \int_\Omega \lambda p(\lambda|\omega)\mathrm{d}\lambda G(\mathrm{d}\omega) < \infty$ to expect $\sum_{k=1}^\infty \lambda_k < \infty$.
- Parameters to the "distribution" (Poisson process rate) for λ, $p(\lambda|\omega)$, would control the expected number of non-zero $x_{i,k}$.

Bernoulli-Beta Process (Indian Buffet Process)

- infinite Boolean vectors $\vec{x}_i$ with a finite number of 1's;
- each parameter λ_k is an independent probability,

$$p(x_{i,k}|\lambda_k) \;=\; \lambda_k^{x_{i,k}}(1-\lambda_k)^{1-x_{i,k}}$$

- to have finite 1's, require $\sum_k \lambda_k < \infty$
- improper prior (Poisson process rate) is the 3-parameter Beta process

$$p(\lambda|\alpha,\beta,\theta) \;=\; \theta\lambda^{-\alpha-1}(1-\lambda)^{\alpha+\beta-1}$$

(some versions add additional constants with θ)

- is in improper Beta because seeing "1" makes it proper:

$$\int_{\lambda=0}^{1} p(x=1|\lambda)p(\lambda)\mathrm{d}\lambda \;=\; \theta\,\mathrm{Beta}(1-\alpha,\alpha+\beta)$$

Outline

Conjugate Discrete Processes

Each conjugate family has a corresponding non-parametric version:

- Uses the improper versions of the prior $p(\lambda|\omega)$
  e.g. for Gamma, Beta, Dirichlet
- Want to generate a countably infinite number of λ but have almost all infinitesimally small.
- Theory done with Poisson processes, see 2014 ArXiv paper by Lancelot James (`http://arxiv.org/abs/1411.2936`).
- Presention here uses the more informal language of "improper priors," but the correct theory is Poisson processes.

Conjugate Discrete Processes, cont.

Non-parametric versions of models for discrete feature vectors:

Process Name	$p(x\|\lambda)$	$p(\lambda)$
Poisson-Gamma	$\frac{1}{x!}\lambda^x e^{-\lambda}$	$\theta\lambda^{-\alpha-1}e^{-\beta\lambda}$
Bernoulli-Beta	$\lambda^x(1-\lambda)^{1-x}$	$\theta\lambda^{-\alpha-1}(1-\lambda)^{\alpha+\beta-1}\delta_{0<\lambda<1}$
negtve-Binomial-Gamma	$\frac{1}{x!}(\lambda)_x\rho^x(1-\rho)^\lambda$	$\theta\lambda^{-\alpha-1}e^{-\beta\lambda}$

$\beta, \theta > 0$
$0 \leq \alpha < 1$

- In common they make the power of λ lie in $(-2, -1]$ to achieve the "improper prior" effect.
- Term θ is just a general proportion to uniformally increase number of λ_k's in any region.
- Whereas α and β control the relative size of the λ_k's.

Conjugate non-Parametric Discrete Families, cont.

- Given λ, probability of I samples with at least one non-zero entry is

$$\left(1 - p(x_{i,k} = 0|\lambda)^I\right) \ .$$

- By Poisson process theory, expectation of this (in general case)

$$\Psi_I = \int_\Omega \int_0^\infty \left(1 - p(x_{i,k} = 0|\lambda)^I\right) \rho(\lambda|\omega) \mathrm{d}\lambda G_0(\mathrm{d}\omega_k)$$

- Call Ψ_I the Poisson non-zero rate, a function of I and the underlying distributions.
- With I vectors, number of non-zero dimensions K is Poisson with rate Ψ_I, having probability

$$\frac{1}{K!} e^{-\Psi_I} \Psi_I^K \ .$$

Posterior Marginal

- With I vectors, number of non-zero dimensions K is Poisson with rate Ψ_I, having probability
$$\frac{1}{K!} e^{-\Psi_I} \Psi_I^K \ .$$
- Take particular dimension ordering (remove $\frac{1}{K!}$) and replace "not all zero" by actual data, $x_{i,1}, ..., x_{i,K}$ to get:
$$e^{-\Psi_I} \prod_{k=1}^{K} p(x_{1,k}, ..., x_{i,k}, \omega_w) \ .$$
- Expand using model to get posterior marginal:
$$p(\vec{x}_1, ..., \vec{x}_I, \vec{\omega}) = e^{-\Psi_I} \prod_{k=1}^{K} \left(\int_0^\infty \left(\prod_{i=1}^{I} p(x_{i,k}|\lambda) \right) \rho(\lambda|\omega) \mathrm{d}\lambda \right) G_0(\mathrm{d}\omega_k)$$

Outline

Bernoulli-Beta Process (Indian Buffet Process)

- The Poisson non-zero rate
  trick: use $1 - y^I = (1-y)\sum_{i=0}^{I} y^i$

$$\Psi_I \;=\; \theta\Gamma(1-\alpha)\sum_{i=0}^{I}\frac{\Gamma(\beta+\alpha+i)}{\Gamma(\beta+1+i)}\;.$$

- The marginal for the k-th dimension

$$\int_0^\infty \left(\prod_{i=1}^{I} p(x_{i,k}|\lambda)\right)\rho(\lambda|\omega)\mathrm{d}\lambda \;=\; \theta\,\mathrm{Beta}(c_k-\alpha, I-c_k+\alpha+\beta)$$

  where c_k is times dimension k is "on," so $c_k = \sum_{i=1}^{I} x_{i,k}$.
- Gibbs sampling $x_{i,k}$ is thus simple.
- Sampling parameters: posterior of θ is Poisson; posterior for β is log-concave so sampling "easier".

Poisson-Gamma Process

- The Poisson non-zero rate
  trick: use the Laplace exponent from Poisson process theory

$$\Psi_I = \theta \frac{\Gamma(1-\alpha)}{\alpha} \left((I+\beta)^\alpha - \beta^\alpha \right) .$$

- The marginal for the k-th dimension

$$\int_0^\infty \left(\prod_{i=1}^I p(x_{i,k}|\lambda) \right) \rho(\lambda|\omega) \mathrm{d}\lambda = \theta \left(\prod_{i=1}^I \frac{1}{x_{i,k}!} \right) \frac{\Gamma(x_{\cdot,k} - \alpha)}{(I+\beta)^{x_{\cdot,k}-\alpha}}$$

  where $x_{\cdot,k} = \sum_{i=1}^I x_{i,k}$.
- Gibbs sampling the $x_{i,k}$ is thus simple.
- Sampling parameters: posterior of θ is Poisson; posterior of β is unimodal (and no other turning points) with simple closed form for MAP.

Negative-Binomial-Gamma Process

- Series of papers for this case by Mingyuan Zhou and colleages.
- The Poisson non-zero rate
  trick: use the Laplace exponent from Poisson process theory

$$\Psi_l = \theta \frac{\Gamma(1-\alpha)}{\alpha} \left(\left(l \log \left(\frac{1}{1-p} \right) + \beta \right)^{\alpha} - \beta^{\alpha} \right) .$$

- The marginal for the k-th dimension

$$\int_0^{\infty} \left(\prod_{i=1}^{I} p(x_{i,k}|\lambda, p) \right) \rho(\lambda|\omega) \mathrm{d}\lambda$$

$$= p^{x_{\cdot,k}} \left(\prod_{i=1}^{I} \frac{1}{x_{i,k}!} \right) \int_0^{\infty} (1-p)^{I\lambda} \left(\prod_{i=1}^{I} (\lambda)_{x_{i,k}} \right) \rho(\lambda) \mathrm{d}\lambda$$

- Gibbs sampling the $x_{i,k}$ is more challenging.
  - keep λ as a latent variable (posterior is log concave);
  - use approximation $(\lambda)_x \approx \lambda^{t^*} S^x_{t^*,0}$ where $t^* = \mathrm{argmax}_{t \in [1,x)} \lambda^t S^x_{t,0}$.

Example Constructed Discrete Family

James' more general theory allows more creativity in construction.

Bernoulli-Beta-Beta process

- model is a hierarchy of Bernoulli-Beta processes
- infinite feature vector $\vec{\lambda}$ is a Beta Process as before;
- these varied with point-wise Beta distributions to create a set of parent nodes $\vec{\psi}_j$, so $\psi_{j,k} \sim \text{Beta}(\alpha\lambda_{j,k}, \alpha(1-\lambda_{j,k}))$
- discrete features ordered in a hierarchy below nodes j so $x_{i,k} \sim \text{Bernoulli}(\psi_{j,k})$ for j the parent of node i.
- Use hierarchical Dirichlet process techniques to implement efficiently.

Outline

Definition: Species Sampling Model

Definition of a species sampling model

Have a probability vector $\vec{p}$ (so $\sum_{k=1}^{\infty} p_k = 1$), and a domain Θ and a countably infinite sequence of elements $\{\theta_1, \theta_2, ...\}$ from Θ.
A species sampling model (SSM) draws a sample θ according to the distribution

$$p(\theta) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\theta) .$$

- sample θ_k with probability p_k
- if $\forall_k \theta \neq \theta_k$, then $p(\theta) = \sum_k p_k 0 = 0$
- if $\forall_{k\,:\,k\neq l} \theta_l \neq \theta_k$, then $p(\theta_l) = \sum_k p_k \delta_{k=l} = p_l$

Species Sampling Model, cont.

SSM defined as:

$$p(\theta) \;=\; \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\theta) \;.$$

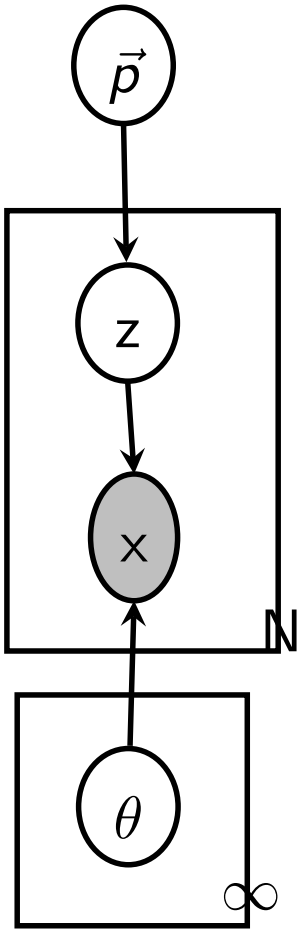
- the indices themselves in $(\sum_{k=1}^{\infty} \cdot)$ are irrelevant, so for any renumbering σ,

$$p(\theta) \;=\; \sum_{k=1}^{\infty} p_{\sigma(k)} \delta_{\theta_{\sigma(k)}}(\theta) \;;$$

- to create an SSM, one needs a sequence of values θ_k
  - usually we generate these independently according to some base distribution (usually $H(\cdot)$) so $\theta_k \sim H(\cdot)$
- to create an SSM, one also needs a vector $\vec{p}$;
  - this construction is where all the work is!

Using an SSM for a Mixture Model

Classic MM



On the left

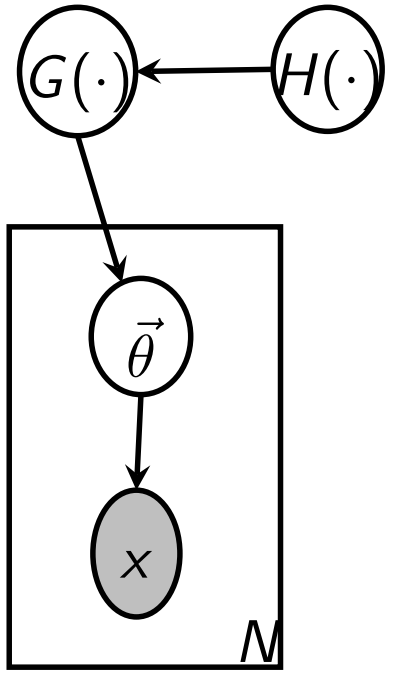
$$
\begin{aligned}
\vec{p} &\sim \text{SSM-p}(\cdot) \\
\vec{\theta}_k &\sim H(\cdot) && \forall_{k=1,\ldots,K} \\
z_n &\sim \vec{p} && \forall_{n=1,\ldots,N} \\
x_n &\sim f\left(\vec{\theta}_{z_n}\right) && \forall_{n=1,\ldots,N}
\end{aligned}
$$

Versus, on the right

$$
\begin{aligned}
G(\cdot) &\sim \text{SSM}\left(H(\cdot)\right) \\
\vec{\theta}_n &\sim G(\cdot) && \forall_{k=1,\ldots,N} \\
x_n &\sim f\left(\vec{\theta}_n\right) && \forall_{n=1,\ldots,N}
\end{aligned}
$$

where $G(\vec{\theta})$ is an SSM, including a vector $\vec{p}$.

SSM MM



Infinite Probability Vectors $\vec{p}$

- The number of $p_k > \delta$ must be less than $(1/\delta)$.

  e.g. there can be no more than 1000 p_k greater than 0.001.
- The value of p_{58153} is almost surely infinitessimal.
  - and for $p_{9356483202}$, *etc.*
- But some of the p_k must be larger and significant.
- It is meaningless to consider a p_k without:
  - defining some kind of ordering on indices,
  - only considering those greater than some δ, **or**
  - ignoring the indices and only considering the partitions of data induced by the indices.

ASIDE: Schemes for Generating $\vec{p}$

There are general schemes (but also more) for sampling infinite probability vectors:

Normalised Random Measures: sample an independent set of weights w_k (a "random measure") using for instance, a Poisson process, and then normalise, $p_k = \frac{w_k}{\sum_{k=1}^{\infty} w_k}$.

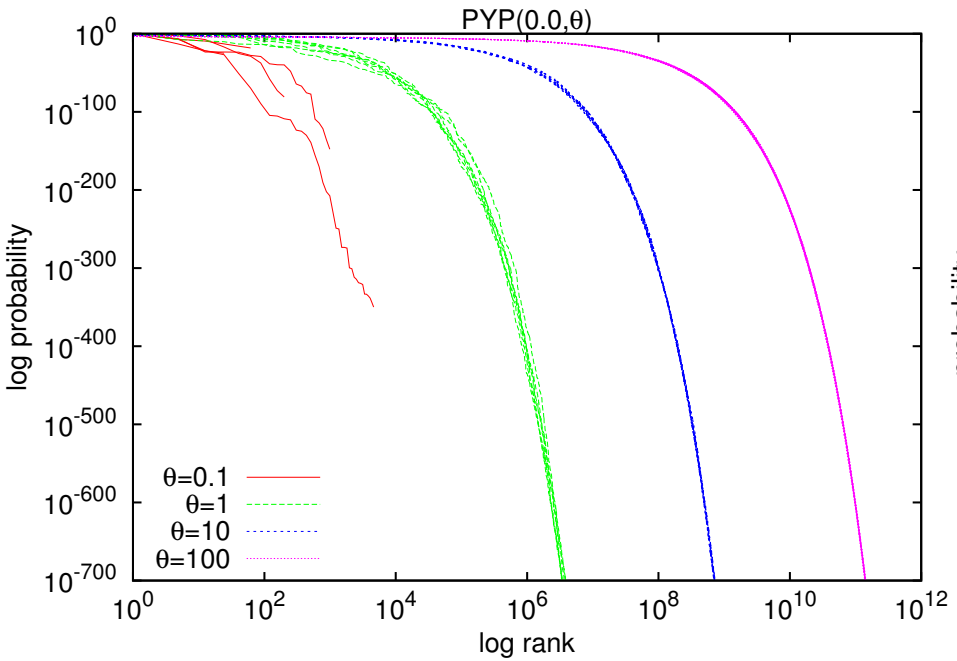
Predictive Probability Functions: generalises the famous "Chinese Restaurant Process" we will cover later. See Lee, Quintana, Müller and Trippa (2013).

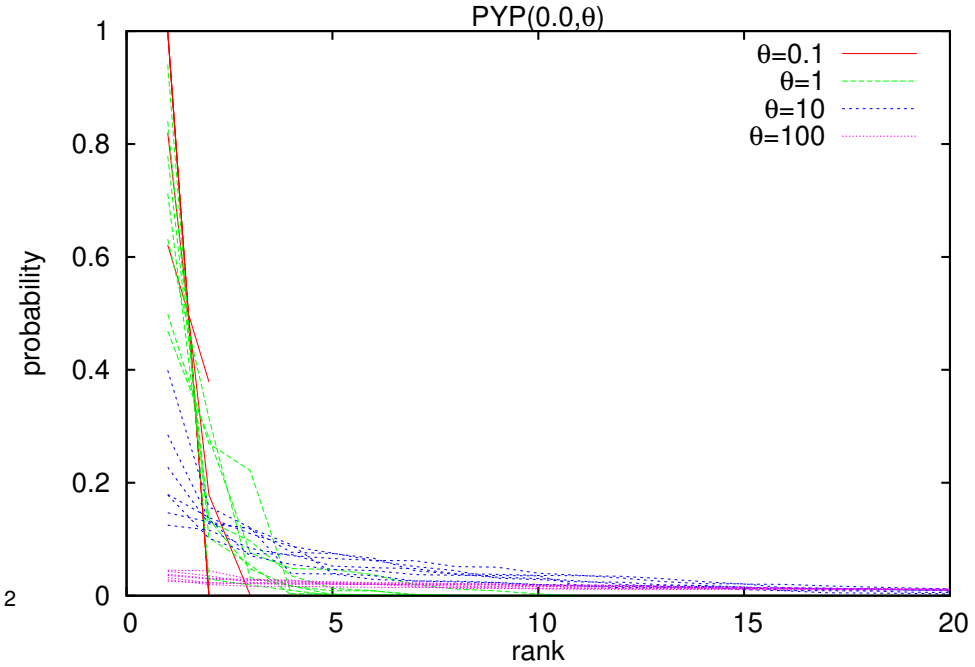
Stick-Breaking Construction: commonly used definition for the Pitman-Yor process we will consider later. See Ishwaran and James (2001).

Outline

Ranked plots of sampled DP probability vectors

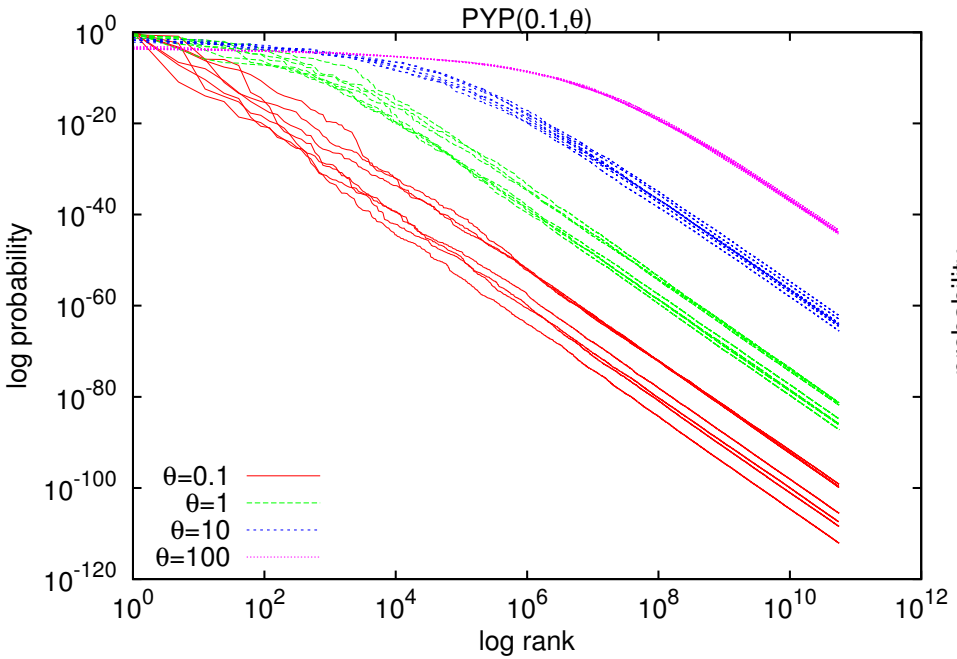
The probabilities in $\vec{p}$ are ranked in decreasing order and then plotted.

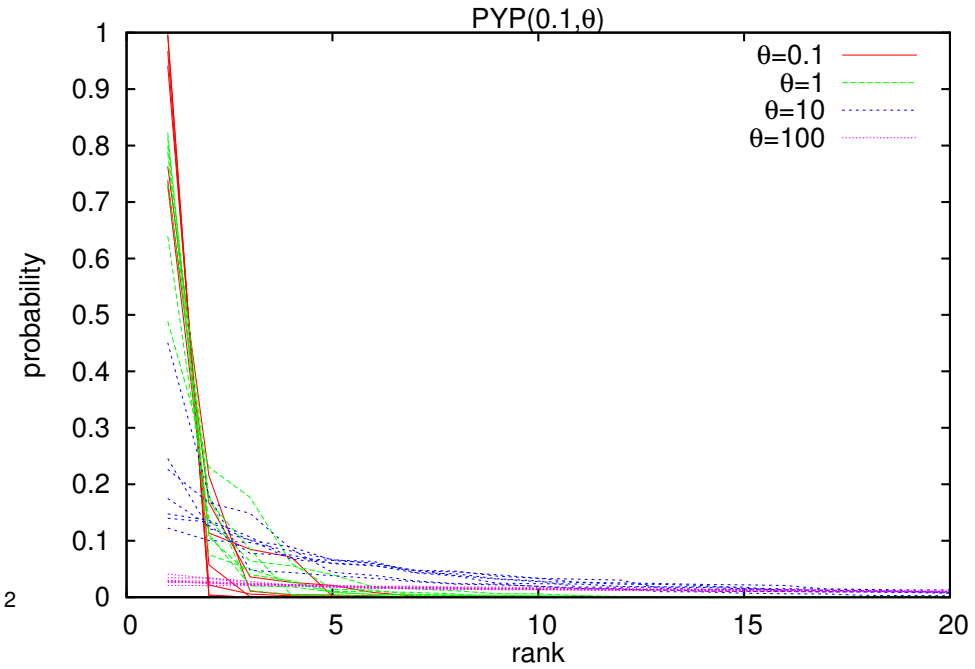




Note: how separated the different plots are for different concentration

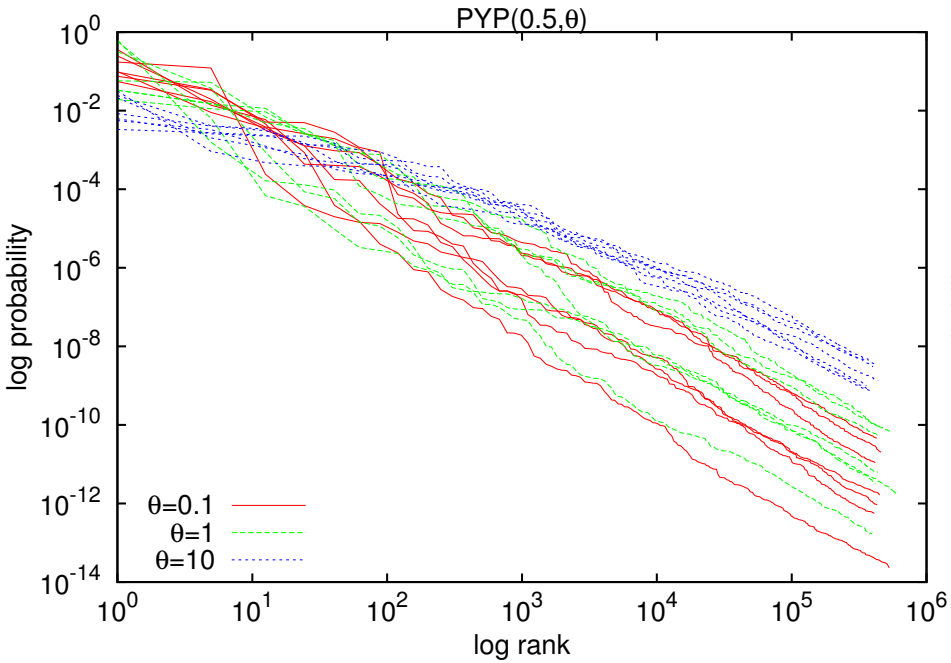
Ranked plots of sampled PYP probability vectors with $d = 0.1$

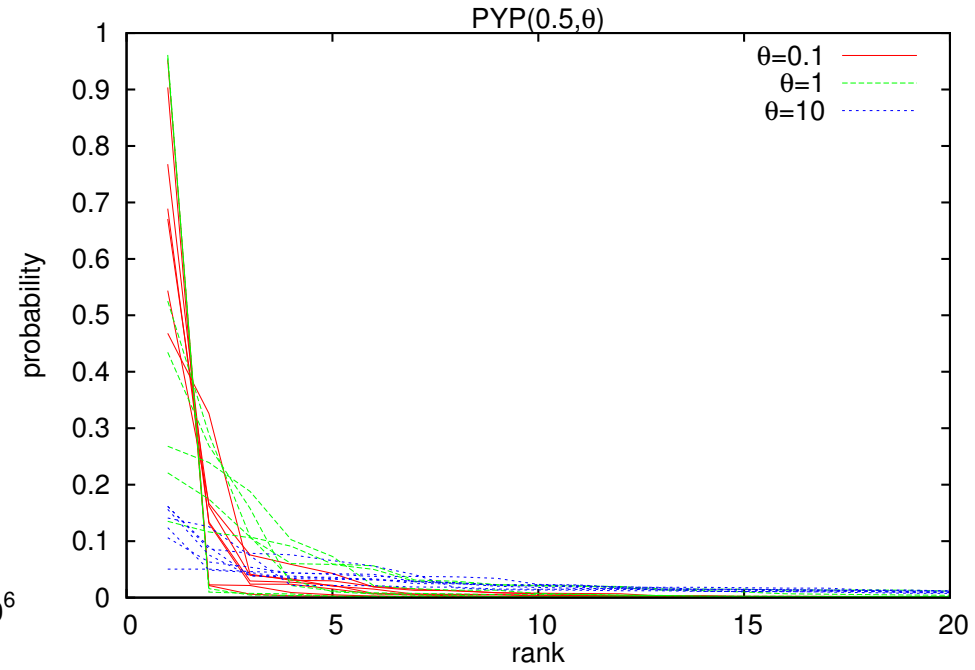




Note: how separated the different plots are for different concentration

Ranked plots of sampled PYP probability vectors with $d = 0.5$





Note: separation not as clear for different plots for different concentration

Sneak Peak: DP and PYP behaviour

The DP and PYP samples look like $\sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\cdot)$.

- with DP$(\alpha, H(\cdot))$, the $\vec{p}$ probabilities, when rank ordered, behave like a geometric series $p_k \tilde{\propto} \frac{1}{e^{k/\alpha}}$.
- with PYP$(d, \alpha, H(\cdot))$, the $\vec{p}$ probabilities, when rank ordered, behave like a Dirichlet series $p_k \tilde{\propto} \frac{1}{k^{1/d}}$.
- probabilities for the PYP are thus Zipfian (somewhat like Zipf's law$^{\mathcal{W}}$) and converge slower.

Outline

Partitions

Definition of partition

A partition of a set[W] P of a countable set X is a mutually exclusive and exhaustive set of non-empty subsets of X. The partition size of P is given by the number of sets $|P|$.

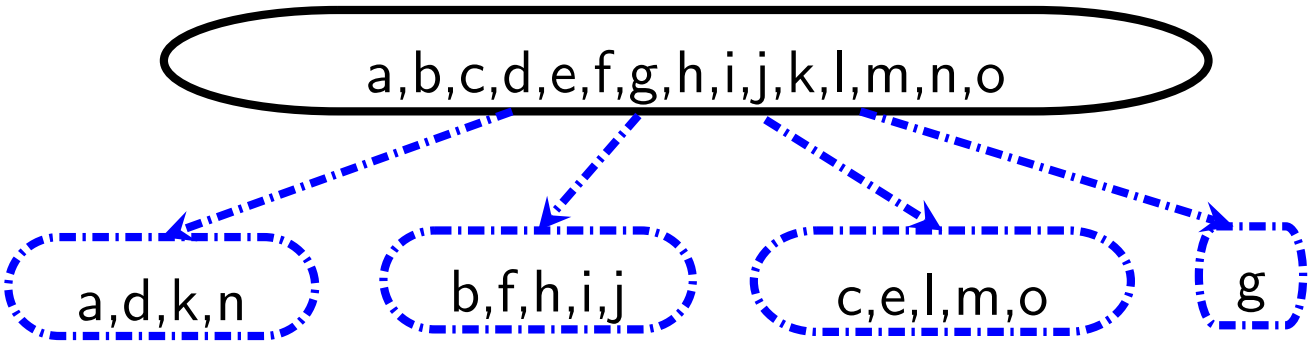
Consider partitions of the set of letters $\{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o\}$:

Candidate	Legality
{a,d,k,n},{b,f,h,i,j},{c,e,l,m,o},{g}	OK
{a,d,k,n},{b,f,h,i,j},{c,e,l,m,o},{g,k}	no, 'k' duped
{a,d,k,n},{b,f,h,i,j},{c,e,l,m,o}	no, not exhaustive
{a,d,k,n},{b,f,h,i,j},{c,e,l,m,o},{g},{}	no, an empty set

Partitions over $\{a, b, c\}$

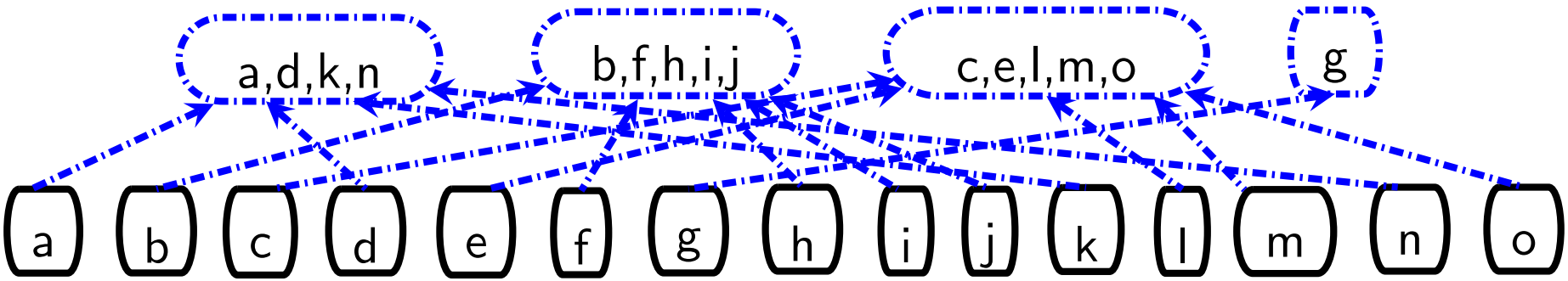
partition P	$\{\{a, b, c\}\}$	$\{\{a, b\}, \{c\}\}$	$\{\{a, c\}, \{b\}\}$	$\{\{a\}, \{b, c\}\}$	$\{\{a\}, \{b\}, \{c\}\}$
indices	$(1, 1, 1)$	$(1, 1, 2)$	$(1, 2, 1)$	$(1, 2, 2)$	$(1, 2, 3)$
size $\|P\|$	1	2	2	2	3
counts $\vec{n}$	(3)	$(2, 1)$	$(2, 1)$	$(1, 2)$	$(1, 1, 1)$

Partitioning a Set in a Node: Fragmentation



We make a new node for each set in the partition.

