Simpler Non-parametric Bayesian Models

Wray Buntine

Monash University
http://topicmodels.org

Thanks to: **Lan Du** of Monash University and **Kar Wai Lim** of ANU for content, **Lancelot James** of HKUST for teaching me some of the material

5th Dec, 2016

Legend

Color coding:

blue phrases: important terms and phrases;

green phrases: new terms;

red phrases: important phrases with negative connotations.

Wikipedia is of mixed quality with non-parametrics, but recommended content is marked so$^{\mathcal{W}}$.

Outline

Outline

What are Parametric Methods?

By statistics: modelling with a fixed number of parameters,
e.g., linear regression

What are Parametric Methods?

By statistics: modelling with a fixed number of parameters,
e.g., linear regression

Common extension: add a dimension parameter and use model selection[$\mathcal{W}$]
to estimate,
e.g., linear regression with polynomials of order K (unknown)

What are Non-parametric Methods?

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

What are Non-parametric Methods?

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

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

What are Non-parametric Methods?

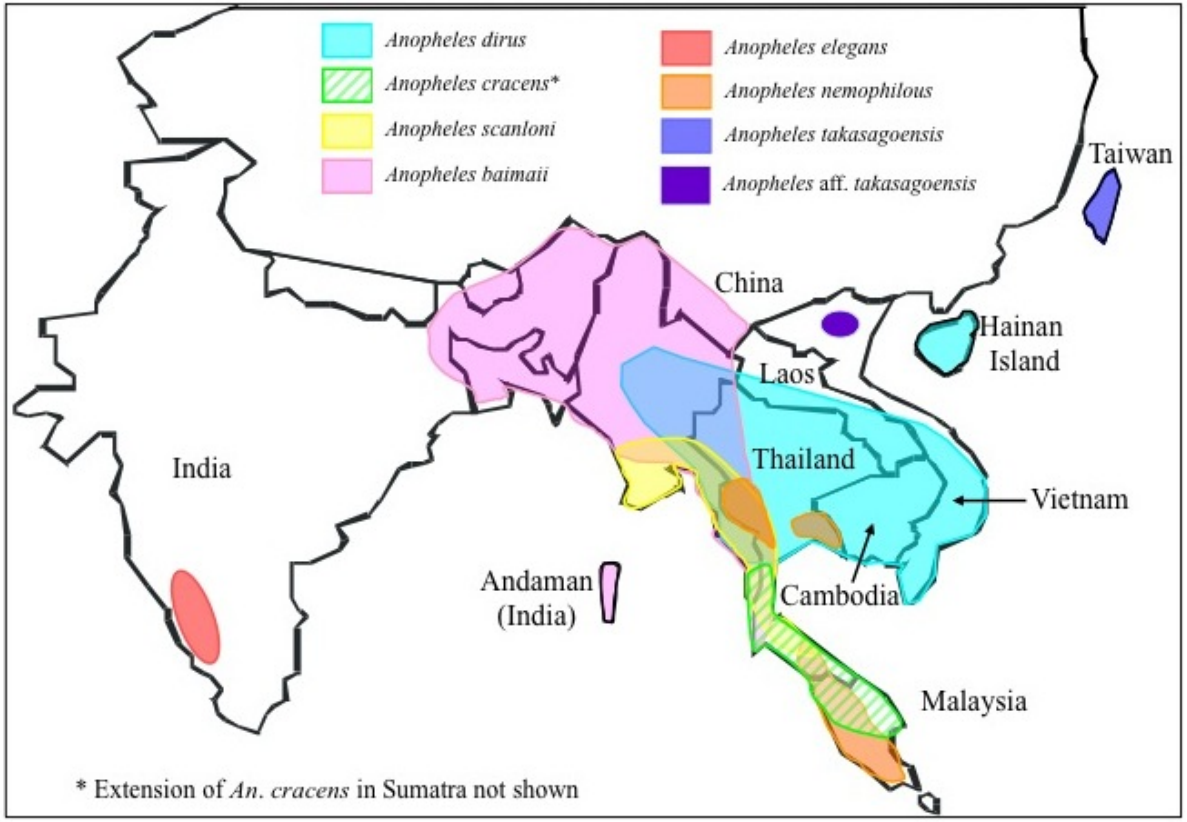
By classical statistics: modelling without parameters,
e.g., nearest neighbour methods

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

More accurately: modelling:

- with a finite but variable number of parameters;
- more parameters are "unfurled" as needed.

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?

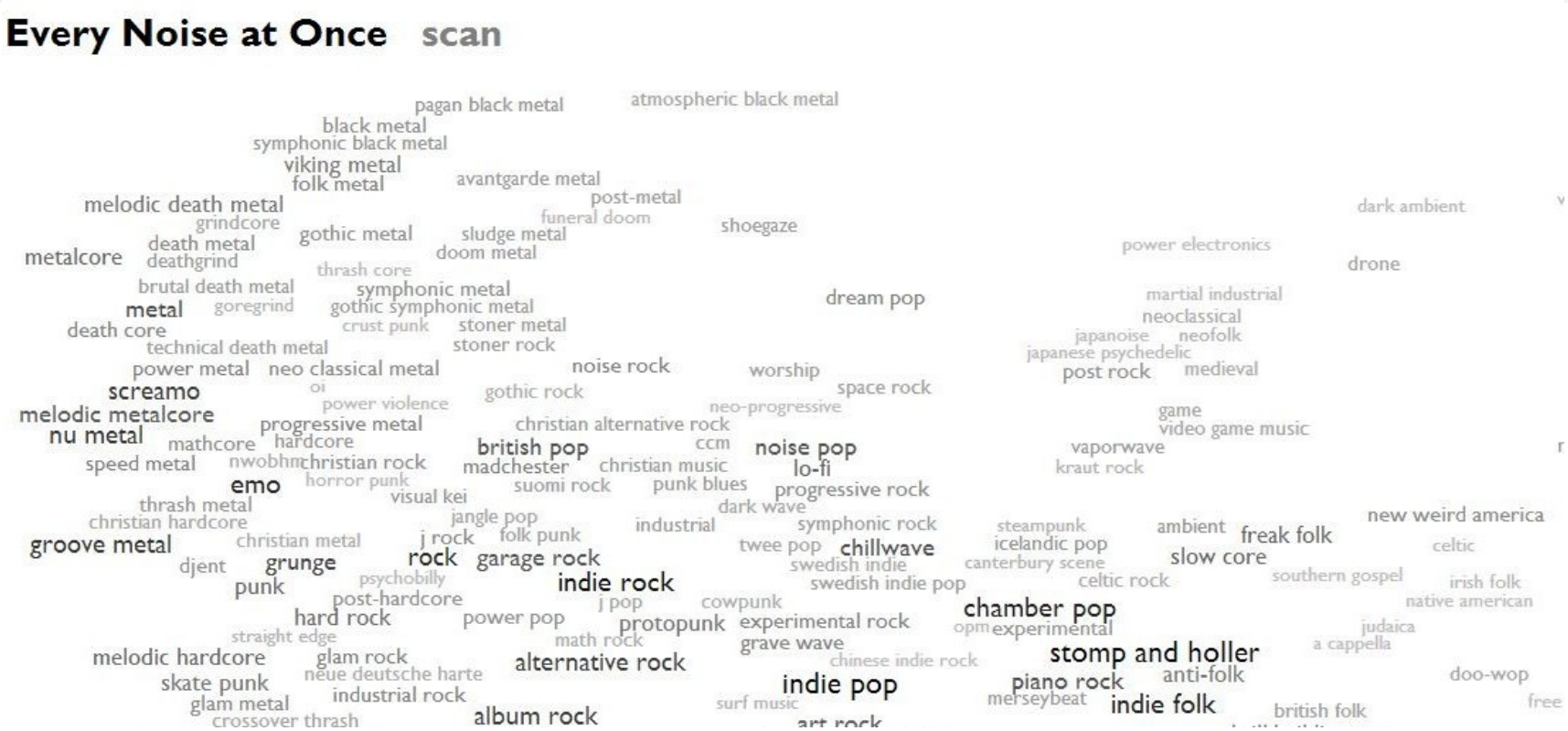
How Many Words in the English Language are There?

> ... 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?

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?

What is an Unknown Dimension?

- Some dimensions are fixed but unknown:
  - this uses parametric statistics
  - this is not Bayesian non-parametrics
- Some dimensions keep on growing as we get more data:
  - this is Bayesian non-parametrics

Modelling – What We Need

- matrices, tensors, graphs

Modelling – What We Need

- matrices, tensors, graphs
- semi-structured data

Modelling – What We Need

- matrices, tensors, graphs
- semi-structured data
- preferences, ratings, connections

Modelling – What We Need

- matrices, tensors, graphs
- semi-structured data
- preferences, ratings, connections
- bioinformatics, social networks, bibliographic data

Modelling – What We Need

- matrices, tensors, graphs
- semi-structured data
- preferences, ratings, connections
- bioinformatics, social networks, bibliographic data
- ever expanding numbers of items/dimensions/nodes, not small but unknown numbers

Probability and Parameter 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,

Probability and Parameter 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,

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

Probability and Parameter 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,

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

- the layers of a deep neural network
- user preferences

Probability and Parameter 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,

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

- the layers of a deep neural network
- user preferences

Likewise for distributions over parameter vectors.

Bayesian Inference – General Methodology

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$^{\mathcal{W}}$.

Bayesian Inference – General Methodology

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$^{\mathcal{W}}$.

Wikipedia coverage of Bayesian non-parametrics is patchy.

I'm not bad, I was just drawn that way

Jessica Rabbit in "Who Shot Roger Rabbit?"

I'm not complex, I was just written that way

For any topological space $\mathcal{T}$, $\mathcal{B}(\mathcal{T})$ will denote the Borel σ-field of subsets of $\mathcal{T}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and $\mathbb{X}$ be complete, separable and endowed with a metric $d_{\mathbb{X}}$. Define on $(\Omega, \mathcal{F}, \mathbb{P})$ a Poisson random measure $\tilde{N}$ on $\mathbb{S} = \mathbb{R}^+ \times \mathbb{X}$ with intensity measure ν. This means that

(i) for any C in $\mathcal{B}(\mathbb{S})$ such that $\nu(C) = \mathbb{E}[\tilde{N}(C)] < \infty$, the probability distribution of the random variable $\tilde{N}(C)$ is Poisson$(\nu(C))$;

(ii) for any finite collection of pairwise disjoint sets, $A_1, \ldots, A_k$, in $\mathcal{B}(\mathbb{S})$, the random variables $\tilde{N}(A_1), \ldots, \tilde{N}(A_k)$ are mutually independent.

Moreover, the measure ν must satisfy the following conditions,

$$\int_{(0,1)} s\nu(\mathrm{d}s, \mathbb{X}) < \infty, \quad \nu([1, \infty) \times \mathbb{X}) < \infty.$$

We refer to Daley & Vere-Jones (1988) for an exhaustive account on Poisson random measures.

From James, Lijoi & Prünster, 2009,

"Posterior analysis for normalized random measures with independent increm

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information
- Non-parametrics lets us deal with variable length parameters vectors and other structures.
  - sometimes is little more complex than parametric methods

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information
- Non-parametrics lets us deal with variable length parameters vectors and other structures.
  - sometimes is little more complex than parametric methods
- Apparent complexity of non-parametrics is largely an artifact of the theoretical literature.
  - oftentimes is little more complex parametric methods

Outline

Discrete Distributions

Name	Domain	$p(x\|\cdots)$
$x \sim$Bernoulli(ρ)	$x \in \{0,1\}$	$\rho^x(1-\rho)^{1-x}$
$n \sim$binomial(N,ρ)	$0 \leq n \leq N$	$\binom{N}{n}\rho^n(1-\rho)^{N-n}$
$x \sim$categorical$_K(\vec{\rho})$	$x \in \{1,...,K\}$	ρ_x
$\vec{n} \sim$multinomial$_K(N,\vec{\rho})$	$\vec{n} \in \mathcal{N}^K, \sum_{k=1}^K n_k = N$	$\binom{N}{\vec{n}}\prod_{k=1}^K \rho_k^{n_k}$
$x \sim$Poisson(λ)	$x \in \mathcal{N}$	$\frac{1}{x!}\lambda^x e^{-\lambda}$

$\mathcal{N} = \{0,1,...,\infty\}$, $\rho \in (0,1)$, $\lambda \in \mathcal{R}^+$, $\rho_k \in (0,1)$, $\sum_{k=1}^K \rho_k = 1$

- a multinomial is an unordered set of categoricals;
- distributions over trees and graphs can be built using Bernoullis and categorical variables.

Dirichlet Distribution

Definition of Dirichlet distribution

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

$$\vec{p} \sim \text{Dirichlet}_K(\vec{\alpha})$$

where $\vec{\alpha}$ is a positive K-dimensional vector.

- used to sample a probability vector "similar" to $\frac{\vec{\alpha}}{\sum_k \alpha_k}$
- like a Gaussian, it has as parameters (inputs)

  mean: probability vector

  concentration: inverse variance parameter

Dirichlet Distribution

Definition of Dirichlet distribution

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

$$\vec{p} \sim \text{Dirichlet}_K(\vec{\alpha})$$

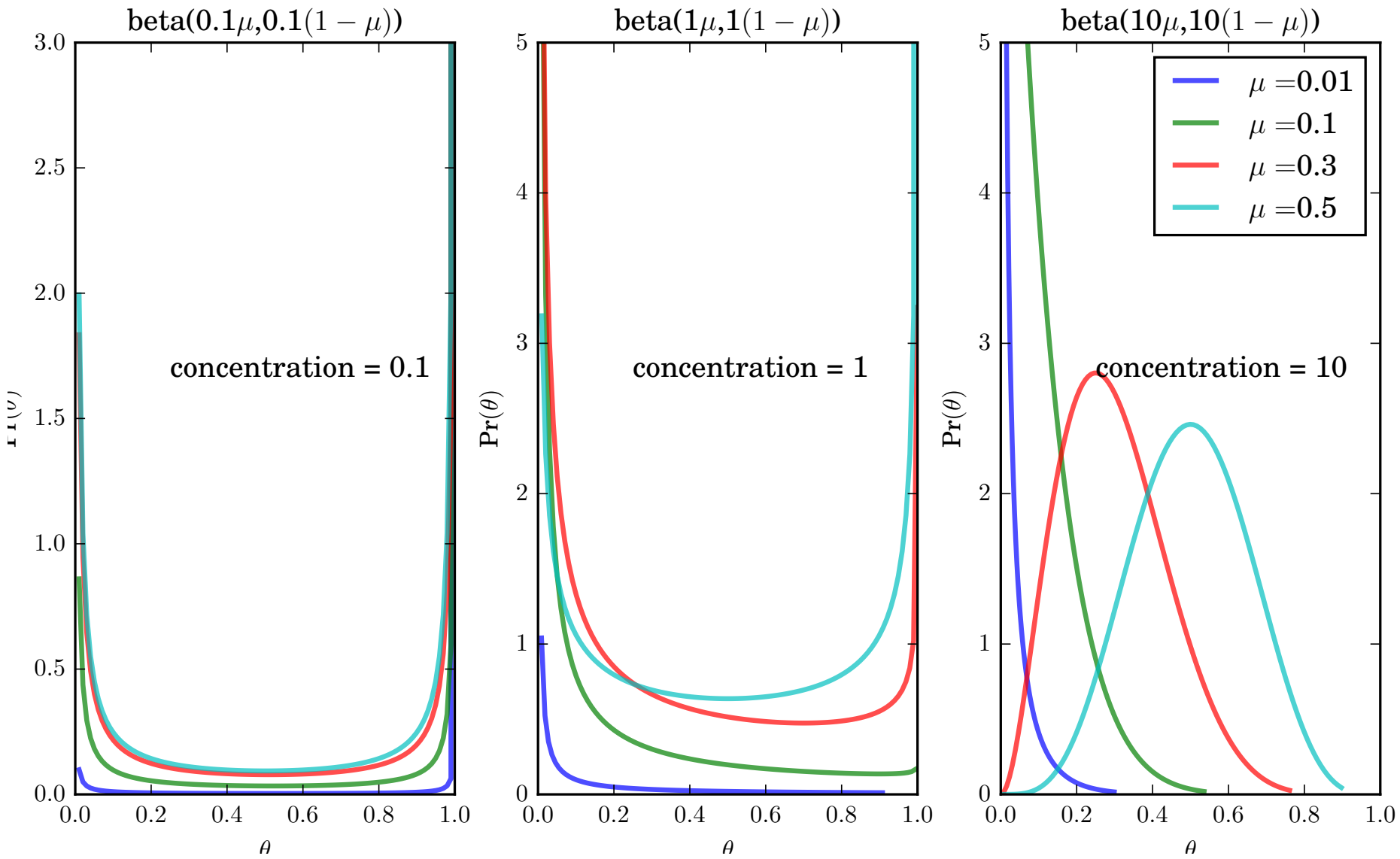
where $\vec{\alpha}$ is a positive K-dimensional vector.

- used to sample a probability vector "similar" to $\frac{\vec{\alpha}}{\sum_k \alpha_k}$
- like a Gaussian, it has as parameters (inputs)

  mean: probability vector

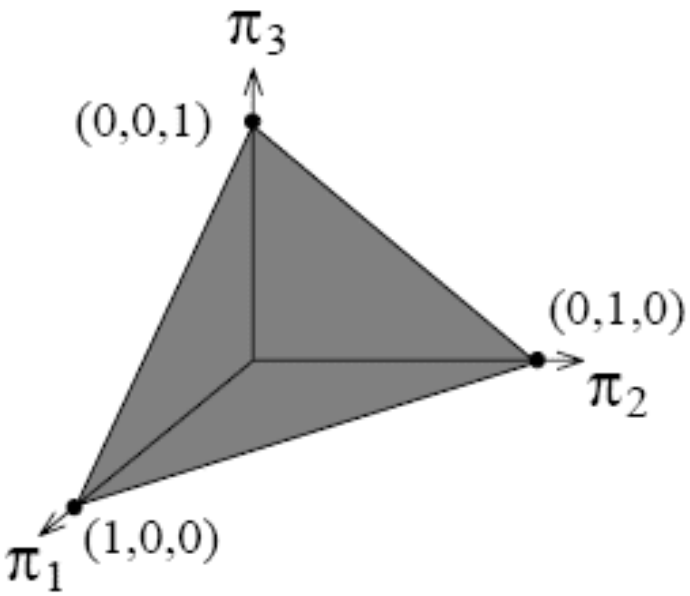
  concentration: inverse variance parameter
- said to be a conjugate prior$^{\mathcal{W}}$ for the multinomial distribution, *i.e.*, makes math easy.

2-D Dirichlet (or Beta) Plots



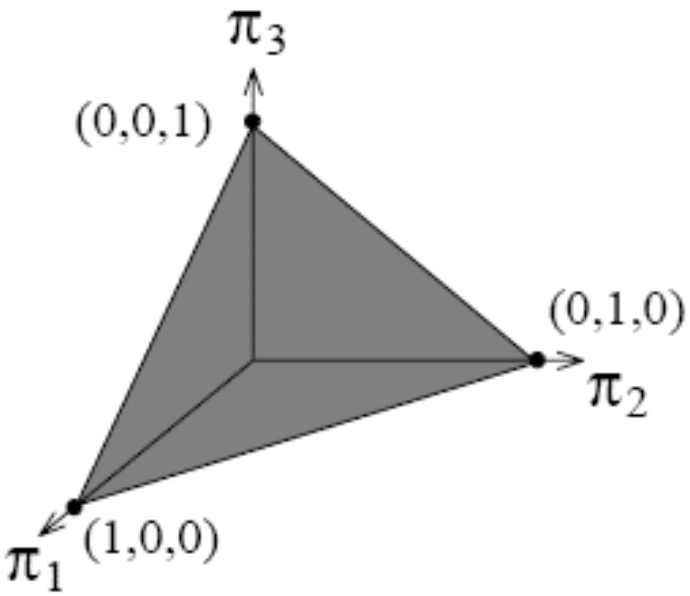
The 3-D Dirichlet

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



The domain of the vector (p_1, p_2, p_3).

The 3-D Dirichlet



The domain of the vector (p_1, p_2, p_3).

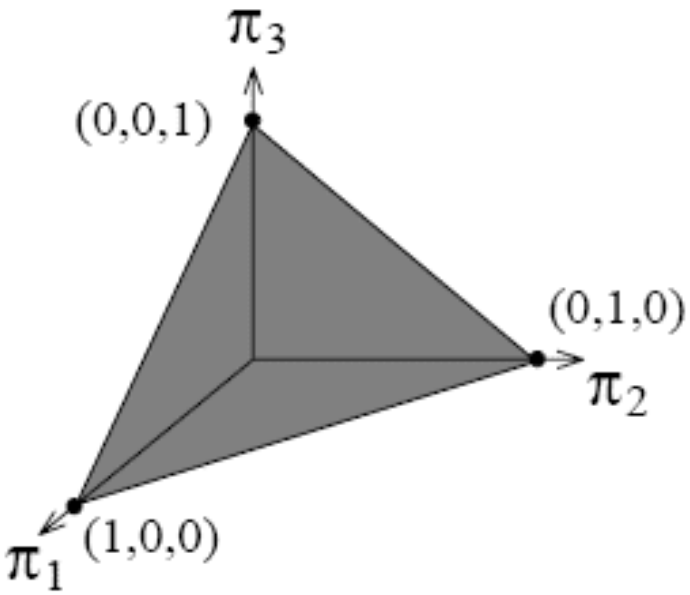
Consider $\vec{p} = (p_1, p_2, p_3)$ where $\sum_k p_k = 1$.

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

$$p(\vec{p}\,|\,\vec{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1}$$

where $\Gamma(\cdot)$ is the gamma function[W].

The 3-D Dirichlet



The domain of the vector (p_1, p_2, p_3).

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

$$p(\vec{p}\,|\,\vec{\alpha}) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} ;$$

where $\Gamma(\cdot)$ is the gamma function[W].