Partitioning a Set of Leaves: Coagulation

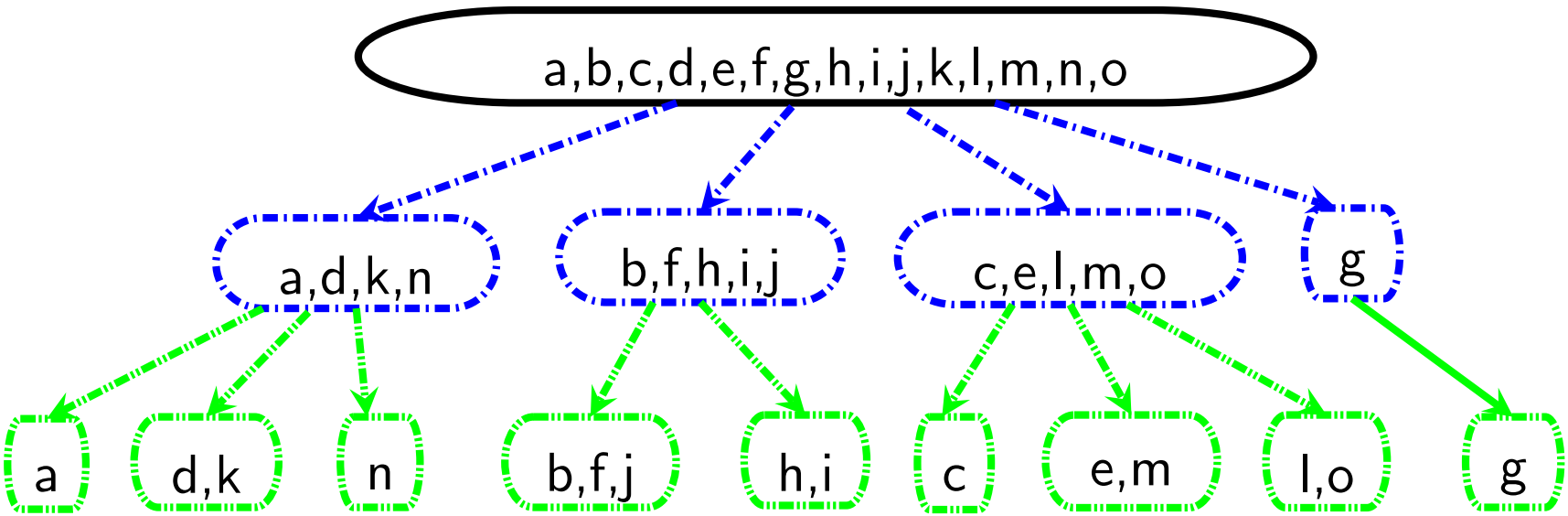


We make a new node for each set in the partition.

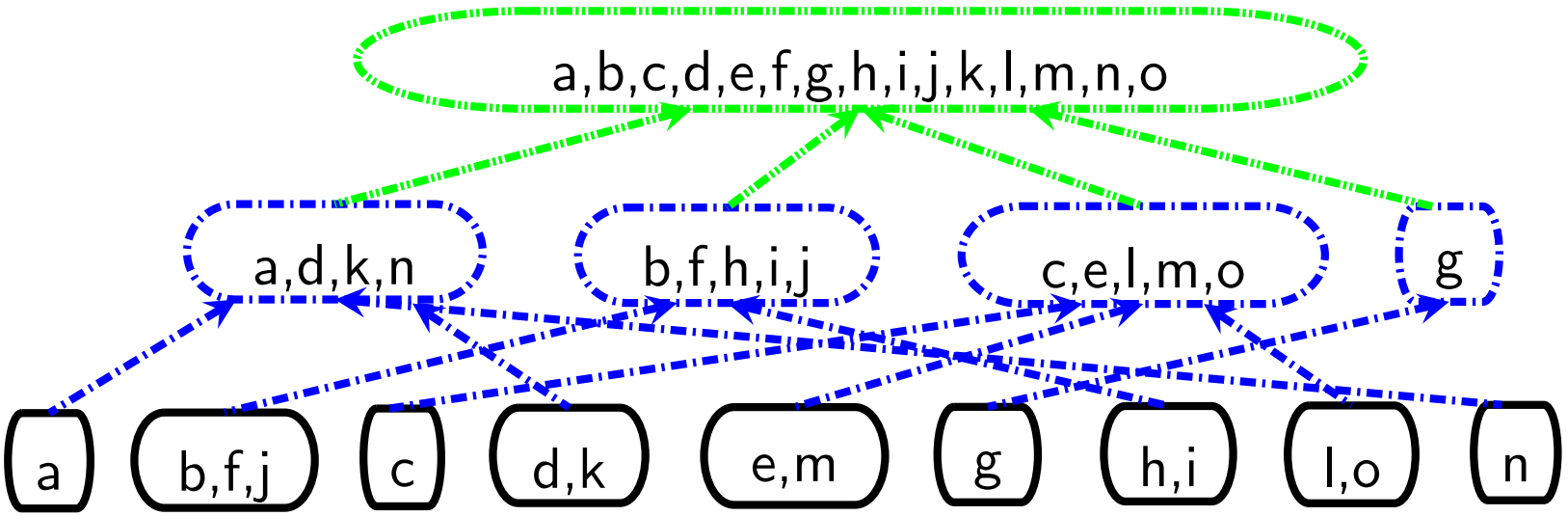
ASIDE: Distributions on Trees

Any probability distribution on partitions can be used to generate trees, top-down (fragmentation) or bottom-up (coagulation).

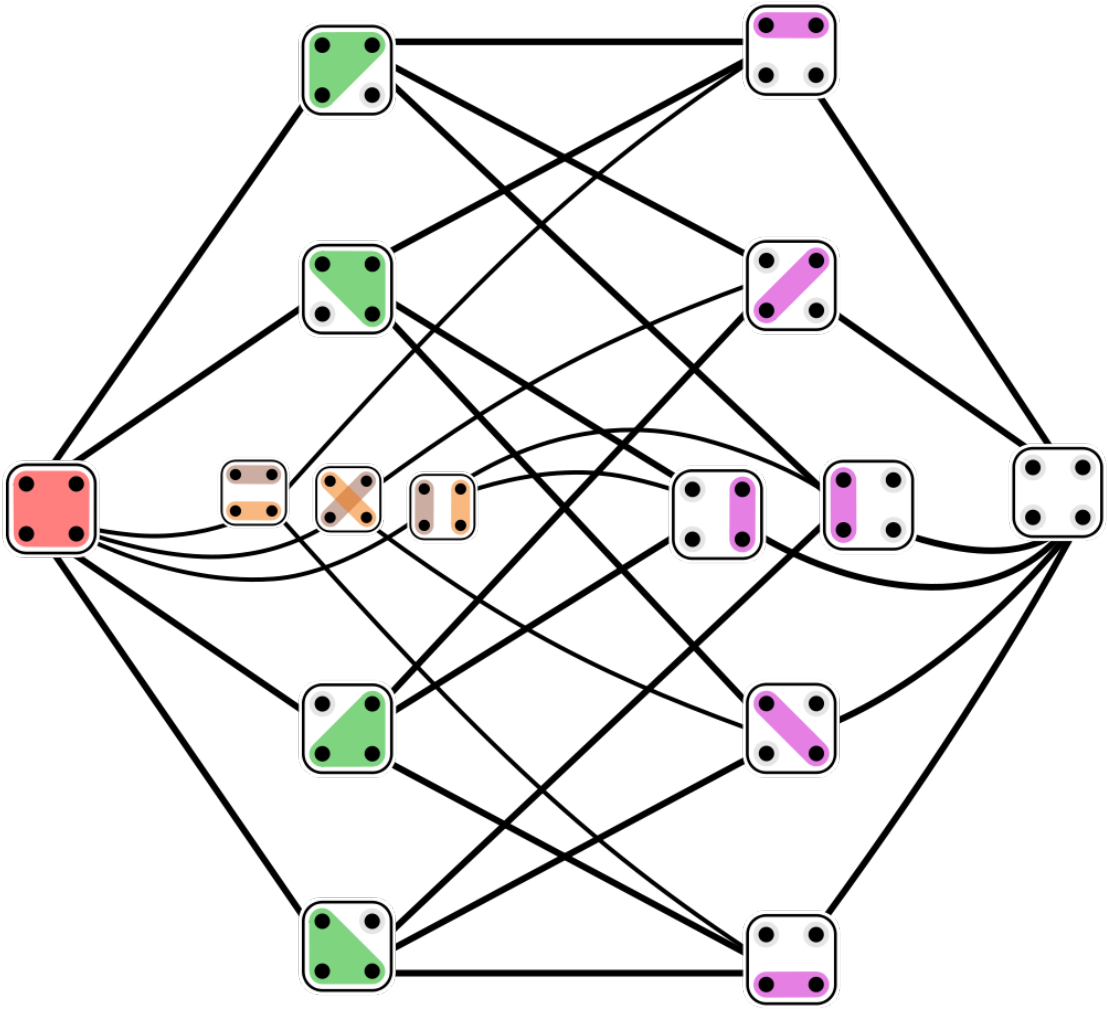
ASIDE: Building a Tree Top-down: Repeated Fragmentation



ASIDE: Building a Tree Bottom-up: Repeated Coagulation



All partitions of 4 objects



Note: space of partitions forms a lattice.

ASIDE: Stirling number of the second kind

Definition of Stirling number of the second kind

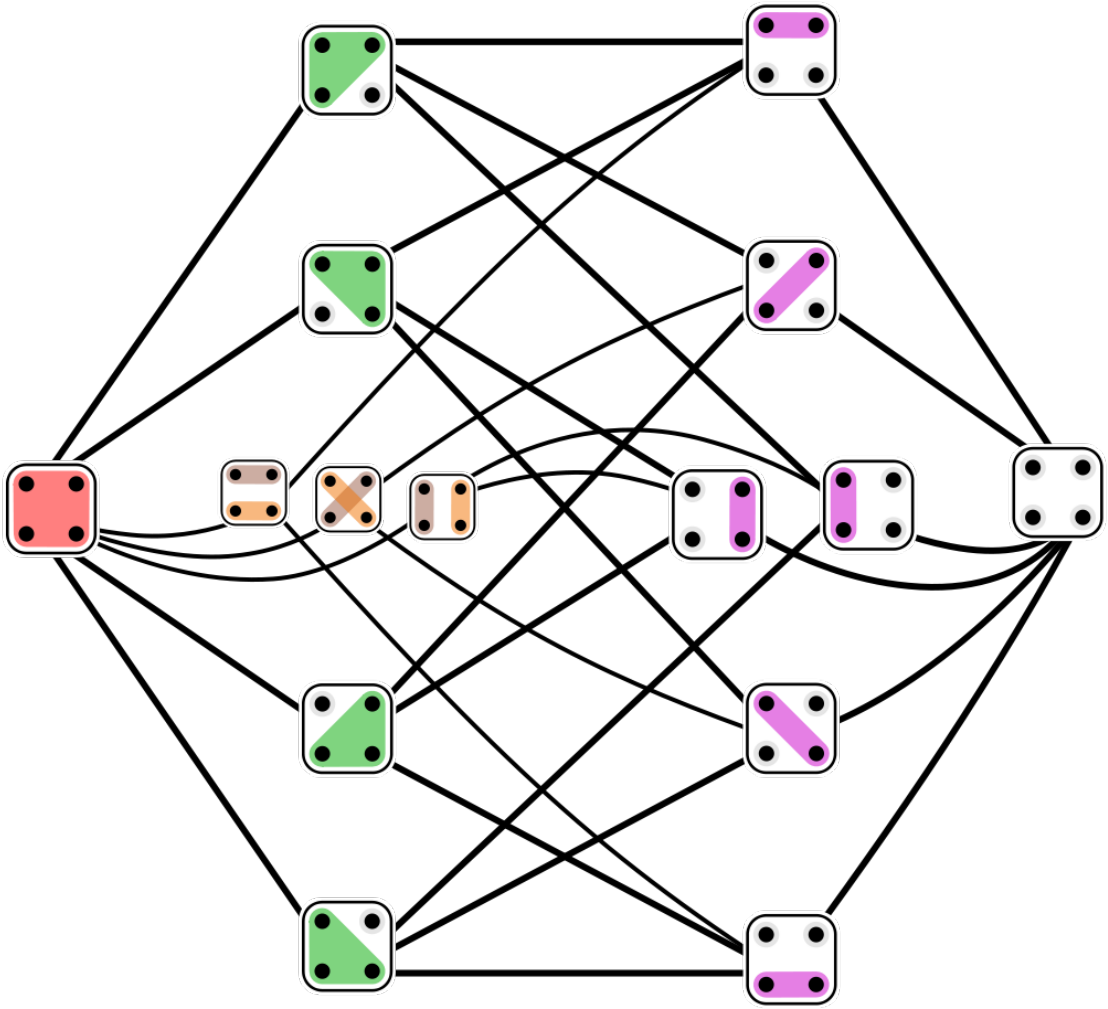
A Stirling number of the second kind[W] is the number of ways of partitioning a set of n objects into k nonempty subsets, denoted by $\left\{ {n \atop k} \right\}$ ("n subset k").

- Stirling numbers of the second kind occur in the field of mathematics called combinatorics and the study of partitions.
- They count the number of different equivalence relations with precisely k equivalence classes definable on an n element set.
- Special cases:

$$\left\{ {n \atop 1} \right\} = \left\{ {n \atop n} \right\} = 1 \qquad \left\{ {n \atop 0} \right\} = \delta(n,0) \qquad \left\{ {n \atop 2} \right\} = \frac{1}{2!}(2^n - 2)$$

$$\left\{ {n \atop 3} \right\} = \frac{1}{3!}(3^n - 3 \times 2^n + 3) \qquad \ldots \qquad \left\{ {n \atop n-1} \right\} = \binom{n}{2}$$

ASIDE: Stirling number of the second kind



$$\left\{ {4 \atop 4} \right\} = 1 \qquad \left\{ {4 \atop 2} \right\} = 7 \qquad \left\{ {4 \atop 3} \right\} = 6 \qquad \left\{ {4 \atop 4} \right\} = 1$$

ASIDE: Stirling number of the second kind, Recurrence

- Recurrence

$$\left\{ {n \atop k} \right\} = \underbrace{k \left\{ {n-1 \atop k} \right\}}_{\text{Case 1}} + \underbrace{\left\{ {n-1 \atop k-1} \right\}}_{\text{Case 2}}$$

  - Case 1: put the last object together with some nonempty subset of the first $n-1$ objects, there are $k\left\{ {n-1 \atop k} \right\}$ ways.
    $\Rightarrow$ note: for each partition, there are k ways to put the last object.
  - Case 2: put the last object into a class by itself, there are $\left\{ {n-1 \atop k-1} \right\}$ ways.
  - Without the factor of k, it would reduce to the addition formula (binomial identity)

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

A Sample of a SSM Induces a Partition

Suppose we have a sample of size $N = 12$ taken from an infinite mixture (for simplicity, we'll label data as 'a', 'b', ...):

a,c,a,d,c,d,a,b,g,g,a,b

This can be represented as follows:

1,2,1,3,2,3,1,4,5,5,1,4

where index mappings are: a=1, c=2, d=3, b=4, g=5.

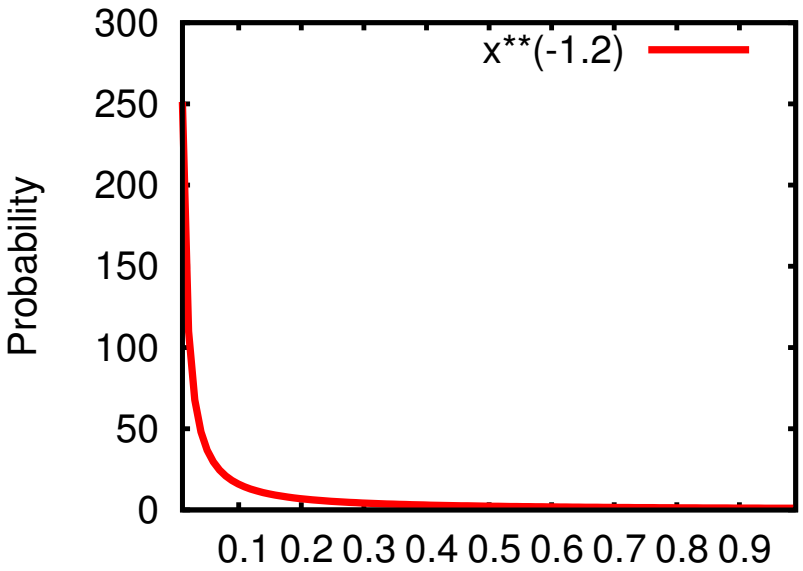
- The sample induces a partition of N objects.
- Index mappings can be arbitrary, but by convention we index data as it is first seen as 1,2,3,...
- This convention gives the size-biased ordering for the partition,
  - because the first data item seen is more likely to have the largest p_k, the second data item seen is more likely to have the second largest p_k, *etc.*

Outline

Simple Improper Dirichlet Prior on 2D

Suppose $\vec{p} \sim \text{Dir}_2(-d, 1)$ for $0 < d < 1$.

$$p(\vec{p}) \propto p_1^{-d-1}(1-p_1)^{1-1} = p_1^{-(d+1)}$$



- cannot be normalised
- goes to infinity at $p_1 = 0$
- so effectively, first data sampled must be first dimension
- posterior is then $\text{Dir}_2(1-d, 1)$
- is an improper prior

Improper (Infinite) Dirichlet Prior

Consider a finite piece of $\vec{p}$ of length K, $\vec{P}_K = (p_{i_1}, p_{i_2}, ...p_{i_K}, q_K)$, where q_K is the remainder, $q_K = 1 - \sum_{k=1}^{K} p_{i_k}$, and a Dirichlet of the form $\vec{P}_K \sim \text{Dir}_{K+1}(-d, -d, ..., -d, \alpha + Kd)$ for $0 < d < 1$ and $\alpha > -d$.

Then

$$p(p_{i_1}, p_{i_2}, ..., p_{i_K}, q_K) \propto \prod_{k=1}^{K} p_{i_k}^{-d-1} \left(1 - \sum_{k=1}^{K} p_{i_k}\right)^{\alpha+K\,d-1}$$

- looks like a Dirichlet but with illegal parameters for p_{i_k}. *i.e.*, the exponent for p_{i_k} is ≤ -1
- is not a probability distribution (cannot be normalised)
- all the usual Dirichlet properties hold: symmetry, aggregation, *etc.*

Improper Dirichlet Prior, cont.

Finite piece of $\vec{p}$ of length K, $\vec{P}_K = (p_{i_1}, p_{i_2}, ...p_{i_K}, q_K)$, where q_K is the remainder, $q_K = 1 - \sum_{k=1}^{K} p_{i_k}$ with a Dirichlet of the form $\vec{P}_K \sim \text{Dir}_{K+1}(-d, -d, ..., -d, \alpha + Kd)$.

- is improper in the sense that:
  - if a single data point is seen in all but the last dimension, the posterior becomes proper: $\text{Dir}_{K+1}(1-d, 1-d, ..., 1-d, \alpha + Kd)$.
  - it is the limit of proper priors (by constraining all $p_{i_k} > \delta$ as $\delta \to 0$).
- is self consistent: any sub-splice has an equivalent form of prior.

  for $\{j_1, ..., j_L\} \subset \{i_1, ., , , , .i_K\}$, change of variables from $\vec{P}_K$ to $\vec{P}_L = (p_{j_1}, p_{j_2}, ...p_{j_L}, q_L)$ then $\vec{P}_L \sim Dir_{L+1}(-d, -d, ..., -d, \alpha + Ld)$.

- is thus well behaved prior — full theory for this presented in Buntine and Hutter 2012.

Posterior Sampling with the Improper Dirichlet

- There is $N = 0$ and $\vec{P}_0 = (q_0) \sim \text{Dir}_1(\alpha)$.
  - Sample "remainder" as only choice. Improper prior does not say which index we get. External forces return i_1, but we don't care.
- There is $N = 1$ and $\vec{P}_1 = (p_{i_1}, q_1) \sim \text{Dir}_2(1-d, \alpha+d)$.
  - Sample"remainder" again, but with probability $\frac{\alpha+d}{\alpha+1}$. External forces return i_2, but again we don't care.
- There is $N = 2$ and $\vec{P}_2 = (p_{i_1}, p_{i_2}, q_2) \sim \text{Dir}_3(1-d, 1-d, \alpha+2d)$.
  - Sampling, get i_1 again, but with probability $\frac{1-d}{\alpha+2}$.
- There is $N = 3$ and $\vec{P}_2 \sim \text{Dir}_3(2-d, 1-d, \alpha+2d)$.
  - Suppose immediately get "remainder", with probability $\frac{\alpha+2d}{\alpha+3}$.
- There is $N = 4$ and $\vec{P}_3 \sim \text{Dir}_4(2-d, 1-d, 1-d, \alpha+3d)$.
  - Sampling, get i_2, then i_1 again, with probability $\frac{1-d}{\alpha+4}$ and $\frac{2-d}{\alpha+5}$.
- There is $N = 6$ and $\vec{P}_3 \sim \text{Dir}_4(3-d, 2-d, 1-d, \alpha+3d)$.

Posterior Sampling with the Improper Dirichlet, cont.

N	K	datum	posterior	probability
0	0	remainder (i_1)	$\mathrm{Dir}_1(\alpha)$	1
1	1	remainder (i_2)	$\mathrm{Dir}_2(1-d, \alpha+d)$	$\frac{\alpha+d}{1+\alpha}$
2	2	i_1	$\mathrm{Dir}_3(1-d, 1-d, \alpha+2d)$	$\frac{1-d}{2+\alpha}$
3	2	remainder (i_3)	$\mathrm{Dir}_3(2-d, 1-d, \alpha+3d)$	$\frac{\alpha+d}{3+\alpha}$
4	3	i_2	$\mathrm{Dir}_4(2-d, 1-d, 1-d, \alpha+3d)$	$\frac{1-d}{4+\alpha}$
5	3	i_1	$\mathrm{Dir}_4(2-d, 2-d, 1-d, \alpha+3d)$	$\frac{2-d}{5+\alpha}$
6	3		$\mathrm{Dir}_4(3-d, 2-d, 1-d, \alpha+3d)$	

posterior on observed indices $\mathrm{Dir}_K(n_1-d, n_2-d, ..., n_K-d, \alpha+K\,d)$

probability of datum $(N+1)$ being k-th index is $\frac{n_k-d}{N+\alpha}$;

probability of datum $(N+1)$ being remainder (a new index) is $\frac{\alpha+K\,d}{N+\alpha}$;

Evidence for the Improper Dirichlet

- only keeps places for observed indices; unobserved indices are kept "unfurled" in the $K+1$-th remainder term;
- is a prior on partitions, indices are irrelevant;
- by induction, with N data and K distinct indices, the posterior for seeing sample counts $(n_{i_1}, n_{i_2}, ..., n_{i_K})$

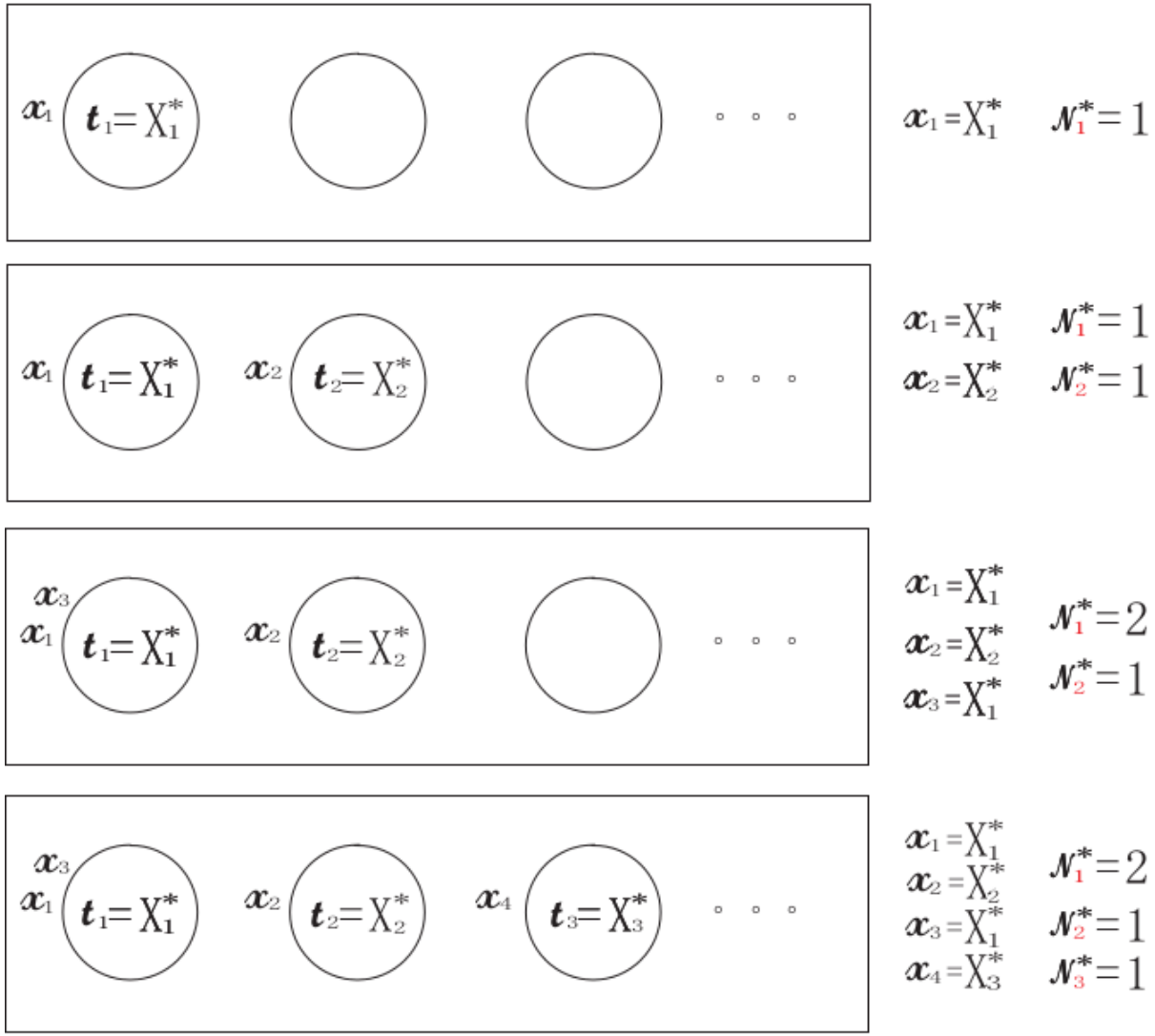
$$\frac{\alpha(\alpha+d)...(\alpha+(K-1)d)}{\alpha(\alpha+1)...(\alpha+(N-1))}\prod_{k=1}^{K}(1-d)...(n_k-1-d)$$

- introducing the rising factorial or Pochhammer symbol[W] $(x|y)_n = x(x+y)...(x+(n-1)y)$, and $(x)_n = (x|1)_n$ giving

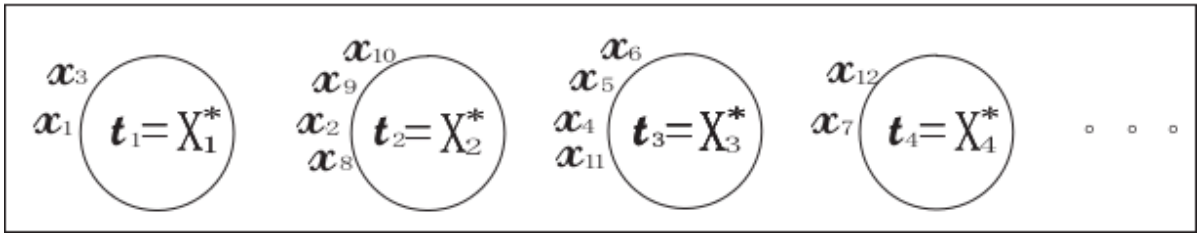
$$p(\text{observed indices}\,|\,N, \text{ImpDir}, d, \alpha) = \frac{(\alpha|d)_K}{(\alpha)_N}\prod_{k=1}^{K}(1-d)_{(n_k-1)}$$

Outline

"Chinese Restaurant" Sampling



Chinese Restaurant Process (CRP)

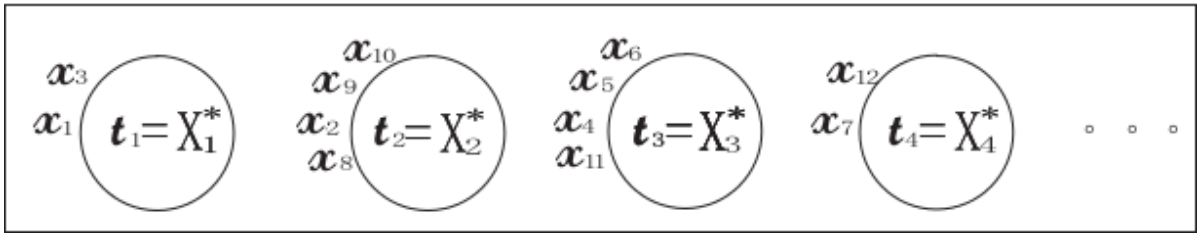


Definition of a Chinese restaurant process

A sample from a Chinese restaurant process$^{\mathcal{W}}$ (CRP) with parameters $(d, \alpha, H(\cdot))$ is generated as follows (we use the two-parameter version):

1. First customer enters the restaurant, sits at the first empty table and picks a dish according to $H(\cdot)$.
2. Subsequent customer numbered $N+1$:
   1. with probability $\frac{\alpha + K\,d}{N+\alpha}$ (K is count of existing tables) sits at an empty table and picks a dish according to $H(\cdot)$;
   2. otherwise, joins an existing table k with probabiliity proportional to $\frac{n_k - d}{N+\alpha}$ (n_k is the count of existing customers at the table) and has the dish served at that table.

CRP Terminology



Restaurant: single instance of a CRP, roughly like a Dirichlet-multinomial distribution.

Customer: one data point.

Table: cluster of data points sharing the one sample from $H(.)$.

Dish: the data value corresponding to a particular table; all customers at the table have this "dish".

Table count: number of customers at the table.

Seating plan: full configuration of tables, dishes and customers.

ASIDE: Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2006: Teh develops hierarchical n-gram models using PYs.

2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes, *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

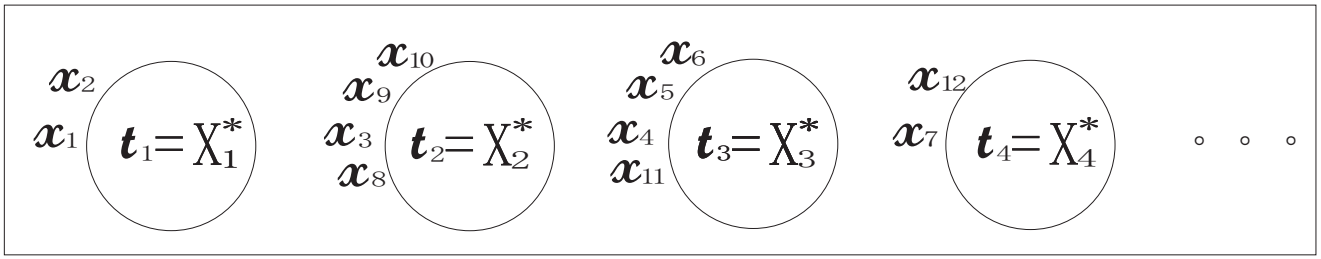
- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process (Wood and Teh, 2009),
- *etc.*

Opened up whole field of non-parametrics, but generally regarded as slow and unrealistic for scaling up

CRP Example

CRP with base distribution $H(\cdot)$:

$$p(x_{N+1} \mid x_{1:N}, d, \alpha, H(\cdot)) = \frac{\alpha + Kd}{N+\alpha} H(x_{N+1}) + \sum_{k=1}^{K} \frac{n_k - d}{N+\alpha} \delta_{X_k^*}(x_{N+1}) ,$$



$$p(x_{13} = X_1^* \mid x_{1:12}, \ldots) = \frac{2-d}{12+\alpha} \qquad p(x_{13} = X_2^* \mid x_{1:12}, \ldots) = \frac{4-d}{12+\alpha}$$

$$p(x_{13} = X_3^* \mid x_{1:12}, \ldots) = \frac{4-d}{12+\alpha} \qquad p(x_{13} = X_4^* \mid x_{1:12}, \ldots) = \frac{2-d}{12+\alpha}$$

$$p(x_{13} = X_5^* \mid x_{1:12}, \ldots) = \frac{\alpha + 4d}{12+\alpha} H(X_5^*)$$

CRP, cont.

$$p(x_{N+1}|x_1,...,x_N,d,\alpha,H(\cdot)) = \frac{K\,d+\alpha}{N+\alpha}H(x_{N+1}) + \sum_{k=1}^{K}\frac{n_k-d}{N+\alpha}\delta_{X_k^*}(x_{N+1})$$

- is like Laplace sampling, but with a discount $(-d)$ instead of $1/2$ offset
- doesn't define a vector $\vec{p}$; the CRP is exchangable, de Finetti's theorem on exchangable observations proves a vector $\vec{p}$ must exist
- exactly the same process as we saw in posterior sampling with the improper Dirichlet

Evidence for the CRP

It does show how to compute evidence only when $H(\cdot)$ is non-discrete.

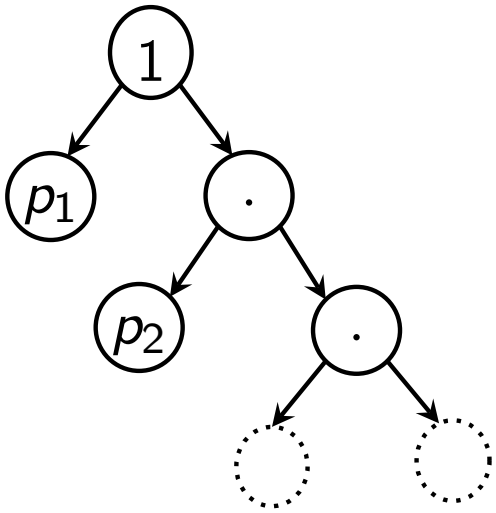
$$
\begin{aligned}
& p(x_1, ..., x_N | N, CRP, d, \alpha, H(\cdot)) \\
&= \frac{\alpha(d+\alpha)...((K-1)d+\alpha)}{\alpha(1+\alpha)...(N-1+\alpha)} \prod_{k=1}^{K} ((1-d)...(n_k-1-d) H(X_k^*)) \\
&= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(X_k^*) \ ,
\end{aligned}
$$

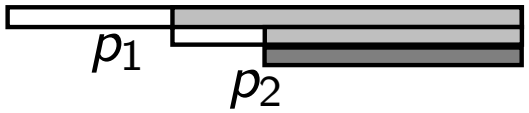
where there are K distinct data values X_1^*, ...X_K^*.

Same as the evidence for the Improper Dirichlet, but with a data probability term $H(X_k^*)$ added.

Outline

ASIDE: Stick Breaking

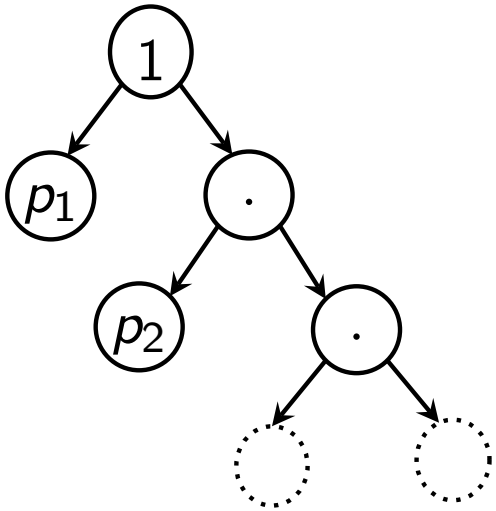


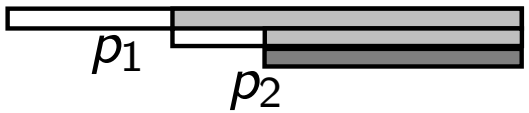


$1 \quad (1 - V_1) \quad (1 - V_1)(1 - V_2)$

- break piece, take remainder, break piece, take remainder, repeat!
- repeatedly sample a proportion using a Beta distribution to break a piece off the remainder stick of length $\left(1 - \sum_{j=1}^{k-1} p_j\right)$
- due to the construction, remainder sticks vanish quickly so p_k usually get vanishingly small as k grows
- see Ishwaran and James (2001)

ASIDE: Stick Breaking, cont





$1 \quad (1 - V_1) \quad (1 - V_1)(1 - V_2)$

The infinite probability vector $\vec{p}$ can be generated incrementally:

$$
\begin{aligned}
p_1 &= V_1 && \text{where } V_1 \sim \text{Beta}(1-d, \alpha+d) \\
p_2 &= V_2\,(1-p_1) && \text{where } V_2 \sim \text{Beta}(1-d, \alpha+2d) \\
p_3 &= V_3\,(1-p_1-p_2) && \text{where } V_3 \sim \text{Beta}(1-d, \alpha+3d) \\
p_k &= V_k \left(1 - \sum_{j=1}^{k-1} p_j\right) && \text{where } V_k \sim \text{Beta}(1-d, \alpha+k\,d)
\end{aligned}
$$

ASIDE: Definition of the GEM

Definition of the GEM distribution

The infinite probability vector $\vec{p}$ generated by the stick breaking construction with discount $d = 0$ and concentration α is said to be from the GEM(α) distribution.

- GEM stands for Griffiths, Engen and McCloskey.
- use GEM(d, α) for the obvious extension with a discount, and it is the "stick breaking" distribution.
- theory shows the GEM has a size-biased ordering

i.e. p_k are sorted according to first occurence in some sample.

ASIDE: Evidence for the GEM

Data for the GEM are in the form of indices for a size-biased ordering.
e.g. 1,1,2,3,2,3,1,4,3,5,2,...

The distribution has a clear form as a set of independent Betas, so one can, with difficulty, obtain the evidence.

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\theta + (k-1)d + \sum_{l=k}^{K} n_l}$$

where n_k is number of occurrences of index k, $n_k > 0$.

Same as the evidence for the Improper Dirichlet, but with a final penalty term.

ASIDE: Evidence for the GEM is Different

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\theta + (k-1)d + \sum_{l=k}^{K} n_l}$$

- final penalty term corresponds to the choice of ordering of indices
- is approximated by:

$$\frac{n_1}{N} \frac{n_2}{N-n_1} \frac{n_3}{N-n_1-n_2} \frac{n_4}{N-n_1-n_2-n_3} \cdots$$

i.e. (probability type 1 will be seen first) *times* (probability type 2 will be seen second given that type 1 is seen first) *times* ...

ASIDE: Historical Context for Stick-breaking + GEM

2001: Ishwaran and James publish the general stick breaking scheme and sampling methods

2008: Teh, Furikawa and Welling develop a variational stick breaking scheme for a HDP version of topic models

2008-2014: used widely in variational approaches with HDPs

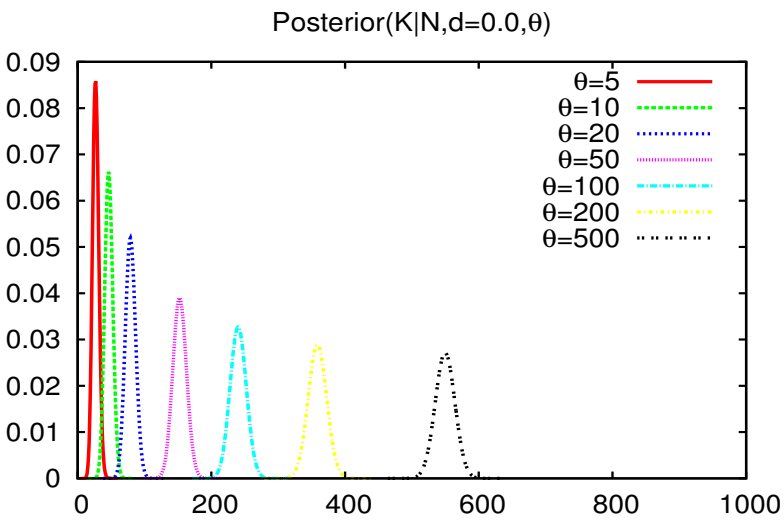
Implications of the difference between the GEM and the CRP unexplored, though some authors have reported improvements by periodically sorting the GEM.

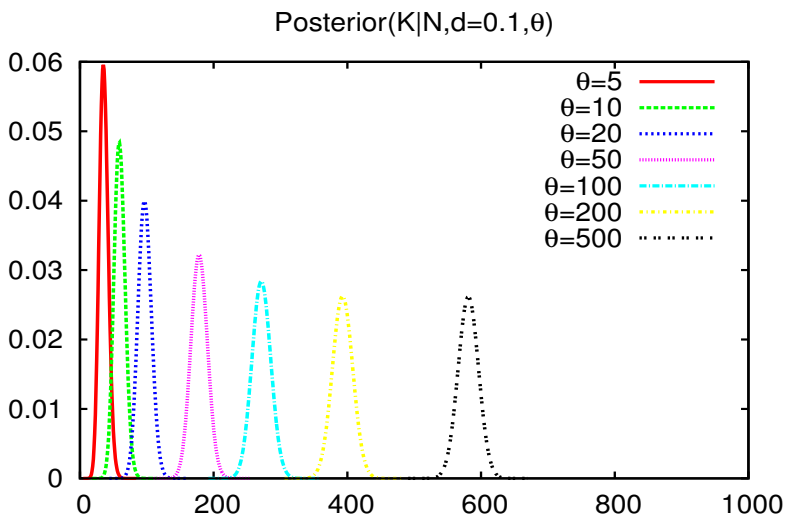
Outline

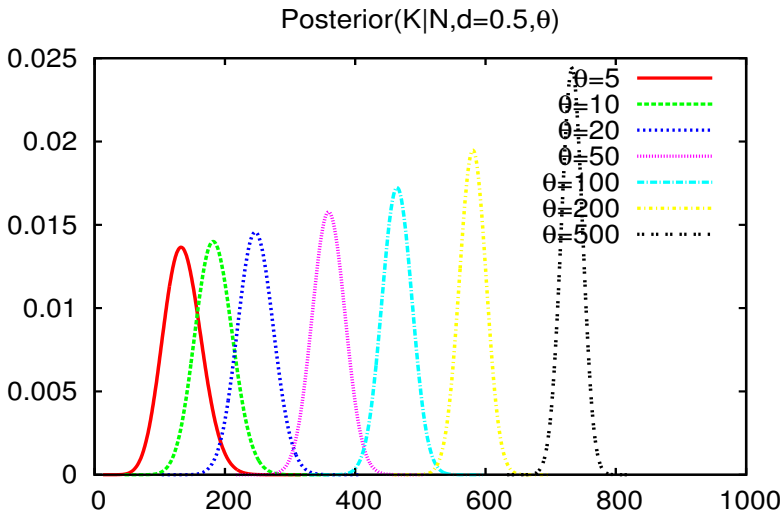
How Many Species/Tables are There?

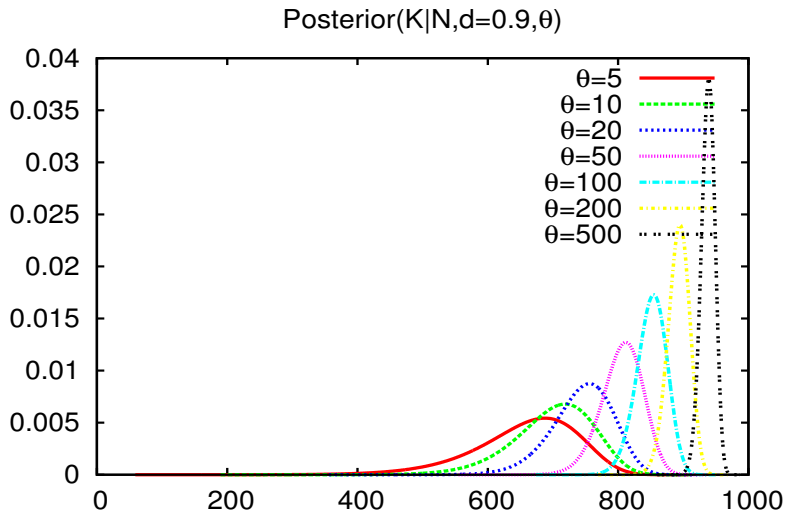
- Consider just the case where the sample $\vec{x}$ of size N has just K species (or tables for the CRP), denoted as $|\vec{x}| = K$.
- What is the expected distribution on K?
- Easily sampled using the CRP.
- Can also be found in closed form.

How Many Species/Tables are There?









Posterior probability on K given $N = 1000$ and different d, θ.

Number of Species/Tables

- The number of species/tables varies dramatically depending on the discount and the concentration.
- Is approximately Gaussian with a smallish standard-deviation.
- In applications, we should probably sample discount and/or the concentration.
- Note concentration has fast effective samplers, but sampling discount is slow.

Improper Dirichlet Posterior

Consider just the case where the sample $\vec{x}$ of size N has just K species, denoted as $|\vec{x}| = K$.

$$p(|\vec{x}| = K|d, \alpha, N, \text{ImpDir}) \quad \propto \quad \frac{(\alpha|d)_K}{(\alpha)_N} \sum_{\vec{x}\,:\,|\vec{x}|=K} \prod_{k=1}^{K} (1-d)_{n_k-1}$$

$$\text{Define} \qquad \mathcal{S}^N_{K,d} \quad := \quad \sum_{\vec{x}\,:\,|\vec{x}|=K} \prod_{k=1}^{K} (1-d)_{n_k-1} \ ,$$

$$\text{then} \quad p(|\vec{x}| = K|d, \alpha, N, \text{ImpDir}) \quad = \quad \frac{(\alpha|d)_K}{(\alpha)_N} \mathcal{S}^N_{K,d} \ .$$

The $\mathcal{S}^N_{K,d}$ is a generalised Stirling number of the second kind, with many nice properties. Is easily tabulated so $O(1)$ to compute.

See Buntine & Hutter 2012, and for code the MLOSS project `libstb`.

ASIDE: Generalised Stirling Number of the Second Kind

Definition for Generalised Stirling numbers of the second kind

Generalised Stirling numbers of the second kind $\mathcal{S}^n_{k,<\alpha,\beta,r>}$ (Hsu and Shiue, 1998) are characterised by

$$[x|\alpha]_n = \sum_{k=0}^{n} \mathcal{S}^n_{k,<\alpha,\beta,r>} [x - r|\beta]_k$$

where $[x|y]_n$ is the generalised falling factorial

$$[x|y]_n = x(x-y)(x-2y)\dots(x-(n-1)y)$$

ASIDE: Generalised Stirling Number of the Second Kind, cont.

- Recurrence formula for the generalised Stirling number (Hsu and Shiue, 1998)

$$\mathcal{S}_k^{n+1} = \mathcal{S}_{k-1}^n + (k\beta - n\alpha + r)\mathcal{S}_k^n$$

- If $\alpha = 0$, $\beta = 1$ and $r = 0$, we have the recurrence formula for the Stirling number of the second kind

$$\mathcal{S}_k^{n+1} = \mathcal{S}_{k-1}^n + k\mathcal{S}_k^n$$

- For the CRD, the Stirling number considered is when $\alpha = -1$, $\beta = -d$ and $r = 0$: $\mathcal{S}_{k,<-1,-d,0>}^n$ $(\mathcal{S}_{k,d}^n)$. The recurrence formula is

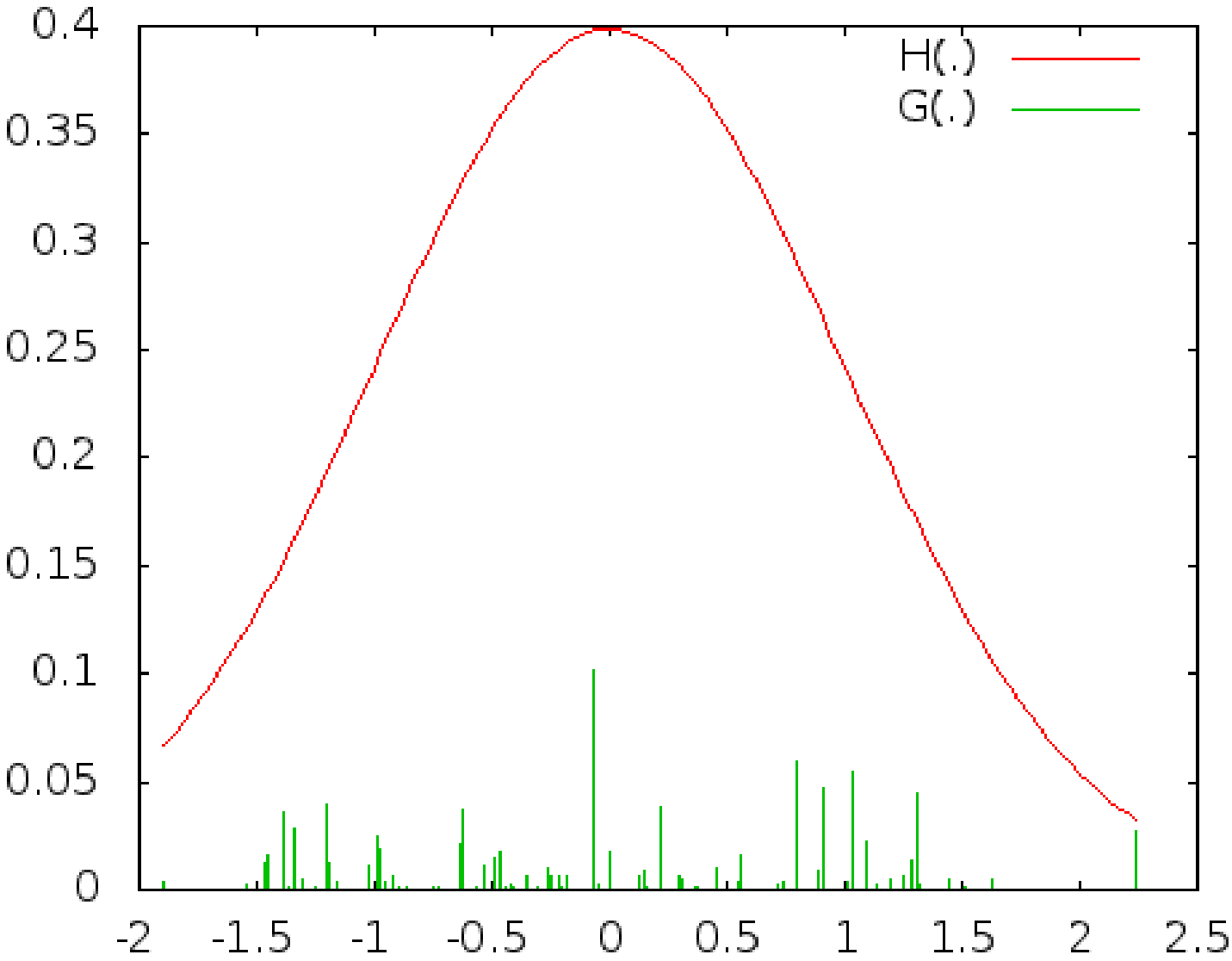
$$\mathcal{S}_{k,d}^{n+1} = \mathcal{S}_{k-1,d}^n + (n - kd)\mathcal{S}_{k,d}^n$$

Outline

PYP and DP

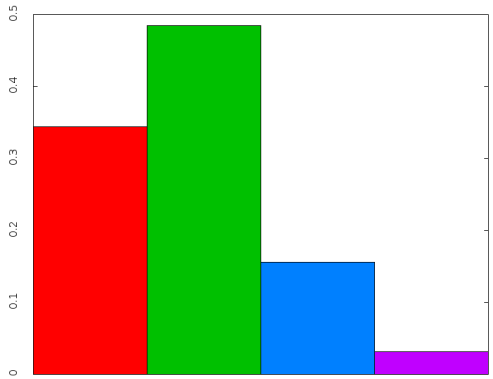
- The Pitman-Yor Process (PYP) has three arguments PYP $(d, \alpha, H(\cdot))$ and the Dirichlet Process (DP) has two arguments DP $(\alpha, H(\cdot))$:
  - discount d is the Zipfian slope for the PYP.
  - Concentration α is inversely proportional to variance.
  - Base distribution $H(\cdot)$ that seeds the distribution and is the mean.
  - *e.g.*, as $\alpha \to \infty$, a sample from them gets closer to $H(\cdot)$.
- They return an SSM, $p(\theta) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\theta)$, where θ_k are independently and identically distributed according to the base distribution $H(\cdot)$.
- They return a distribution on the same space as the base distribution (hence are a functional).
  - fundamentally different depending on whether $H(\cdot)$ is discrete or not.
- PYP originally called "two-parameters Poisson-Dirichlet process" (Ishwaran and James, 2003).

Example: $G(\cdot) \sim \mathrm{DP}(1, \mathrm{Gaussian}(0, 1))$

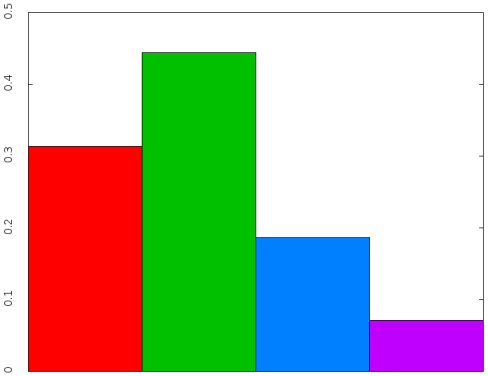


Example: DP on a 4-D vector

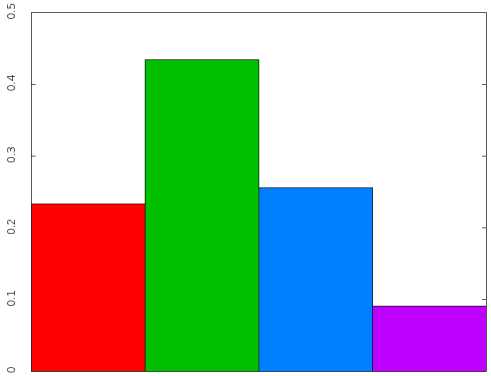
$\vec{p}_0$



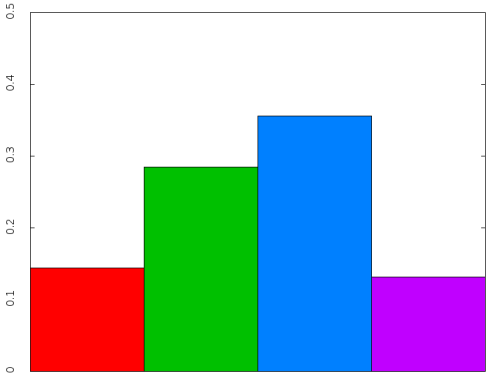
$\vec{p}_1 \sim \text{DP}(500, \vec{p}_0)$



$\vec{p}_2 \sim \text{DP}(5, \vec{p}_0)$



$\vec{p}_3 \sim \text{DP}(0.5, \vec{p}_0)$



Definition

There are several ways of defining a Pitman-Yor Process$^{\mathcal{W}}$ of the form PYP$(d, \alpha, H(\cdot))$.

- Generate a $\vec{p}$ with a GEM(d, α), then form an SSM by independently sampling $\theta_k \sim H(\cdot)$ (Pitman and Yor, 1997).
- Generate a $\vec{p}$ with an ImpDir(d, α), then form an SSM by independently sampling $\theta_k \sim H(\cdot)$.
- Proport its existence by saying it has posterior sampler given by a CRP$(d, \alpha, H(\cdot))$.

There is another way of defining a Dirichlet Process$^{\mathcal{W}}$ of the form DP$(\alpha, H(\cdot))$.

- As a natural extension to the Dirichlet in non-discrete or countably infinite domains (see "formal definition" in Wikipedia).

Dirichlet Process

- When applied to a finite probability vector $\vec{\mu}$ of dimension K, the DP and the Dirichlet are identical:

$$\text{Dirichlet}_K(\alpha, \vec{\mu}) \;=\; \text{DP}(\alpha, \vec{\mu}) \;.$$

- Thus in many applications, the use of a DP is equivalent to the use of a Dirichlet.
- Why use the DP then?
  - Hierarchical Dirichlets have fixed point MAP solutions, but more sophisticated reasoning is not always possible.
  - Hierarchical DPs have fairly fast samplers (as we shall see).
  - MAP solutions for hierarchical Dirichlets could be a good way to "burn in" samplers.