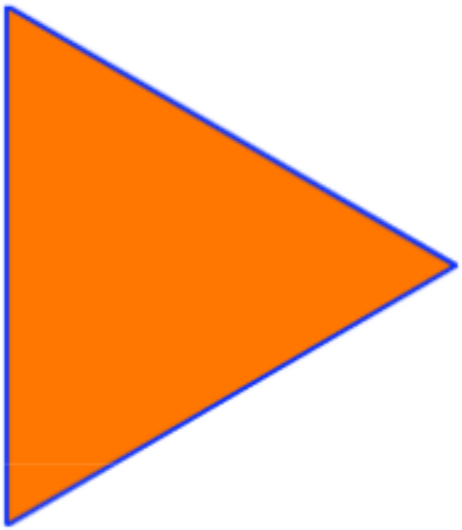
Normalising constant for Dirichlet is called a beta function[W]:

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

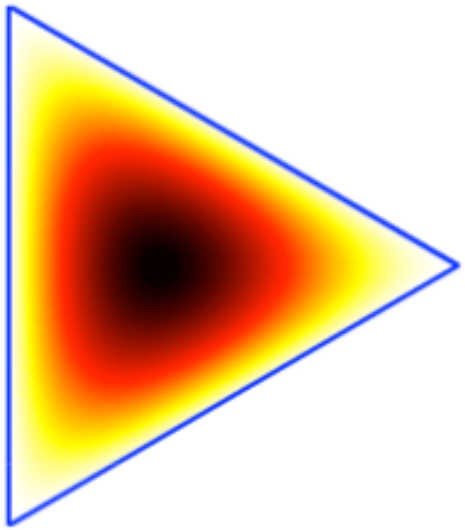
NB. also have 2-D and k-D version

3-D Dirichlet Plots

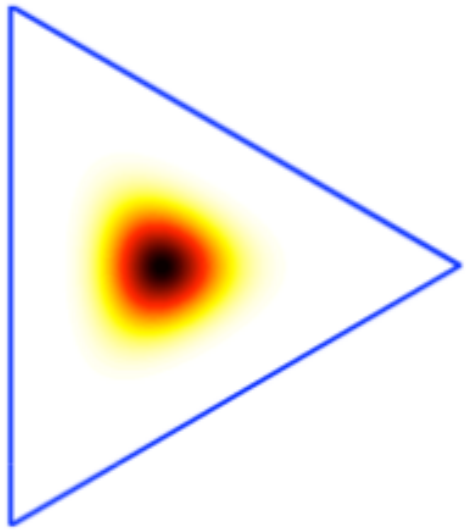
Dirichlet(1,1,1)



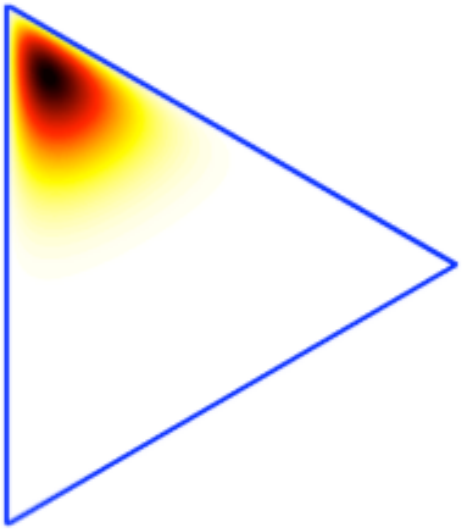
Dirichlet(2,2,2)



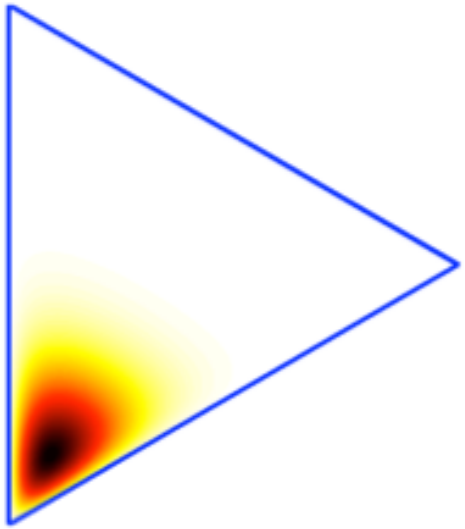
Dirichlet(10,10,10)



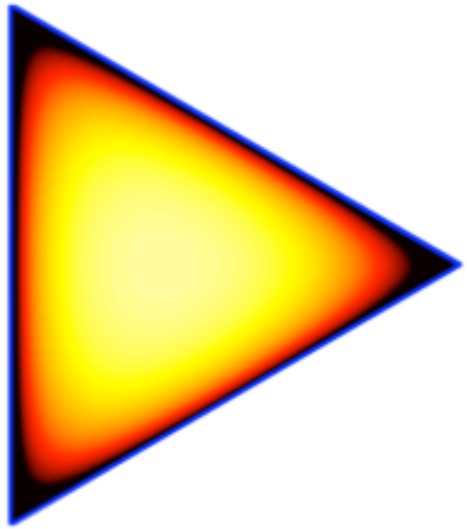
Dirichlet(2,2,10)



Dirichlet(2,10,2)

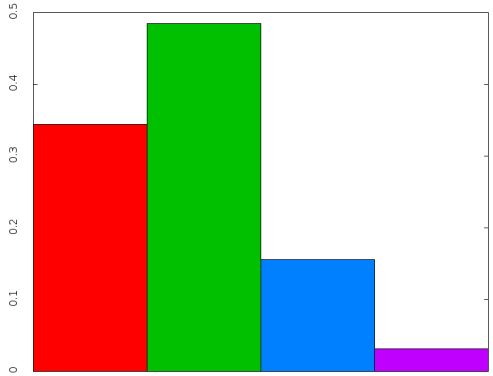


Dirichlet(0.8,0.8,0.8)

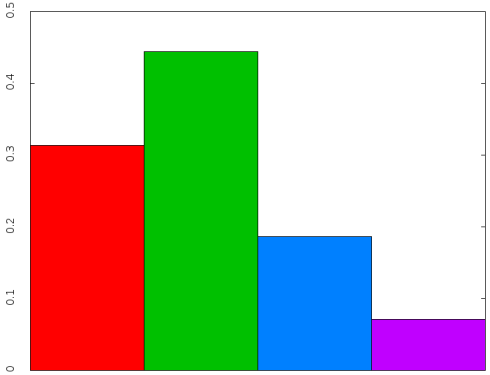


4-D Dirichlet samples

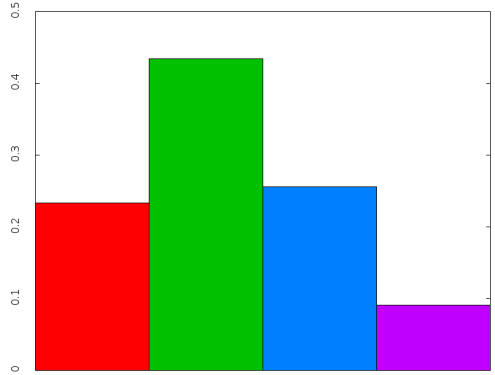
$\vec{p}_0$



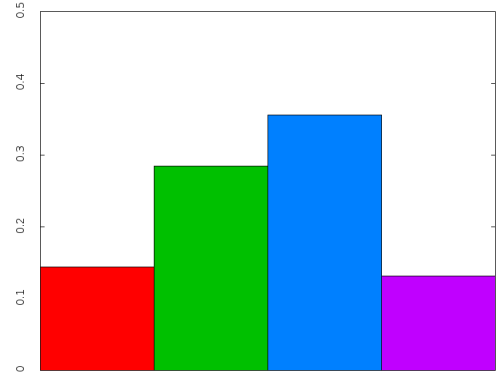
$\vec{p}_1 \sim \text{Dirichlet}_4(500\vec{p}_0)$



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



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



Conjugate Distributions

Name	Domain	$p(\lambda \mid \cdots)$
$\lambda \sim$beta(α, β)	$\lambda \in (0,1)$	$\frac{1}{\text{beta}_2(\alpha,\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$
$\vec{\lambda} \sim$Dirichlet$(\vec{\alpha})$	$\lambda_k \in (0,1), \sum_{k=1}^{K} \lambda_k = 1$	$\frac{1}{\text{beta}(\vec{\alpha})} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}$
$\lambda \sim$gamma(α, β)	$\lambda \in (0, \infty)$	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\beta\lambda}$

$\alpha, \beta > 0$, $\alpha_k > 0$

- these distributions are used as conjugate priors$^{\mathcal{W}}$
- conjugate pairs are: beta-binomial; Dirichlet-multinomial; gamma-Poisson
- beta distribution is a 2-D case of the K-dimensional Dirichlet distribution

Normalising Random Variables

- A multinomial comes from normalising Poissons:

$$x_k \sim Poisson(\lambda_k) \text{ for } k \in \{1, ..., K\} \text{ is equivalent to}$$
$$\left(\sum_{k=1}^{K} x_k\right) \sim Poisson\left(\sum_{k=1}^{K} \lambda_k\right) \text{ and independently}$$
$$\vec{x} \sim multinomial_K\left(\sum_{k=1}^{K} x_k, \vec{\lambda}\right)$$

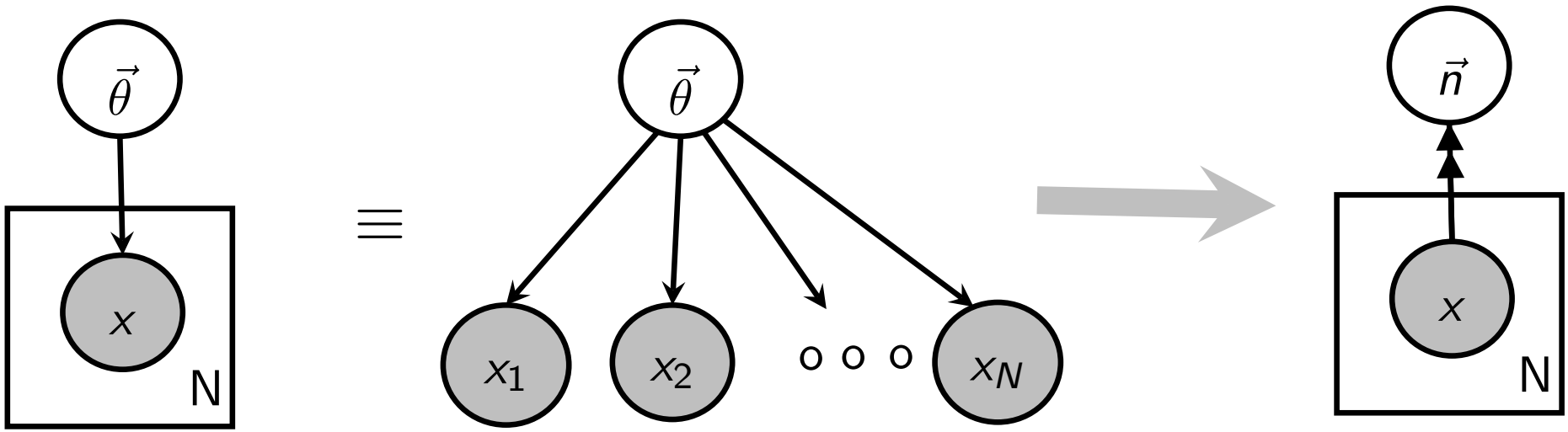
- A Dirichlet comes from normalising gammas with same scale β. Let $\lambda_0 = \sum_{k=1}^{K} \lambda_k$, then:

$$\lambda_k \sim gamma(\alpha_k, \beta) \text{ for } k \in \{1, ..., K\} \text{ is equivalent to}$$
$$\lambda_0 \sim gamma\left(\sum_{k=1}^{K} \alpha_k, \beta\right) \text{ and independently}$$
$$\frac{1}{\lambda_0}\vec{\lambda} \sim Dirichlet_K(\vec{\alpha})$$

These results extend to some stochastic processes.

Outline

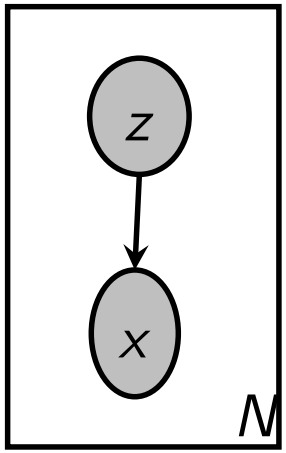
Reading a Graphical Model



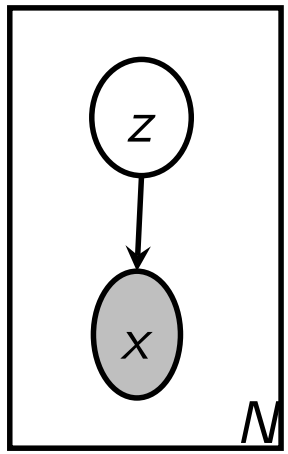
arcs	=	"depends on"
double headed arcs	=	"deterministically computed from"
shaded nodes	=	"supplied variable/data"
unshaded nodes	=	"unknown variable/data"
boxes	=	"replication"

Models in Graphical Form

Supervised learning or Prediction model

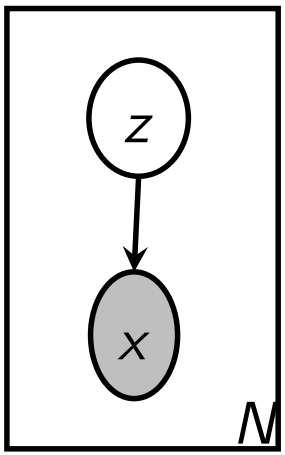


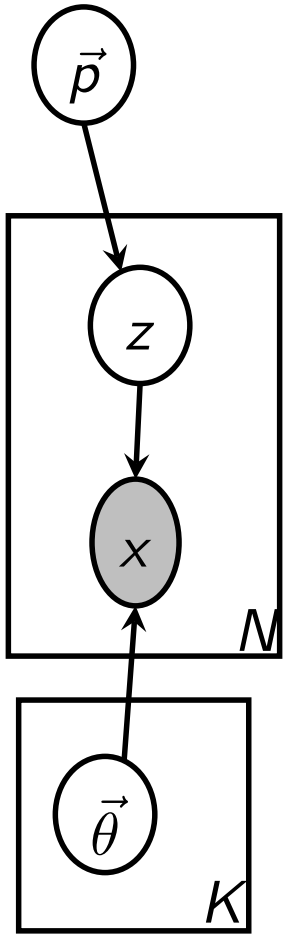
Clustering or Mixture model

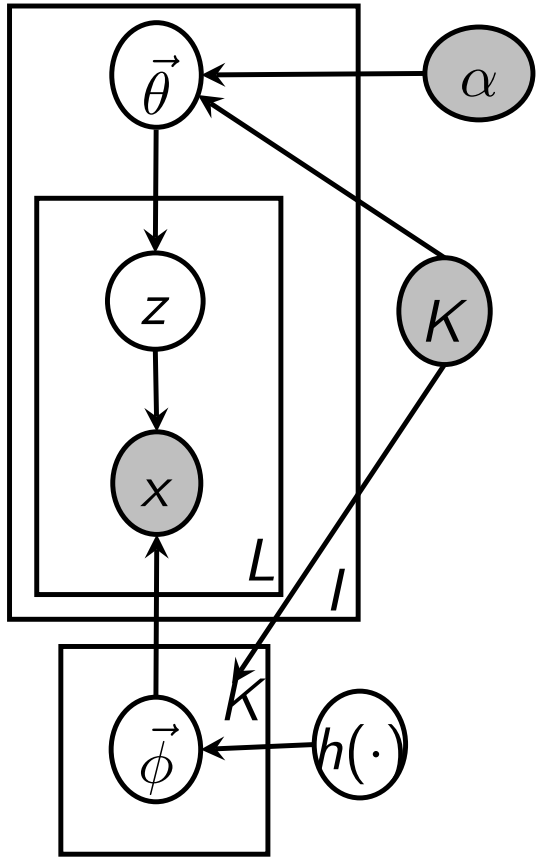


Mixture Models in Graphical Form

Building up the parts:



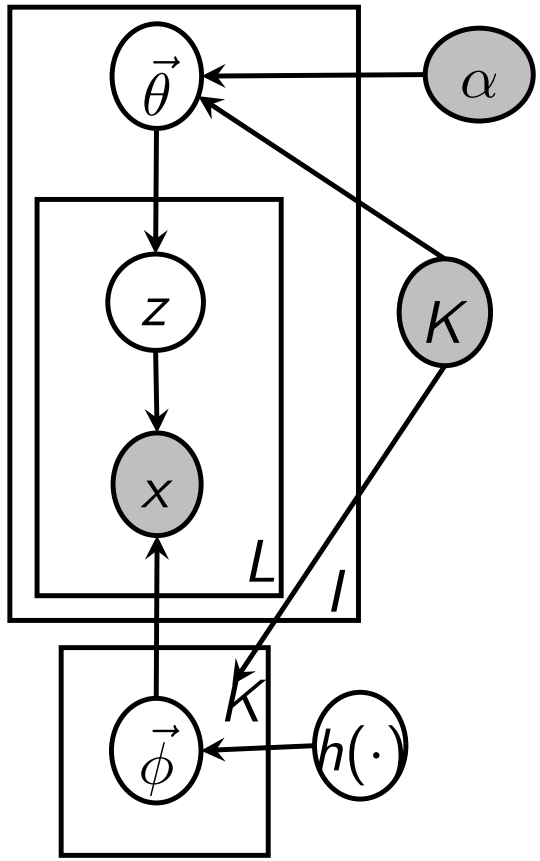




The Classic Mixture Model

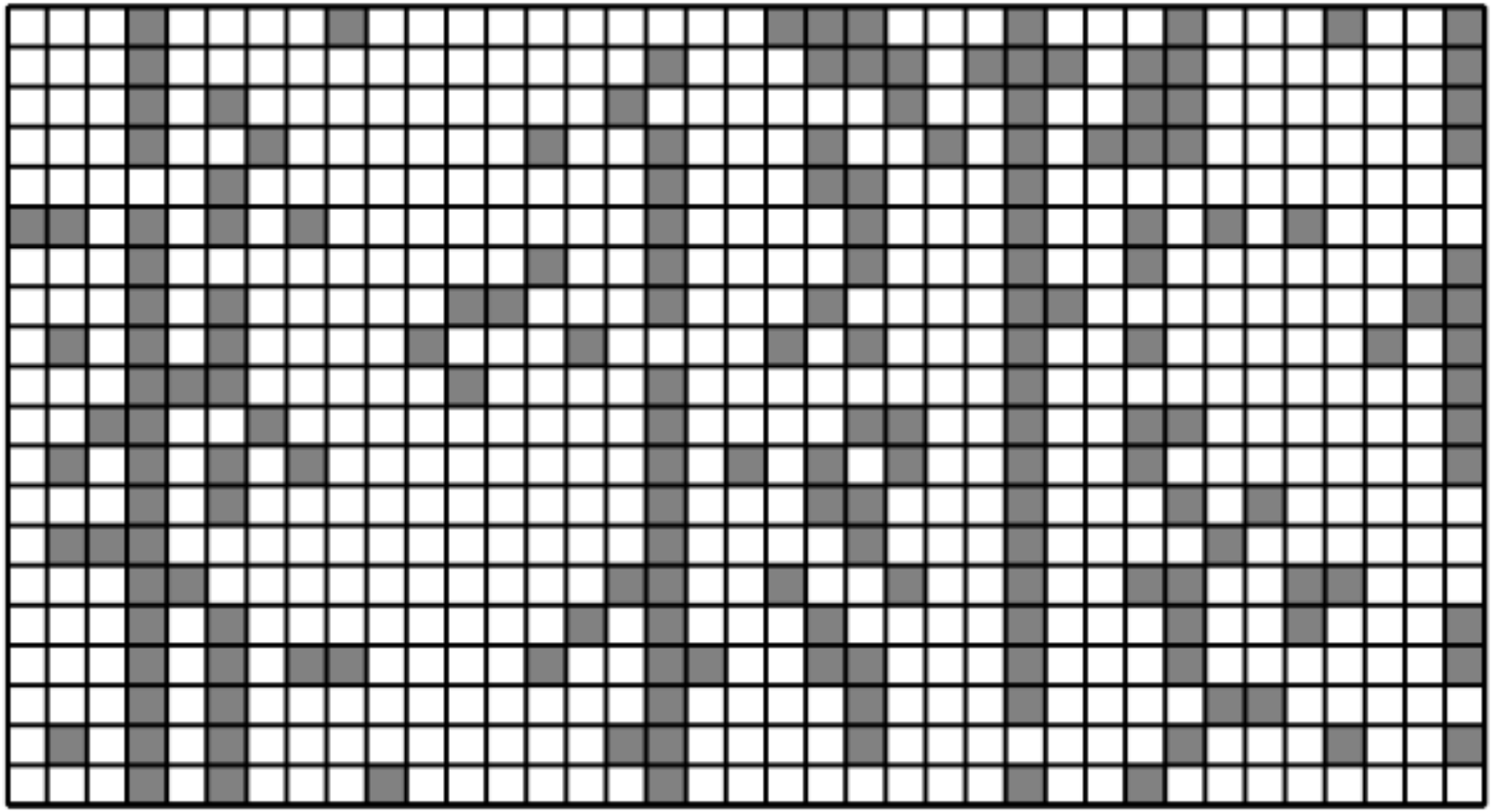
Data is a mixture of unknown dimension K and base distribution $h(\cdot)$ generating mixture entries with proportions $\vec{p}$.