Summary: What You Need to Know

Species Sampling Model: SSM returns a discrete number of points from a domain

Partition: mixture models partition data, indexes are irrelevant

Improper Dirichlet: is a Dirichlet type distribution on partitions

Chinese Restaurant Process: CRP is the posterior sampler for the improper Dirichlet

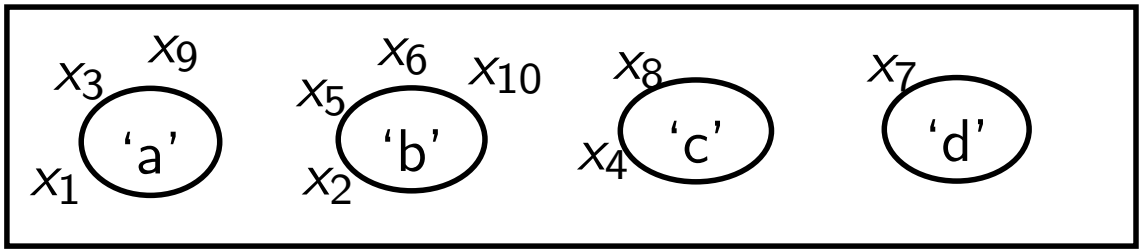
GEMs and Stick Breaking: Similar to an improper Dirichlet but in sized-based order and with an ordering penalty

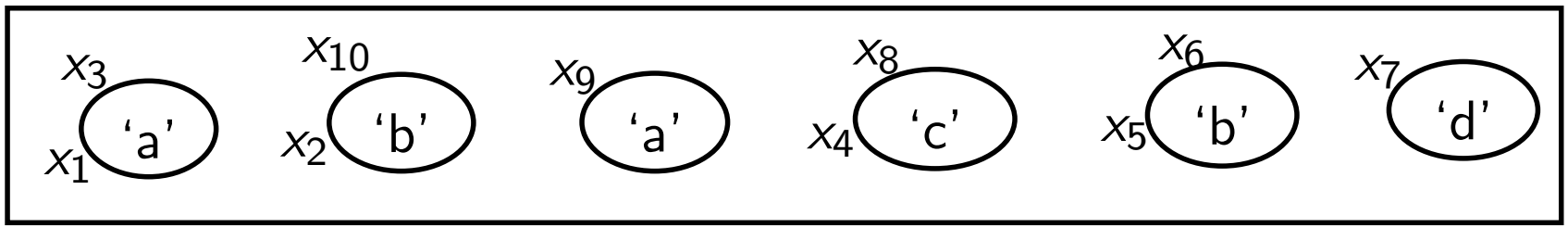
Stirling Numbers: distribution on the number of tables/species given by generalised Stirling number of second kind

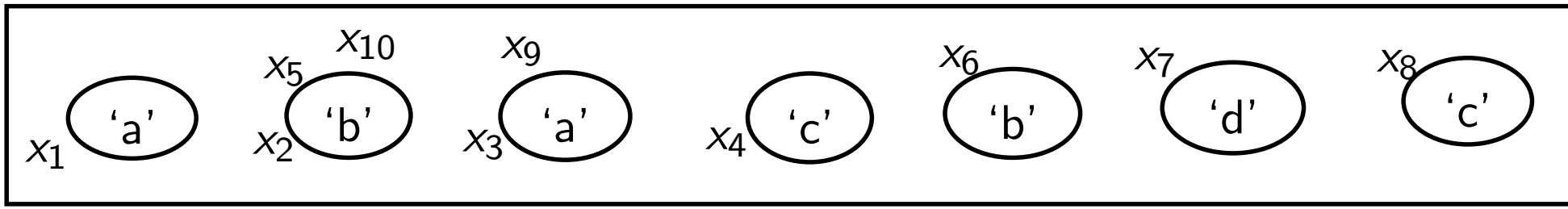
PYP and DP: are an SSM using an Improper Dirichlet to give the probability vector

Outline

CRPs on Discrete Data







The above three table configurations all match the data stream:

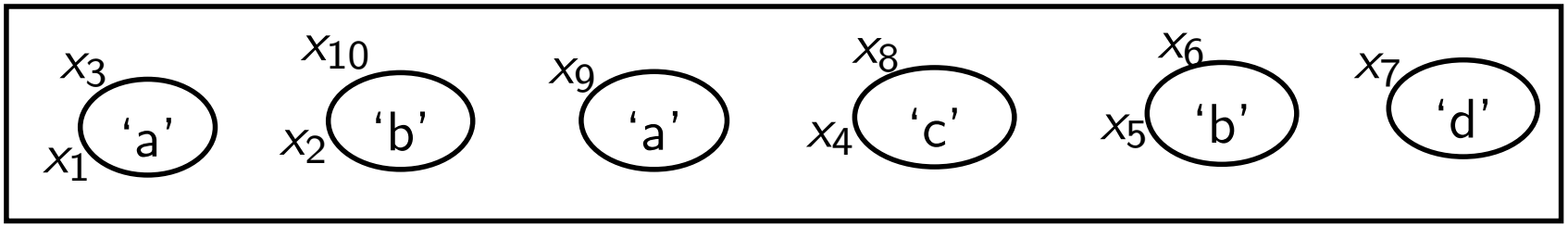
$a, b, a, c, b, b, d, c, a, b$

CRPs on Discrete Data, cont.

Different configurations for the data stream:

a, b, a, c, b, b, d, c, a, b

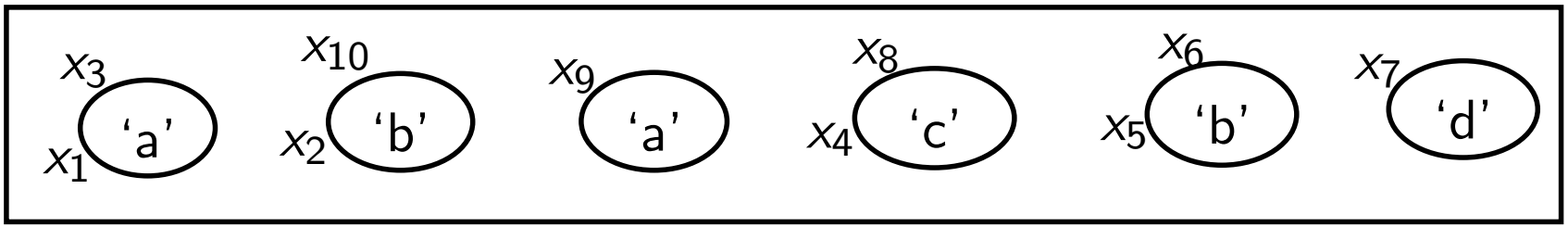
- $n_a = 3$, $n_b = 4$, $n_c = 2$, $n_d = 1$
- So the 3 data points with 'a' could be spread over 1,2 or 3 tables!
- Thus, inference will need to know the particular configuration/assignment of data points to tables.



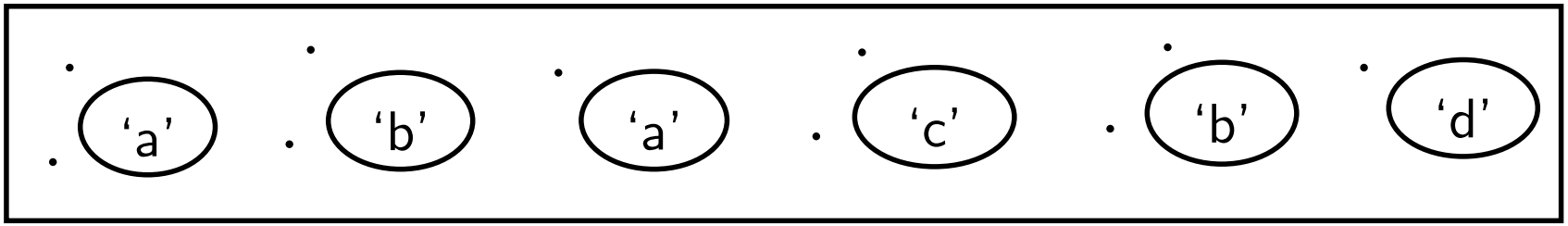
The configuration here has: number of tables $t_a = 2$ with table counts $(2, 1)$, $t_b = 2$ with counts $(2, 2)$, $t_c = 1$ with counts (2) and $t_d = 1$ with counts (1).

CRPs on Discrete Data, cont

We don't need to store the full table configuration:



We just need to store the counts: $t_a = 2$ with table counts $(2, 1)$, $t_b = 2$ with counts $(2, 2)$, $t_c = 1$ with counts (2) and $t_d = 1$ with counts (1).



The full configuration can be reconstructed (upto some statistical variation) by uniform sampling at any stage.

Evidence of PYP for Probability Vectors

Notation: Tables numbered $t = 1, ..., T$. Data types numbered $k = 1, ..., K$. Full table configuration denoted $\mathcal{T}$. Count of data type k is n_k, and number of tables t_k. Table t has data value (dish) X_t^* with m_t customers.

$$n_k = \sum_{t\,:\,X_t^*=k} m_t \ , \qquad t_k = \sum_{t\,:\,X_t^*=k} 1 \ , \qquad N = \sum_{k=1}^{K} n_k \ , \qquad T = \sum_{k=1}^{K} t_k$$

with constraints $t_k \leq n_k$ and $n_k > 0 \rightarrow t_k > 0$.

Evidence:

$$\begin{aligned} p(\vec{x}, \mathcal{T} | N, \text{PYP}, d, \alpha) &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{t=1}^{T} ((1-d)_{m_t-1} H(X_t^*)) \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \left(\prod_{t\,:\,X_t^*=k} (1-d)_{m_t-1} \right) H(k)^{t_k} \end{aligned}$$

Evidence of PYP for Probability Vectors, cont.

Consider configurations of tables with data type k, and denote them by $\mathcal{T}_k$.

$$\mathcal{T} \;=\; \mathcal{T}_1 + \mathcal{T}_2 + \ldots + \mathcal{T}_K$$

Let the function tables$(\cdot)$ return the number of tables in a configuration.

$$\begin{aligned} & p\left(\vec{x}, \vec{t} \,\middle|\, N, \text{PYP}, d, \alpha\right) \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \sum_{\text{tables}(\mathcal{T}_k)=t_k} \left(\prod_{t\,:\,X_t^*=k} (1-d)_{m_t-1} \right) H(k)^{t_k} \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} H(k)^{t_k} \end{aligned}$$

The simplification uses the definition of $\mathcal{S}^n_{t,d}$.

The Pitman-Yor-Multinomial

Definition of Pitman-Yor-Multinomial

Given a discount d and concentration parameter α, a probability vector $\vec{\theta}$ of dimension L, and a count N, the Pitman-Yor-multinomial creates count vector samples $\vec{n}$ of dimension K, and auxiliary counts $\vec{t}$ (constrained by $\vec{n}$). Now $(\vec{n}, \vec{t}) \sim \text{MultPY}\left(d, \alpha, \vec{\theta}, N\right)$ denotes

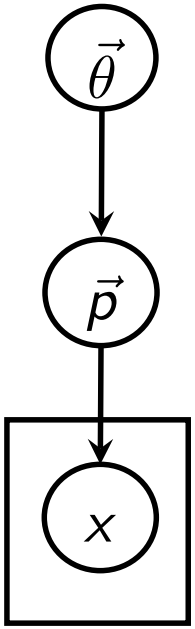
$$p\left(\vec{n}, \vec{t} \middle| N, \text{MultDP}, d, \alpha, \vec{\theta}\right) = \binom{N}{\vec{n}} \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}_{t_k,d}^{n_k} \theta_k^{t_k}$$

where $T = \sum_{k=1}^{K} t_k$.

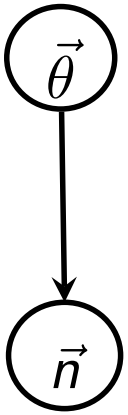
This is a form of evidence for the PYP.

The Ideal Hierarchical Component?

We want a magic distribution that looks like a multinomial likelihood in $\vec{\theta}$.



$$\begin{aligned} \vec{p} &\sim \text{Magic}\left(\alpha, \vec{\theta}\right) \\ x_n &\sim \text{Discrete}(\vec{p}) \quad \forall_n \end{aligned}$$

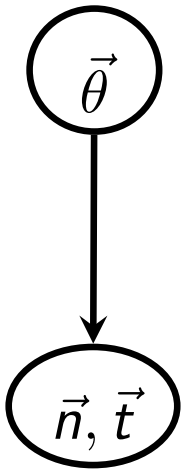


$$\begin{aligned} p\left(\vec{n} \middle| \alpha, \vec{\theta}, N\right) \\ = \quad F_\alpha(\vec{n}) \prod_k \theta_k^{t_k} \end{aligned}$$

$$\text{where } \sum_k n_k = N$$

The PYP/DP is the Magic

- The PYP/DP plays the role of the magic distribution.
- However, the exponent t_k for the θ now becomes a latent variable, so needs to be sampled as well.
- The t_k are constrainted:
  - $t_k \leq n_k$
  - $t_k > 0$ iff $n_k > 0$
- The $\vec{t}$ act like data for the next level up involving $\vec{\theta}$.



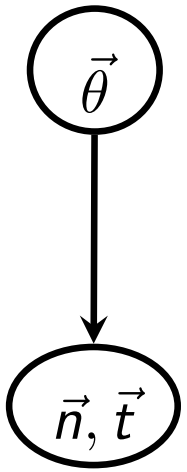
$$p\left(\vec{n}, \vec{t} \,\middle|\, d, \alpha, \vec{\theta}, N\right) = F_{d,\alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

where $\sum_k n_k = N$

Interpreting the Auxiliary Counts

Interpretation: t_k is how much of the count n_k that affects the parent probability (*i.e.* $\vec{\theta}$).

- If $\vec{t} = \vec{n}$ then the sample $\vec{n}$ affects $\vec{\theta}$ 100%.
- When $n_k = 0$ then $t_k = 0$, no effect.
- If $t_k = 1$, then the sample of n_k affects $\vec{\theta}$ minimally.



$$p\left(\vec{n}, \vec{t} \,\middle|\, d, \alpha, \vec{\theta}, N\right) = F_{d,\alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

where $\sum_k n_k = N$

Why We Prefer DPs and PYPs over Dirichlets!

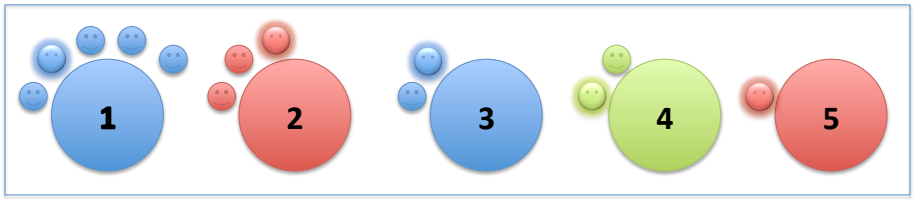
$$p\left(\vec{x}, \vec{t} \,\middle|\, N, \text{MultPY}, d, \alpha, \vec{\theta}\right) \propto \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} {\theta_k}^{t_k} \ ,$$

$$p\left(\vec{x} \,\middle|\, N, \text{MultDir}, \alpha, \vec{\theta}\right) \propto \frac{1}{(\alpha)_N} \prod_{k=1}^{K} (\alpha\theta_k)_{n_k} \ .$$

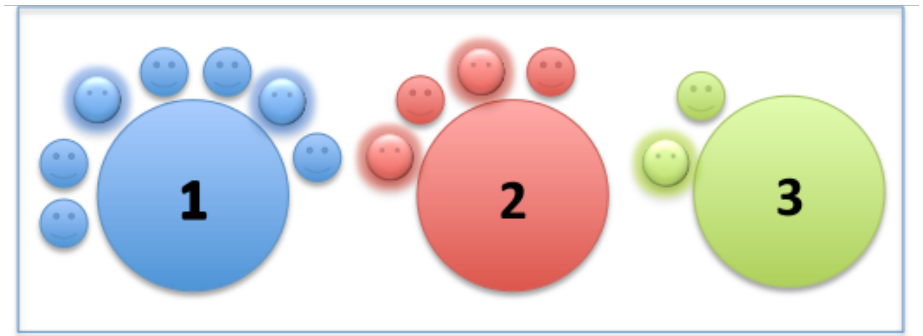
For the PYP, the θ_k just look like multinomial data, but you have to introduce a discrete latent variable $\vec{t}$.
For the Dirichlet, the θ_k are in a complex gamma function.

CRP Samplers versus MultPY Samplers

CRP sampling needs to keep track of full seating plan, such as counts per table (thus dynamic memory).



Sampling using the MultPY formula only needs to keep the number of tables. So rearrange configuration, only one table per dish and mark customers to indicate how many tables the CRP would have had.



CRP Samplers versus MultPY Samplers, cont.

CRP samplers sample configurations $\mathcal{T}$ consisting of (m_t, X_t^*) for $t = 1, ..., T$.

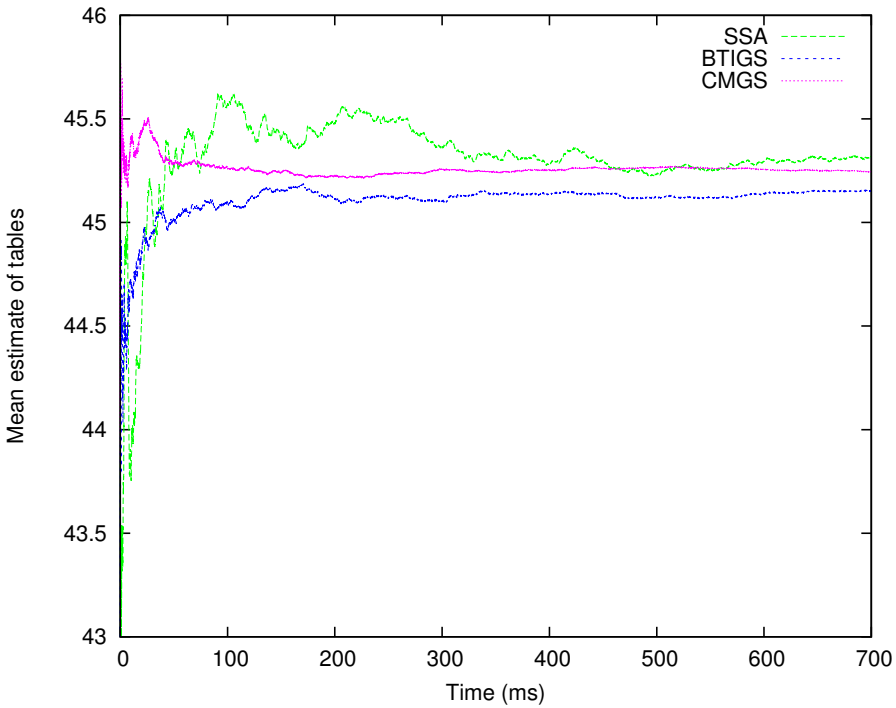
$$p(\vec{x}, \mathcal{T} | N, \text{PYP}, d, \alpha) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{t=1}^{T} ((1-d)_{m_t-1} H(X_t^*))$$

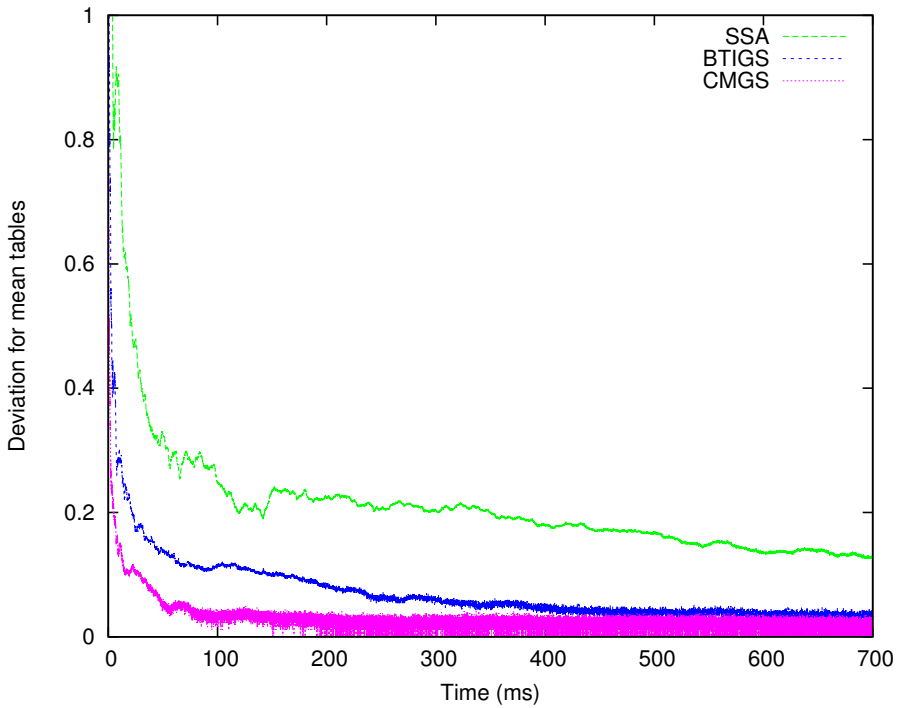
MultPY samplers sample the number of tables t_k for $k = 1, ..., K$. This is a collapsed version of a CRP sampler.

$$p\left(\vec{x}, \vec{t} \middle| N, \text{PYP}, d, \alpha\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}_{t_k,d}^{n_k} H(X_t^*)^{t_k}$$

Requires O(1) access to $\mathcal{S}_{t,d}^{n}$.

Comparing Samplers for the Pitman-Yor-Multinomial





Legend: SSA = "standard CRP sampler of Teh *et al.*"
CMGS = "Gibbs sampler using MultPY posterior"

Mean estimates of the total number of tables T for one of the 20 Gibbs runs (left) and the standard deviation of the 20 mean estimates (right) with $d = 0$, $\alpha = 10$, $K = 50$ and $N = 500$.

Outline

N-grams

To model a sequence of words $p(w_1, w_2, ..., w_N)$ we can use:

Unigram (1-gram) model: $p(w_n|\vec{\theta}_1)$

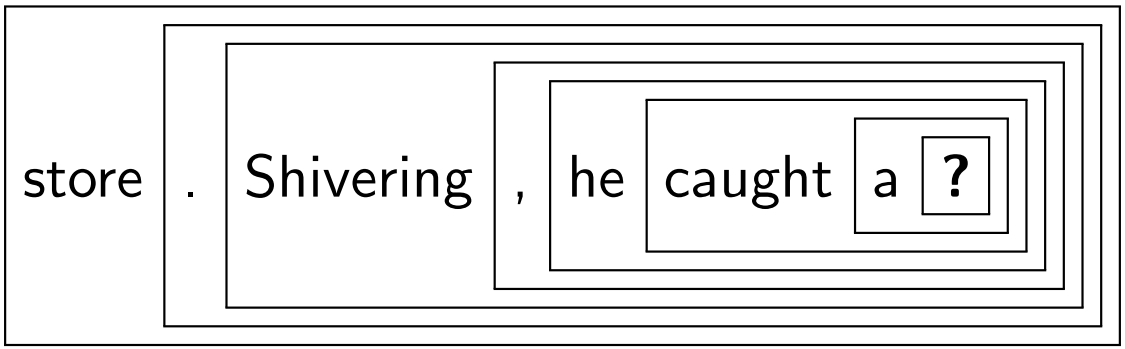
Bigram (2-gram) model: $p(w_n|w_{n-1}, \vec{\theta}_2)$

Trigram (3-gram) model: $p(w_n|w_{n-1}, w_{n-2}, \vec{\theta}_3)$

Or we can interpolate them somehow:

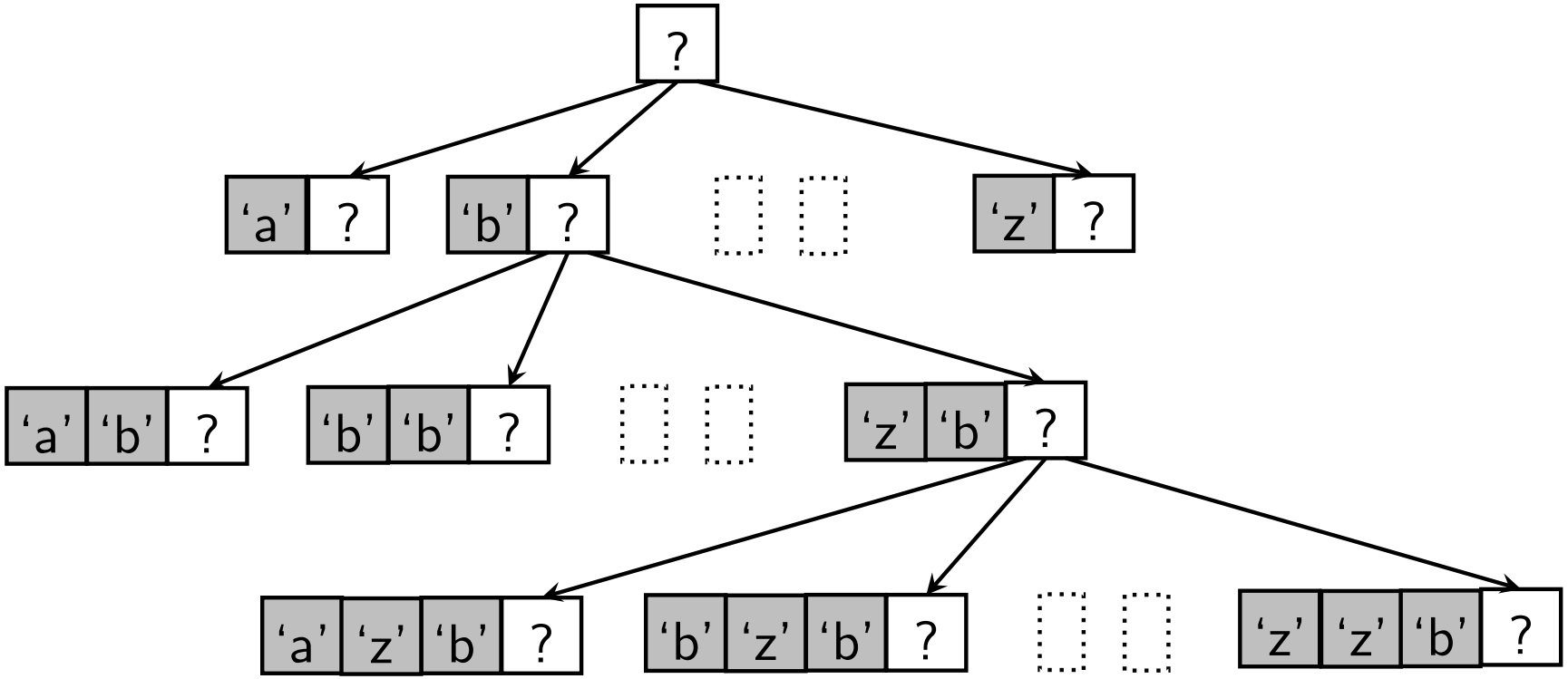
$$\hat{p}(w_n|w_{n-1}, w_{n-2}, \vec{\theta}_3) + \lambda_2 \hat{p}(w_n|w_{n-1}, \vec{\theta}_2)$$

Bayesian Idea: Similar Context Means Similar Word



- Words in [a [?]] should be like words in [?]
  - though no plural nouns
- Words in [caught [a [?]]] should be like words in [a [?]]
  - though a suitable object for "caught"
- Words in [he [caught [a [?]]]] be *very like* words in [caught [a [?]]]
  - "he" shouldn't change things much

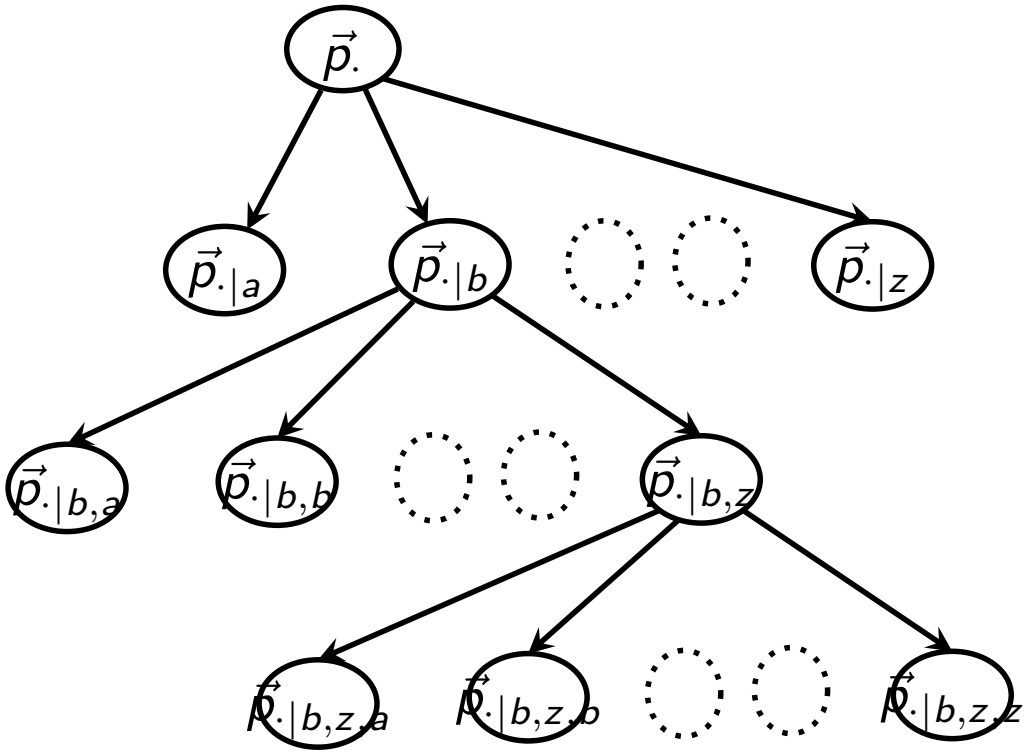
Bayesian N-grams



Build all three together, making the 1-gram a prior for the 2-grams, and the 2-grams a prior for the 3-grams, *etc.*

For this, we need to say each probability vector in the hierarchy is a variant of its parent.

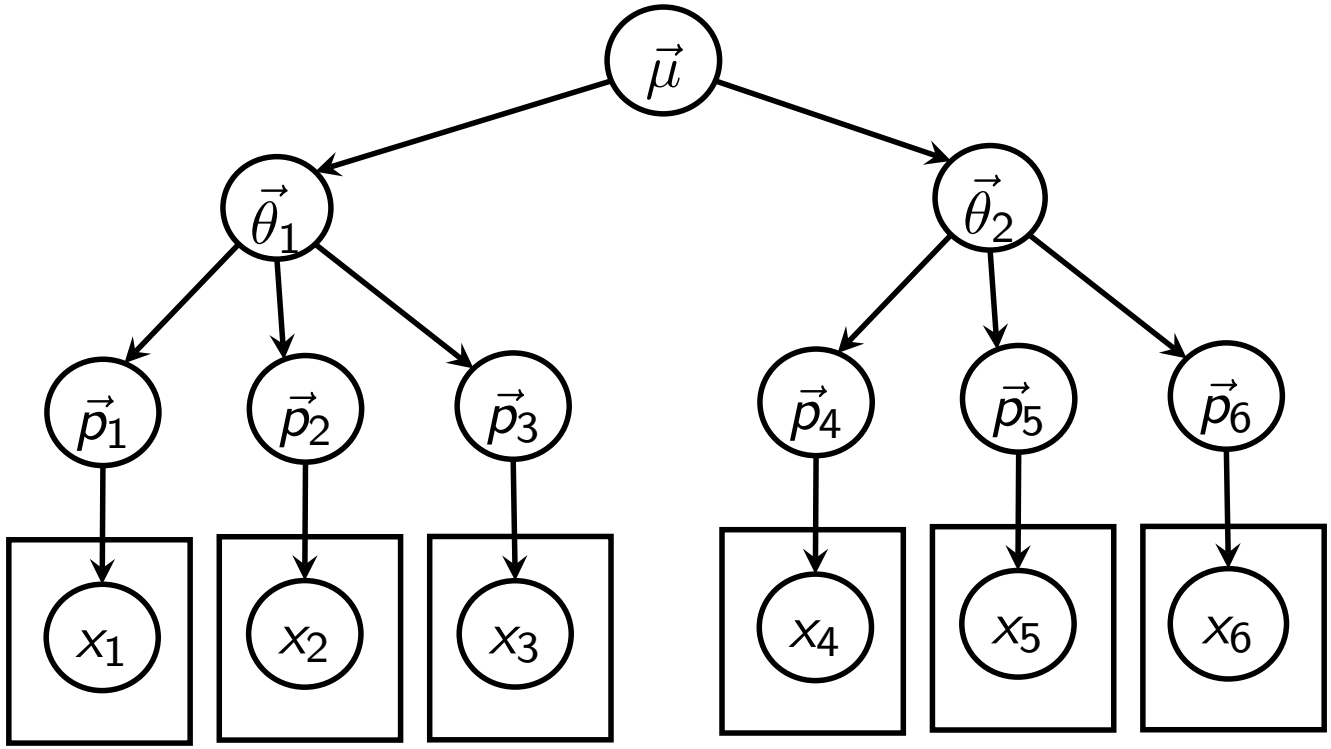
Bayesian N-grams, cont.



$$\begin{aligned}
\mathcal{S} &= \text{symbol set, fixed or possibly countably infinite} \\
\vec{p}_{\cdot} &\sim \text{prior on prob. vectors (initial vocabulary)} \\
\vec{p}_{\cdot|x_1} &\sim \text{dist. on prob. vectors with mean } \vec{p}_{\cdot} \qquad \forall_{x_1 \in \mathcal{S}} \\
\vec{p}_{\cdot|x_1,x_2} &\sim \text{dist. on prob. vectors with mean } \vec{p}_{\cdot|x_1} \qquad \forall_{x_1,x_2 \in \mathcal{S}}
\end{aligned}$$

Outline

A Simple N-gram Style Model



$$p(\vec{\mu})p\left(\vec{\theta}_1 \middle| \vec{\mu}\right) p\left(\vec{\theta}_2 \middle| \vec{\mu}\right)$$

$$p\left(\vec{p}_1 \middle| \vec{\theta}_1\right) p\left(\vec{p}_2 \middle| \vec{\theta}_1\right) p\left(\vec{p}_3 \middle| \vec{\theta}_1\right) p\left(\vec{p}_4 \middle| \vec{\theta}_2\right) p\left(\vec{p}_5 \middle| \vec{\theta}_2\right) p\left(\vec{p}_6 \middle| \vec{\theta}_2\right)$$

$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}}$$

Using the Evidence Formula

We will repeatedly apply the evidence formula

$$\begin{aligned} p\left(\vec{x}, \vec{t} \middle| N, \mathrm{DP}, \alpha\right) &= \frac{\alpha^T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} H(k)^{t_k} \\ &= F_\alpha(\vec{n}, \vec{t}) \prod_{k=1}^{K} H(k)^{t_k} \end{aligned}$$

to marginalise out all the probability vectors.

Apply Evidence Formula to Bottom Level

Start with the full posterior:

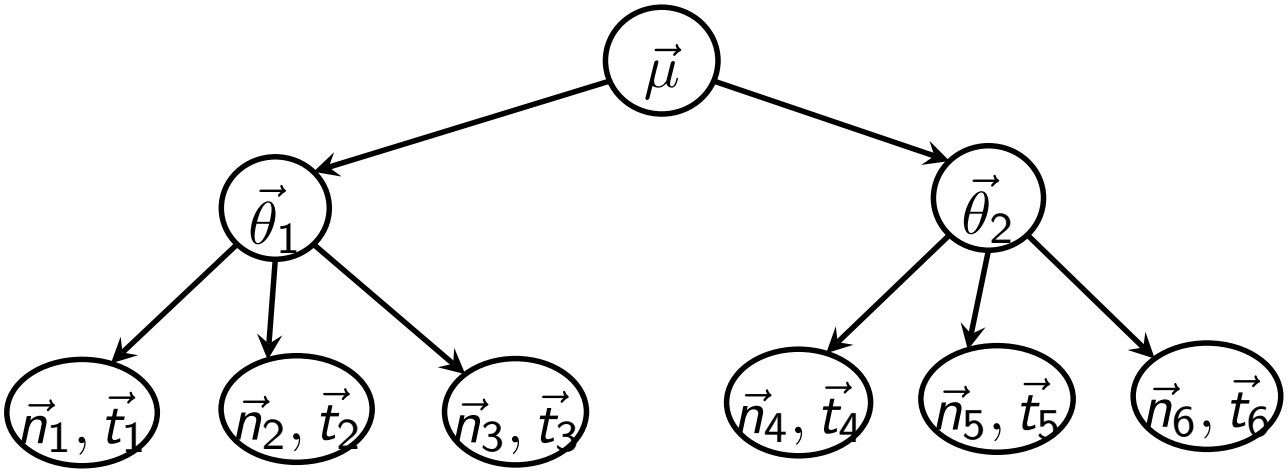
$$p(\vec{\mu})p\left(\vec{\theta}_1 \middle| \vec{\mu}\right) p\left(\vec{\theta}_2 \middle| \vec{\mu}\right)$$
$$p\left(\vec{p}_1 \middle| \vec{\theta}_1\right) p\left(\vec{p}_2 \middle| \vec{\theta}_1\right) p\left(\vec{p}_3 \middle| \vec{\theta}_1\right) p\left(\vec{p}_4 \middle| \vec{\theta}_2\right) p\left(\vec{p}_5 \middle| \vec{\theta}_2\right) p\left(\vec{p}_6 \middle| \vec{\theta}_2\right)$$
$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}} \ .$$

Marginalise out each $\vec{p}_k$ but introducing new auxiliaries $\vec{t}_k$

$$p(\vec{\mu})p\left(\vec{\theta}_1 \middle| \vec{\mu}\right) p\left(\vec{\theta}_2 \middle| \vec{\mu}\right)$$
$$F_\alpha(\vec{n}_1, \vec{t}_1)F_\alpha(\vec{n}_2, \vec{t}_2)F_\alpha(\vec{n}_3, \vec{t}_3) \prod_l \theta_{1,l}^{t_{1,l}+t_{2,l}+t_{3,l}}$$
$$F_\alpha(\vec{n}_4, \vec{t}_4)F_\alpha(\vec{n}_5, \vec{t}_5)F_\alpha(\vec{n}_6, \vec{t}_6) \prod_l \theta_{2,l}^{t_{4,l}+t_{5,l}+t_{6,l}} \ .$$

Thus $\vec{t}_1 + \vec{t}_2 + \vec{t}_3$ looks like data for $\vec{\theta}_1$ and $\vec{t}_4 + \vec{t}_5 + \vec{t}_6$ looks like data for $\vec{\theta}_2$

Apply Evidence Formula, cont.



Terms left in $\vec{n}_k$ and $\vec{t}_k$, and passing up

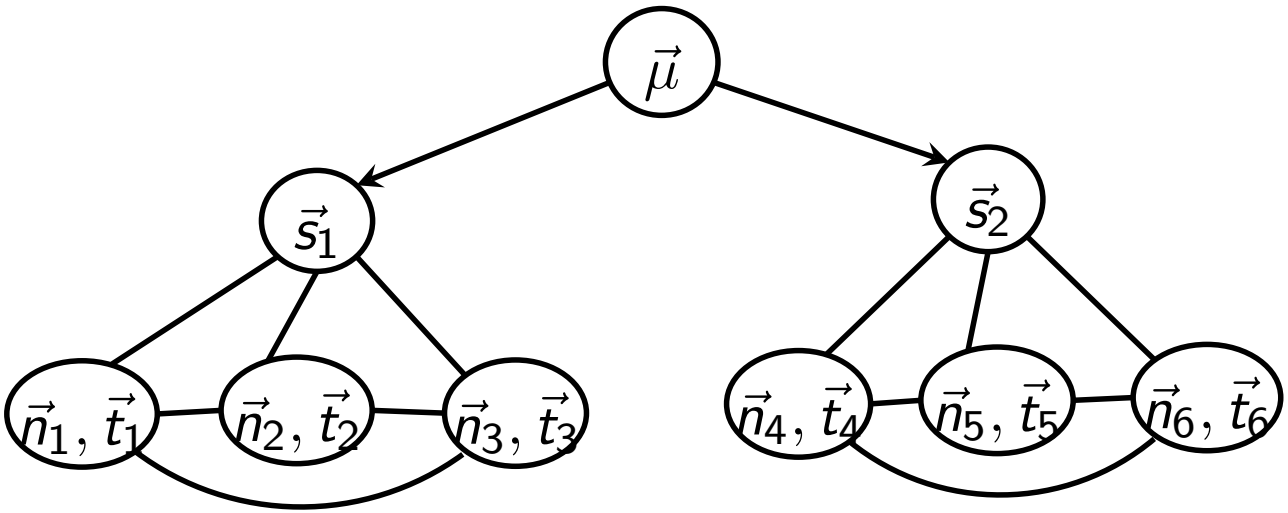
$$\prod_l \theta_{1,l}^{t_{1,l}+t_{2,l}+t_{3,l}} \prod_l \theta_{2,l}^{t_{4,l}+t_{5,l}+t_{6,l}} ,$$

as pseudo-data to the prior on $\vec{\theta}_1$ and $\vec{\theta}_2$.

Apply Evidence Formula, cont.

Repeat the same trick up a level; marginalising out $\vec{\theta}_1$ and $\vec{\theta}_1$ but introducing new auxiliaries $\vec{s}_1$ and $\vec{s}_2$

$$p(\vec{\mu})F_\alpha(\vec{t}_1+\vec{t}_2+\vec{t}_3,\vec{s}_1)F_\alpha(\vec{t}_4+\vec{t}_5+\vec{t}_6,\vec{s}_2)\prod_l \mu_l^{s_{1,l}+s_{2,l}}$$
$$F_\alpha(\vec{n}_1,\vec{t}_1)F_\alpha(\vec{n}_2,\vec{t}_2)F_\alpha(\vec{n}_3,\vec{t}_3)F_\alpha(\vec{n}_4,\vec{t}_4)F_\alpha(\vec{n}_5,\vec{t}_5)F_\alpha(\vec{n}_6,\vec{t}_6)\ .$$



Again left with pseudo-data to the prior on $\vec{\mu}$.

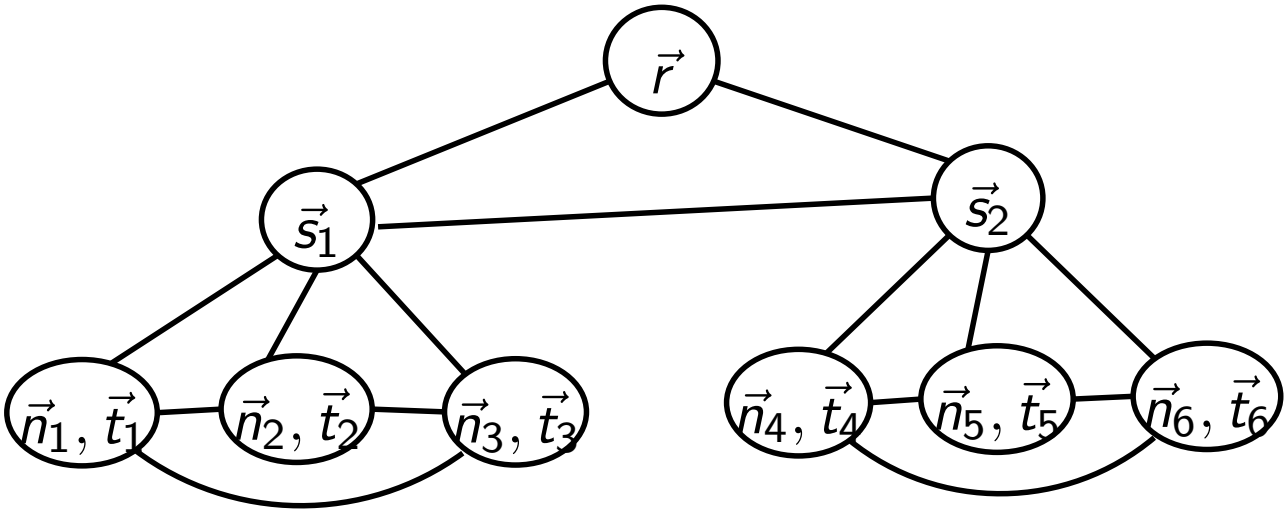
Apply Evidence Formula, cont.

Finally repeat at the top level with new auxiliary $\vec{r}$

$$F_\alpha(\vec{s}_1 + \vec{s}_2, \vec{r})F_\alpha(\vec{t}_1 + \vec{t}_2 + \vec{t}_3, \vec{s}_1)F_\alpha(\vec{t}_4 + \vec{t}_5 + \vec{t}_6, \vec{s}_2)$$
$$F_\alpha(\vec{n}_1, \vec{t}_1)F_\alpha(\vec{n}_2, \vec{t}_2)F_\alpha(\vec{n}_3, \vec{t}_3)F_\alpha(\vec{n}_4, \vec{t}_4)F_\alpha(\vec{n}_5, \vec{t}_5)F_\alpha(\vec{n}_6, \vec{t}_6)$$

where

- $\vec{n}_1$, $\vec{n}_2$,... are the data at the leaf nodes, $\vec{t}_1$, $\vec{t}_2$,... their auxiliary counts
- $\vec{s}_1$ are auxiliary counts constrained by $\vec{t}_1 + \vec{t}_2 + \vec{t}_3$,
- $\vec{s}_2$ are auxiliary counts constrained by $\vec{t}_4 + \vec{t}_5 + \vec{t}_6$,
- $\vec{r}$ are auxiliary counts constrained by $\vec{s}_1 + \vec{s}_2$,



The Worked N-gram Style Model

Original posterior in the form:

$$p(\vec{\mu})p\left(\vec{\theta}_1 \middle| \vec{\mu}\right) p\left(\vec{\theta}_2 \middle| \vec{\mu}\right)$$
$$p\left(\vec{p}_1 \middle| \vec{\theta}_1\right) p\left(\vec{p}_2 \middle| \vec{\theta}_1\right) p\left(\vec{p}_3 \middle| \vec{\theta}_1\right) p\left(\vec{p}_4 \middle| \vec{\theta}_2\right) p\left(\vec{p}_5 \middle| \vec{\theta}_2\right) p\left(\vec{p}_6 \middle| \vec{\theta}_2\right)$$
$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}}$$

Collapsed posterior in the form:

$$F_\alpha(\vec{s}_1 + \vec{s}_2, \vec{r})F_\alpha(\vec{t}_1 + \vec{t}_2 + \vec{t}_3, \vec{s}_1)F_\alpha(\vec{t}_4 + \vec{t}_5 + \vec{t}_6, \vec{s}_2)$$
$$F_\alpha(\vec{n}_1, \vec{t}_1)F_\alpha(\vec{n}_2, \vec{t}_2)F_\alpha(\vec{n}_3, \vec{t}_3)F_\alpha(\vec{n}_4, \vec{t}_4)F_\alpha(\vec{n}_5, \vec{t}_5)F_\alpha(\vec{n}_6, \vec{t}_6)$$

where

- $\vec{n}_1$, $\vec{n}_2$,... are the data at the leaf nodes, $\vec{t}_1$, $\vec{t}_2$,... their auxiliary counts
- $\vec{s}_1$ are auxiliary counts constrained by $\vec{t}_1 + \vec{t}_2 + \vec{t}_3$,
- $\vec{s}_2$ are auxiliary counts constrained by $\vec{t}_4 + \vec{t}_5 + \vec{t}_6$,
- $\vec{r}$ are auxiliary counts constrained by $\vec{s}_1 + \vec{s}_2$,

The Worked N-gram Style Model, cont.

Note the probabilities are then estimated from the auxiliary counts during MCMC. This is the standard recursive CRP formula.

$$
\begin{aligned}
\widehat{\vec{\mu}} &= \frac{\vec{s}_1 + \vec{s}_2}{S_1 + S_2 + \alpha} + \frac{\alpha}{S_1 + S_2 + \alpha}\left(\frac{\vec{r}}{R + \alpha} + \frac{R}{R + \alpha}\frac{1}{L}\right) \\
\widehat{\vec{\theta}}_1 &= \frac{\vec{t}_1 + \vec{t}_2 + \vec{t}_3}{T_1 + T_2 + T_3 + \alpha} + \frac{\alpha}{T_1 + T_2 + T_3 + \alpha}\widehat{\vec{\mu}} \\
\widehat{\vec{p}}_1 &= \frac{\vec{n}_1}{N_1 + \alpha} + \frac{\alpha}{N_1 + \alpha}\widehat{\vec{\theta}}_1
\end{aligned}
$$

Note in practice:

- the α is varied at every level of the tree and sampled as well,
- the PYP is used instead because words are often Zipfian

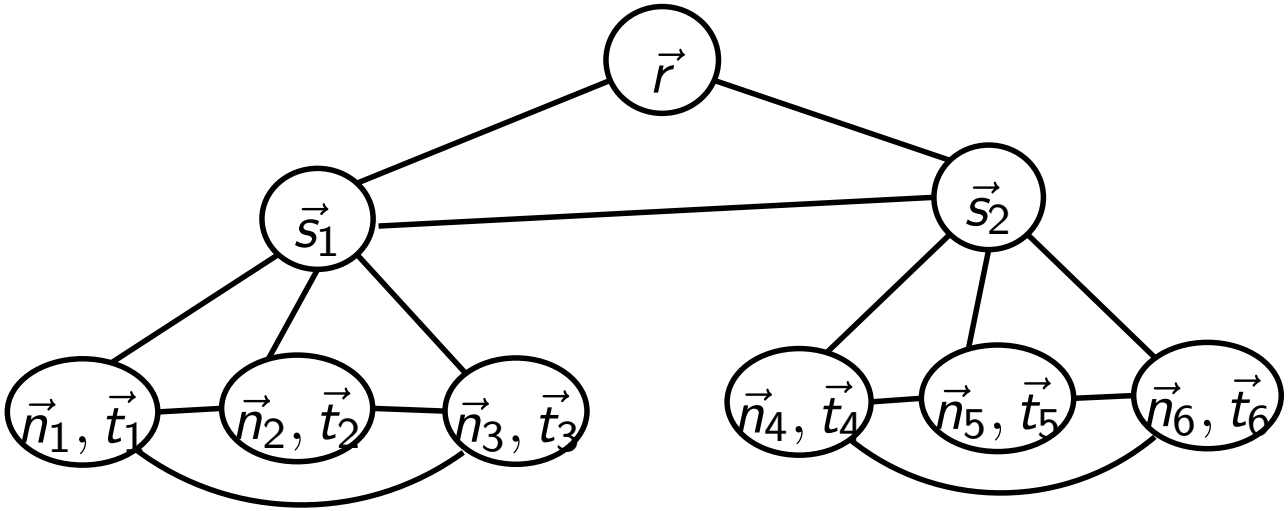
The Worked N-gram Style Model, cont.

What have we achieved:

- Bottom level probabilities $(\vec{p}_1, \vec{p}_2, \ldots)$ marginalised away.
- Each non-leaf probability vector $(\vec{\mu}, \vec{\theta}_1, \ldots)$ replaced by corresponding constrained auxiliary count vector $(\vec{r}, \vec{s}_1, \ldots)$ as psuedo-data.
- The auxiliary counts correspond to how much of the counts get inherited up the hierarchy.
- This allows a collapsed sampler in a discrete (versus continuous) space.

MCMC Problem Specification for N-grams

Build a Gibbs/MCMC sampler for:



$$(\vec{\mu}) \quad \cdots \quad \frac{\alpha^R}{(\alpha)_{S_1+S_2}} \prod_{k=1}^{K} \left(\mathcal{S}^{s_{1,k}+s_{2,k}}_{r_k,0} \frac{1}{K^{r_k}} \right)$$

$$(\vec{\theta}_1, \vec{\theta}_2) \quad \cdots \quad \frac{\alpha^{S_1}}{(\alpha)_{T_1+T_2+T_3}} \prod_{k=1}^{K} \mathcal{S}^{t_{1,k}+t_{2,k}+t_{3,k}}_{s_{1,k},0} \frac{\alpha^{S_2}}{(\alpha)_{T_4+T_5+T_6}} \prod_{k=1}^{K} \mathcal{S}^{t_{4,k}+t_{5,k}+t_{6,k}}_{s_{2,k},0}$$

$$(\forall_k \vec{p}_k) \quad \cdots \quad \prod_{l=1}^{6} \left(\prod_{k=1}^{K} \mathcal{S}^{n_{l,k}}_{t_{l,k},0} \right)$$

Sampling Ideas

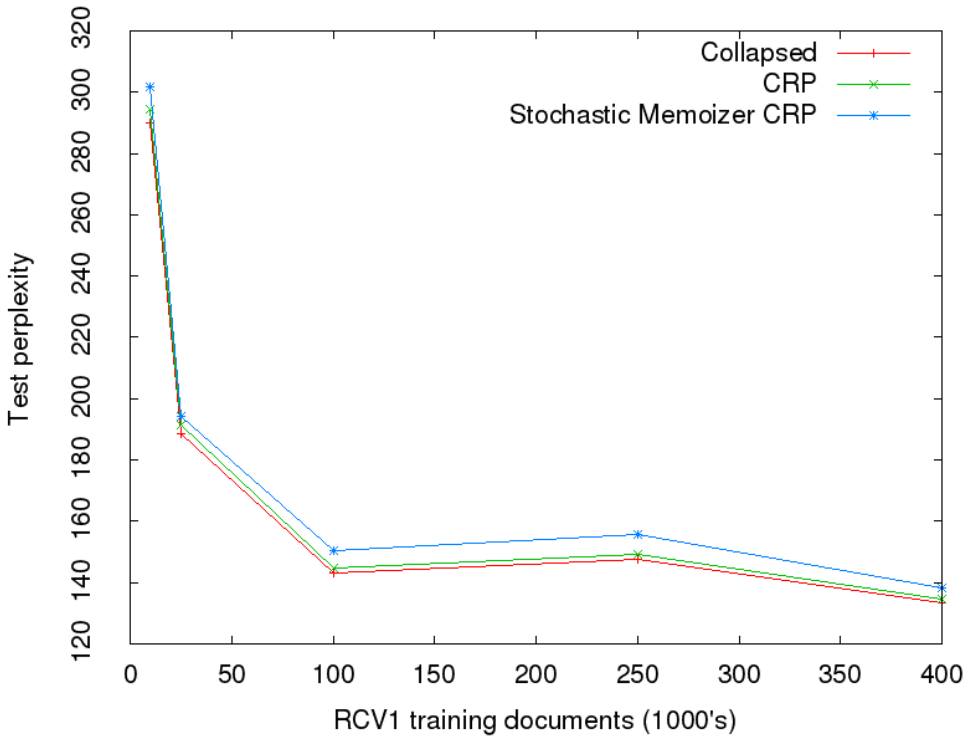
Consider the term in $s_{1,k}$ where $s_{1,k} \leq t_{1,k} + t_{2,k} + t_{3,k}$.

$$\frac{1}{(\alpha)_{S_1'+S_2+s_{1,k}}} \mathcal{S}_{r_k,0}^{s_{1,k}+s_{2,k}} \alpha^{S_1'+s_{1,k}} \mathcal{S}_{s_{1,k},0}^{t_{1,k}+t_{2,k}+t_{3,k}}$$

- Gibbs by sampling $s_{1,k}$ proportional to this for all $1 \leq s_{1,k} \leq t_{1,k} + t_{2,k} + t_{3,k}$.
- Approximate Gibbs by sampling $s_{1,k}$ proportional to this for $s_{1,k}$ in a window of size 21 around the current: $\max(1, s_{1,k} - 10) \leq s_{1,k} \leq \min(s_{1,k} + 10, t_{1,k} + t_{2,k} + t_{3,k})$.
- Metropolis-Hastings by sampling $s_{1,k}$ proportional to this for $s_{1,k}$ in a window of size 3 or 5 or 11.

Note: have used the second in implementations.

Some Results on RCV1 with 5-grams



Collapsed is the method here using PYPs with both discount and concentration sampled level-wise.

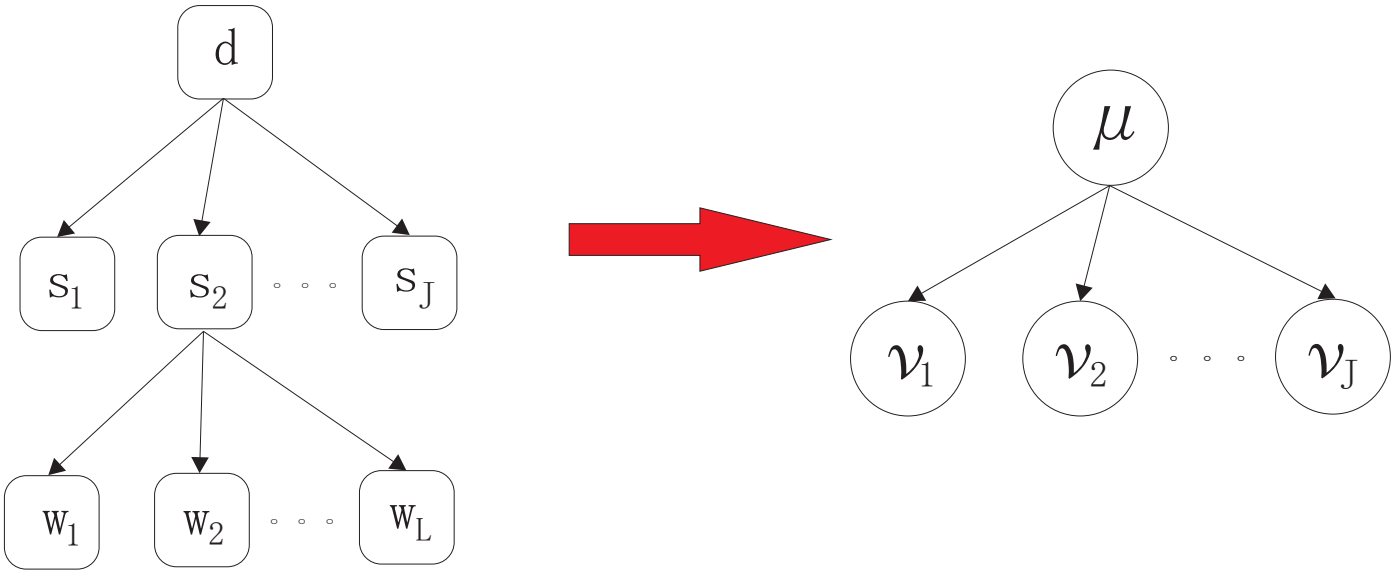
CRP is the CRP method of Teh (ACL 2006) with both discount and concentration sampled level-wise.

Memoizer fixes the CRP parameters to that Wood *et al.* (ICML 2009).

- Used Reuters RCV1 collection with 400k documents (about 190M words), and following 5k for test.
- Gave methods equal time.
- `Collapsed` and `CRP` exhibit similar convergence.
- `Collapsed` requires no dynamic memory so takes about 1/2 of the space.
- `Collapsed` improves with 10-grams on full RCV1.