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



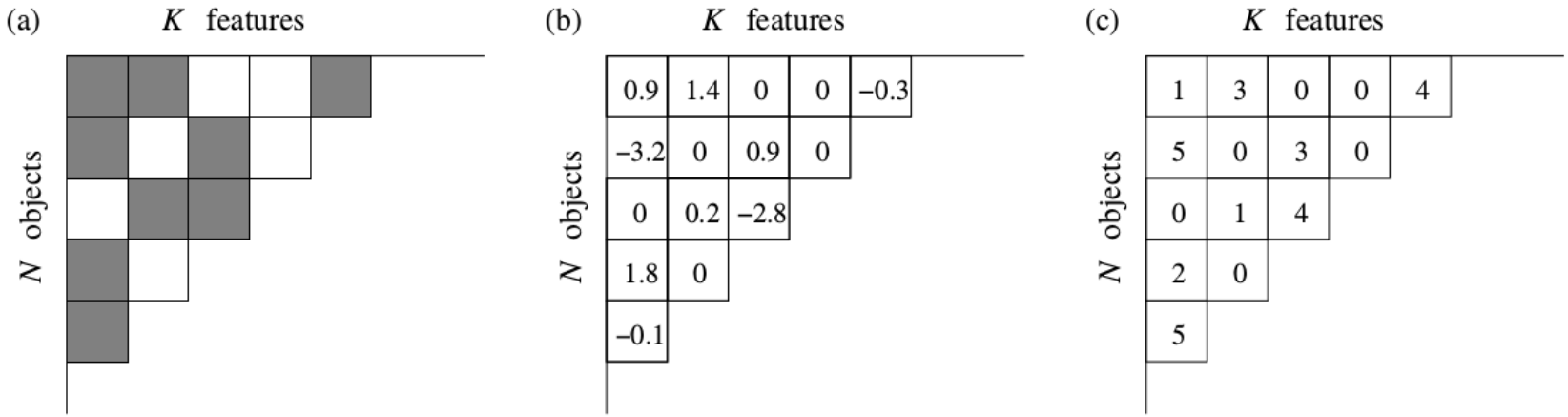
Outline

Example Feature Matrix



e.g., rows = documents/data/images, columns = words/features

Other Example Feature Matrices



(a) Boolean features

(b) Gaussian features multiplied by Boolean (like spike and slab[W] model)

(c) count features (for Poisson factorisation model, like non-negative matrix factorization[W])

figure taken from Griffiths and Ghahramani, JMLR, 2011

Example Feature Matrix, Left-Ordered Form

can order features (columns) in the order of first occurrence in data

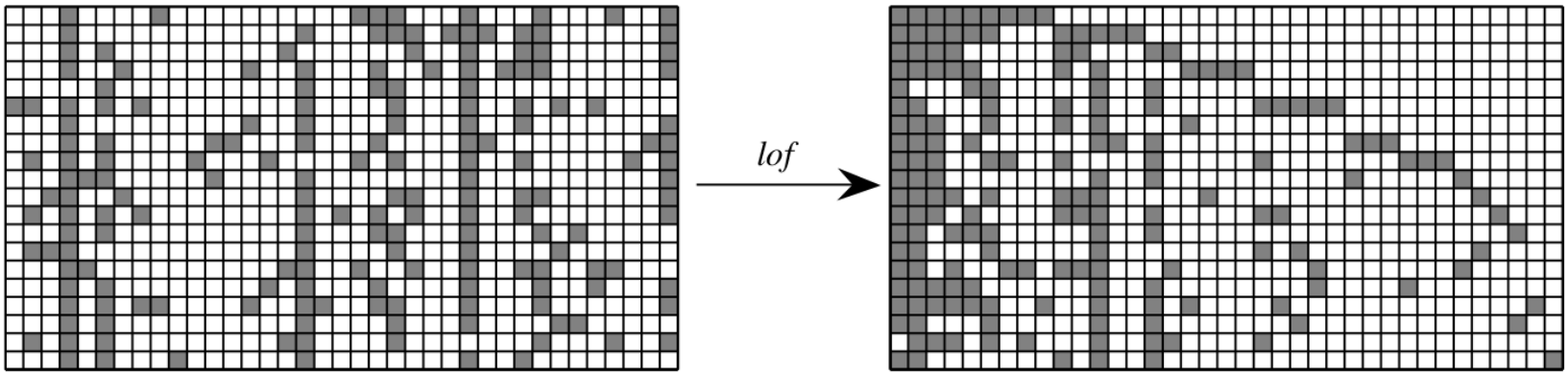
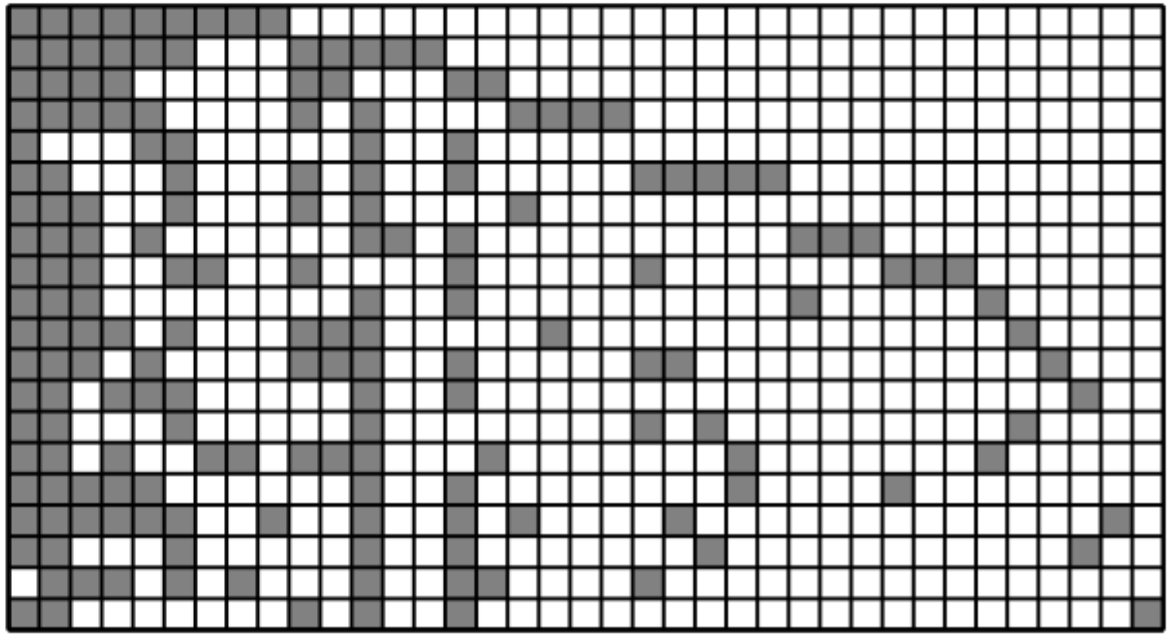


Figure 5: Binary matrices and the left-ordered form. The binary matrix on the left is transformed into the left-ordered binary matrix on the right by the function $lof(\cdot)$. This left-ordered matrix was generated from the exchangeable Indian buffet process with $\alpha = 10$. Empty columns are omitted from both matrices.

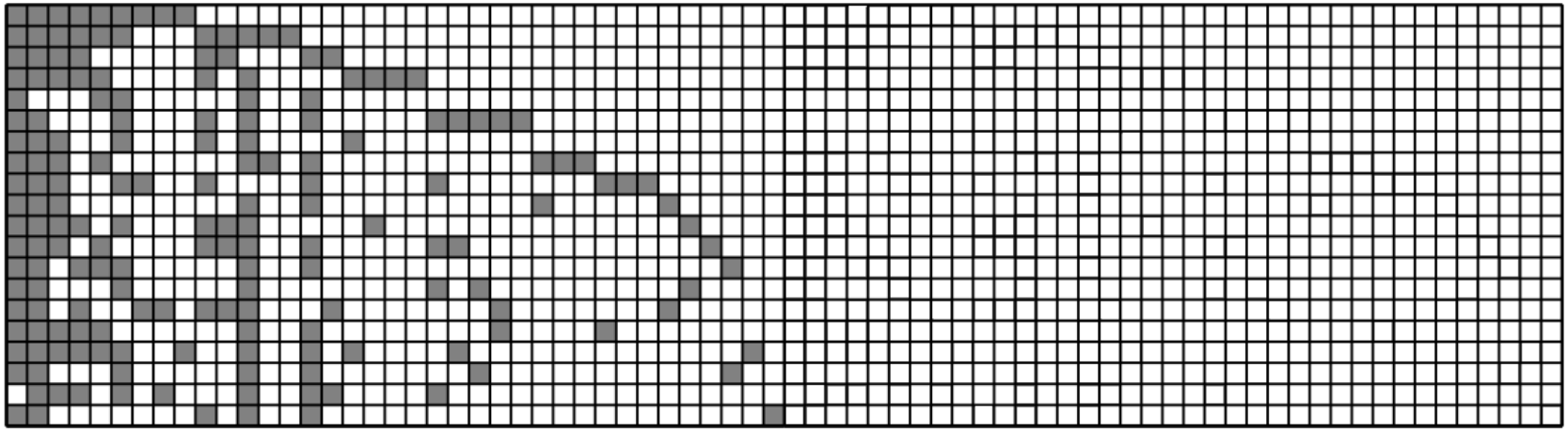
figure taken from Griffiths and Ghahramani, JMLR, 2011

Simple Independence Model



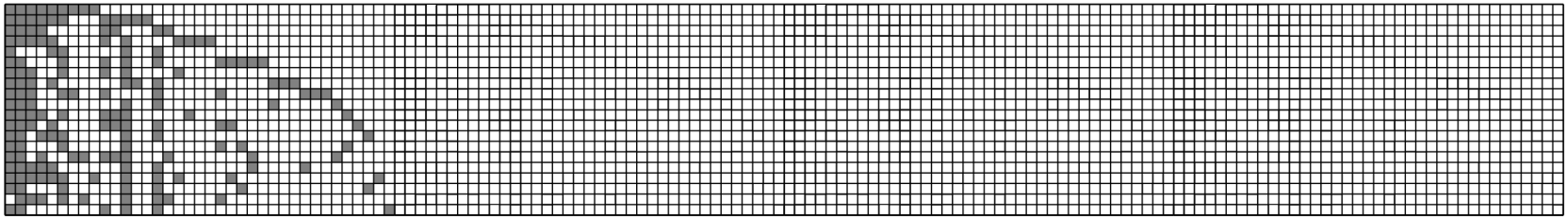
- assume rows and columns exchangeable (random variables)$^{\mathcal{W}}$
- denote $b_{i,k}$ = Boolean for row i, column k,
- model $b_{i,k} \sim$ Bernoulli(θ_k) for column k,
- what is a suitable prior distribution on θ_k assuming infinite columns?
  NB., on average θ_k might be about 0.3!

Simple Independence Model, cont



- now what is a suitable prior distribution on θ_k assuming columns exchangeable? **NB.**, on average θ_k might now be about 0.1!

Simple Independence Model, cont



- now what is the prior distribution on θ_k assuming columns exchangeable? **NB.**, on average θ_k might now be about 0.02!

Need to rethink how to model potentially unlimited features.

Outline

Outline

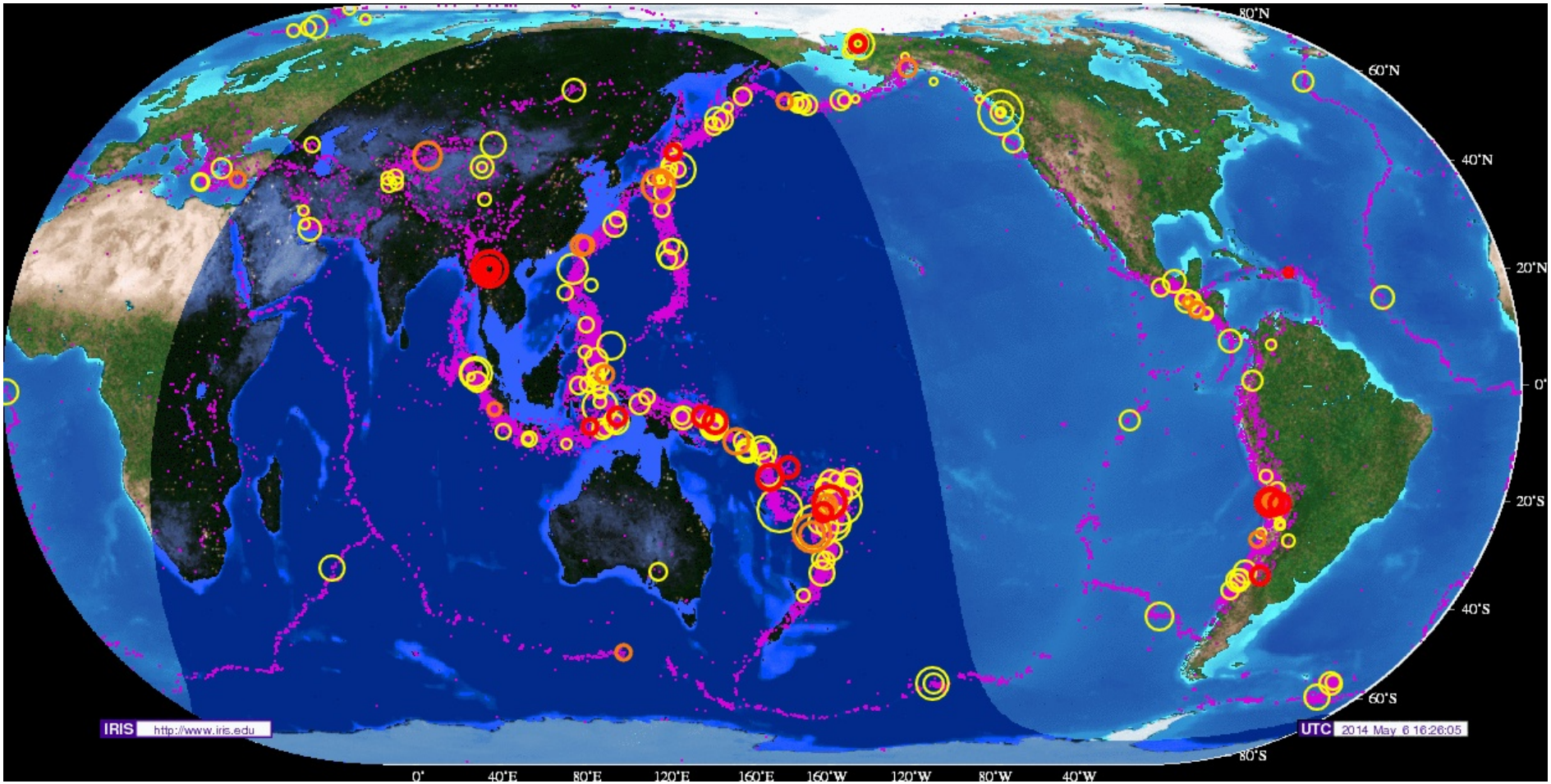
Poisson Point Processes – Motivation

Poisson Point Processes$^{\mathcal{W}}$ (PPPs) are a special class of spatial processes that behave like Poisson distributions on finite subsets.

Used to:

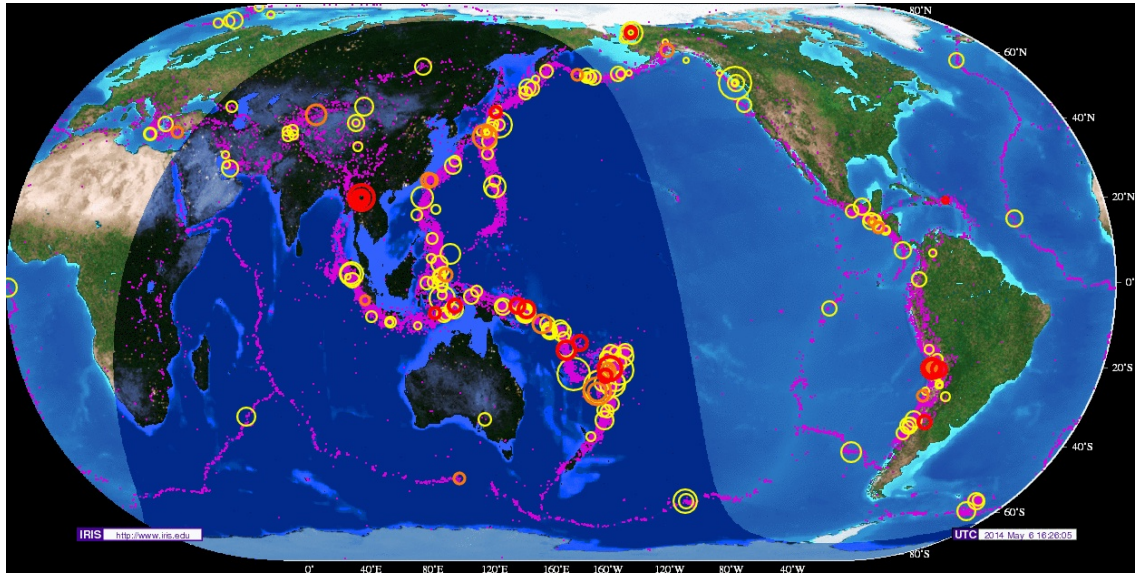
- analyse variable length parameter vectors and other structures with varying dimensions and elements
  - e.g. stochastic graphs and trees
- analyse time/space varying affects
  - e.g. event bursts in Twitter data
  - e.g. crime incidents in a city

Earthquakes in 2014: A Spatial Process



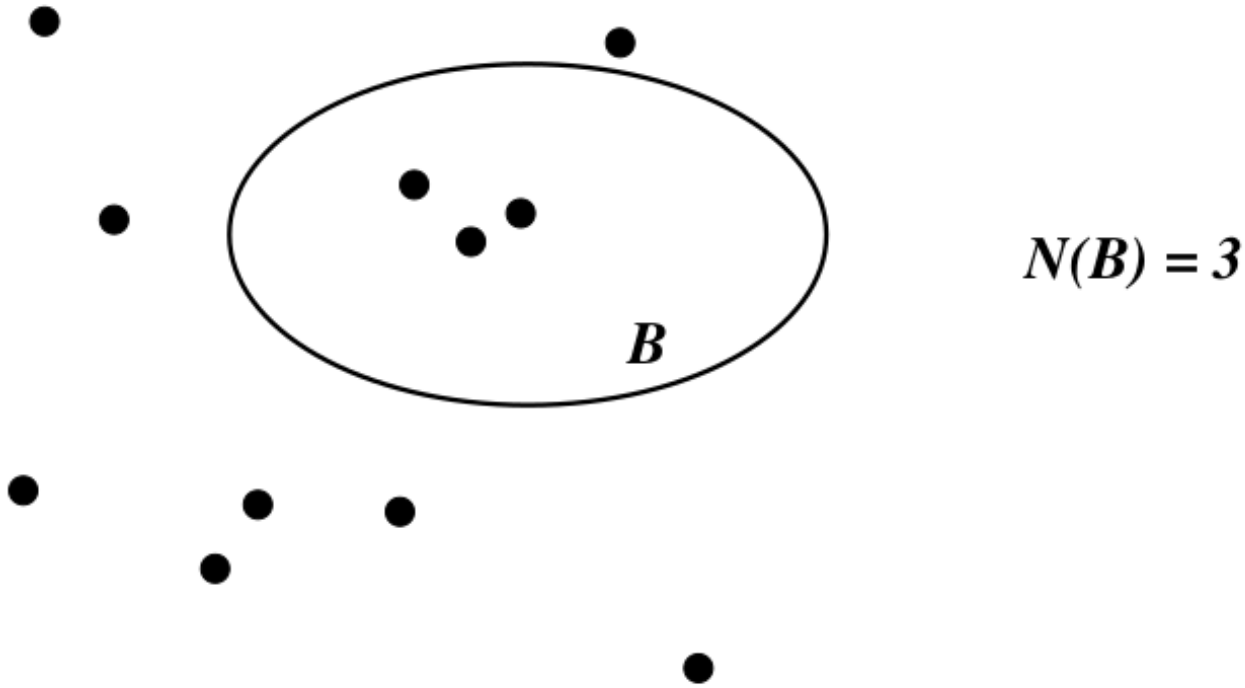
- we have points from a process given by (*latitude*, *longitude*)
- each point marked with the *magnitude*

Earthquakes in 2014, cont.



- The process part, points generated by a rate, denoted $\rho(\mathit{latitude}, \mathit{longitude})$:
  - number of earthquakes in 2014 per unit area, *i.e.* this rate is spatial
  - rate is not the same as a probability
  - can generate a countably infinite number of points too
- The marked part attaches auxiliary variables to the points in the process generated by a probability $p(\mathit{magnitude} \mid \mathit{latitude}, \mathit{longitude})$.

What is a Spatial Process?



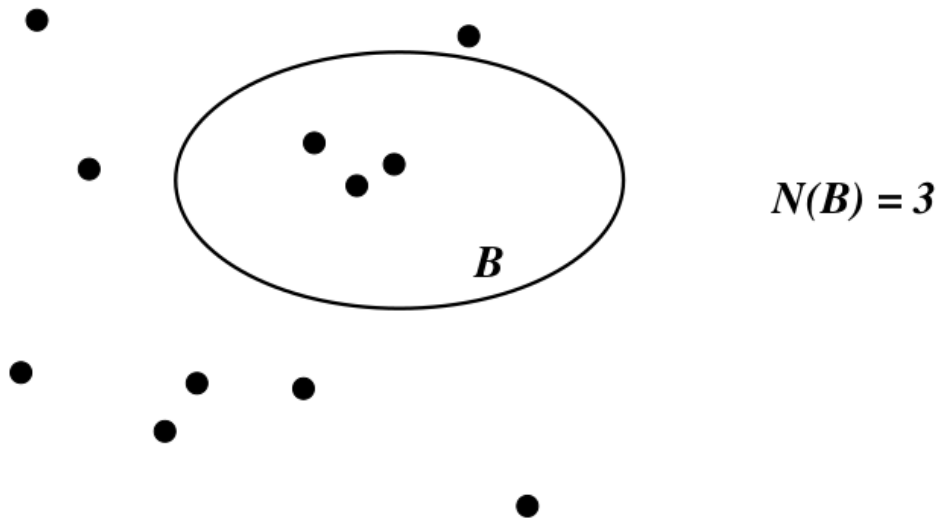
- spatial process gives a sample of points on a space $\mathcal{X}$
- useful statistic: $N(B)$ is the number of points inside subset $B \subseteq \mathcal{X}$
- we can characterise spatial processes by properties of $N(B)$ for different subsets B

few following examples from Adrian Badderly, "Spatial Point Processes and their

What is a Spatial Process?

Definition of Spatial Process

A spatial process$^{\mathcal{W}}$ on $\mathcal{X}$ is a collection of random variables, representing a countable set of random values in the space $\mathcal{X}$.



- a single sample from a spatial process depicted in the figure
- $N(B)$ is the number of points inside subset $B \subseteq \mathcal{X}$

Note: points in the process can also be marked, by attaching auxiliary variables generated by a probability.

Additivity for Poissons

For Poissons, data and parameters are jointly additive:

- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;

Additivity for Poissons

For Poissons, data and parameters are jointly additive:

- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;
- related property referred to as infinite divisibility (probability)$^{\mathcal{W}}$.