Outline

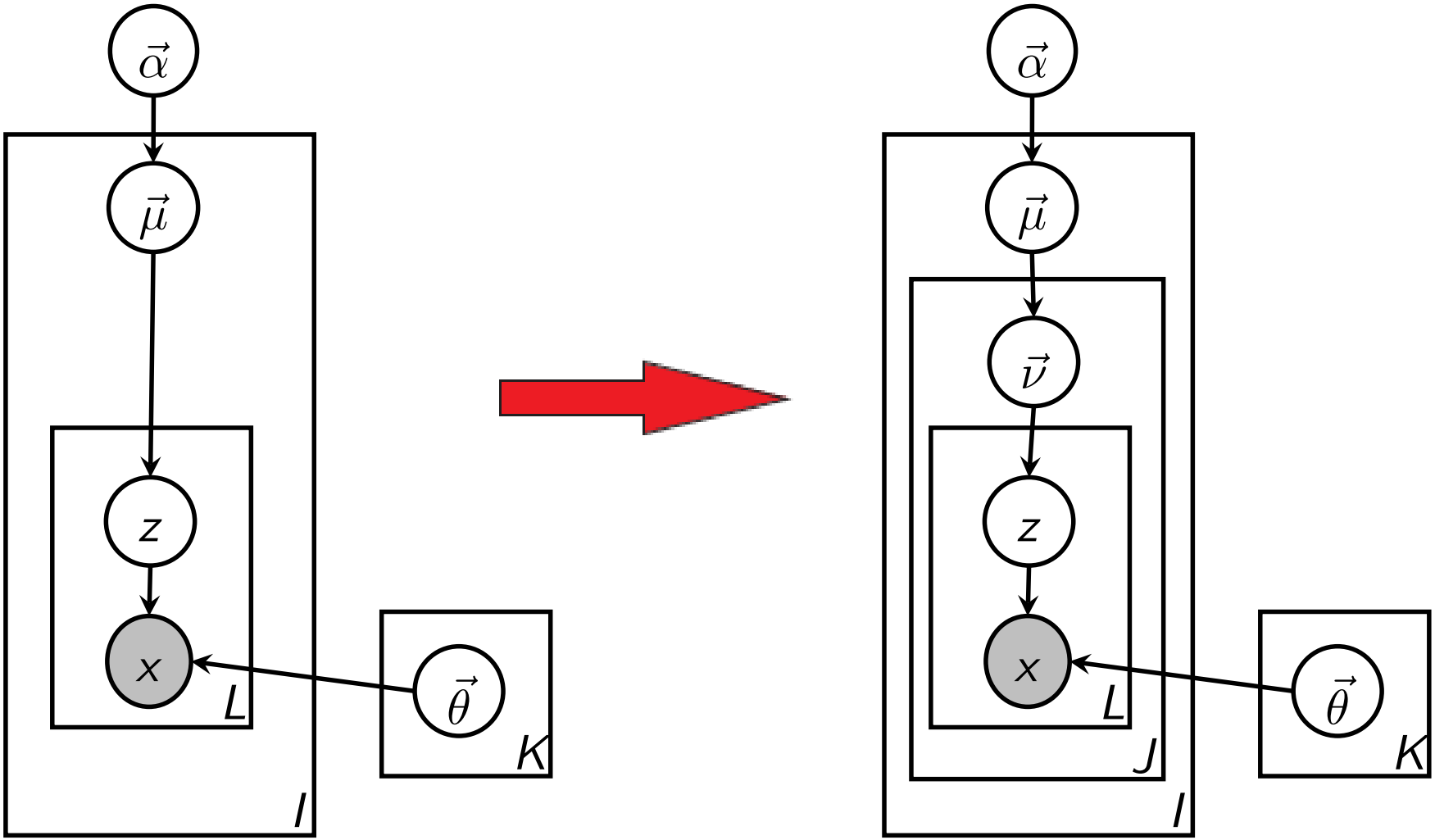
Structured Documents



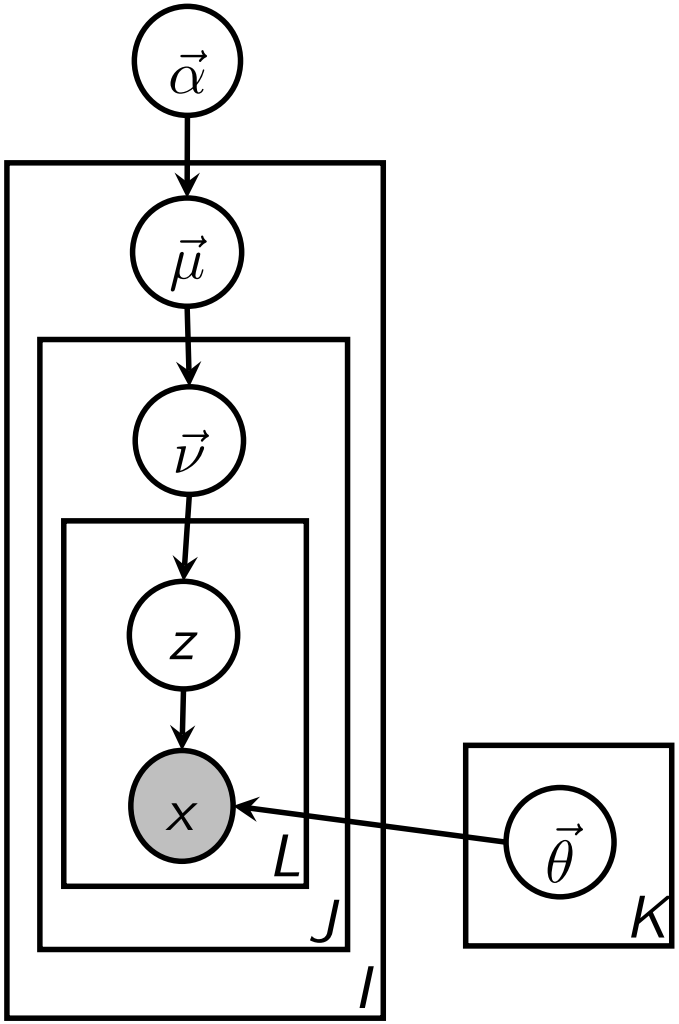
- A document contains sections which contains words.
- This implies the graphical model between the topics of the document and its sections.

Structured Topic Model (STM)

We add this "structure" to the standard topic model.



Structured Topic Model, cont.



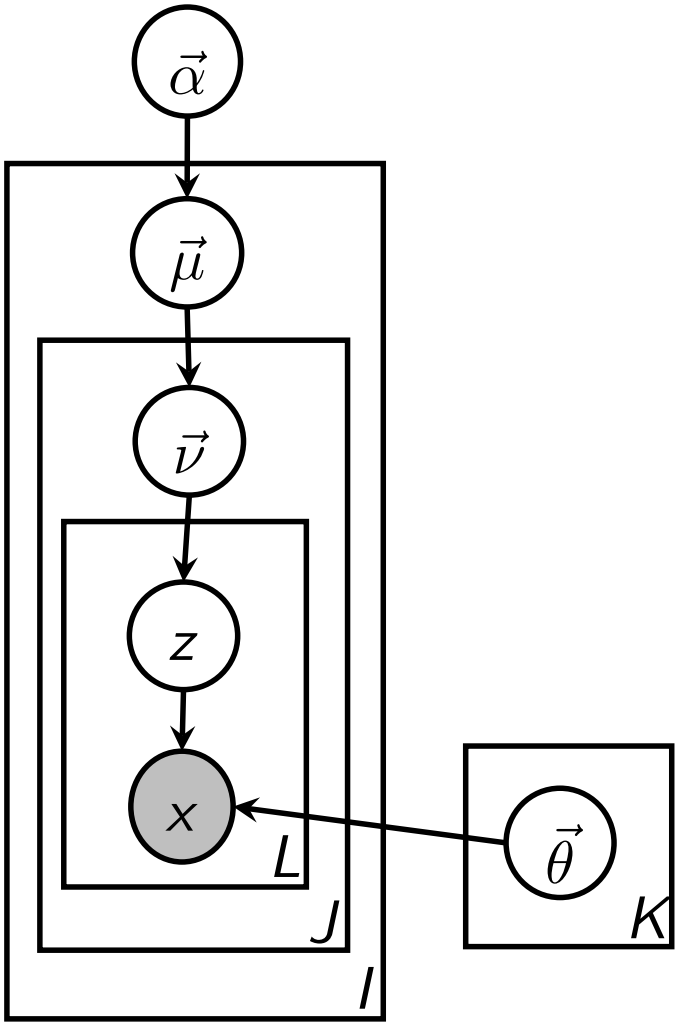
$$
\begin{aligned}
\vec{\theta}_k &\sim \text{Dirichlet}_V(\vec{\gamma}) \quad \forall_{k=1}^{K} , \\
\vec{\mu}_i &\sim \text{Dirichlet}_K(\vec{\alpha}) \quad \forall_{i=1}^{I} , \\
\vec{\nu}_{i,j} &\sim \text{Dirichlet}_K(\beta, \vec{\mu}_i) \quad \forall_{i=1}^{I} \forall_{j=1}^{J_i} , \\
z_{i,j,l} &\sim \text{Discrete}(\vec{\nu}_{i,j}) \quad \forall_{i=1}^{I} \forall_{j=1}^{J_i} \forall_{l=1}^{L_{i,j}} , \\
x_{i,j,l} &\sim \text{Discrete}\left(\vec{\theta}_{z_{i,j,l}}\right) \quad \forall_{i=1}^{I} \forall_{j=1}^{J_i} \forall_{l=1}^{L_{i,j}} .
\end{aligned}
$$

where

$$
\begin{aligned}
K &:= \# \text{ topics}, \\
V &:= \# \text{ words}, \\
I &:= \# \text{ documents}, \\
J_i &:= \# \text{ segments in doc } i, \\
L_{i,j} &:= \# \text{ words in seg } j \text{ of doc } i
\end{aligned}
$$

Extending LDA, Strategy

Structure Topic Model



Consider the collapsed LDA posterior:

$$\prod_{i=1}^{I} \frac{\text{Beta}_K\left(\vec{\alpha}+\vec{m}_i\right)}{\text{Beta}\left(\vec{\alpha}\right)} \prod_{k=1}^{K} \frac{\text{Beta}_J\left(\vec{\gamma}+\vec{n}_k\right)}{\text{Beta}\left(\vec{\gamma}\right)}$$

We can extend LDA in all sorts of ways by replacing the Dirichlet-multinomial parts with Dirichlet-multinomial processes or Pitman-Yor-multinomial.

- expanding vocabulary;
- expanding topics (called HDP-LDA);
- also, Structured topic models.

Structured Topic Model Posterior

Full posterior:

$$= \prod_{k=1}^{K} p(\vec{\theta}_k|\vec{\gamma}) \prod_{i=1}^{I} p(\vec{\mu}_i|\vec{\alpha}) \prod_{i,j=1}^{I,J_i} p(\vec{\nu}_{i,j}|\beta,\vec{\mu}_i) \prod_{i,j,l=1}^{I,J_i,L_{i,j}} \nu_{i,j,z_{i,j,l}} \theta_{z_{i,j,l},x_{i,j,l}}$$

$$= \overbrace{\prod_{i=1}^{I} p(\vec{\mu}_i|\vec{\alpha})}^{\text{terms in } \vec{\mu}_i} \overbrace{\prod_{i,j=1}^{I,J_i} p(\vec{\nu}_{i,j}|\beta,\vec{\mu}_i) \prod_{i,j,k=1}^{I,J_i,K} \nu_{i,j,k}^{m_{i,j,k}}}^{\text{terms in } \vec{\mu}_i + \vec{\nu}_{i,j}} \overbrace{\prod_{k=1}^{K} p(\vec{\theta}_k|\vec{\gamma}) \prod_{k,v=1}^{K,V} \theta_{k,v}^{n_{k,v}}}^{\text{terms in } \vec{\theta}_k}$$

$$= \overbrace{\prod_{i=1}^{I} F_{\alpha_0}(\vec{t}_{i,\cdot},\vec{s}_i) \prod_{i,k=1}^{I,K} \left(\frac{\alpha_k}{\alpha_0}\right)^{s_{i,k}}}^{\text{marginalising } \vec{\mu}_i} \overbrace{\prod_{i,j=1}^{I,J_i} F_{\beta}(\vec{m}_{i,j},\vec{t}_{i,j})}^{\text{marginalising } \vec{\nu}_{i,j}} \overbrace{\prod_{k=1}^{K} \frac{\text{Beta}_J(\vec{\gamma}+\vec{n}_k)}{\text{Beta}(\vec{\gamma})}}^{\text{marginalising } \vec{\theta}_k}$$

Marginalise using the same methods as before.

Structured Topic Model Posterior, cont.

$$\overbrace{\prod_{i=1}^{I} \frac{1}{(\alpha_0)_{T_{i,\cdot}}} \prod_{i,k=1}^{I,K} \mathcal{S}^{t_{i,\cdot,k}}_{s_{i,k},0} \alpha_k^{s_{i,k}}}^{\text{marginalising } \vec{\mu}_i} \overbrace{\prod_{i,j=1}^{I,J_i} \frac{\beta^{T_{i,j}}}{(\beta)_{M_{i,j}}} \prod_{i,j,k=1}^{I,J_i,K} \mathcal{S}^{m_{i,j,k}}_{t_{i,j,k},0}}^{\text{marginalising } \vec{\nu}_{i,j}} \overbrace{\prod_{k=1}^{K} \frac{\text{Beta}_J(\vec{\gamma} + \vec{n}_k)}{\text{Beta}(\vec{\gamma})}}^{\text{marginalising } \vec{\theta}_k}$$

with statistics

$\vec{m}_{i,j}$:= dim(K) data counts of topics for seg j in doc i, given by $m_{i,j,k} = \sum_{l=1}^{L_{i,j}} 1_{z_{i,j,l}=k}$

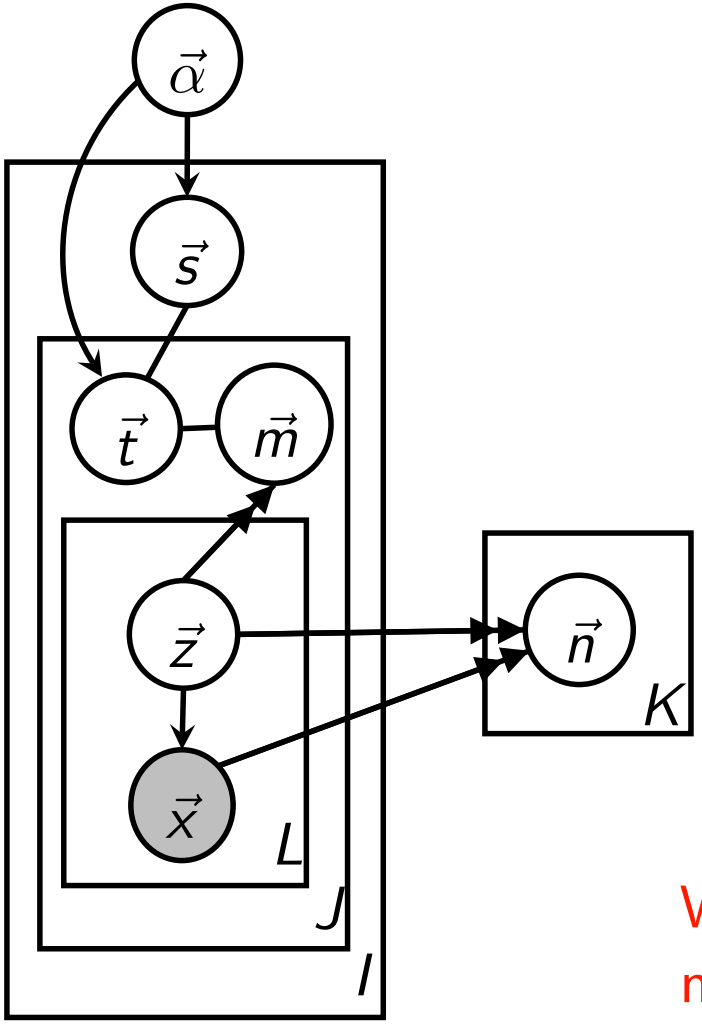
$\vec{n}_k$:= dim(V) data counts of words for topic k, given by $n_{k,v} = \sum_{i,j,l=1}^{I,J_i,L_{i,j}} 1_{z_{i,j,l}=k} 1_{x_{i,j,l}=v}$

$\vec{t}_{i,j}$:= dim(K) auxiliary counts for $\vec{\mu}_i$ from seg j in doc i, constrained by $\vec{m}_{i,j}$,

$\vec{s}_i$:= dim(K) auxiliary counts for $\vec{\alpha}$ from doc i, constrained by $\vec{t}_{i,\cdot}$.

and totals: $\vec{t}_{i,\cdot} = \sum_{j=1}^{J_i} \vec{t}_{i,j}$, $T_{i,j} = \sum_{k=1}^{K} t_{i,j,k}$, $M_{i,j} = \sum_{k=1}^{K} m_{i,j,k}$.

Structured Topic Model Posterior, cont.



$\vec{m}_{i,j}$:= $\dim(K)$ data counts of topics for seg j in doc i,
given by $m_{i,j,k} = \sum_{l=1}^{L_{i,j}} 1_{z_{i,j,l}=k}$

$\vec{n}_k$:= $\dim(V)$ data counts of words for topic k, given by
$n_{k,v} = \sum_{i,j,l=1}^{I,J_i,L_{i,j}} 1_{z_{i,j,l}=k} 1_{x_{i,j,l}=v}$

$\vec{t}_{i,j}$:= $\dim(K)$ auxiliary counts for $\vec{\mu}_i$ from seg j in doc i,
constrained by $\vec{m}_{i,j}$,

$\vec{s}_i$:= $\dim(K)$ auxiliary counts for $\vec{\alpha}$ from doc i,
constrained by $\vec{t}_{i,\cdot} = \sum_j \vec{t}_{i,j}$

We need to sample the topics $z_{i,j,l}$, all the while maintaining the counts $\vec{n}_k$ and $\vec{m}_{i,j}$, and concurrently resampling $\vec{t}_{i,j}$ and $\vec{s}_i$.

Sampling Notes

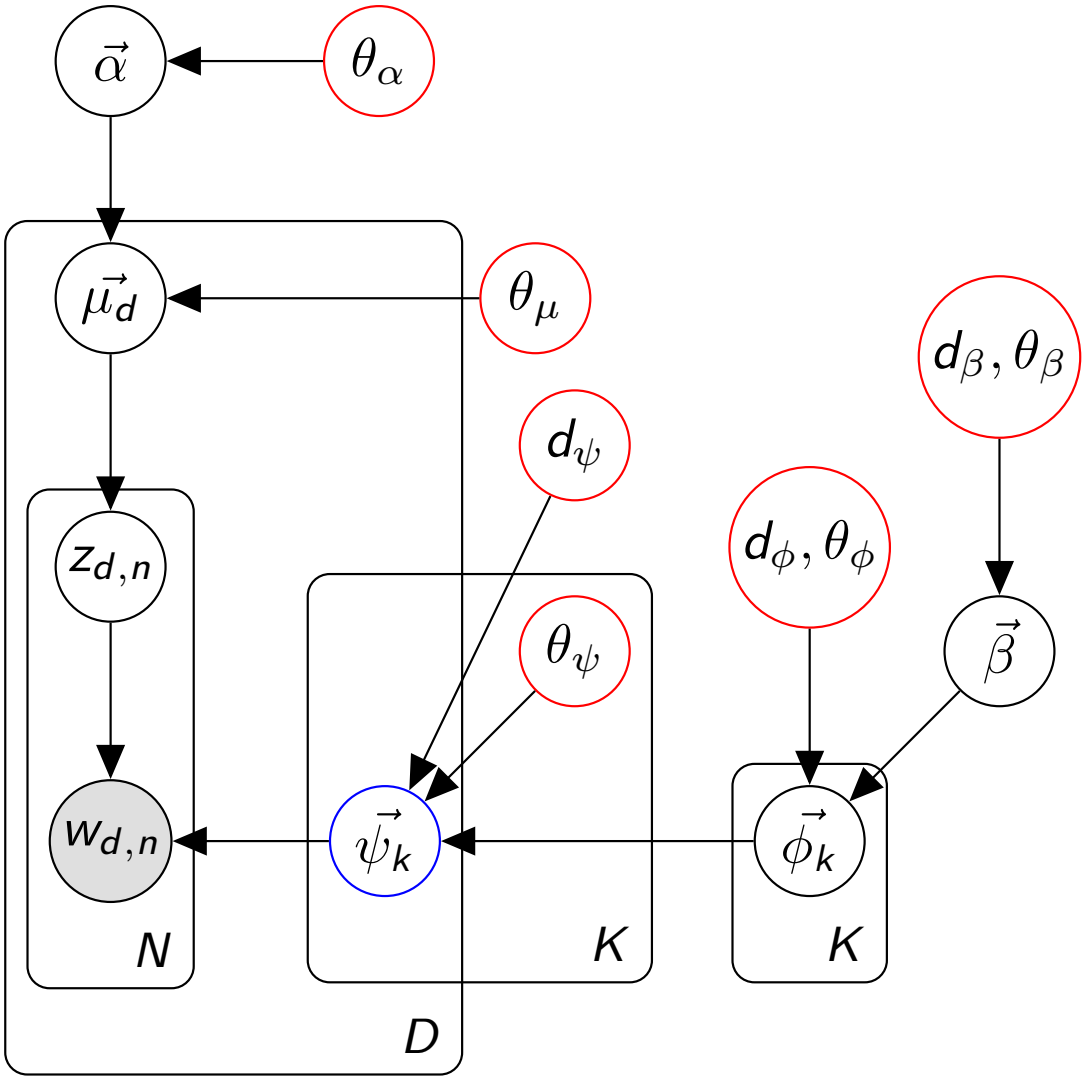
The key variables being sampled and their relevant terms are:

$$\overbrace{\mathcal{S}^{m_{i,j,k}}_{t_{i,j,k},0}\frac{\gamma_{x_{i,j,l}}+n_{k,x_{i,j,l}}}{\sum_v(\gamma_v+n_{k,v})}}^{z_{i,j,l}=k}\quad\overbrace{\frac{\beta^{t_{i,j,k}}}{(\alpha_0)_{T_{i,\cdot}}}\mathcal{S}^{t_{i,\cdot,k}}_{s_{i,k},0}\mathcal{S}^{m_{i,j,k}}_{t_{i,j,k},0}}^{t_{i,j,k}}\quad\overbrace{\beta^{s_{i,k}}\mathcal{S}^{t_{i,\cdot,k}}_{s_{i,k},0}}^{s_{i,k}}$$

- Note $t_{i,j,k}$ is correlated with $m_{i,j,k}$, and $s_{i,k}$ is correlated with $t_{i,j,k}$.
- Option is to sample sequentially $z_{i,j,l}$, $t_{i,j,k}$ and $s_{i,k}$ (*i.e.*, sweeping up the hierarchy) in turn,
  - can be expensive if full sampling ranges done, *e.g.*, $s_{i,k} : 1 \leq s_{i,k} \leq t_{i,\cdot,k}$.
- In practice, works OK, but is not great: mixing is poor!

Outline

Bursty Non-parametric Topic Model



- from "Experiments with Non-parametric Topic Models," Buntine and Mishra, KDD 2014
- extends LDA with hierarchical PYP on word side and a hierarchical DP on the document side
- hyperparameters (for PYPs and DPs) are red nodes
- also makes each topic bursty, see the blue $\vec{\psi}_k$

Bursty Non-parametric Topic Model, cont.

Read off the marginalised posterior as follows:

$$\frac{\text{Beta}\left(\vec{n}^{\alpha} + \theta_\alpha \vec{1}/K\right)}{\text{Beta}\left(\theta_\alpha \vec{1}/K\right)} \prod_{d=1}^{D} F_{\theta_\mu}(\vec{n_d}^{\mu}, \vec{t_d}^{\mu}) \qquad \text{(document side)}$$

$$F_{d_\beta,\theta_\beta}(\vec{n}^{\beta}, \vec{t}^{\beta}) \prod_{k=1}^{K} F_{d_\psi,\theta_\psi}(\vec{n_k}^{\psi}, \vec{t_k}^{\psi}) F_{d_\phi,\theta_\phi}(\vec{n_k}^{\phi}, \vec{t_k}^{\phi}) \qquad \text{(word side)}$$

where

$$\vec{n}^{\alpha} = \sum_{d=1}^{D} \vec{t_d}^{\mu} \ , \qquad \vec{n_k}^{\phi} = \vec{t_k}^{\psi} \ , \qquad \vec{n}^{\beta} = \sum_{k=1}^{K} \vec{t_k}^{\phi} \ ,$$

plus all the constraints hold, such as

$$\forall_{k,w} \left(n^{\phi}_{k,w} \geq t^{\phi}_{k,w} \ \ \& \ \ n^{\phi}_{k,w} > 0 \text{ iff } t^{\phi}_{k,w} > 0 \right)$$

Summary: Simple PYP Sampling

- probabilities in each PYP hierarchy are marginalised out from the bottom up.
- simple sampling strategy $\equiv$ sample the numbers of tables vectors $(\vec{t})$
- with the n-gram, the leaves of the PYP hierarchy are observed, and a simple sampling strategy works well
  - Teh tried this (2006a, p16) but says "it is expensive to compute the generalized Stirling numbers."
- with unsupervised models generally, like STM, the leaves of the PYP hierarchy are unobserved and the simple sampling strategy gives poor mixing
- on more complex models, not clear simple sampling strategy is any better than hierarchical CRP sampling

Summary: What You Need to Know

PYPs on discrete domains: samples from the SSM get duplicates

PItan-Yor-Multinomial: the PYP variant of the Dirichlet-Multinomial

N-grams: simple example of a hierarchical PYP/DP model

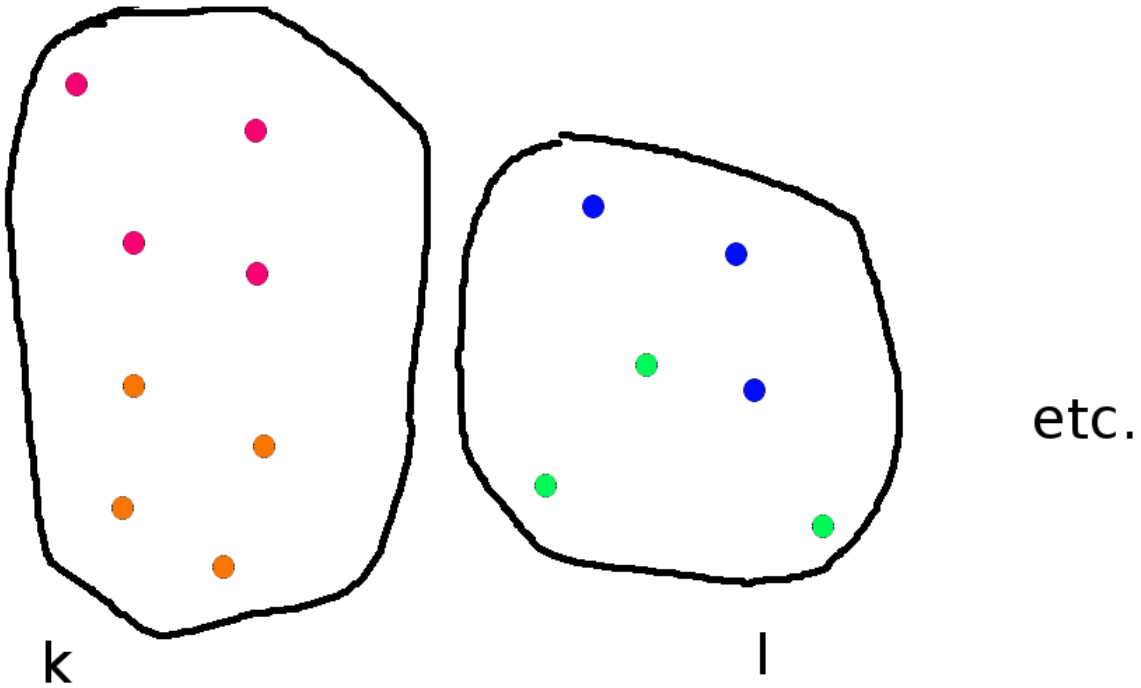
Hierarchical PYPs: the hierarchy of probability vectors are marginalised out leaving a hierarchy of number of tables vectors corresponding to the count vectors.

Structured topic model: STM is a simple extension to LDA showing hierarchical PYPs, see Du, Buntine and Jin (2012)

Simple Gibbs sampling: sampling number of tables vectors individually is poor due to poor mixing.