Additivity for Poissons

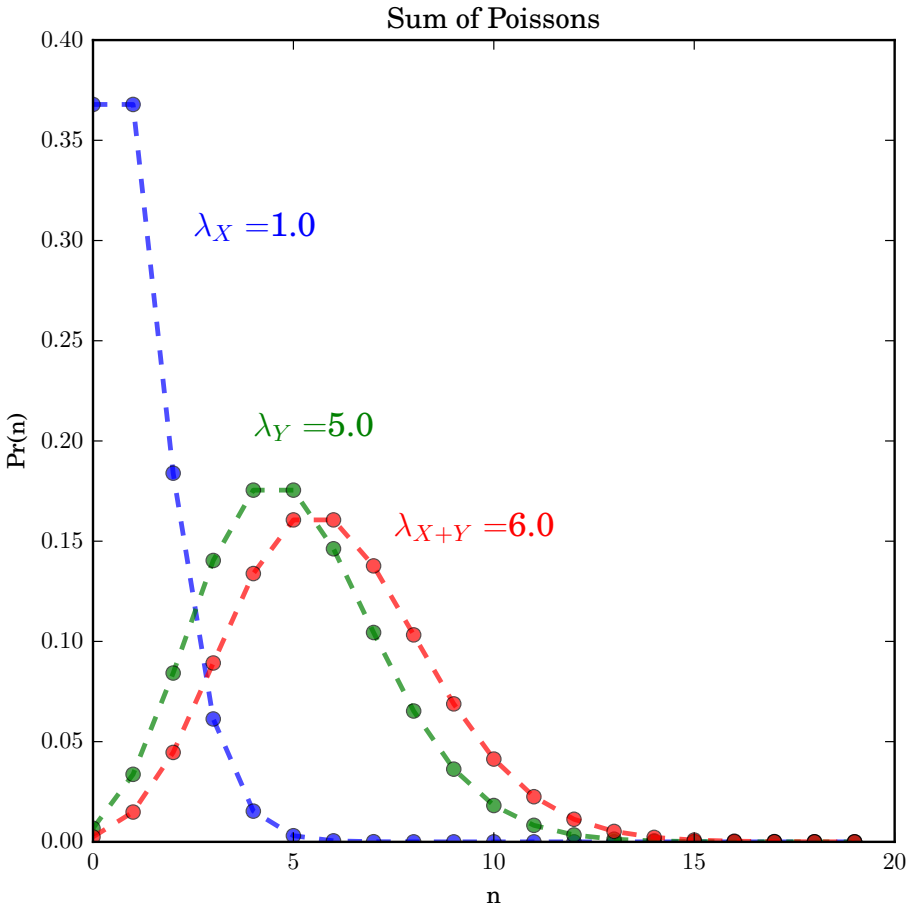
For Poissons, data and parameters are jointly additive:

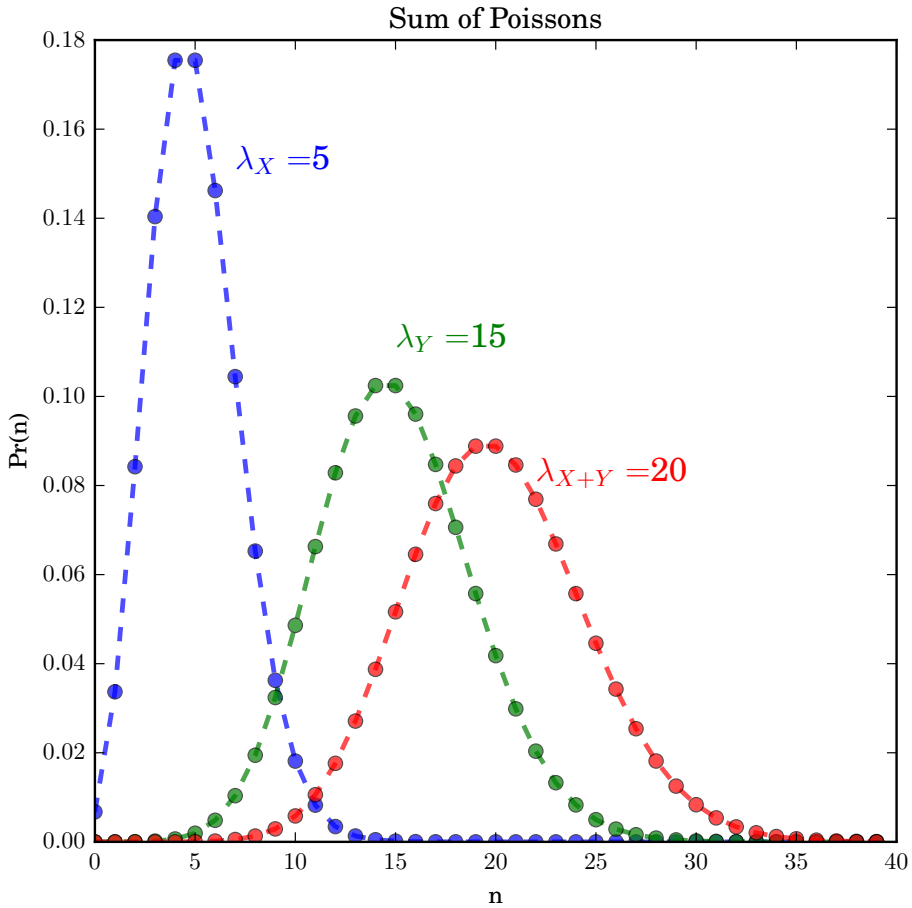
- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;
- related property referred to as infinite divisibility (probability)$^{\mathcal{W}}$.

Same property holds for the gamma, negative binomial, Gaussian, stable distributions (and their key parameter).

More on this later ....

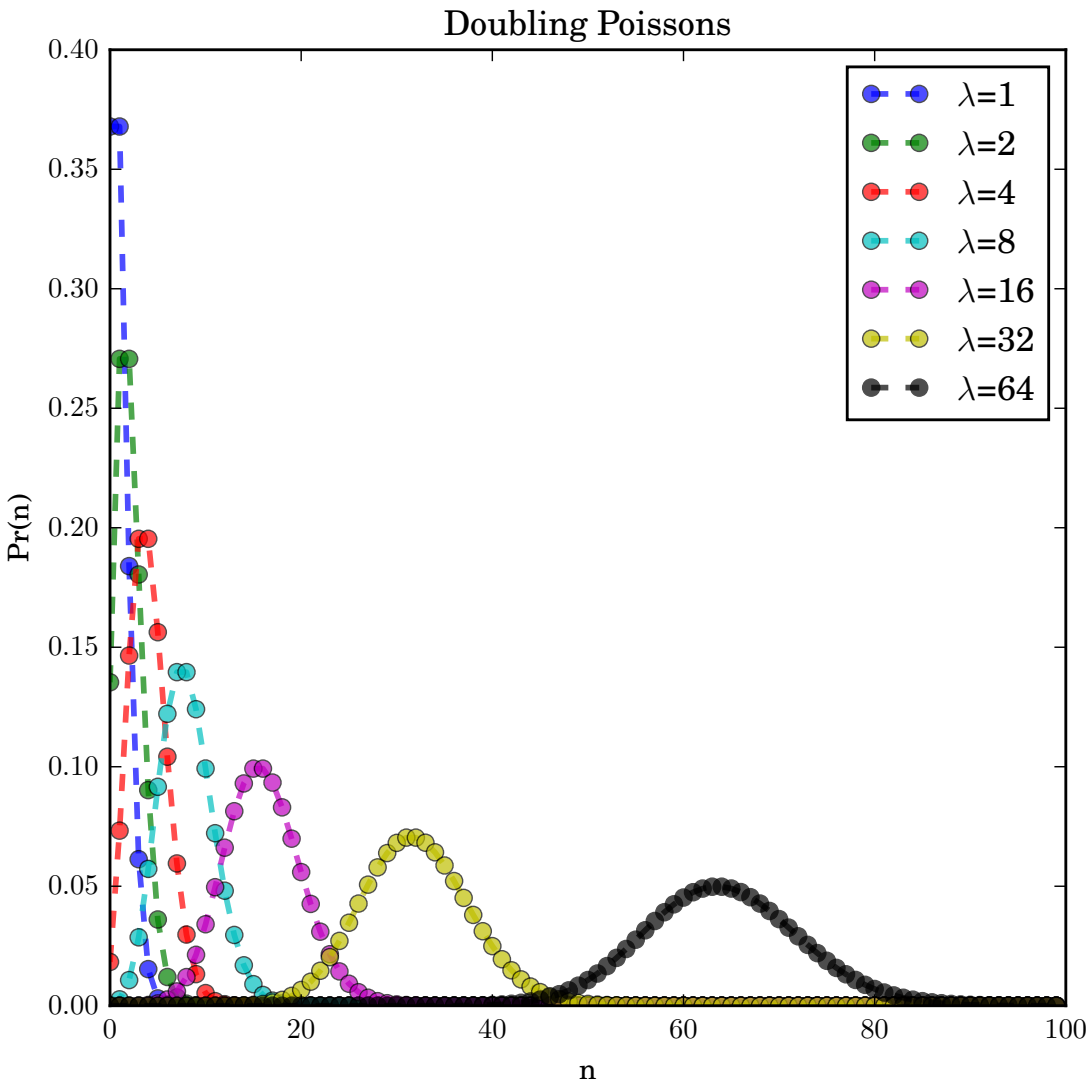
Adding Poissons



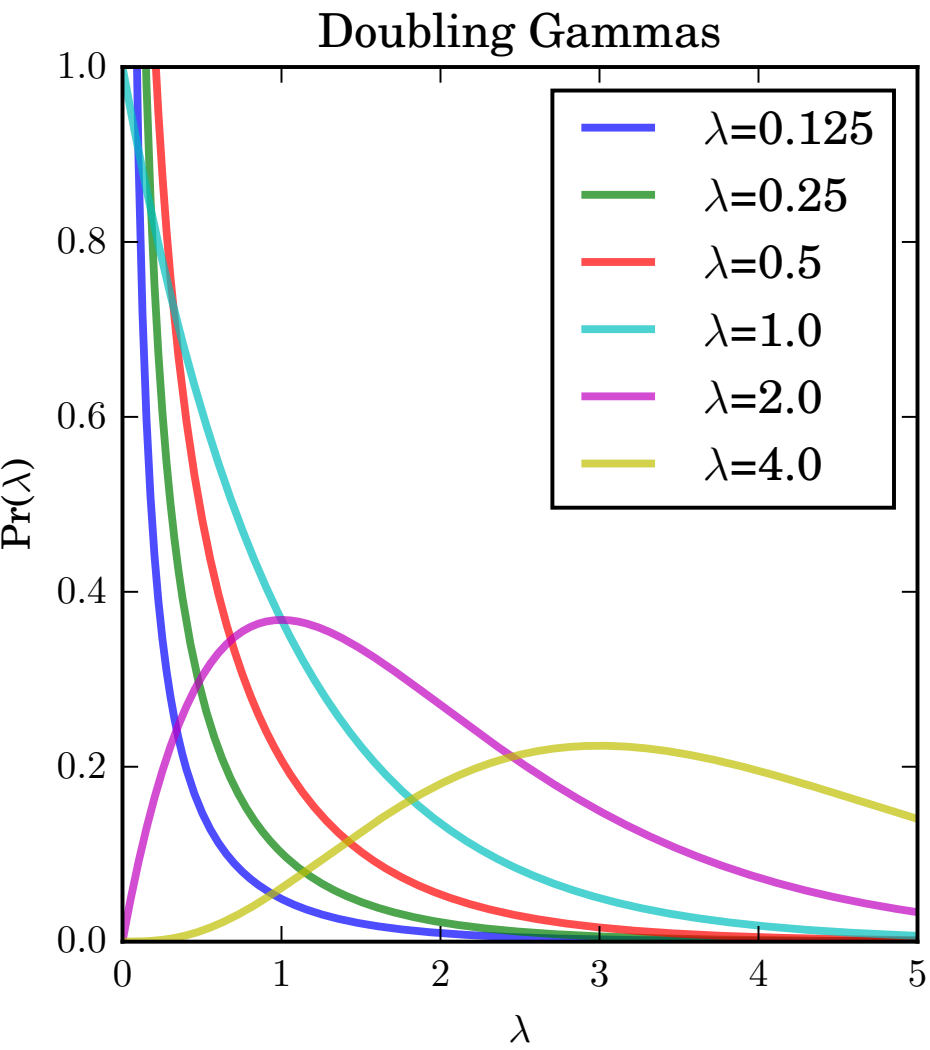


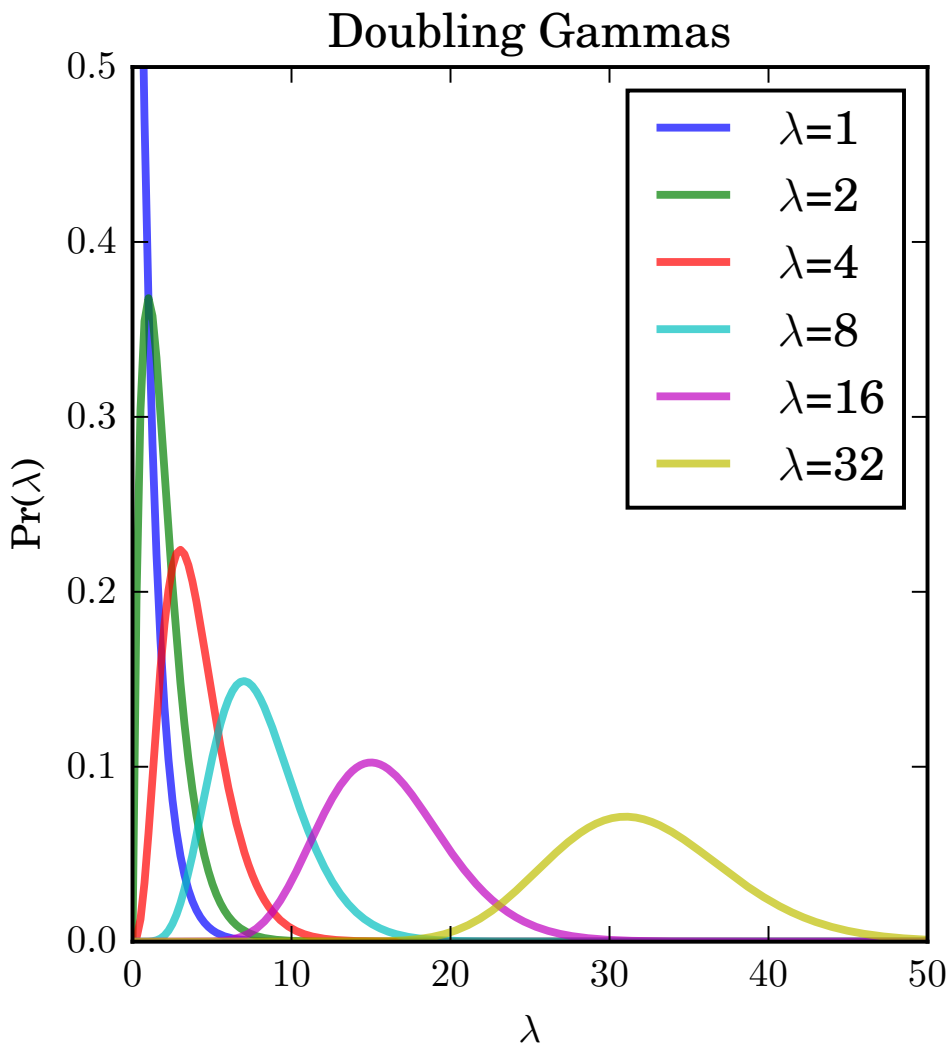
if $X \sim \text{Poisson}(\lambda_X)$ and $Y \sim \text{Poisson}(\lambda_Y)$
then $(X + Y) \sim \text{Poisson}(\lambda_{X+Y})$

Doubling Poissons



Doubling Gammas





Measures

A measure corresponds to the notion of the size or area.

Definition of Additivity for Functions

A function $m(\cdot)$ on domain $\mathcal{X}$ is additive if whenever $A, B \subset \mathcal{X}$ and $A \cap B = \emptyset$ then $m(A \cup B) = m(A) + m(B)$.

Definition (roughly) of a Measure

A measure$^{\mathcal{W}}$ on domain $\mathcal{X}$ is a non-negative additive function $m(\cdot)$ on (certain well behaved) subsets of X so that $m(\emptyset) = 0$.

Measures

A measure corresponds to the notion of the size or area.

Definition of Additivity for Functions

A function $m(\cdot)$ on domain $\mathcal{X}$ is additive if whenever $A, B \subset \mathcal{X}$ and $A \cap B = \emptyset$ then $m(A \cup B) = m(A) + m(B)$.

Definition (roughly) of a Measure

A measure$^{\mathcal{W}}$ on domain $\mathcal{X}$ is a non-negative additive function $m(\cdot)$ on (certain well behaved) subsets of X so that $m(\emptyset) = 0$.

- If $m(\mathcal{X}) = 1$ then the measure is a probability measure$^{\mathcal{W}}$
  - generalising a PDF (density) and PMF (mass).
- Measure can be defined in terms of an integral on the space.
- In formal theory, the space $\mathcal{X}$ is made to be closed under finite union and complement, *i.e.* Borel σ-field.

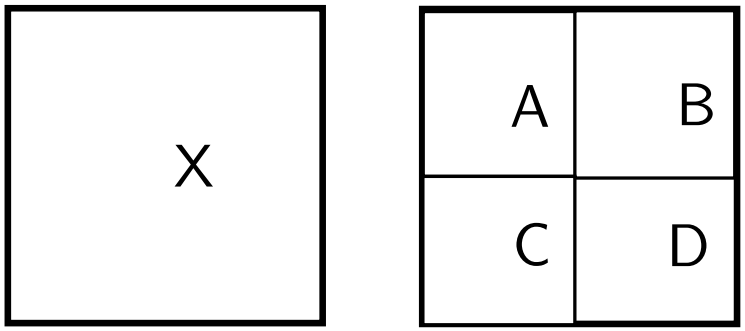
Measures and Poissons

- A measure corresponds to the notion of size or area.
- Note:
  - measures are additive functions, and
  - Poisson distribution is additive.

Measures and Poissons

- A measure corresponds to the notion of size or area.
- Note:
  - measures are additive functions, and
  - Poisson distribution is additive.
- **Idea:** use a measure to define a Poisson-based spatial process,
  - so the measure gives the Poisson rate for count of points in any subset of the space,
  - and make non-overlapping subspaces independent.
- Called a Poisson Point Process.

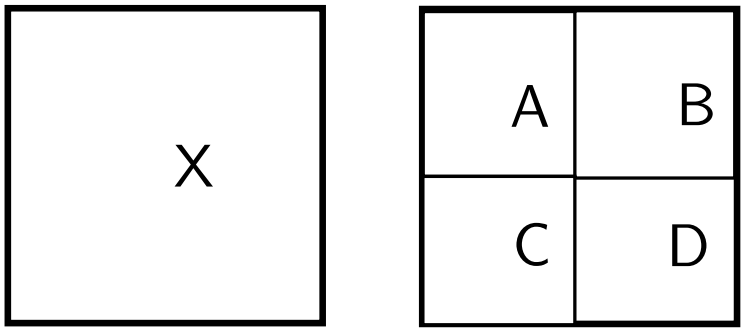
A Spatial Interpretation of PPPs



partition $\mathcal{X}$ into A, B, C, D

- number of points sampled in $\mathcal{X}$, $N(\mathcal{X}) \sim \text{Poisson}(m(\mathcal{X}))$, $N(A) \sim \text{Poisson}(m(A))$, $N(B) \sim \text{Poisson}(m(B))$, *etc.*
  - by definition of the PPP

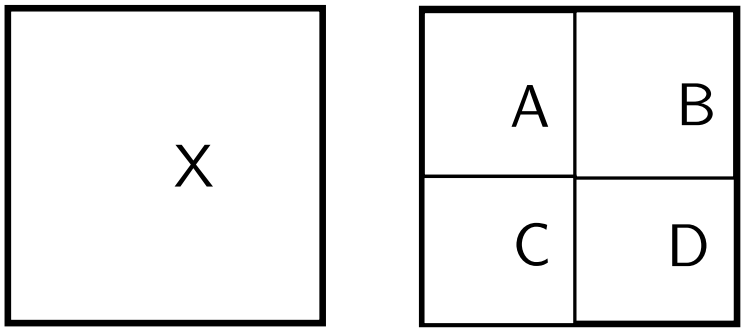
A Spatial Interpretation of PPPs



partition $\mathcal{X}$ into A, B, C, D

- number of points sampled in $\mathcal{X}$, $N(\mathcal{X}) \sim \text{Poisson}(m(\mathcal{X}))$, $N(A) \sim \text{Poisson}(m(A))$, $N(B) \sim \text{Poisson}(m(B))$, *etc.*
  - by definition of the PPP
- $N(\mathcal{X}) \sim \text{Poisson}(m(A) + m(B) + m(C) + m(D))$
  - by additivity of measure $m(\cdot)$

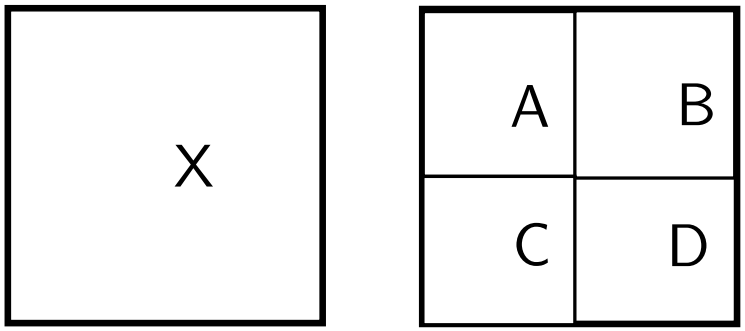
A Spatial Interpretation of PPPs



partition $\mathcal{X}$ into A, B, C, D

- number of points sampled in $\mathcal{X}$, $N(\mathcal{X}) \sim \text{Poisson}(m(\mathcal{X}))$, $N(A) \sim \text{Poisson}(m(A))$, $N(B) \sim \text{Poisson}(m(B))$, *etc.*
  - by definition of the PPP
- $N(\mathcal{X}) \sim \text{Poisson}(m(A) + m(B) + m(C) + m(D))$
  - by additivity of measure $m(\cdot)$
- $N(A) + N(B) + N(C) + N(D) \sim \text{Poisson}(m(A) + m(B) + m(C) + m(D))$
  - by additivity of Poisson

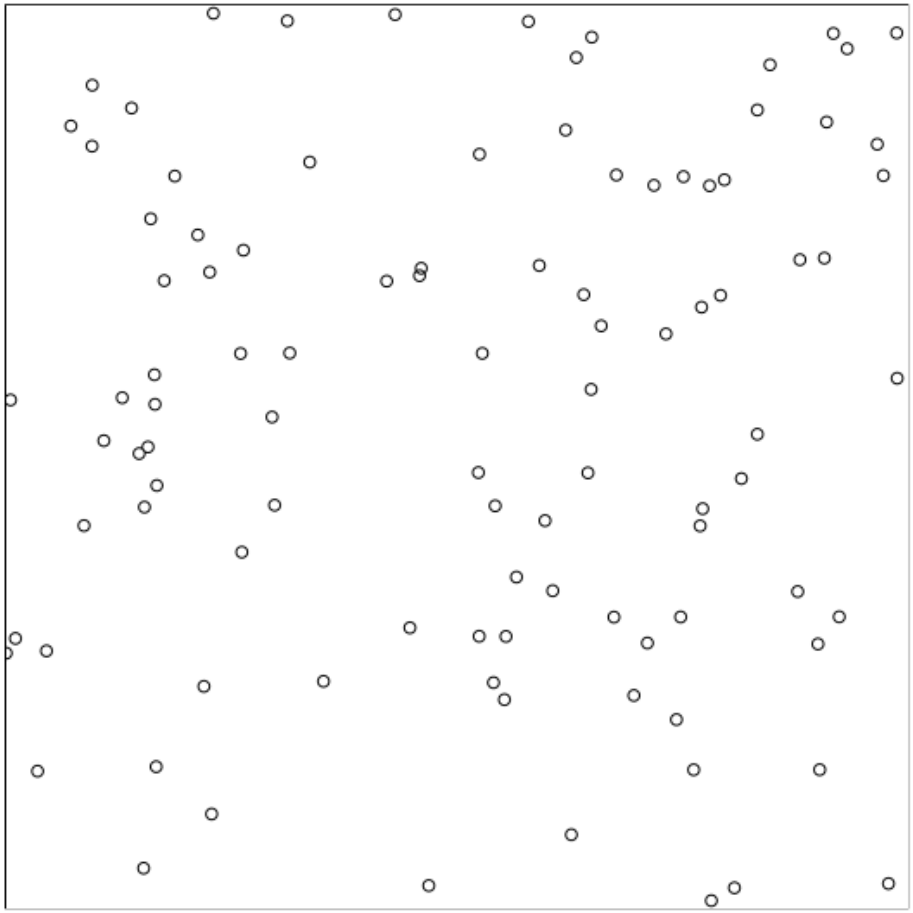
A Spatial Interpretation of PPPs



partition $\mathcal{X}$ into A, B, C, D

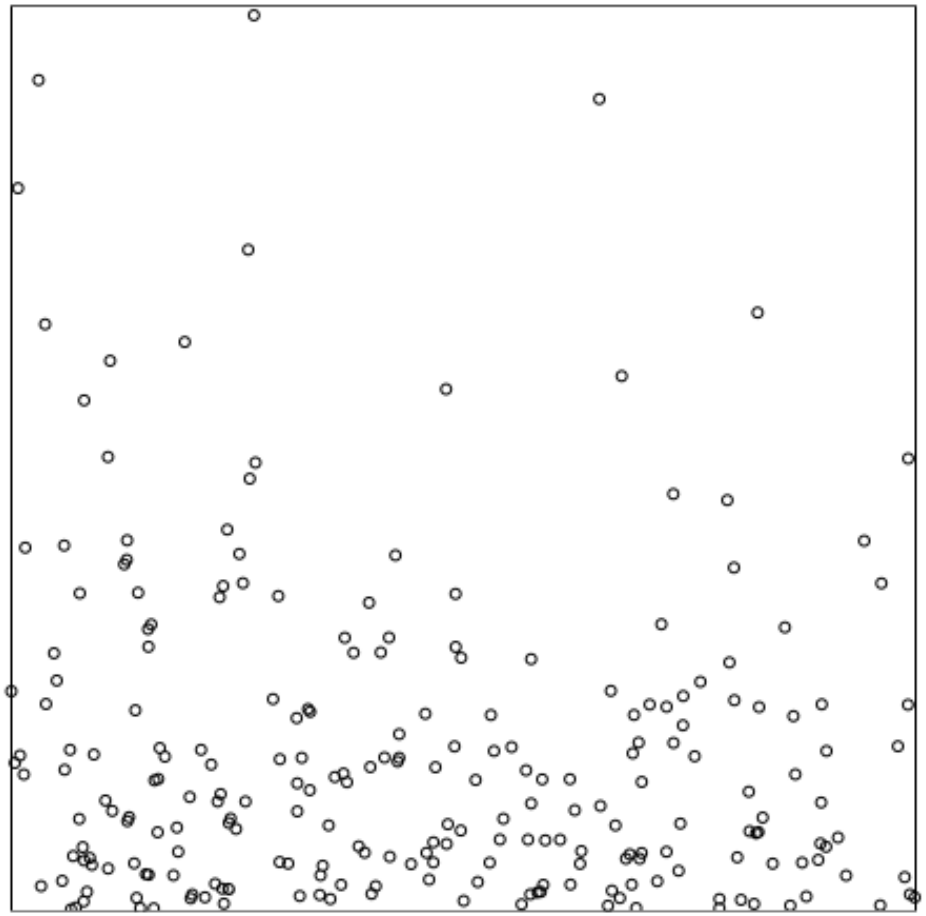
- number of points sampled in $\mathcal{X}$, $N(\mathcal{X}) \sim \text{Poisson}(m(\mathcal{X}))$, $N(A) \sim \text{Poisson}(m(A))$, $N(B) \sim \text{Poisson}(m(B))$, *etc.*
  - by definition of the PPP
- $N(\mathcal{X}) \sim \text{Poisson}(m(A) + m(B) + m(C) + m(D))$
  - by additivity of measure $m(\cdot)$
- $N(A) + N(B) + N(C) + N(D) \sim \text{Poisson}(m(A) + m(B) + m(C) + m(D))$
  - by additivity of Poisson
- so $N(\mathcal{X})$ equivalent in probability to $N(A) + N(B) + N(C) + N(D)$

Example of a PPP



- PPP is a spatial process
- domain $\mathcal{X} = [0, 1]^2$ with constant rate $\rho(x, y) = 100$
- total rate is $\int_0^1 \int_0^1 \rho(x, y) \mathrm{d}\, x, y = 100$
- generate $N \sim \text{Poisson}(100)$ points uniformly on unit square
- both the number of points N and their locations are sampled

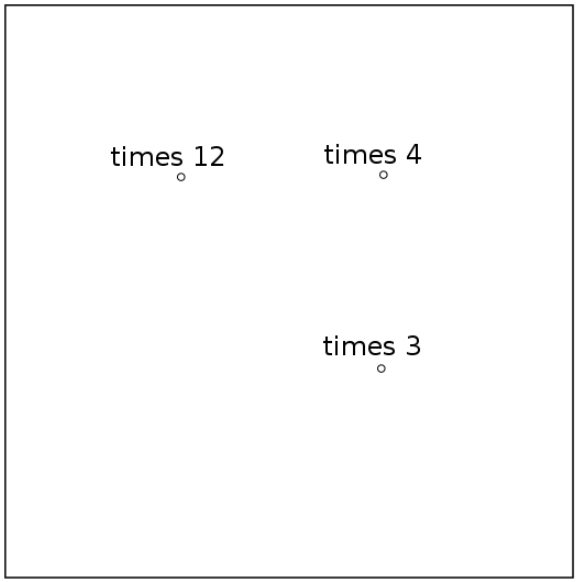
Example of a PPP, cont.



- PPP on unit square with rate function $\rho(x, y) = e^{7-5y}$
- total rate is $\int_0^1 \int_0^1 \rho(x, y) \mathrm{d}\, x, y = 217.85$
- generate $N \sim \text{Poisson}(217.85)$ points according to PDF

$$p(x, y) = \frac{1}{217.85} e^{7-5y} \ .$$

Example of a PPP, cont



- PPP is a unit square with rate function

$$\rho(x,y) = 10\,\delta_{x=1/3,y=2/3} + 5\,\delta_{x=2/3,y=2/3} + 2\,\delta_{x=2/3,y=1/3} + 1\,\delta_{x=1/3,y=1/3}$$

- total rate is $\int_0^1 \int_0^1 \rho(x,y)\mathrm{d}\,x,y = 18$
- but rate is discrete, so to sample:
  1. sample Poisson(10) points at $(1/3, 2/3)$,
  2. sample Poisson(5) points at $(2/3, 2/3)$,
  3. sample Poisson(2) points at $(2/3, 1/3)$,
  4. sample Poisson(1) points at $(1/3, 1/3)$,
- so behaves like a Poisson at the individual points

The Formal Definition

For any topological space $\mathcal{T}$, $\mathscr{B}(\mathcal{T})$ will denote the Borel σ-field of subsets of $\mathcal{T}$. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be some probability space and $\mathbb{X}$ be complete, separable and endowed with a metric $d_{\mathbb{X}}$. Define on $(\Omega, \mathscr{F}, \mathbb{P})$ a Poisson random measure $\tilde{N}$ on $\mathbb{S} = \mathbb{R}^+ \times \mathbb{X}$ with intensity measure ν. This means that

(i) for any C in $\mathscr{B}(\mathbb{S})$ such that $\nu(C) = \mathbb{E}[\tilde{N}(C)] < \infty$, the probability distribution of the random variable $\tilde{N}(C)$ is Poisson$(\nu(C))$;

(ii) for any finite collection of pairwise disjoint sets, $A_1, \ldots, A_k$, in $\mathscr{B}(\mathbb{S})$, the random variables $\tilde{N}(A_1), \ldots, \tilde{N}(A_k)$ are mutually independent.

Moreover, the measure ν must satisfy the following conditions,

$$\int_{(0,1)} s\nu(\mathrm{d}s, \mathbb{X}) < \infty, \quad \nu([1,\infty) \times \mathbb{X}) < \infty.$$

We refer to Daley & Vere-Jones (1988) for an exhaustive account on Poisson random measures.

What is a Poisson Point Process?

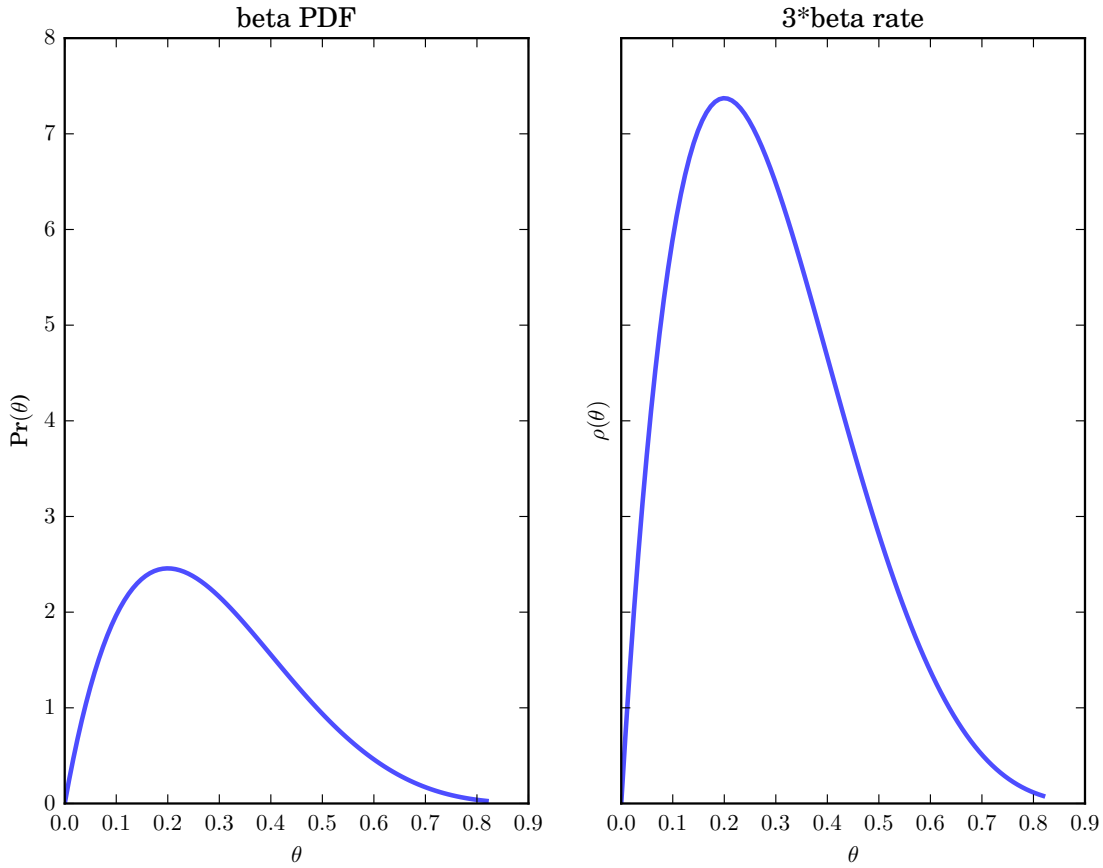
Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Complex ... so look at more examples.

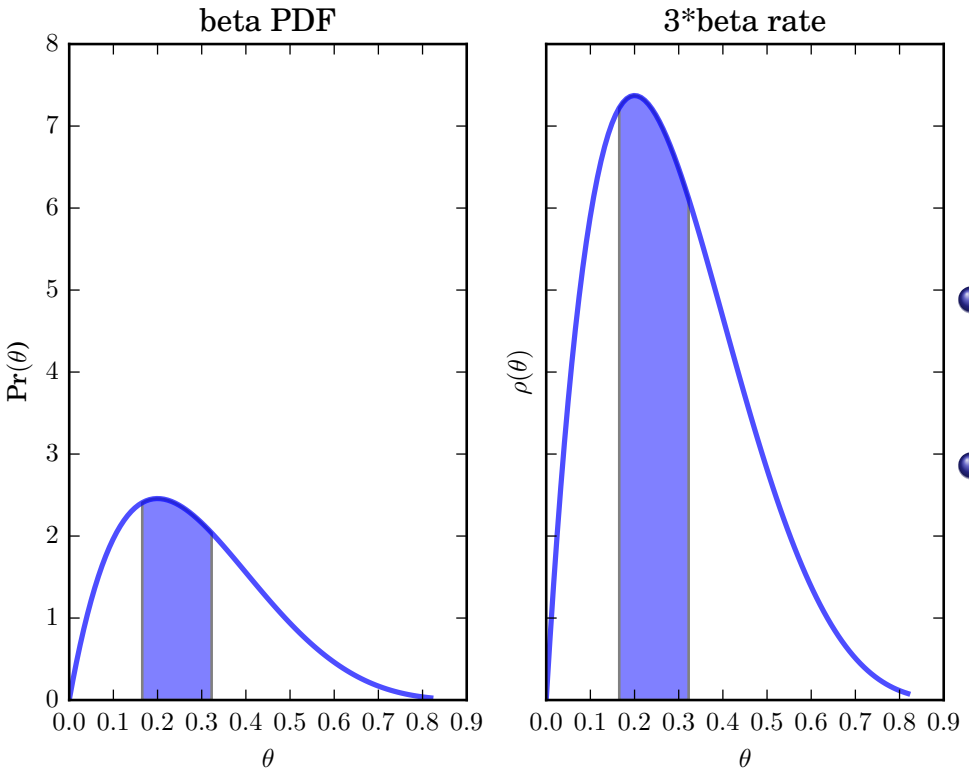
Poisson Point Process versus Probability Density Function



domain $\mathcal{X} = (0, 1)$

- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$
  - note the rate does not sum/integrate to 1

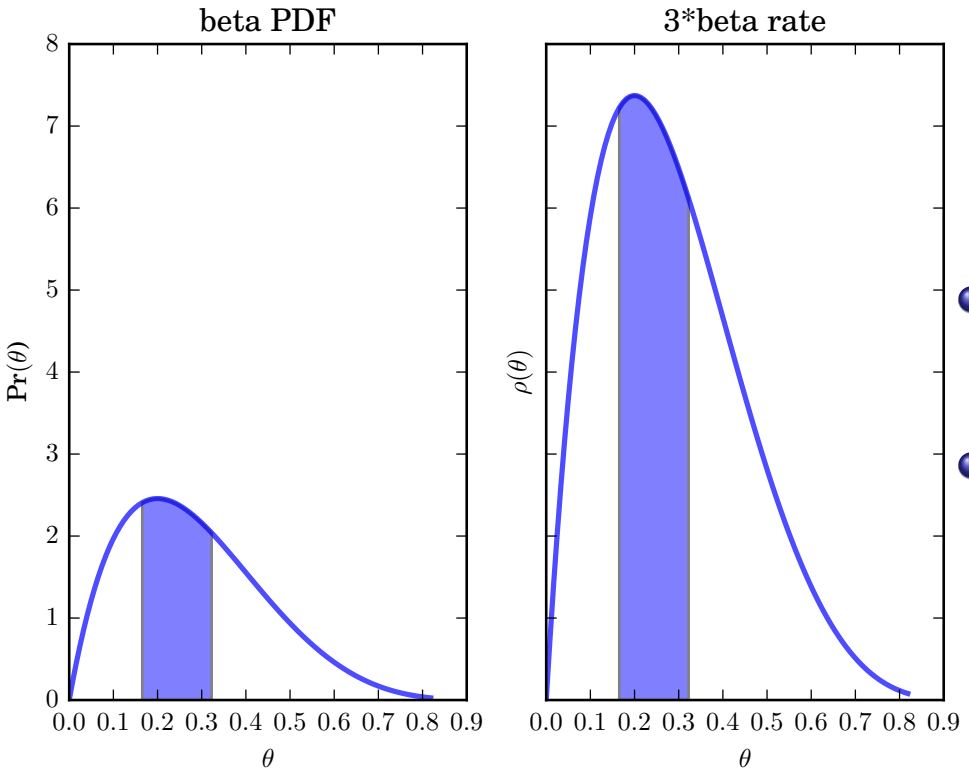
PPP versus PDF



- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$

- shaded part on left has $0.165 < x < 0.32$ and area 0.385

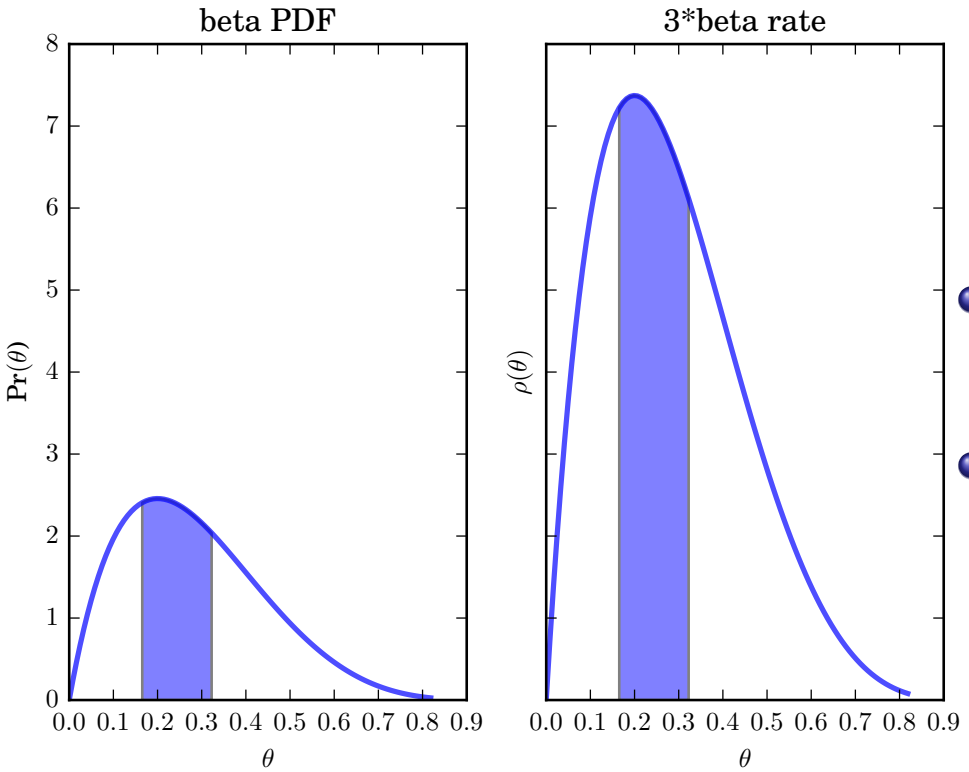
PPP versus PDF



- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$

- shaded part on left has $0.165 < x < 0.32$ and area 0.385
- PDF says we have 0.385 chance of getting a point in shaded area

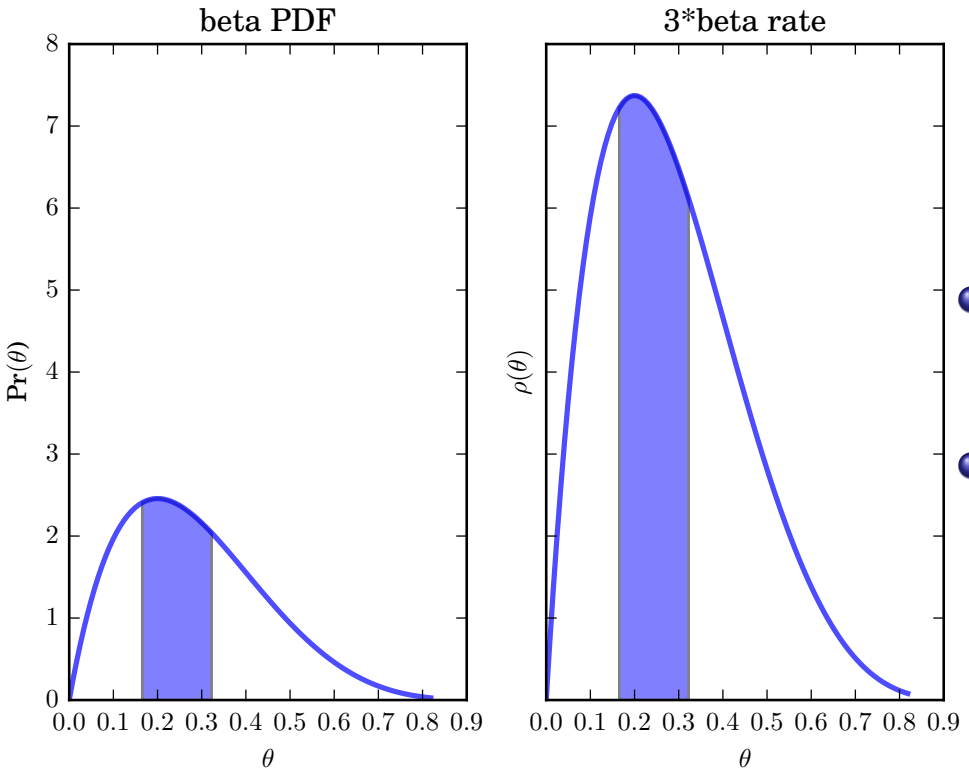
PPP versus PDF



- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$

- shaded part on left has $0.165 < x < 0.32$ and area 0.385
- PDF says we have 0.385 chance of getting a point in shaded area
- PPP says we will get $n \sim \text{Poisson}(3 * 0.385)$ points in shaded area

PPP versus PDF



- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$

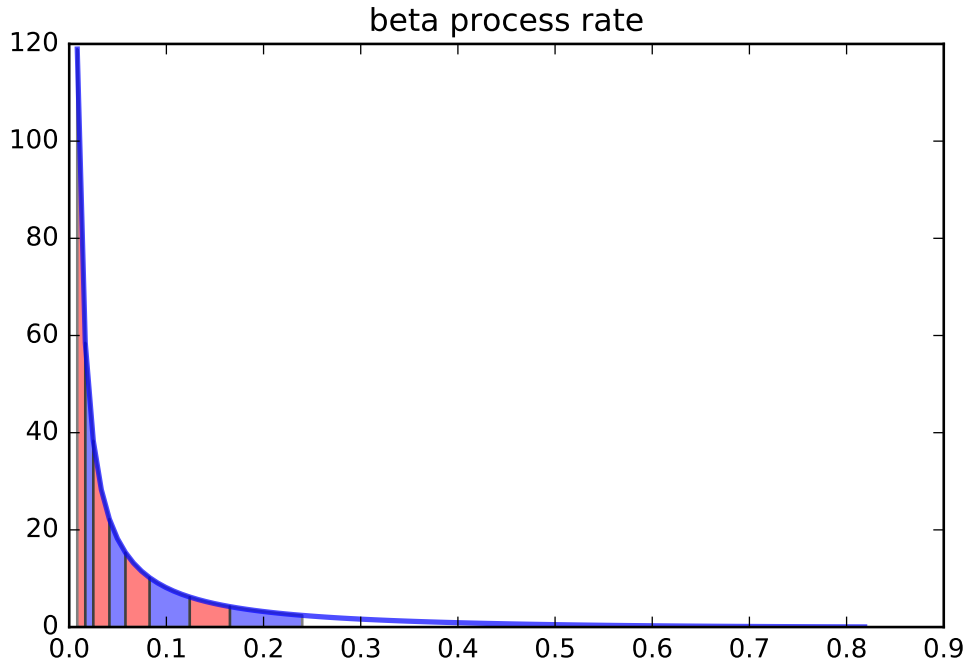
- shaded part on left has $0.165 < x < 0.32$ and area 0.385
- PDF says we have 0.385 chance of getting a point in shaded area
- PPP says we will get $n \sim \text{Poisson}(3 * 0.385)$ points in shaded area
- for PDF and PPP, probability of generating a single point in the shaded part is the same, $p(x)/0.385$

PPP versus PDF, cont.