Outline

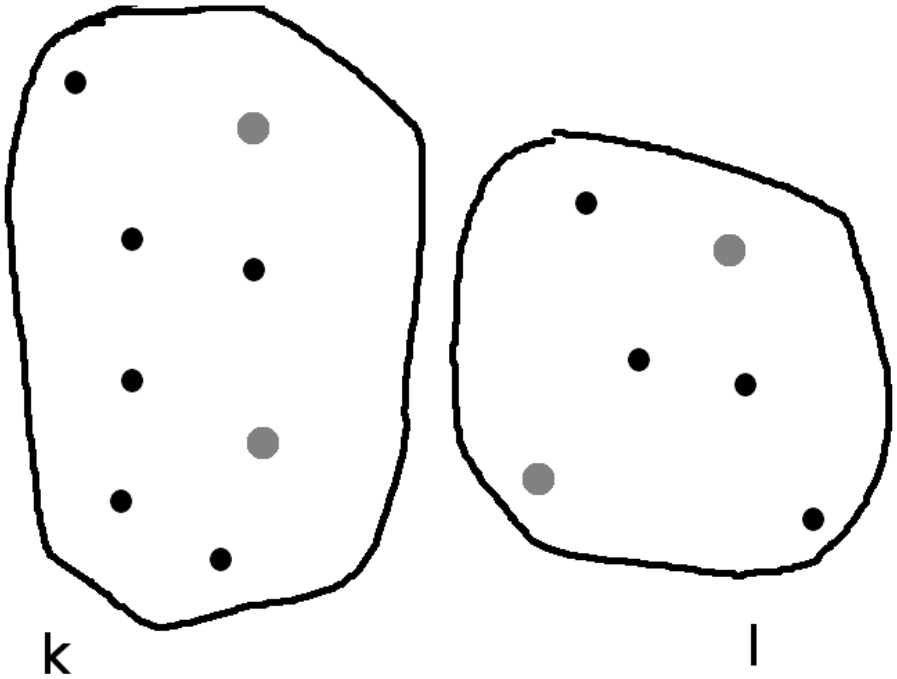
Species with Subspecies



Within species there are separate *sub-species*, pink and orange for type k, blue and green for type l.

Chinese restaurant samplers work in this space, keeping track of all counts for sub-species.

Species with New Species

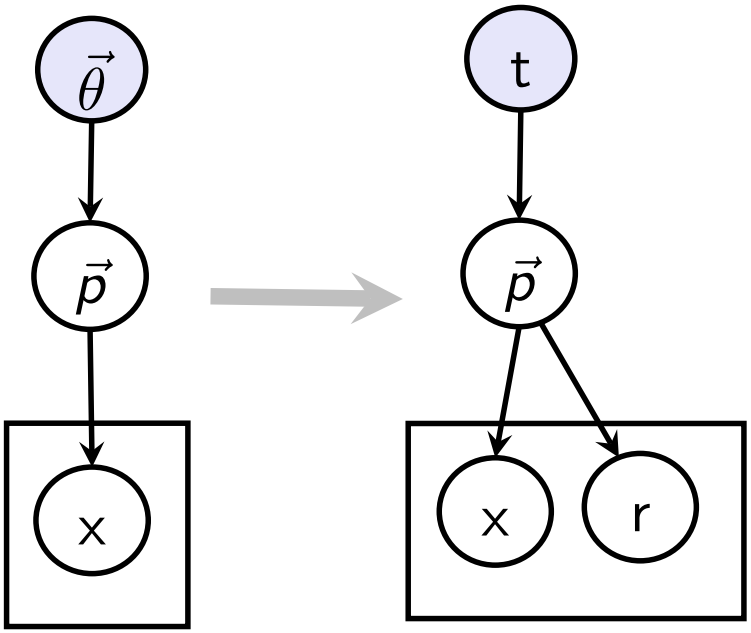


etc.

Within species there are separate *sub-species*, but we only know which data is the first of a new sub-species.

Block table indicator samplers work in this space, where each datum has a Boolean indicator.

Categorical Data plus Table Indicators



LHS = *categorical* form with sample of discrete values $x_1, ..., x_N$ drawn from categorical distribution $\vec{p}$ which in turn has mean $\vec{\theta}$

RHS = *species sampling* form where data is now pairs $(x_1, r_1)..., (x_N, r_N)$ were r_n is a Boolean indicator saying "is new subspecies"

$r_n = 1$ then the sample x_n was drawn from the parent node with probability θ_{x_n}, otherwise is existing subspecies

Table Indicators

Definition of table indicator

Instead of considering the Pitman-Yor-multinomial with counts $(\vec{n}, \vec{t})$, work with sequential data with individual values $(x_1, r_1), (x_2, r_2), ..., (x_N, r_N)$. The table indicator r_n indicates that the data contributes one count up to the parent probability.

So the data is treated sequentially, and taking statistics of $\vec{x}$ and $\vec{r}$ yields:

$$\begin{aligned} n_k &:= \text{counts of } k\text{'s in } \vec{x}, \\ &= \sum_{n=1}^{N} 1_{x_n=k}, \\ t_k &:= \text{counts of } k\text{'s in } \vec{x} \text{ co-occuring with an indicator}, \\ &= \sum_{n=1}^{N} 1_{x_n=k} 1_{r_n}. \end{aligned}$$

The Pitman-Yor-Categorical

Definition of Pitman-Yor-Categorical

Given a concentration parameter α, a discount parameter d, a probability vector $\vec{\theta}$ of dimension L, and a count N, the Pitman-Yor-categorical distribution creates a sequence of discrete class assignments and indicators $(x_1, r_1), ...(x_N, r_N)$. Now $(\vec{x}, \vec{r}) \sim \text{CatPY}\left(d, \alpha, \vec{\theta}, N\right)$ denotes

$$p\left(\vec{x}, \vec{r} \middle| N, \text{CatPY}, d, \alpha, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{l=1}^{L} \mathcal{S}_{t_l,d}^{n_l} \theta_l^{t_l} \binom{n_l}{t_l}^{-1}$$

where the counts are derived, $t_l = \sum_{n=1}^{N} 1_{x_n=l} 1_{r_l}$, $n_l = \sum_{n=1}^{N} 1_{x_n=l}$, $T = \sum_{l=1}^{L} t_l$.

The Categorical- versus Pitman-Yor-Multinomial

Pitman-Yor-Multinomial: working off counts $\vec{n}, \vec{t}$,

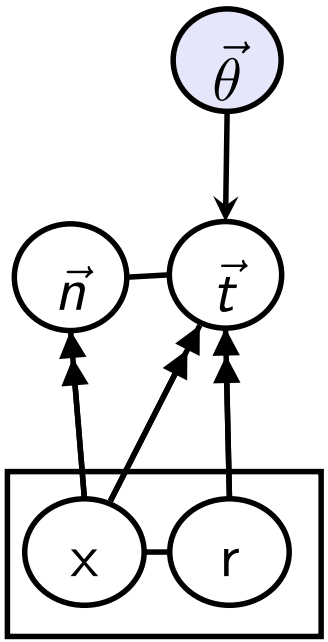
$$p\left(\vec{n}, \vec{t} \middle| N, \text{MultPY}, d, \alpha, \vec{\theta}\right) = \binom{N}{\vec{n}} \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}_{t_k,d}^{n_k} \theta_k^{t_k}$$

Pitman-Yor-Categorical: working off sequential data $\vec{x}, \vec{r}$, the counts $\vec{n}, \vec{t}$ are now derived,

$$p\left(\vec{x}, \vec{r} \middle| N, \text{CatPY}, d, \alpha, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}_{t_k,d}^{n_k} \theta_k^{t_k} \binom{n_k}{t_k}^{-1}$$

- remove the $\binom{N}{\vec{n}}$ term because sequential order now matters
- divide by $\binom{n_k}{t_k}$ because this is the number of ways of distributing the t_k indicators that are on amongst n_k places

Pitman-Yor-Categorical Marginalisation

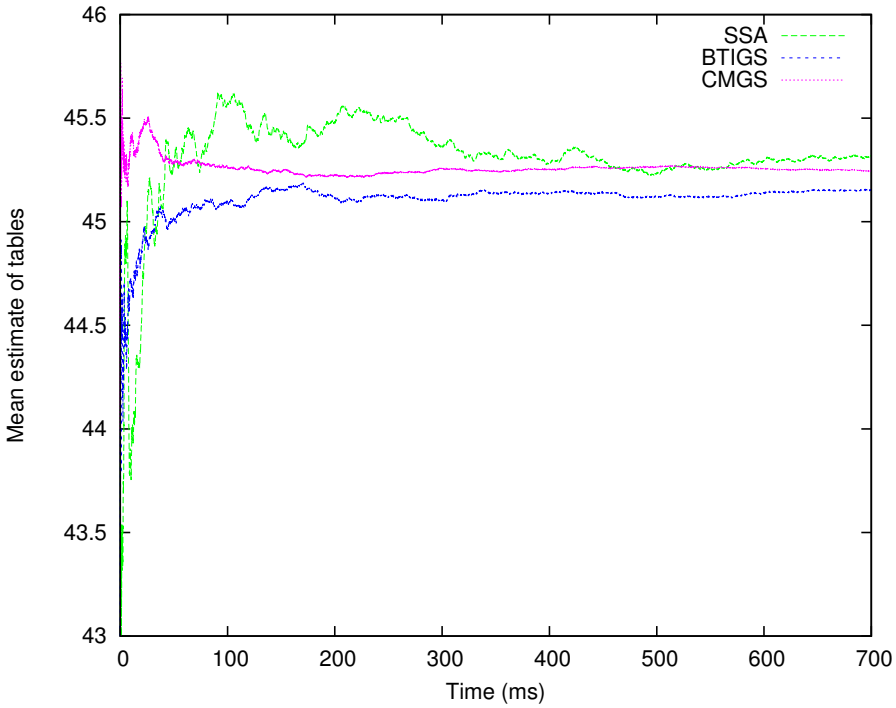


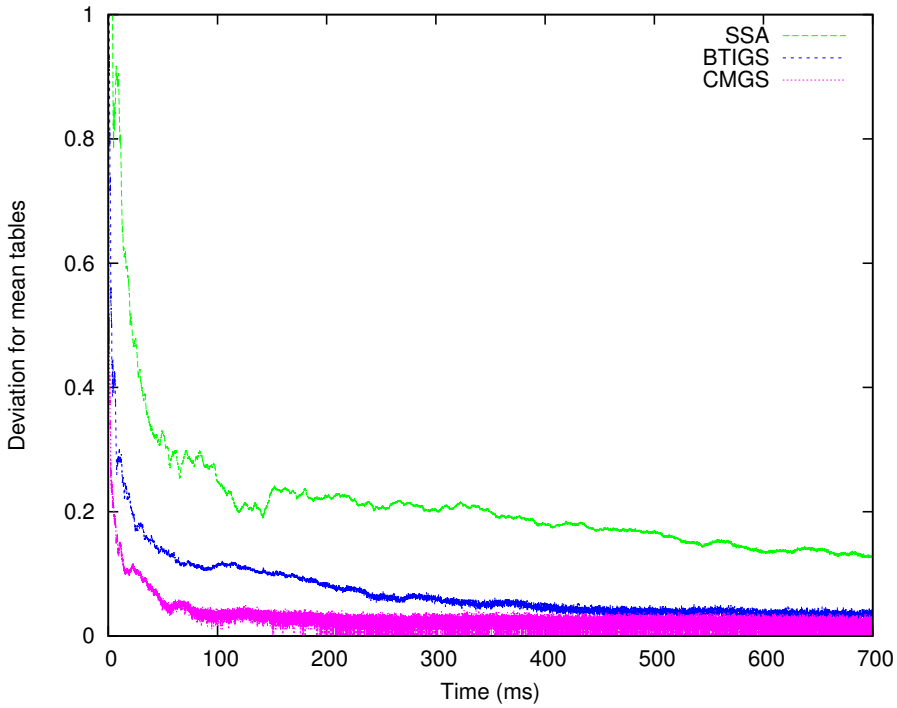
$\vec{n} =$ *vector of counts* of different species (how much data of each species);
computed from the data $\vec{x}$

$\vec{t} =$ count vector giving how many different subspecies; computed from the paired data $\vec{x}, \vec{r}$;
called *number of tables*

$$p\left(\vec{x}, \vec{r} \middle| d, \alpha, \text{PYP}, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \theta_k^{t_k} \mathcal{S}_{t_k,d}^{n_k} \binom{n_k}{t_k}^{-1}$$

Comparing Samplers for CatPY versus MultPY

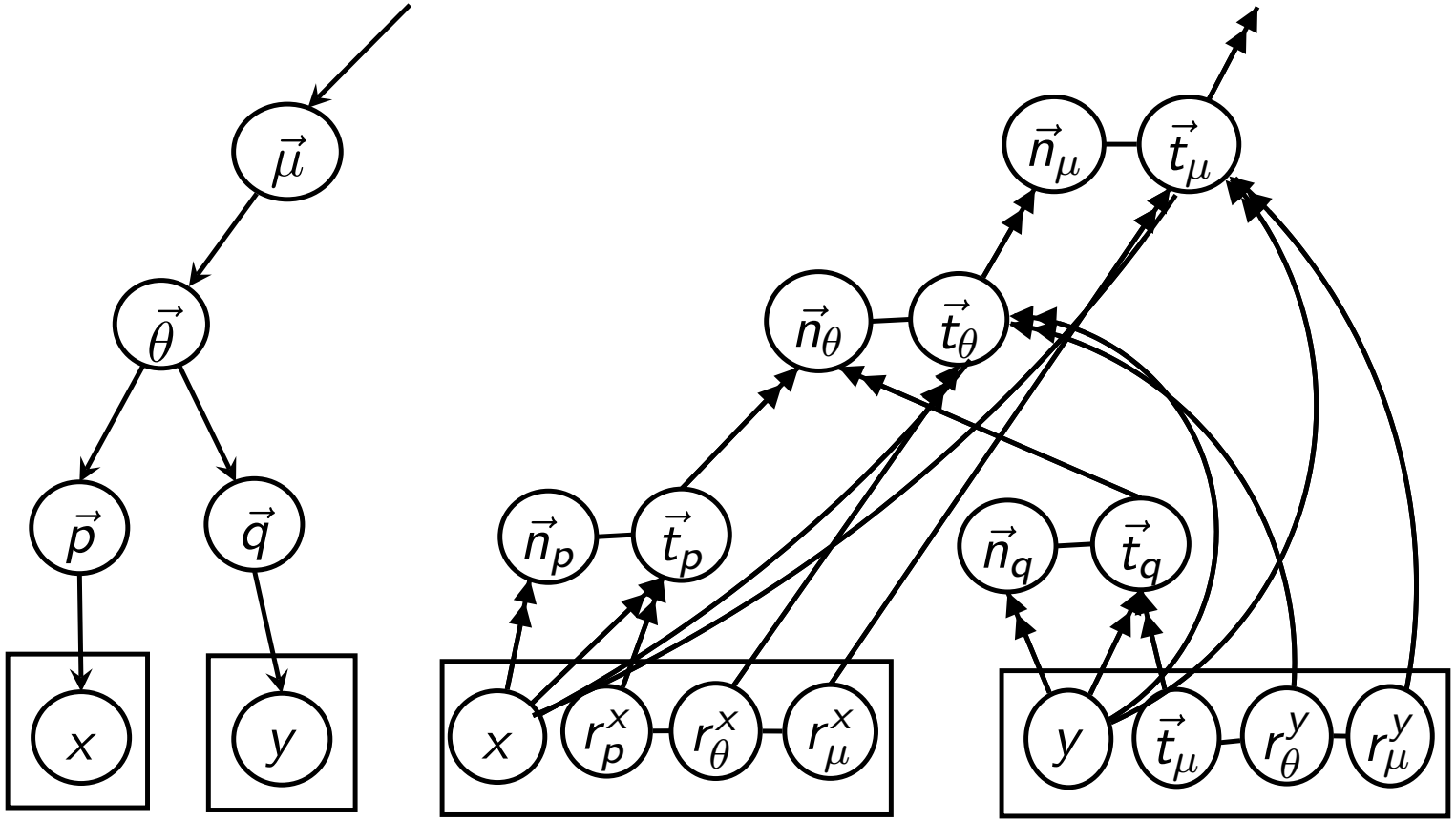




Legend: SSA = "standard CRP sampler of Teh *et al.*"
BTIGS = "Gibbs sampler using CatPY posterior"
CMGS = "Gibbs sampler using MultPY posterior"

Mean estimates of the total number of tables T for one of the 20 Gibbs runs (left) and the standard deviation of the 20 mean estimates (right) with $d = 0$, $\alpha = 10$, $K = 50$ and $N = 500$.

Hierarchical Marginalisation



left is the original probability vector hierarchy, **right** is the result of marginalising out probability vectors then

- indicators are attached to their originating data as a set
- all $\vec{n}$ and $\vec{t}$ counts up the hierarchy are computed from these

Outline

Using the (Categorical) Evidence Formula

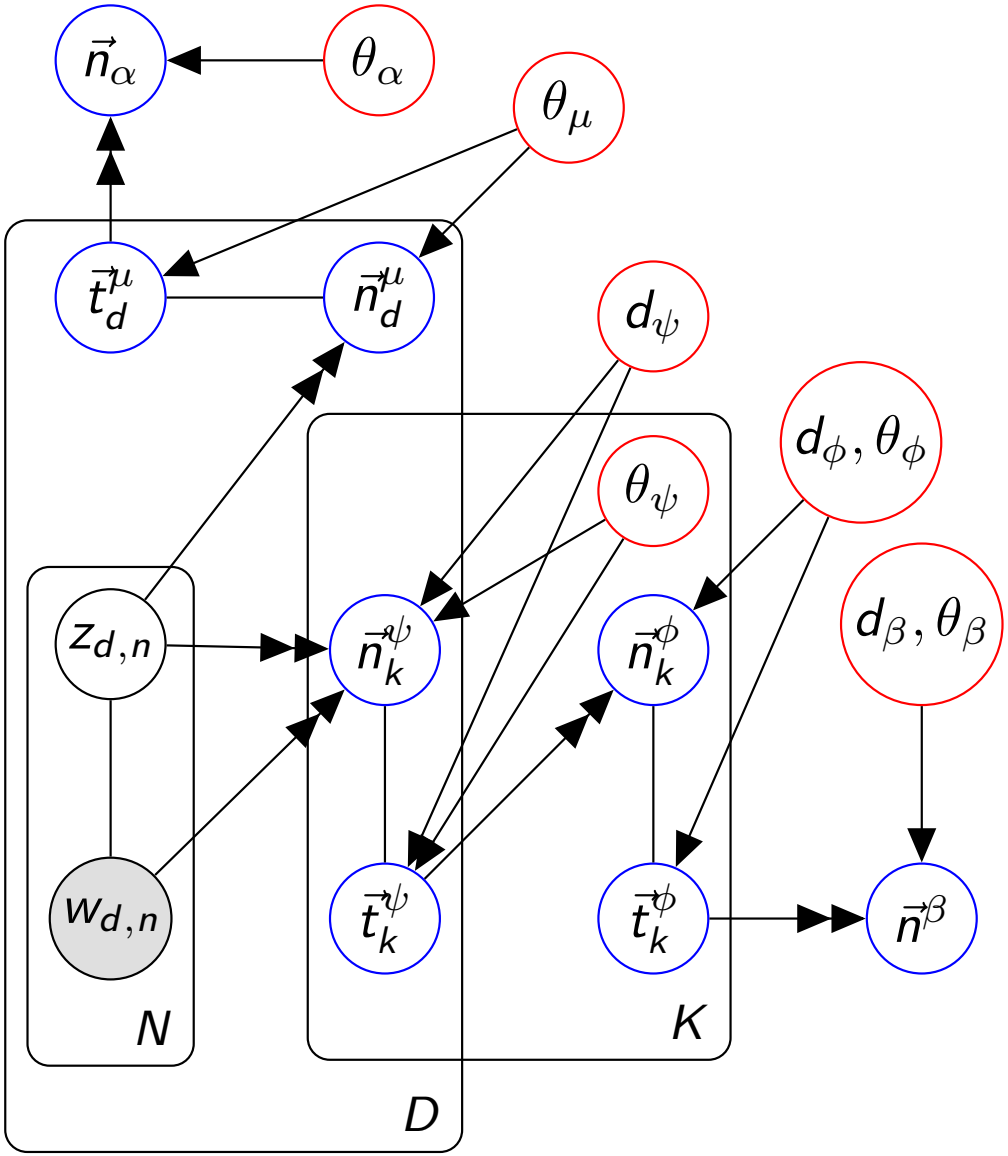
We will repeatedly apply the evidence formula

$$\begin{aligned} p\left(\vec{x}, \vec{t} \,\middle|\, N, \text{CatPY}, d, \alpha\right) &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} \binom{n_k}{t_k}^{-1} H(k)^{t_k} \\ &= F'_{d,\alpha}(\vec{n}, \vec{t}) \prod_{k=1}^{K} H(k)^{t_k} \end{aligned}$$

to marginalise out all the probability vectors where

$$F'_{d,\alpha}(\vec{n}, \vec{t}) = F_{d,\alpha}(\vec{n}, \vec{t}) \prod_{k=1}^{K} \binom{n_k}{t_k}^{-1}$$

Marginalised Bursty Non-parametric Topic Model



- started with two hierarchies $\vec{\mu}_d \to \vec{\alpha}$ and $\vec{\psi}_k \to \vec{\phi}_k \to \vec{\beta}$
- counts (in blue) $\vec{n}_d^{\,\mu}$, $\vec{n}^\alpha$, $\vec{n}_k^{\,\psi}$, $\vec{n}_k^{\,\phi}$ and $\vec{n}^\beta$ introduced, and their numbers of tables $\vec{t}_d^{\,\mu}$, *etc.*
- root of each hierarchy modelled with an improper Dirichlet so no $\vec{t}^{\,\alpha}$ or $\vec{t}^{\,\beta}$
- table indicators, not shown, are $r_{d,n}^{\mu}$, $r_{d,n}^{\psi}$, and $r_{d,n}^{\phi}$
- all counts and numbers of tables can be derived from topic $z_{d,n}$ and indicators

Bursty Non-parametric Topic Model, cont.

Modify the evidence to add choose terms to get:

$$E = \frac{\text{Beta}\left(\vec{n}^{\alpha} + \theta_\alpha \vec{1}/K\right)}{\text{Beta}\left(\theta_\alpha \vec{1}/K\right)} \prod_{d=1}^{D} F'_{\theta_\mu}(\vec{n_d}^{\mu}, \vec{t_d}^{\mu}) \qquad \text{(document side)}$$

$$F'_{d_\beta,\theta_\beta}(\vec{n}^{\beta}, \vec{t}^{\beta}) \prod_{k=1}^{K} F'_{d_\psi,\theta_\psi}(\vec{n_k}^{\psi}, \vec{t_k}^{\psi}) F'_{d_\phi,\theta_\phi}(\vec{n_k}^{\phi}, \vec{t_k}^{\phi}) \qquad \text{(word side)}$$

where totals and constraints hold as before, derived as

$$n^{\mu}_{d,k} = \sum_{n=1}^{N} 1_{z_{d,n}=k} \qquad t^{\mu}_{d,k} = \sum_{n=1}^{N} 1_{z_{d,n}=l} 1_{r^{\mu}_{d,n}}$$

$$n^{\psi}_{k,w} = \sum_{n=1}^{N} 1_{z_{d,n}=k} 1_{w_{d,n}=w} \qquad t^{\psi}_{k,w} = \sum_{n=1}^{N} 1_{z_{d,n}=k} 1_{w_{d,n}=w} 1_{r^{\psi}_{d,n}}$$

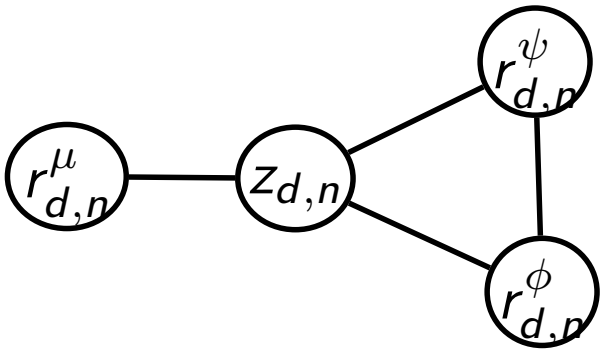
$$n^{\phi}_{k,w} = t^{\psi}_{k,w} \qquad t^{\phi}_{k,w} = \sum_{n=1}^{N} 1_{z_{d,n}=k} 1_{w_{d,n}=w} 1_{r^{\psi}_{d,n}} 1_{r^{\phi}_{d,n}}$$

$$n^{\beta}_{w} = \sum_{k} t^{\phi}_{k,w} \qquad t^{\beta}_{w} = 1_{n^{\beta}_{w}>0}$$

Block Sampler for Bursty Non-parametric Topic Model

At the core of the block Gibbs sample, we need to reestimate $(z_{d,n}, r^{\mu}_{d,n}, r^{\psi}_{d,n}, r^{\phi}_{d,n})$ for all documents d and words n.

Considering the evidence (previous slide) as a function of these, $E(z_{d,n}, r^{\mu}_{d,n}, r^{\psi}_{d,n}, r^{\phi}_{d,n})$, we get the graphical model below left:



With belief propagation algorithms, it is easy to:

- compute the marginal contribution for $z_{d,n}$,

$$\sum_{r^{\mu}_{d,n}, r^{\psi}_{d,n}, r^{\phi}_{d,n}} E(z_{d,n}, r^{\mu}_{d,n}, r^{\psi}_{d,n}, r^{\phi}_{d,n})$$

  needed for a block Gibbs sampler
- sample $(r^{\mu}_{d,n}, r^{\psi}_{d,n}, r^{\phi}_{d,n})$ for given $z_{d,n}$

LDA Versus NP-LDA Samplers

	LDA	NP-LDA with table indicators
latent vars	word topics $\vec{z}$	word topics $\vec{z}$, and boolean table indicators $r^{\vec{\mu}}$, $r^{\vec{\psi}}$, $r^{\vec{\phi}}$
derived vectors	topic count $\vec{n}_d^{\mu}$ and word count $\vec{n}_k^{\phi}$	topic count $\vec{n}_d^{\mu}, \vec{t}_d^{\mu}, \vec{n}^{\alpha}$ word count $\vec{n}_k^{\phi}, \vec{n}_k^{\psi}, \vec{t}_k^{\psi}, \vec{n}^{\beta}$
totals kept	$\vec{N}_d^{\mu}$, $\vec{N}_k^{\phi}$	$\vec{N}_d^{\mu}$, $\vec{T}_d^{\mu}$, $\vec{N}_k^{\phi}$, $\vec{N}_k^{\psi}$, $\vec{T}_k^{\psi}$
Gibbs method	on each $z_{d,n}$	blockwise on $(z_{d,n}, r_{d,n}^{\mu}, r_{d,n}^{\psi}, r_{d,n}^{\phi})$

Notes:

- table indicators don't have to be stored but can be resampled as needed by uniform assignment.
- block sampler and posterior form with table indicators are more complex!

Outline

The Hierarchy Algorithm: Context

- a general purpose scheme for performing block Gibbs sampling on a probability network with one or more embedded PYP hierarchies
- assume the probability vectors are marginalised out to introduce counts and numbers of tables, as in previous examples
- all the inverse choose terms, $\binom{n_k}{t_k}^{-1}$, are added to get evidence E
- the assumptions required are as follows:
  - a single latent variable z (such as "topic" or "cluster") is to be sampled
  - z only affects a single leaf in each affected PYP hierarchy,
  - the affected PYP hierarchies otherwise are independent

e.g. formula for E and the undirected graph on $z_{d,n}$ and table indicators just given for the bursty NP-LDA model

e.g. Blunsom and Cohn's unsupervised part-of-speech model fails assumptions

The Block Gibbs Sampler

An instance of the latent variable z, indexed by I is to be sampled.
e.g., $I = (d, n)$ for the given bursty NP-LDA model

- the table indicators associated with z_I are removed from the statistics, using `RemoveSample()` which does the following:
  - the table indicators are not stored as they can be resampled as needed
  - if removing the indicators from the statistics yields an inconsistent state, then the whole resampling step for z_I is skipped
  - otherwise, the indicators are removed from the statistics
- sample the latent variable z_I with the help of `PosteriorWeight()`
  - sampling is done by computing E for z_I summed over possible associated table indicators
  - for each value k of the latent z_I, compute `PosteriorWeight`(I, z) and sample z_I proportionally
- once z_I is sampled, we sample its corresponding table indicators for each affected PYP hierarchy according to E

Resample and Remove Table Indicators: `RemoveSample`(I)

This will be called for an instance of latent variable z to resample its table indicators. If their removal yields a consistent state, then remove them, otherwise return false and cancel sampling z_I.

Input: index set I to uniquely identify variable to sample z_I

$\mathcal{S} \leftarrow$ set of probability vector nodes in un-marginalised graph neighbouring z_I
if any pair of nodes are in same PYP hierarchy **then**
  return false {*model inappropriate*}
end if
$\mathcal{I} = \{(\theta, \texttt{Resample}(\theta, \text{value}(z_I))) : \theta \in \mathcal{S}\}$
if $(\cdot, \textbf{false}) \in \mathcal{I}$ **then**
  return false {*a* `Resample()` *call returned false*}
end if
foreach $(\theta.\mu) \in \mathcal{I}$ **do**
  $\texttt{Remove}(\theta, \text{value}(z_I), \mu)$ {*remove from stats*}
end for
return true

Resampling Table Indicators: Resample(θ, k)

This will be called for every PYP hierarchy the latent variable z affects. It returns a node μ below which all indicators are sampled to be on. Too be used in Remove(θ, k, μ),

Input: probability vector node θ and type k

1: **if** parent$(\theta) \equiv$ undefined **then**
2: **return** θ {*no parent*}
3: **end if**
4: $(n_k^\theta, t_k^\theta) \leftarrow$ the (count,number of tables) of type k for $\vec{\theta}$
5: **if** $t_k^\theta <$ rand$(0,1) n_k^\theta$ **then**
6: **return** θ {*recursion stops because indicator off*}
7: **else if** $n_k^\theta > 1$ and $t_k^\theta \equiv 1$ **then**
8: **return false** {*illegal state so abandon sampling*}
9: **end if**
10: **return** Resample(parent$(\theta), k)$)

Output: return node below which all indicators are on, else return **false**

The root node counts may need to be treated separately; not done here for simplicity.