- A PPP may generate a countably infinite number of points in the full space:
  - is OK as long as you can partition the space into parts with finite measure (so they have finite number of points).
- This is exactly what we need for non-parametrics:
  - to generate an infinite number of possibilities to work with.

PPP versus PDF, cont.

- A PPP may generate a countably infinite number of points in the full space:
  - is OK as long as you can partition the space into parts with finite measure (so they have finite number of points).
- This is exactly what we need for non-parametrics:
  - to generate an infinite number of possibilities to work with.



- PDF for "improper" beta$(0, 3)$ (*i.e.*, does not normalise)
- integral diverges as $\theta \to 0$
- so PPP get infinite number of points near zero

Poisson Point Process in the Discrete Case

Interpretation of the *discrete case*:

- $\rho(x) = \sum_k \lambda_k \delta_{x_k}$,
- generate $N_k \sim \text{Poisson}(\lambda_k)$ occurrences of x_k
- superimpose the sets of points

Poisson Point Process in the Finite Case

Interpretation of the *finite case*:

- assume $\rho(\mathcal{X})$ is finite,
- generate count of points $N \sim \text{Poisson}\left(\rho(\mathcal{X})\right)$
- generate points X_n for $n = 1, ..., N$ by sampling according to the PDF $\rho(x)/\rho(\mathcal{X})$.

Poisson Point Process in the Infinite Case

Interpretation of the *infinite case*:

- partition up $\mathcal{X}$ into sets A_k for $k = 1, ..., \infty$ so that each $\rho(A_k)$ is finite,
- apply the finite case to the Poisson process on A_k to get samples X_k,
- merge the resulting samples to get sample $X = \bigcup_k X_k$,

Poisson Point Process in the Infinite Case

Interpretation of the *infinite case*:

- partition up $\mathcal{X}$ into sets A_k for $k = 1, ..., \infty$ so that each $\rho(A_k)$ is finite,
- apply the finite case to the Poisson process on A_k to get samples X_k,
- merge the resulting samples to get sample $X = \bigcup_k X_k$,

Method of combining independent PPPs called superposition.

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process[W]. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process[W]. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- All works due to additivity.

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process[W]. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- All works due to additivity.
- Rate $\rho(x)$ also called intensity, or Lévy measure (in some contexts).

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- All works due to additivity.
- Rate $\rho(x)$ also called intensity, or Lévy measure (in some contexts).
- Points in sample sometimes called "events".

A Poisson Point Process

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process[W]. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $A \cap B =$ then $N(A) \perp\!\!\!\perp N(B)$.

- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- All works due to additivity.
- Rate $\rho(x)$ also called intensity, or Lévy measure (in some contexts).
- Points in sample sometimes called "events".
- Just need to know the measure, not the rate.

Processes other than Poisson

- PPPs are defined axiomatically from the Poisson distribution.
- Works because it is additive.
- But, so are negative binomials, gammas, stable distributions, *etc.*.
  - look up infinite divisibility (probability)$^{\mathcal{W}}$ for more!

Processes other than Poisson

- PPPs are defined axiomatically from the Poisson distribution.
- Works because it is additive.
- But, so are negative binomials, gammas, stable distributions, *etc.*.
  - look up infinite divisibility (probability)[W] for more!
- We can similarly define:
  - negative binomial process,
  - gamma process,
    - no longer integer counts but now a discrete measure (weights on points)
  - stable process,
  - Dirichlet process
    - now normalised weights, a probability vector
    - by normalising a gamma process.

Processes other than Poisson

- PPPs are defined axiomatically from the Poisson distribution.
- Works because it is additive.
- But, so are negative binomials, gammas, stable distributions, *etc.*.
  - look up infinite divisibility (probability)$^{\mathcal{W}}$ for more!
- We can similarly define:
  - negative binomial process,
  - gamma process,
    - no longer integer counts but now a discrete measure (weights on points)
  - stable process,
  - Dirichlet process
    - now normalised weights, a probability vector
    - by normalising a gamma process.
- These definitions are not constructive.
- We will define these later in a constructive way!

Example: A Gamma Process

Definition of gamma process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$ with finite integral $(\int_{\mathcal{X}} \rho(x) \mathrm{d}\, x < \infty)$ define a gamma process$^{\mathcal{W}}$ with scale β. The gamma process generates a random measure $\Gamma(\cdot)$ on $\mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$. The random measure satisfies:

- $\Gamma(A) \sim \text{Gamma}\,(\rho(A), \beta)$, and
- if $A \cap B =$ then $\Gamma(A) \perp\!\!\!\perp \Gamma(B)$.

Some modifications needed from the PPP:

- total measure cannot be infinite, usually
- no longer returning number of points $N(\cdot)$ but a measure $\Gamma(\cdot)$.

Outline

Introduction to Hierarchical Dirichlet Processes

- Early 2000's Prof. Michael Jordan introduced the Dirichlet Process[W] to the Machine Learning community as a non-parametric method.
- Good tutorials and reviews can be found at:
  - *"Tutorial on Dirichlet Processes and Hierarchical Dirichlet Processes"* by Yeh Whye Teh, 2007
  - *"Completely Random Measures, Hierarchies and Nesting"* by Michael Jordan, 2010

Dirichlet Process Definition

- A Dirichlet Process[W] (DP) is a distribution over probability measures (densities, masses, distributions).
- A DP has two parameters (inputs):

  mean: called the base distribution, denoted $H(\cdot)$ below

  concentration: which is like an *inverse variance*, denoted α

  or alternatively one parameter (input)

  measure: denoted $m(\cdot)$

  so $m(A) = \alpha H(A)$ and $\alpha = m(\mathcal{X})$.

Dirichlet Process Definition

- A Dirichlet Process[W] (DP) is a distribution over probability measures (densities, masses, distributions).
- A DP has two parameters (inputs):

  mean: called the base distribution, denoted $H(\cdot)$ below

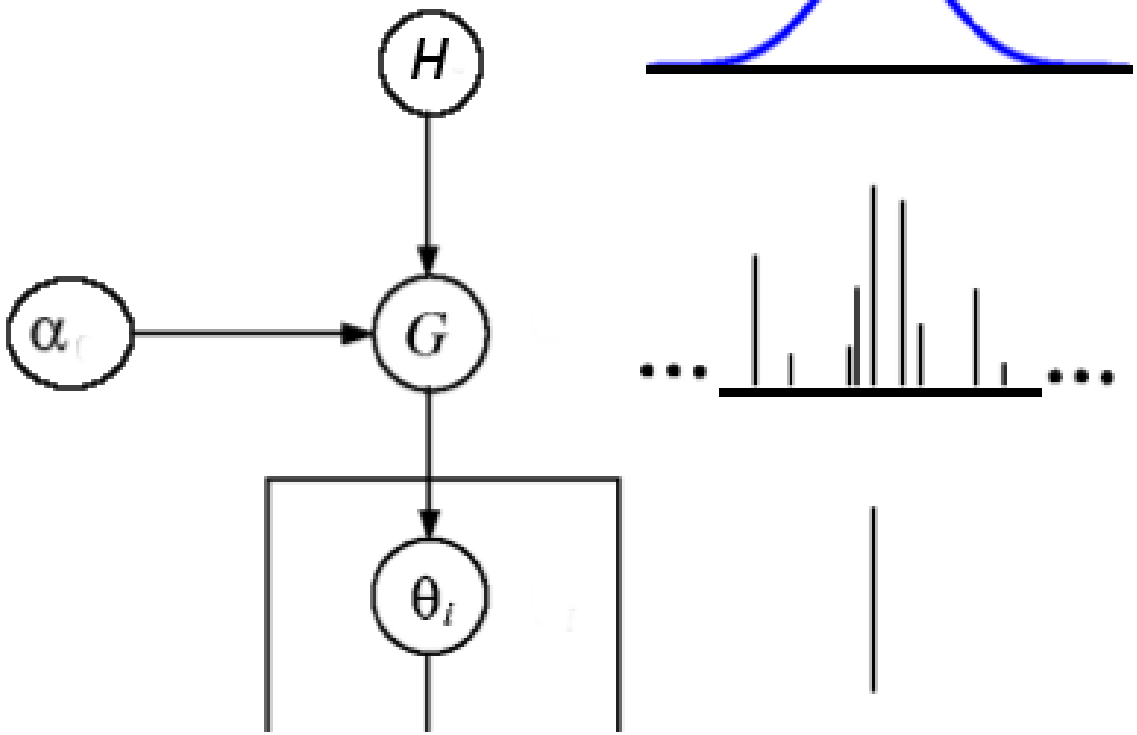
  concentration: which is like an *inverse variance*, denoted α

  or alternatively one parameter (input)

  measure: denoted $m(\cdot)$

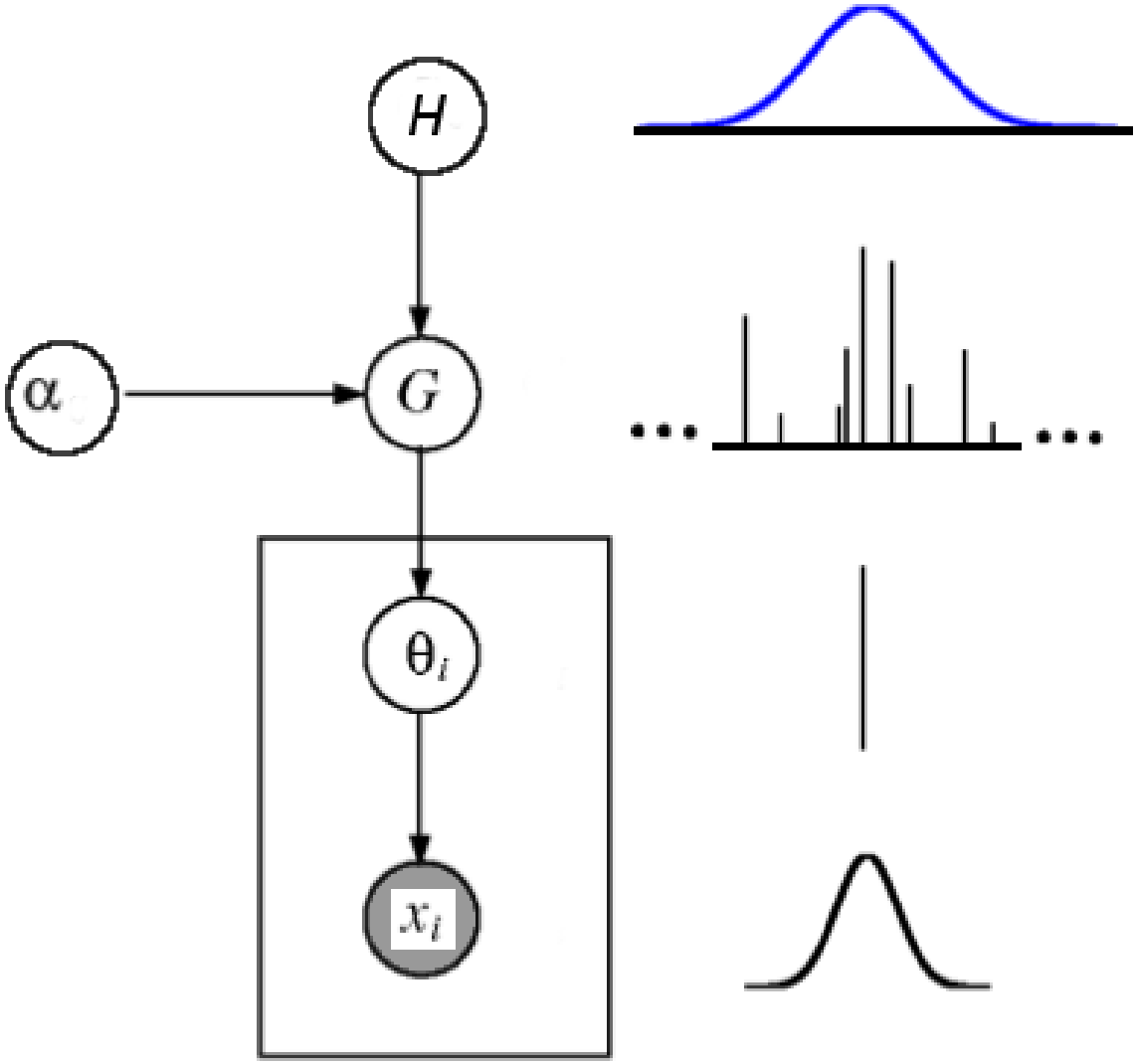
  so $m(A) = \alpha H(A)$ and $\alpha = m(\mathcal{X})$.
- DP defined as the natural extension of a Dirichlet to arbitrary probability measures.
- Details not important here: but related to the extension of the gamma process using additivity mentioned previously.

Dirichlet Process on Gaussian



- $H(\cdot)$ is a Gaussian
- G is DP sample of points from a Gaussian weighted according to a DP probability vector
- θ_i is a sample from G, originally from Gaussian $H(\cdot)$

Dirichlet Process for Gaussian Mixture



- θ_i is a sample from G, so have repeats
- $x_i \sim \text{Gaussian}(\theta_i, \sigma)$ (so x_i are clustered)

Dirichlet Process, cont.

- On a continuous and non-discrete domain, output/sample of a DP always looks like

$$\mu(x) = \sum_{i=1}^{\infty} \mu_i \delta_{\theta_i}(x)$$

  where each $\theta_i \sim H(\cdot)$ and $\mu(\mathcal{X}) = \sum_{\rangle=\infty}^{\infty} \mu_\rangle = \infty$.
- We refer to this as an an infinite probability vector indexed by points $\theta_i \in \mathcal{X}$.

Dirichlet Process, cont.

- On a continuous and non-discrete domain, output/sample of a DP always looks like

$$\mu(x) = \sum_{i=1}^{\infty} \mu_i \delta_{\theta_i}(x)$$

  where each $\theta_i \sim H(\cdot)$ and $\mu(\mathcal{X}) = \sum_{\rangle=\infty}^{\infty} \mu_\rangle = \infty$.
- We refer to this as an an infinite probability vector indexed by points $\theta_i \in \mathcal{X}$.
- The samples from the DP are not continuous distributions, even if the input $H(\cdot)$ is.

Dirichlet Process Review

- When the base (input) distribution is continous on domain $\mathcal{X}$, the DP produces a probability vector over a countable number of items in the domain $\mathcal{X}$.

Dirichlet Process Review

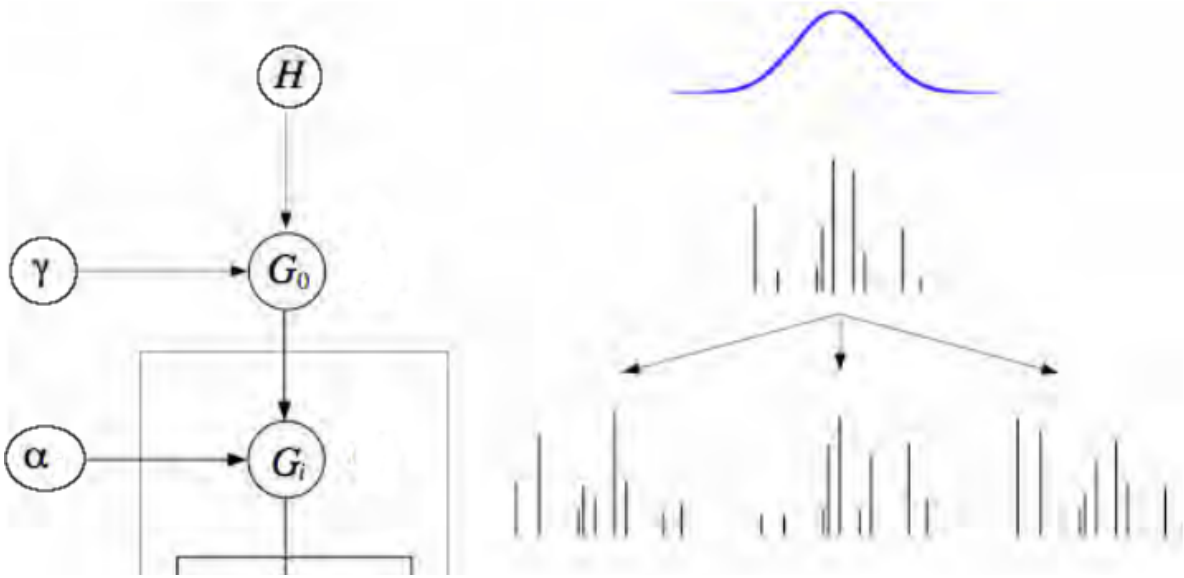
- When the base (input) distribution is continous on domain $\mathcal{X}$, the DP produces a probability vector over a countable number of items in the domain $\mathcal{X}$.
- DP can be used to build "infinite" mixture models.
- DP is ideal for non-parametrics!

Hierarchical Dirichlet Process Outline

Idea: If we take the output of a DP and supply it as input to another DP, what do we get?

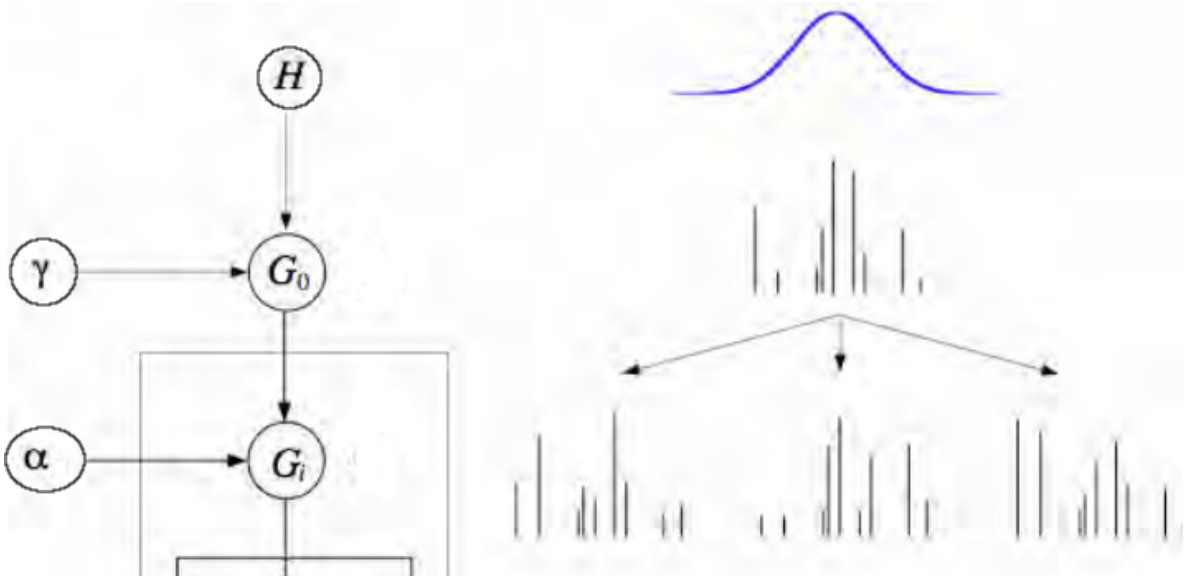
Hierarchical Dirichlet Process Outline

Idea: If we take the output of a DP and supply it as input to another DP, what do we get?



Hierarchical Dirichlet Process Outline

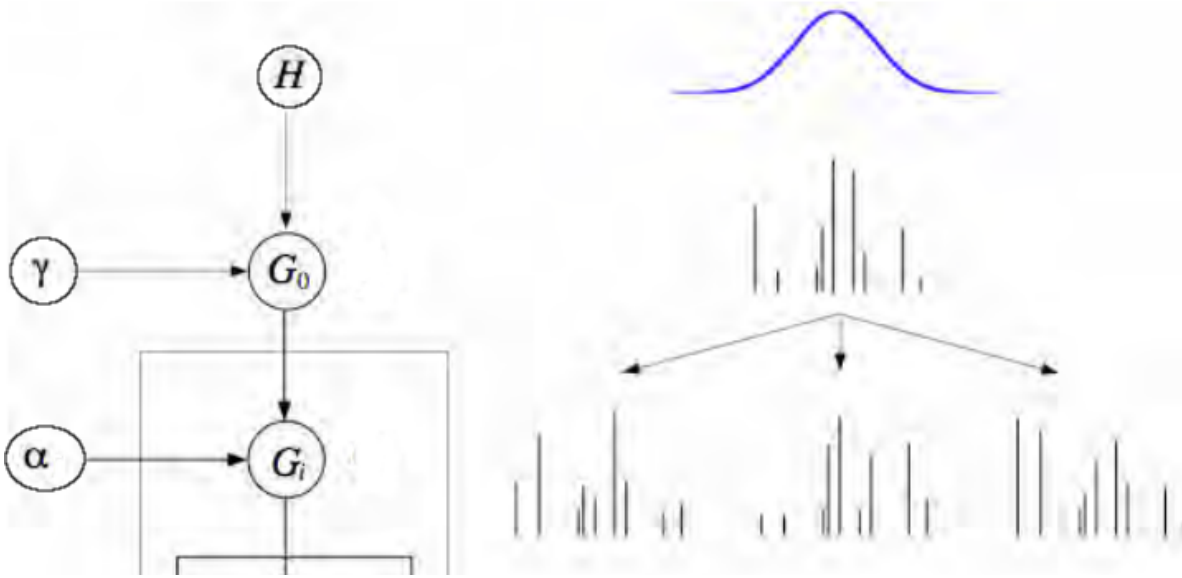
Idea: If we take the output of a DP and supply it as input to another DP, what do we get?



- Since the input measure is discrete, the DP will behave like a Dirichlet distribution.

Hierarchical Dirichlet Process Outline

Idea: If we take the output of a DP and supply it as input to another DP, what do we get?



- Since the input measure is discrete, the DP will behave like a Dirichlet distribution.
- What is the advantage of dealing with hierarchical DPs?
  - DP gives us alternative computational models to the Dirichlet.

Outline

Bayesian Non-parametrics: Critical Review

- How is the state of the field?
- Let us reconsider some famous work
- Hindsight is a wonderful thing

Bayesian Non-parametrics: Critical Review

- How is the state of the field?
- Let us reconsider some famous work
- Hindsight is a wonderful thing
- With all respect for my co-researchers 

The Hierarchical Dirichlet Process – HDP

Teh, Jordan, Beal & Blei, 2006 (2640 citations)

4.2 The Chinese restaurant franchise

In this section we describe an analog of the Chinese restaurant process for hierarchical Dirichlet processes that we refer to as the *Chinese restaurant franchise*. In the Chinese restaurant franchise, the metaphor of the Chinese restaurant process is extended to allow multiple restaurants which share a set of dishes.

The metaphor is as follows (see Figure 2). We have a restaurant franchise with a shared menu across the restaurants. At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers who sit at that table. Multiple tables in multiple restaurants can serve the same dish.

In this setup, the restaurants correspond to groups and the customers correspond to the factors θ_{ji}. We also let $\phi_1, \ldots, \phi_K$ denote K i.i.d. random variables distributed according to H; this is the global menu of dishes. We also introduce variables ψ_{jt} which represent the table-specific choice of dishes; in particular, ψ_{jt} is the dish served at table t in restaurant j.

Note that each θ_{ji} is associated with one ψ_{jt}, while each ψ_{jt} is associated with one ϕ_k. We introduce indicators to denote these associations. In particular, let t_{ji} be the index of the ψ_{jt} associated with θ_{ji}, and let k_{jt} be the index of ϕ_k associated with ψ_{jt}. In the Chinese restaurant franchise metaphor, customer i in restaurant j sat at table t_{ji} while table t in restaurant j serves dish k_{jt}.

We also need a notation for counts. In particular, we need to maintain counts of customers and counts of tables. We use the notation n_{jtk} to denote the number of customers in restaurant j at table t eating dish k. Marginal counts are represented with dots. Thus, $n_{jt\cdot}$ represents the number of customers in restaurant j at table t and $n_{j\cdot k}$ represents the number of customers in restaurant j eating dish k. The notation m_{jk} denotes the number of tables in restaurant j serving dish k. Thus, $m_{j\cdot}$ represents the number of tables in restaurant j, $m_{\cdot k}$ represents the number of tables serving dish k, and $m_{\cdot\cdot}$ the total number of tables occupied.

Let us now compute marginals under a hierarchical Dirichlet process when G_0 and G_j are

see also the Wikipedia *Hierarchical Dirichlet process*

Online Variational Inference for the HDP

Wang, Paisley & Jordan, 2011 (194 citations)

Online Variational Inference for the Hierarchical Dirichlet Process

Chong Wang **John Paisley** **David M. Blei**
Computer Science Department, Princeton University
{chongw,jpaisley,blei}@cs.princeton.edu

Abstract

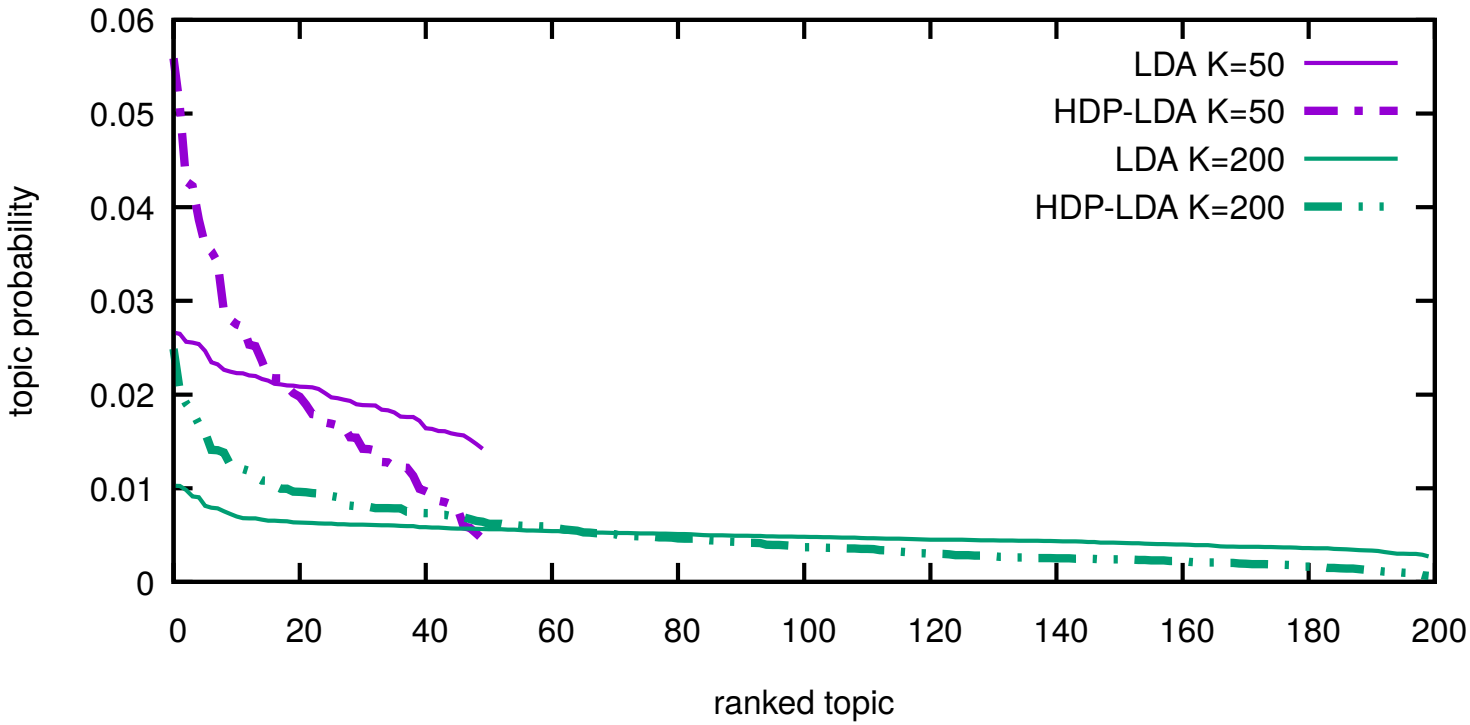
The hierarchical Dirichlet process (HDP) is a Bayesian nonparametric model that can be used to model mixed-membership data with a potentially infinite number of components. It has been applied widely in probabilistic topic modeling, where the data are documents and the compo-

like classification, exploration, and summarization. Unlike Our motivation for extending this algorithm to the HDP is that LDA requires choosing the number of topics in advance. In a traditional setting, where fitting multiple models might be viable, the number of topics can be determined with cross validation or held-out likelihood. However, these techniques become impractical when the data set size is large, and they passes through the data and are not easily applicable to

See *Online Variational Inference for the Hierarchical Dirichlet Process*

Choosing the "right" number of topics is the wrong emphasis.

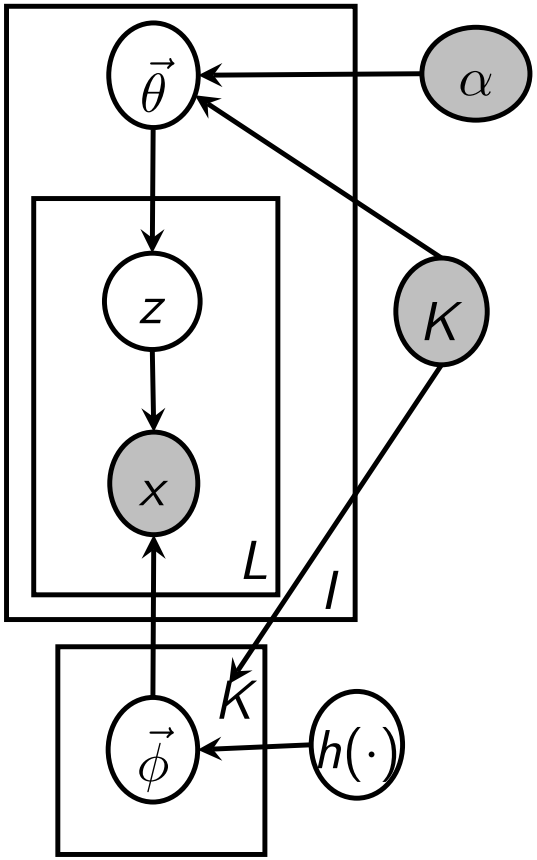
Topic Proportions in LDA versus HDP-LDA



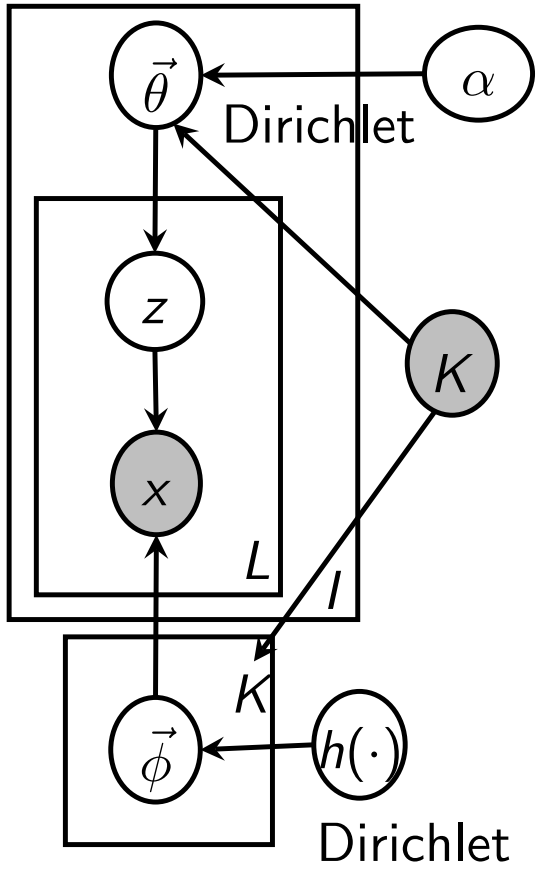
- vanilla LDA aprior expects topics to be equally likely
- HDP-LDA also "learns" the right topic proportions
- for any large collection, the "right" number of topics is huge
- getting the "right" number of topics is not so important

LDA Models Graphically

LDA (2003)

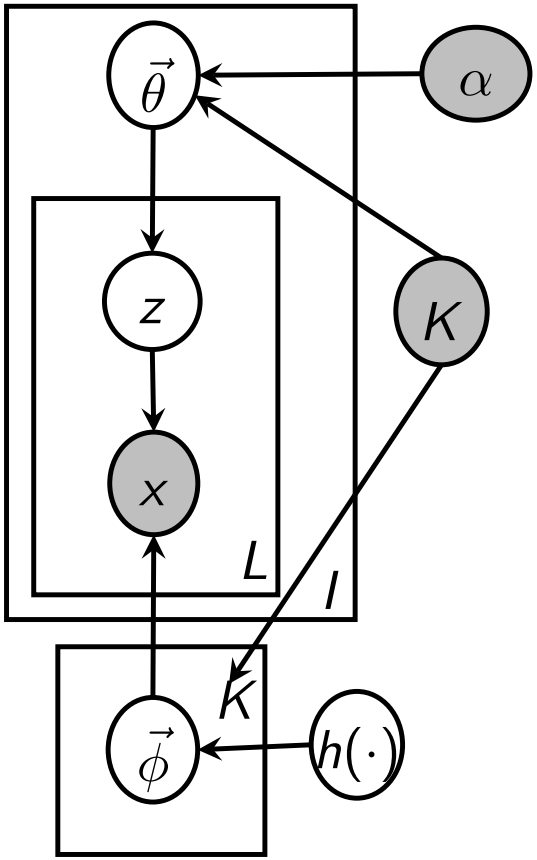


LDA (2005)

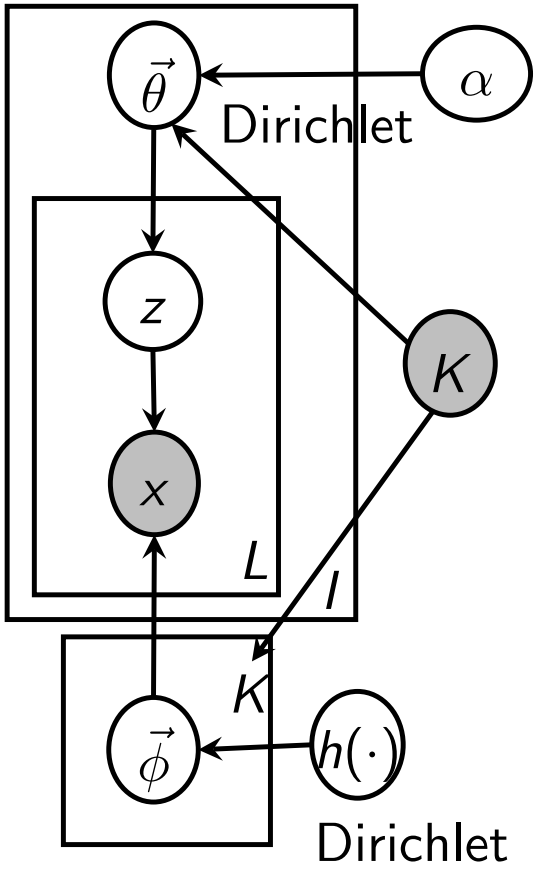


LDA Models Graphically

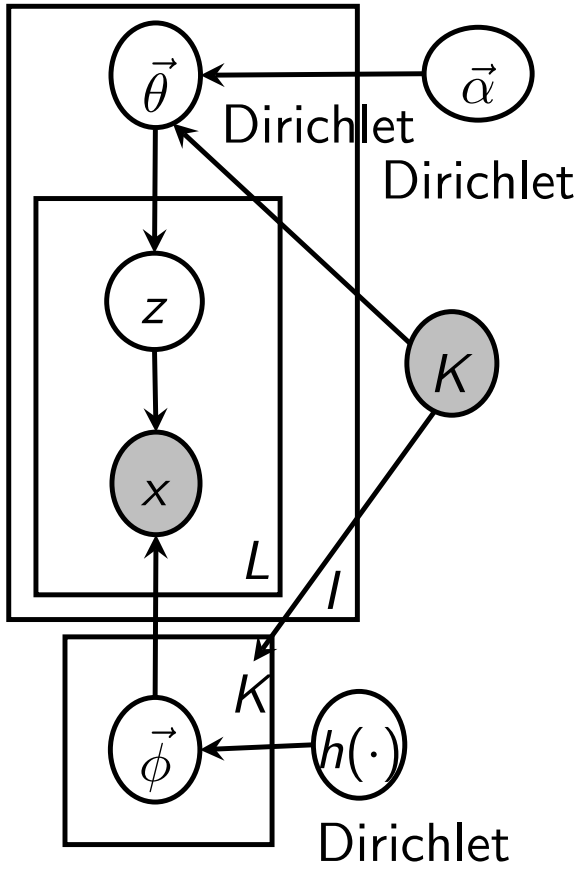
LDA (2003)



LDA (2005)

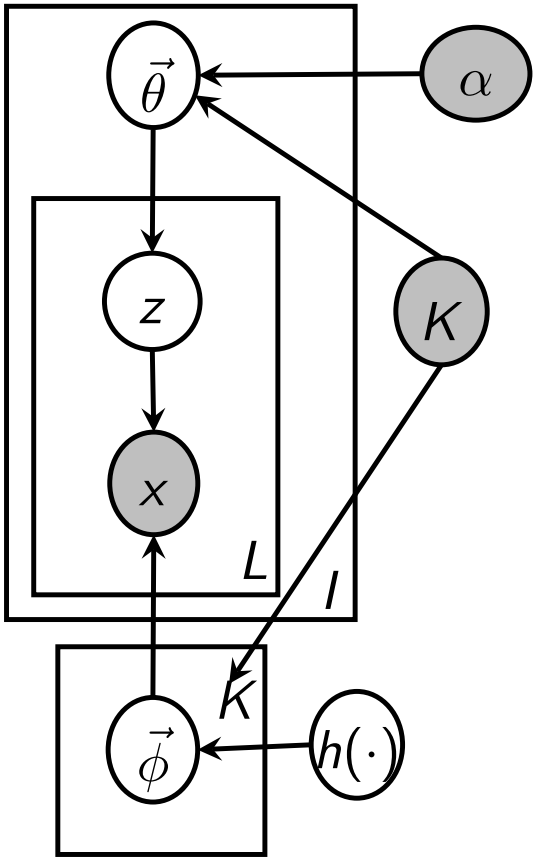


LDA (2008)

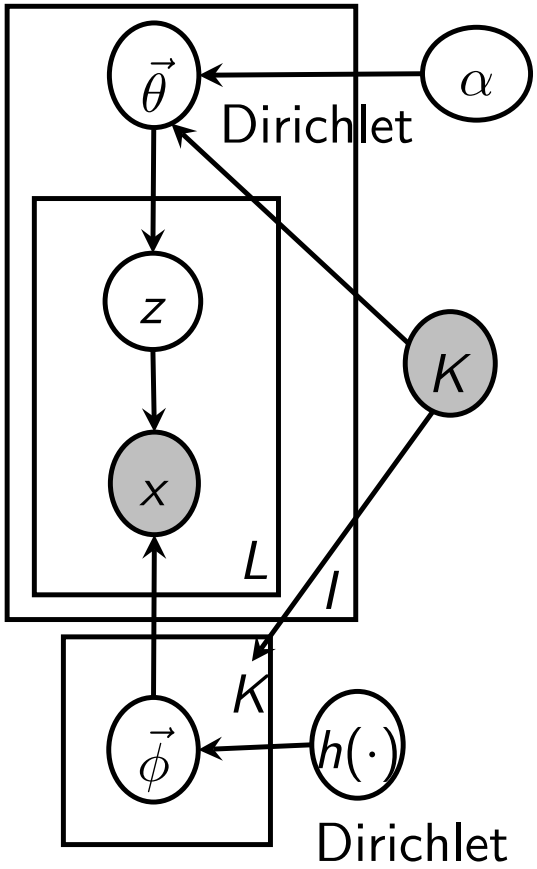


LDA Models Graphically

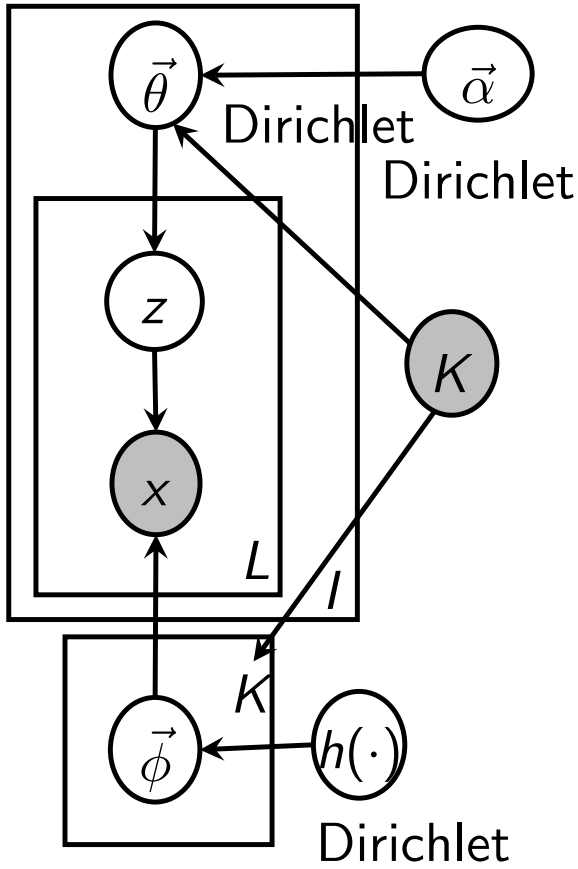
LDA (2003)



LDA (2005)