Remove Table Indicators from Stats: $\texttt{Remove}(\theta, k, \mu)$

All nodes ψ from θ up its parents to the child of μ currently have their table indicator r_l^ψ set to 1. So remove these from corresponding counts.

Input: probability vector node θ, type k and stopping node μ

1: $n_k^\theta \leftarrow n_k^\theta - 1$
2: **if** $\theta \equiv \mu$ **then**
3: **return** {*recursion stops here*}
4: **end if**
5: $t_k^\theta \leftarrow t_k^\theta - 1$
6: $\texttt{Remove}(\text{parent}(\theta), k, \mu)$ {*recursion*}

Compute Bayes Factor: PosteriorWeight(I, k)

This will be called for an instance of latent variable, $z_i = k$ to compute its relative weight, so z_i can be sampled. Let $\vec{r}_I$ denote the table indicators associated with z_I and $E(z_i, \vec{r}_I)$ is the evidence with choose terms added, extended to include the affect of $z_i, \vec{r}_I$.

PosteriorWeight(I, k) :

Input: index set I to uniquely identify variable z to sample, and value k

1: **return** $\sum_{\vec{r}_I} E(z_i = k, \vec{r}_I)$.

To understand this summation, consider Graph(I, k) :

1: $\mathcal{S} \leftarrow$ set of probability vector nodes in un-marginalised graph neighbouring z_I
2: $\mathcal{G} \leftarrow \bigcup_{\theta \in \mathcal{S}}$ (the ancestral set of θ connected in single chain)
3: add z_I to $\mathcal{G}$ and connect it with an undirected arc to every other node
4: **return** $\mathcal{G}$.

Note that Graph(I, k) is chordal and each term in $E(z_i, \vec{r}_I)$ can be assigned to one of the cliques in the graph.

Other Variations

- Hidden Markov models with higher-order n-grams on the hidden topics have multiple parts of the n-gram hierarchy affected by the hidden state at a given position. See Blunsom and Cohn (2013).
- Splitting leaf data in a pre-determined way requires segmentation of the indicator variables too; for instance, document segmentation, see Du, Buntine and Johnson (2013),
- PYP hierarchies with cross links (so the PYPs no longer occur in a tree, so nodes can have multiple parents) require extra book-keeping. See Wood and Teh (2009) and Du, Buntine and Jin (2012).

Outline

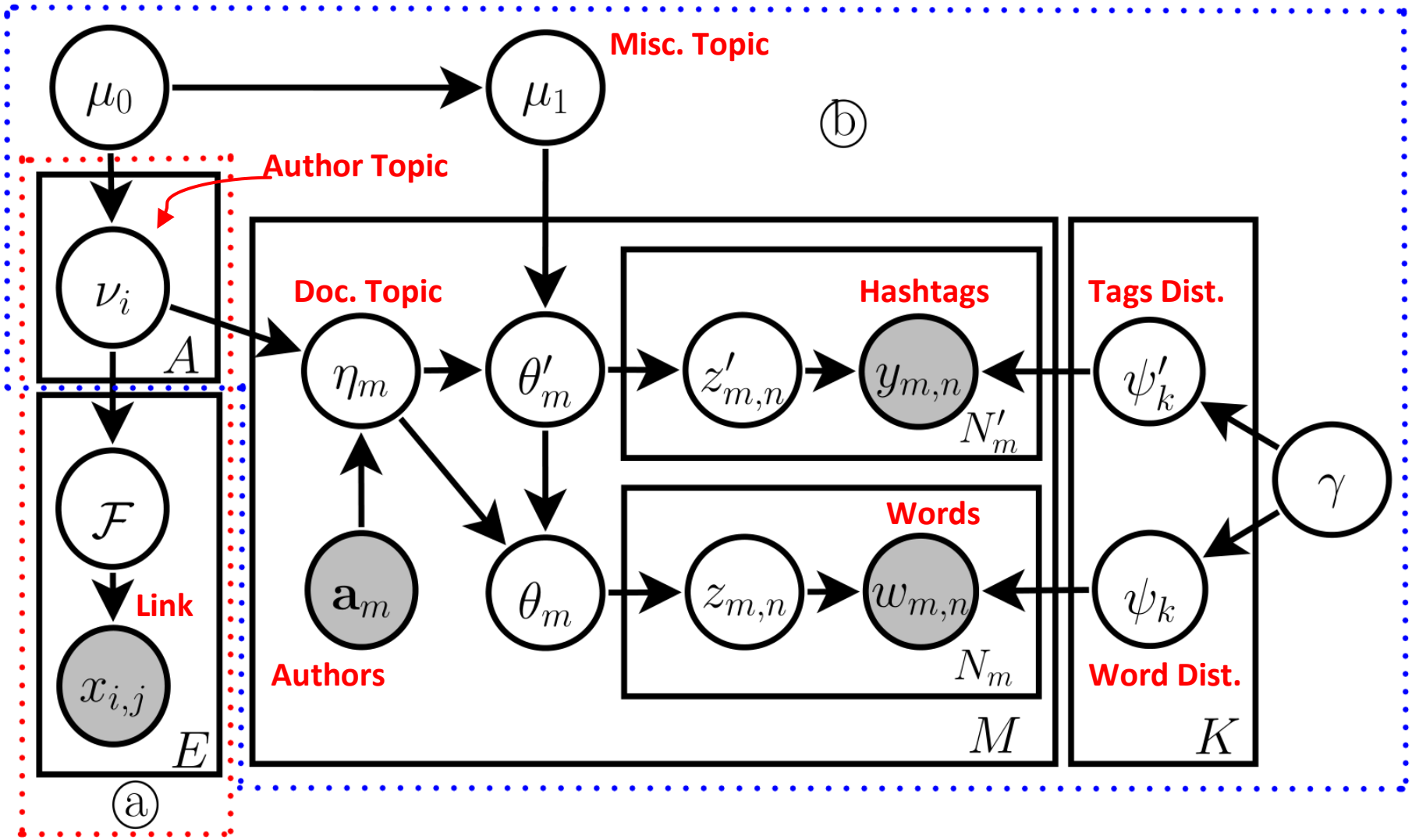
Analysing Tweets

Tweets have a number of facts that make them novel/challenging to study:

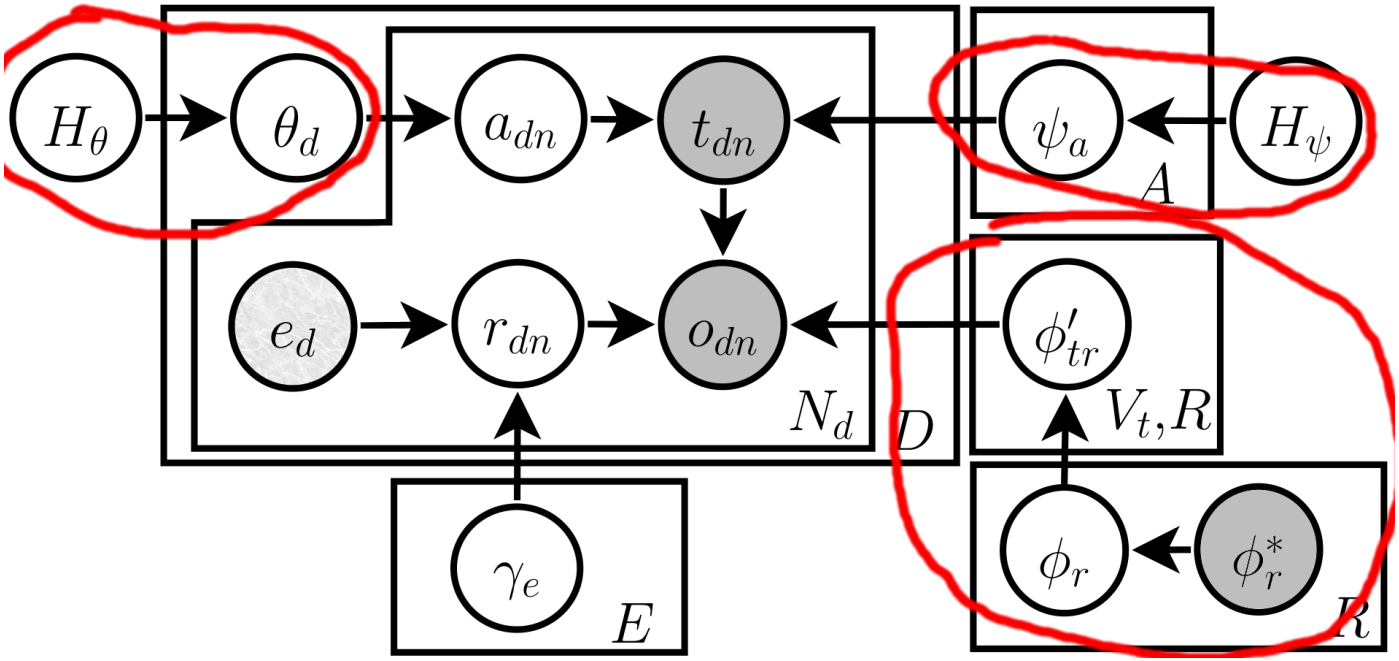
- hashtags, embedded URLs, and retweets,
- small size, informal language and emoticons.
- authors and follower networks,
- frequency in time.

Twitter-Network Topic Model

Kar Wai Lim *et al.*, 2014, submitted



Twitter Opinion Topic Model



"Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon," Lim and Buntine, CIKM 2014

(probability vector hierarchies circled in red)

Unsupervised Part of Speech

A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction

Phil Blunsom
Department of Computer Science
University of Oxford

Trevor Cohn
Department of Computer Science
University of Sheffield

We develop a trigram hidden Markov model which models the joint probability of a sequence of latent tags, $\mathbf{t}$, and words, $\mathbf{w}$, as

$$P_\theta(\mathbf{t}, \mathbf{w}) = \prod_{l=1}^{L+1} P_\theta(t_l|t_{l-1}, t_{l-2}) P_\theta(w_l|t_l) \,,$$

Unsupervised Part of Speech, cont.

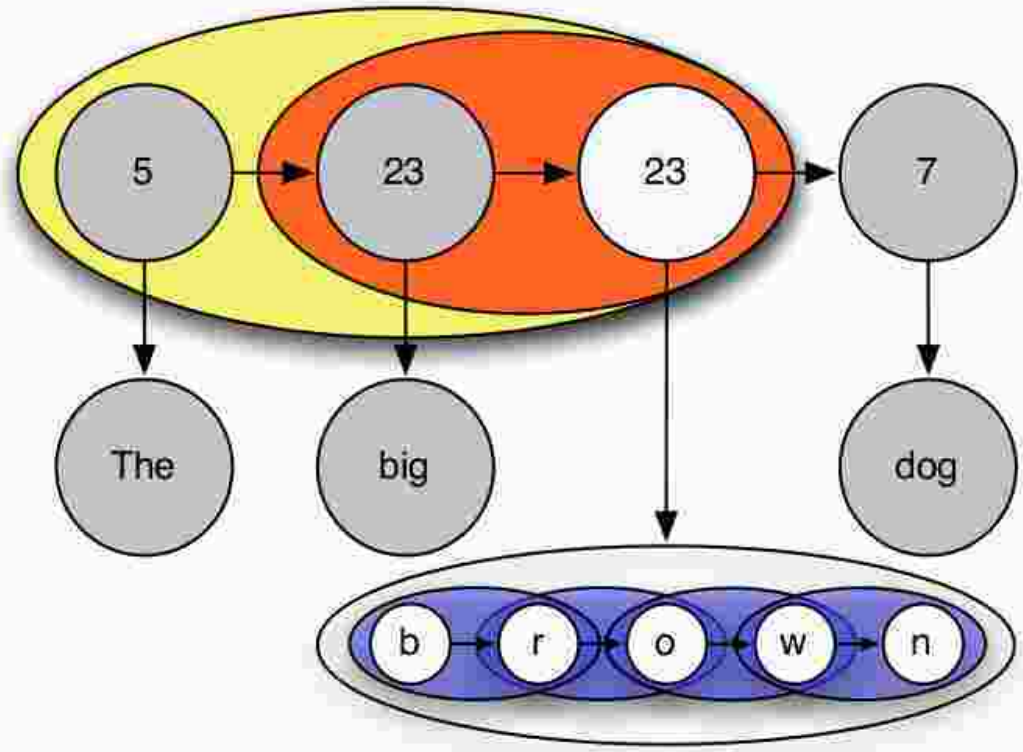


Figure 2: The conditioning structure of the hierarchical PYP with an embedded character language models.

Unsupervised Part of Speech, cont.

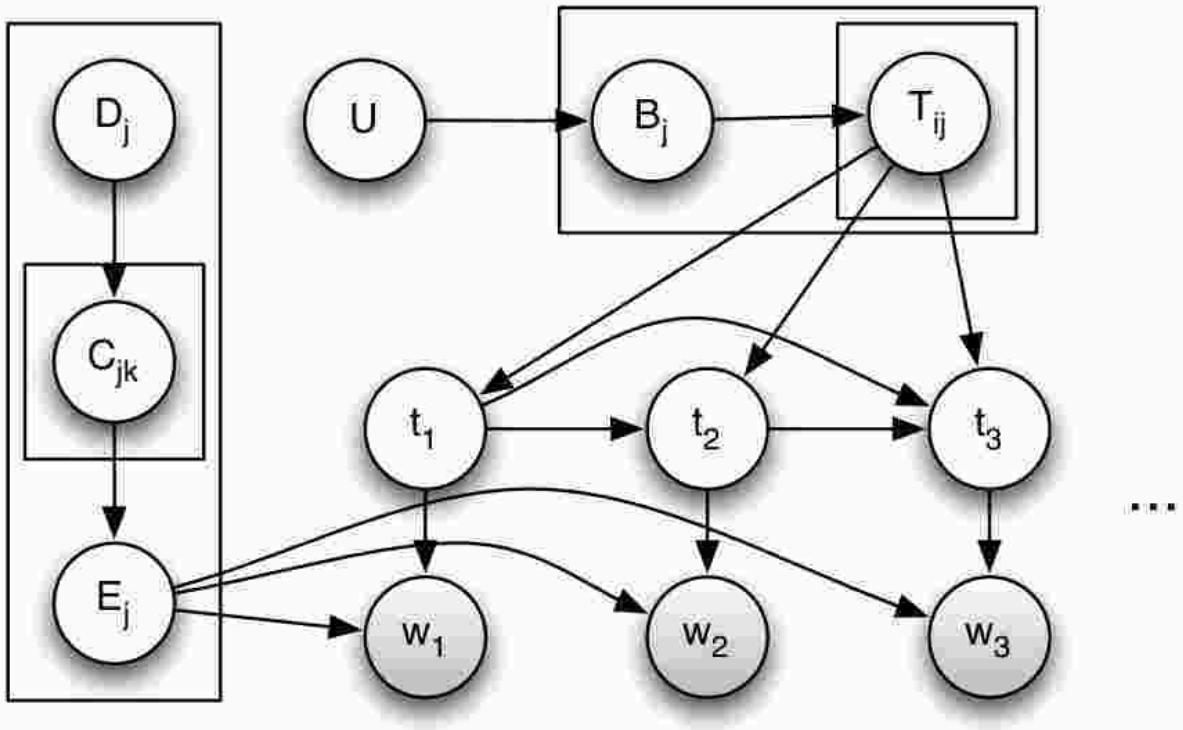
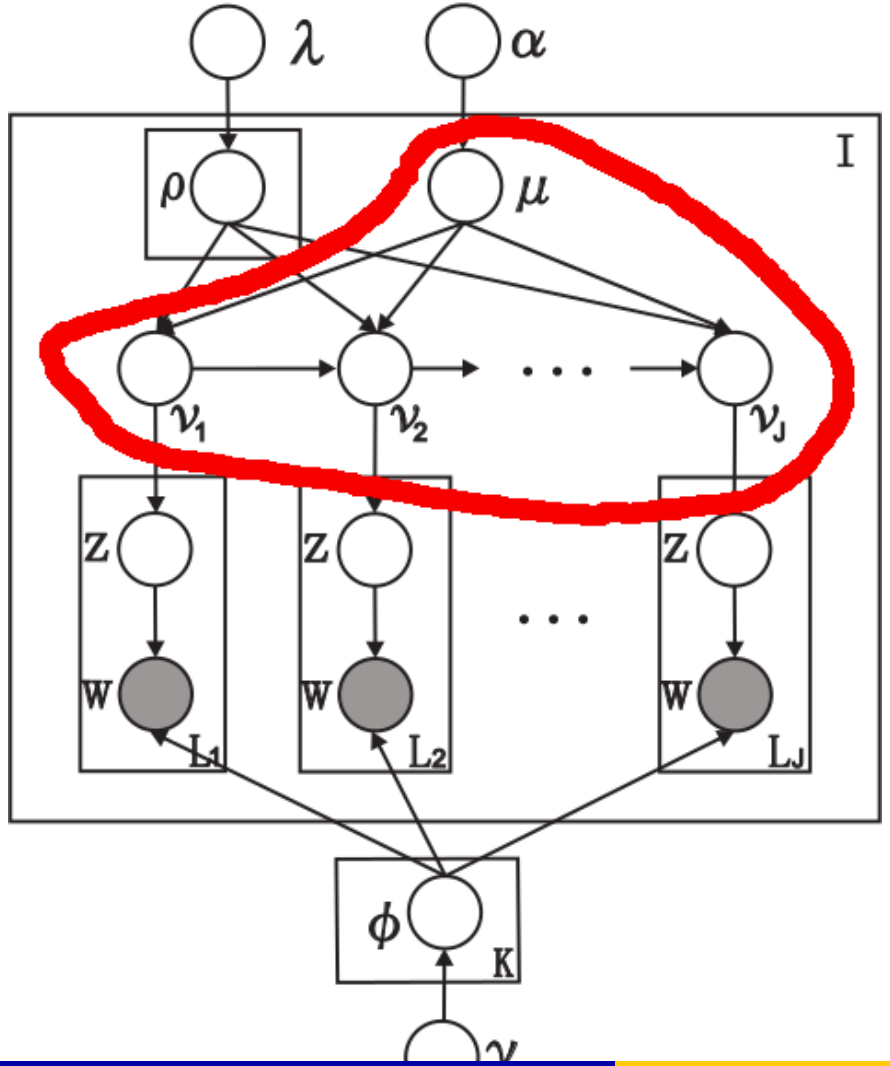


Figure 1: Plate diagram representation of the trigram HMM. The indexes i and j range over the set of tags and k ranges over the set of characters. Hyper-parameters have been omitted from the figure for clarity.

Adaptive Sequential Topic Model

A more complex (sequential) document model.



- The PYPs exist in long chains $(\vec{\nu}_1, \vec{\nu}_2, ..., \vec{\nu}_J)$.
- A single probability vector $\vec{\nu}_j$ can have two parents, $\vec{\nu}_{j-1}$ and $\vec{\mu}$.
- More complex chains of table indicators and block sampling.
- See Du *et al.*, EMNLP 2012.

Dynamic Topic Model

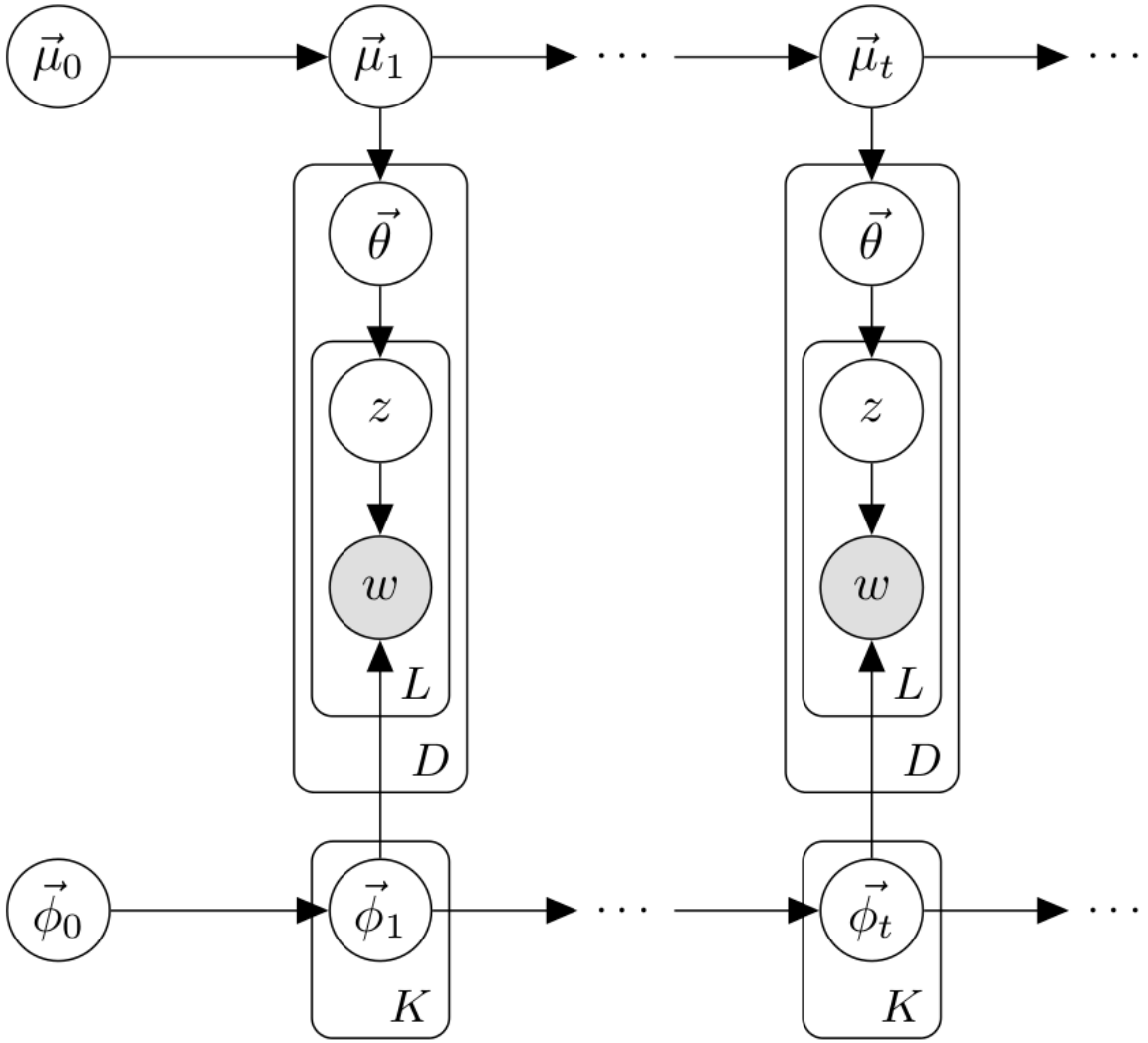
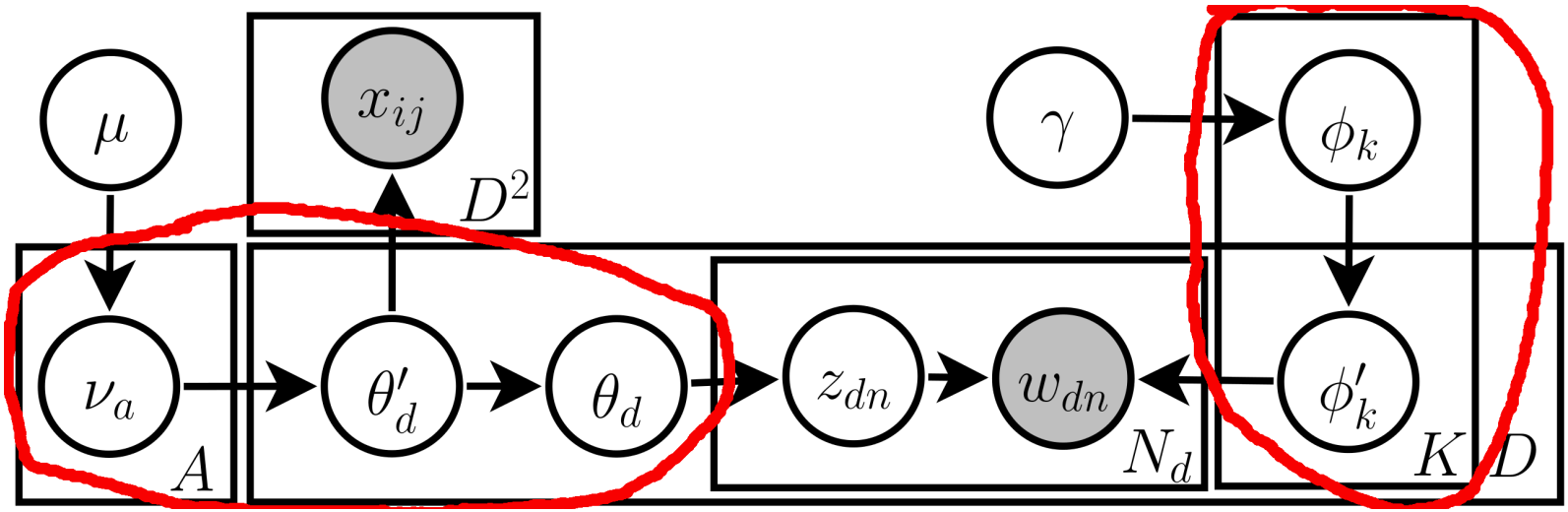


Figure 1: Graphical representation of the proposed model (for epoch 1 and t)

- model is a *sequence* of LDA style topic models chained together
- block table indicator sampling uses caching to work efficiently

Author-Citation Topic Model



"Bibliographic Analysis with the Citation Network Topic Model," Lim and Buntine, ACML, 2014

(probability vector hierarchies circled in red)

Other Regular Extended Models

All of these other related models can be made non-parametric using probability network hierarchies.

Stochastic block models: finding community structure in networks; mixed membership models; bi-clustering;

Infinite hidden relational models: tensor/multi-table extension of stochastic block models;

Tensor component models: tensor extension of component models;

Outline

Sampling Concentration and Discount

- In some cases the concentration and/or discount can be well chosen apriori:
  - the Stochastic Memoizer (Wood *et al.*, ICML 2009) uses a particular set for a hierarchy on text,
  - text is known to work well with discount $d \approx 0.7$,
  - topic proportions in LDA known to work well with discount $d = 0.0$.
- The concentration samples very nicely using a number of schemes.
  - slice sampling or adaptive rejection sampling,
  - auxiliary variable sampling (Teh *et al.*, 2006)

  $\longrightarrow$ usually improves performance, so do by default.
- The discount is expensive to sample (the generalised Stirling number tables need to be recomputed), but can be done with slice sampling or adaptive rejection sampling.

Latent Semantic Modelling

- Variety of component and network models in NLP and social networks can be made non-parametric with deep probability vector networks.
- New fast methods for training deep probability vector networks.
- Allows modelling of latent semantics:
  - semantic resources to integrate, (WordNet, sentiment dictionaries, *etc.*),
  - inheritance and shared learning across multiple instances,
  - hierarchical modelling,
  - deep latent semantics,
  - integrating semi-structured and networked content,

i.e. Same as deep neural networks!

Outline

Alphabetic References

D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, **3**, 2003.

W.L. Buntine and M. Hutter, "A Bayesian View of the Poisson-Dirichlet Process," *arXiv*, arXiv:1007.0296, 2012.

W.L. Buntine and S. Mishra, "Experiments with Non-parametric Topic Models," *KDD*, New York, 2014.

C. Chen, L. Du, L. and W. Buntine, "Sampling table configurations for the hierarchical Poisson-Dirichlet process," *ECML-PKDD*, Athens, 2011.

L. Du, W. Buntine and H. Jin, "A segmented topic model based on the two-parameter Poisson-Dirichlet processh," *Machine Learning Journal*, **81**(1), 2010.

L. Du, W. Buntine, H. Jin, "Modelling sequential text with an adaptive topic model," *EMNLP*, 2012.

L. Du, W. Buntine and M. Johnson, "Topic segmentation with a structured topic model," *NAACL-HLT*, Atlanta, 2013.

L.C. Hsua and P.J-S. Shiueb, "A Unified Approach to Generalized Stirling Numbers," *Advances in Applied Mathematics*, **20**(5), 1998.

Alphabetic References

H. Ishwaran and L.F. James", "Gibbs sampling methods for stick-breaking priors," *Journal of ASA*, **96**(453), 2001.
H. Ishwaran and L.F. James", "Generalized weighted Chinese restaurant processes for species sampling mixture models," *Statistica Sinica*, **13**, 2003.
J. Lee, F. Quintana, P. Müller, and L. Trippa, "Defining Predictive Probability Functions for Species Sampling Models," Statistical Science, **28**(2), 2013.
J. Pitman, "Exchangable and partially exchangeable random partitions", *Probab. Theory Relat. Fields*, **102**, 1995.
J. Pitman and M. Yor, "The two-parameter Poisson-Diriclet Distribution derived from a stable subordinator", *Annals of Probability*, **25** (2), 1997.
Y.W. Teh, "A Bayesian Interpretation of Interpolated Kneser-Ney," Technical Report TRA2/06, School of Computing, National University of Singapore, 2006a.
Y.W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," *ACL*, Sydney, 2006b.
Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei, "Hierarchical Dirichlet Processes," *JASA*, **101**(476), 2006.
F. Wood and Y. W. Teh, "A Hierarchical Nonparametric Bayesian Approach to Statistical Language Model Domain Adaptation," *12th AI and Stats*, 2009.