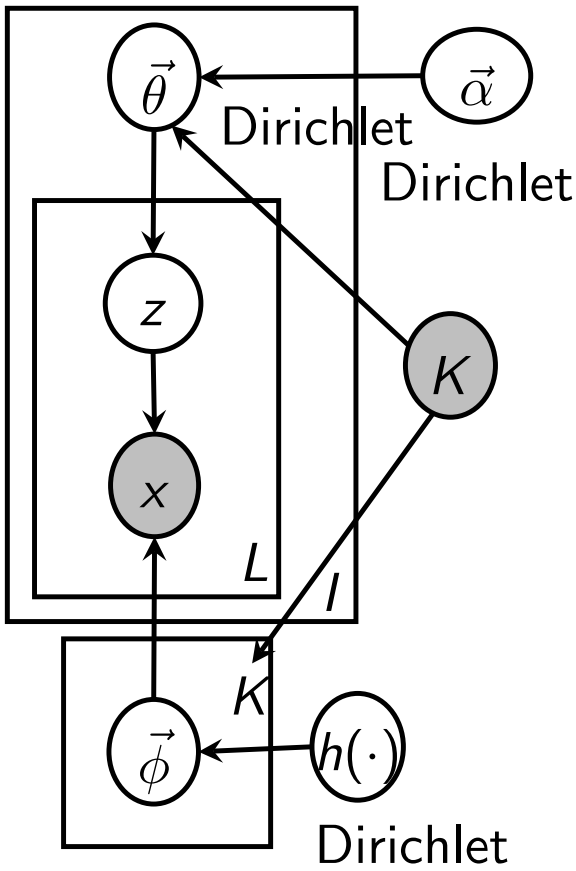
LDA (2008)



- shows the progress of LDA algorithms in the 00s
- 2008 version called asymmetric-symmetric LDA, in Mallet

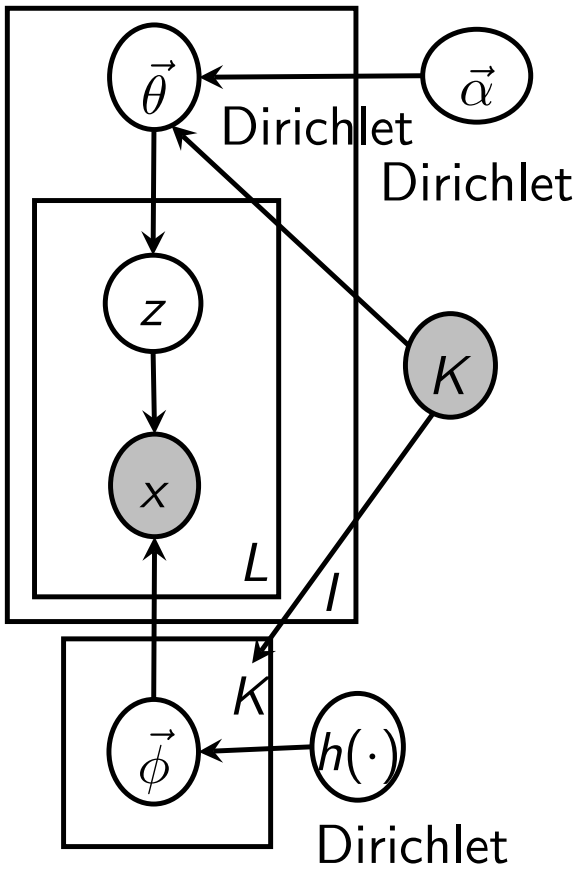
Representing Hierarchical Models Graphically

LDA (2008)

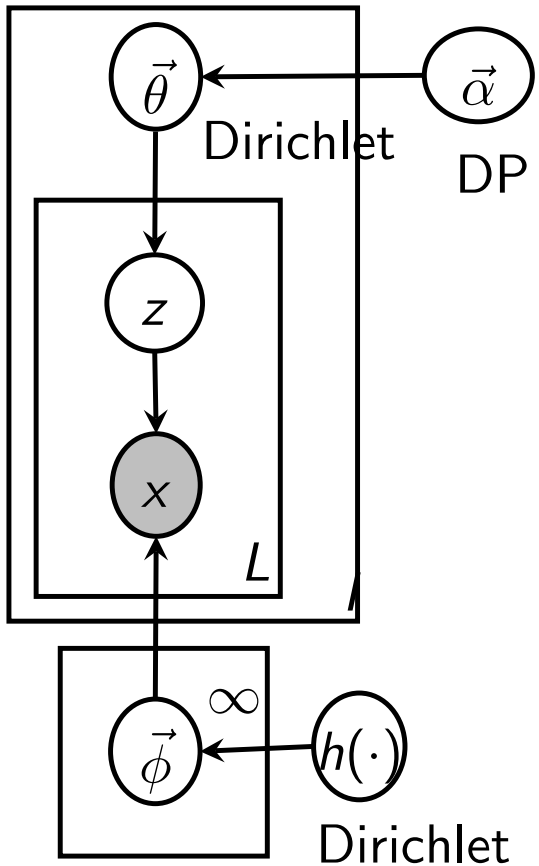


Representing Hierarchical Models Graphically

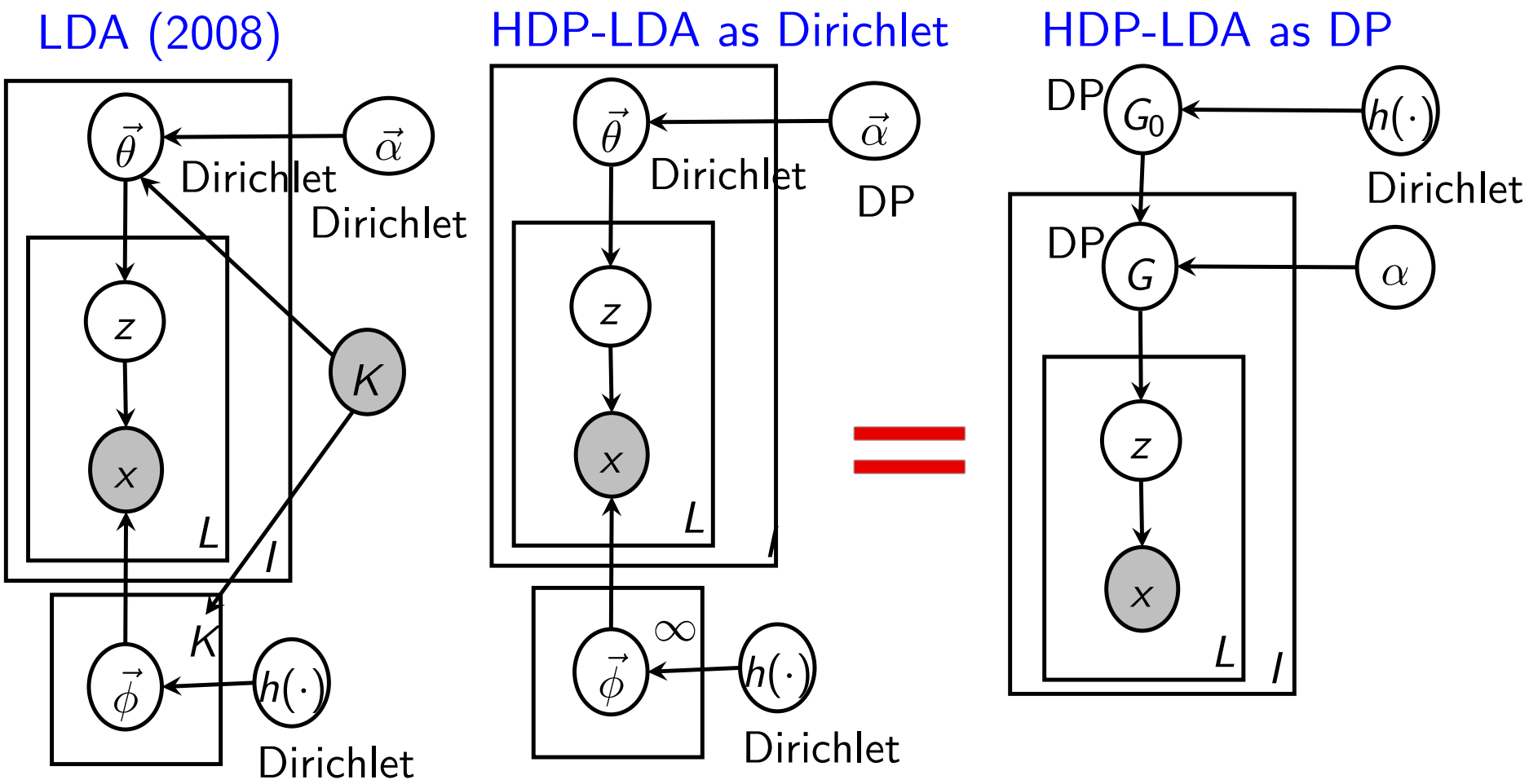
LDA (2008)



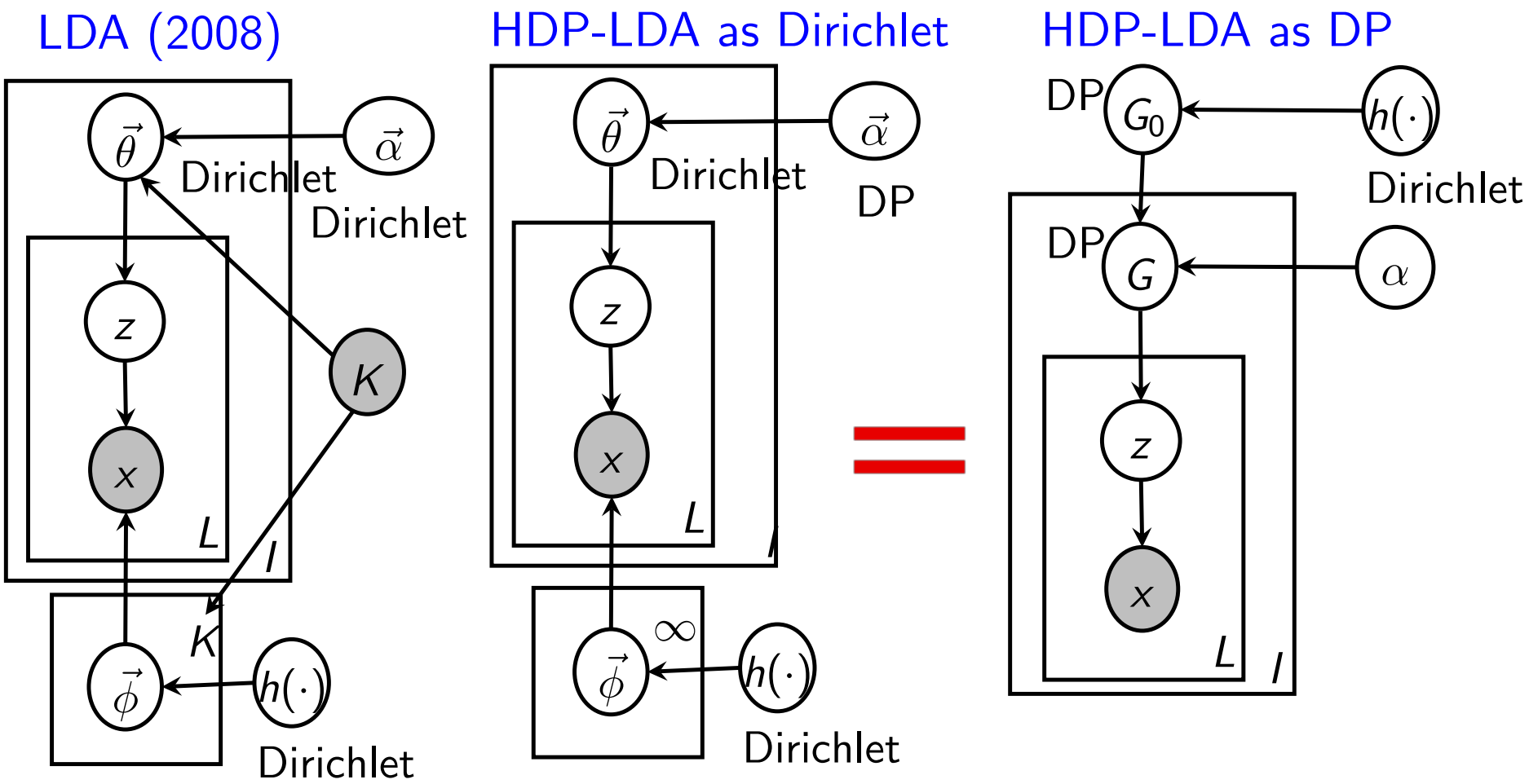
HDP-LDA as Dirichlet



Representing Hierarchical Models Graphically



Representing Hierarchical Models Graphically



- hierarchical part of a HDP *is just* a Dirichlet
- conceptually simpler, broader range of algorithms

Algorithms for HDP LDA

- Wang, Paisley and Jordan's
  Online Variational Inference for the Hierarchical Dirichlet Process,
  2011.
- Wang and Blei's
  Truncation-free online variational inference for Bayesian nonparametric m
  2012
- Bryant and Sudderth's
  Truly nonparametric online variational inference for hierarchical Dirichlet
  2012

Algorithms for HDP LDA

- Wang, Paisley and Jordan's
  Online Variational Inference for the Hierarchical Dirichlet Process,
  2011.
- Wang and Blei's
  Truncation-free online variational inference for Bayesian nonparametric m
  2012
- Bryant and Sudderth's
  Truly nonparametric online variational inference for hierarchical Dirichlet
  2012
- ...

Algorithms for HDP LDA

- Wang, Paisley and Jordan's
  Online Variational Inference for the Hierarchical Dirichlet Process,
  2011.
- Wang and Blei's
  Truncation-free online variational inference for Bayesian nonparametric n
  2012
- Bryant and Sudderth's
  Truly nonparametric online variational inference for hierarchical Dirichlet
  2012
- ...
- Wallach, Mimno and McCallum's
  Rethinking LDA: Why priors matter, 2009
  has a parallel truncated implementation of HDP-LDA that beats all-comers in speed and perplexity since 2008

Differing Models?

- **HDP-LDA**: Wang, Paisley and Jordan's *Online Variational Inference for the Hierarchical Dirichlet Process*, 2011.
- **HPF**: Gopalan, Ruiz, Ranganath and Blei's *Bayesian Nonparametric Poisson Factorization for Recommendation Syst* 2014
- virtually the same model (differ by a few independent gammas)
- related algorithms (variational Bayes on stick-breaking representations)

The Hierarchical Beta Process

Thibaux & Jordan, 2007 (276 citations)

Definition. A *beta process* $B \sim \mathrm{BP}(c, B_0)$ is a positive Lévy process whose Lévy measure depends on two parameters: c is a positive function over Ω that we call the *concentration function*, and B_0 is a fixed measure on Ω, called the *base measure*. In the special case where c is a constant it will be called the *concentration parameter*. We also call $\gamma = B_0(\Omega)$ the *mass parameter*.

If B_0 is continuous, the Lévy measure of the beta process is

$$\nu(d\omega, dp) \quad = \quad c(\omega)p^{-1}(1-p)^{c(\omega)-1}dpB_0(d\omega) \quad (1)$$

on $\Omega \times [0,1]$. As a function of p, it is a degenerate beta distribution, justifying the name. ν has the following elegant interpretation. To draw $B \sim \mathrm{BP}(c, B_0)$, draw a set of points $(\omega_i, p_i) \in \Omega \times [0,1]$ from a Poisson process with base measure ν (see Fig. 1), and let:

$$B = \sum_i p_i \delta_{\omega_i} \quad (2)$$

where δ_ω is a unit point mass (or *atom*) at ω. This implies $B(S) = \sum_{i:\omega_i \in S} p_i$ for all $S \subset \Omega$.

As this representation shows, B is discrete and the pairs (ω_i, p_i) correspond to the location $\omega_i \in \Omega$ and weight $p_i \in [0,1]$ of its atoms. Since $\nu(\Omega \times [0,1]) = \infty$ the Poisson process generates infinitely many points, making (2) a countably infinite sum. Nonetheless, as shown in the appendix, its expectation is finite if B_0 is finite.

If B_0 is discrete[2], of the form $B_0 = \sum_i q_i \delta_{\omega_i}$, then B has atoms at the same locations $B = \sum_i p_i \delta_{\omega_i}$ with

$$p_i \sim \mathrm{Beta}(c(\omega_i)q_i, c(\omega_i)(1-q_i)). \quad (3)$$

This requires $q_i \in [0,1]$. If B_0 is mixed discrete-continuous, B is the sum of the two independent contributions.

A tale of two propositions

"The two-parameter Poisson-Dirichlet distribution derived from a stable subo

Pitman & Yor, *Annals of Probability* **25**(2), 1997 (806 citations)

PROPOSITION 21. *Fix α with $0 < \alpha < 1$ and $C > 0$. Let $(\tau_s, s \geq 0)$ be a subordinator with Lévy measure $\alpha C x^{-\alpha-1} e^{-x}\, dx$. Independent of $(\tau_s, s \geq 0)$, let $(\gamma(t), t \geq 0)$ be a gamma subordinator as defined below (8). For $\theta > 0$ let*

$$S_{\alpha,\theta} = \frac{\gamma(\theta/\alpha)}{C\Gamma(1-\alpha)}. \tag{61}$$

Then for $T = \tau(S_{\alpha,\theta})$ the sequence (60) has PD(α, θ) distribution, independently of T, which has the same gamma(θ) distribution as $\gamma(\theta)$.

A tale of two propositions

"The two-parameter Poisson-Dirichlet distribution derived from a stable subo

Pitman & Yor, *Annals of Probability* **25**(2), 1997 (806 citations)

> PROPOSITION 21. *Fix α with $0 < \alpha < 1$ and $C > 0$. Let $(\tau_s, s \geq 0)$ be a subordinator with Lévy measure $\alpha C x^{-\alpha-1} e^{-x}\, dx$. Independent of $(\tau_s, s \geq 0)$, let $(\gamma(t), t \geq 0)$ be a gamma subordinator as defined below (8). For $\theta > 0$ let*
>
> (61) $$S_{\alpha,\theta} = \frac{\gamma(\theta/\alpha)}{C\Gamma(1-\alpha)}.$$
>
> *Then for $T = \tau(S_{\alpha,\theta})$ the sequence (60) has PD(α, θ) distribution, independently of T, which has the same gamma(θ) distribution as $\gamma(\theta)$.*

From "Theory of dependent hierarchical normalized random measures," by Chen, Buntine & Ding, *arXiv preprints*, arXiv:1205.4159, 2012.

> **Corollary 1** *Let $\vec{\mu} \sim NormalisedGeneralisedGamma(a, M, H(\cdot))$ and suppose $M \sim Gamma(b/a, 1)$ then it follows that $\vec{\mu} \sim Pitman\text{-}YorProcess(a, b, H(\cdot))$*

Bayesian Non-parametrics

- Some important material is inaccessible to the average machine learning researcher.

Bayesian Non-parametrics

- Some important material is inaccessible to the average machine learning researcher.
- Hierarchical stochastic processes are poorly presented:
  - hierarchical part of a HDP model is just a Dirichlet
  - good researchers know this ...
  - look at the Wikipedia on *hierarchical Dirichlet process*
  - What sort of distribution is the hierarchical part of a hierarchical Pitman-Yor process? ...

Bayesian Non-parametrics

- Some important material is inaccessible to the average machine learning researcher.
- Hierarchical stochastic processes are poorly presented:
  - hierarchical part of a HDP model is just a Dirichlet
  - good researchers know this ...
  - look at the Wikipedia on *hierarchical Dirichlet process*
  - What sort of distribution is the hierarchical part of a hierarchical Pitman-Yor process? ...
- Getting the "right" number of clusters/topics is not so significant when using HDP/HPYP;
  - they also learn the relative sizes of clusters/topics.

Bayesian Non-parametrics

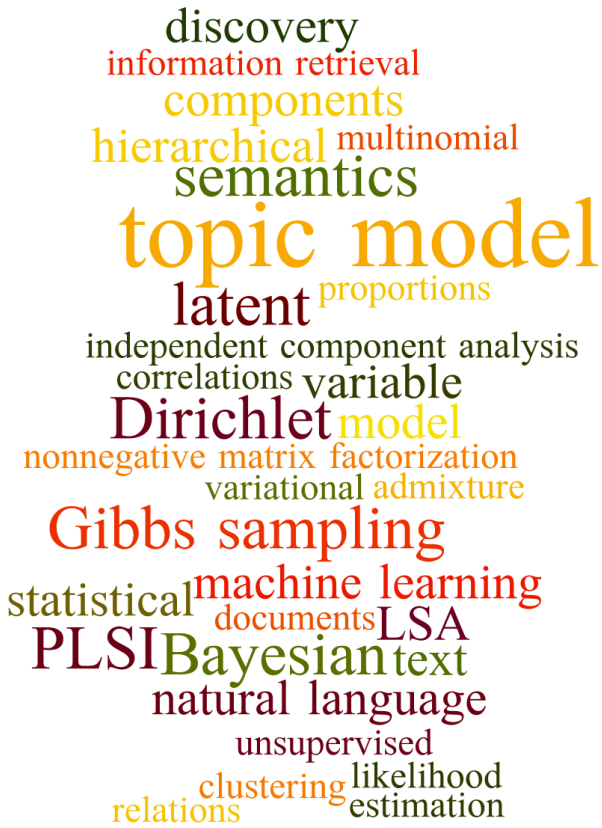
- Some important material is inaccessible to the average machine learning researcher.
- Hierarchical stochastic processes are poorly presented:
  - hierarchical part of a HDP model is just a Dirichlet
  - good researchers know this ...
  - look at the Wikipedia on *hierarchical Dirichlet process*
  - What sort of distribution is the hierarchical part of a hierarchical Pitman-Yor process? ...
- Getting the "right" number of clusters/topics is not so significant when using HDP/HPYP;
  - they also learn the relative sizes of clusters/topics.
- Equivalences and relationships between models not well understood.

Bayesian Non-parametrics

- Some important material is inaccessible to the average machine learning researcher.
- Hierarchical stochastic processes are poorly presented:
  - hierarchical part of a HDP model is just a Dirichlet
  - good researchers know this ...
  - look at the Wikipedia on *hierarchical Dirichlet process*
  - What sort of distribution is the hierarchical part of a hierarchical Pitman-Yor process? ...
- Getting the "right" number of clusters/topics is not so significant when using HDP/HPYP;
  - they also learn the relative sizes of clusters/topics.
- Equivalences and relationships between models not well understood.

Note: have been working on the source theory of Lancelot James, 2009, 2013, 2016.

Outline



Outline

Coverage of this Outline

- Demonstrate the theory for a vector of gamma parameters used to generate a matrix of Poisson data.

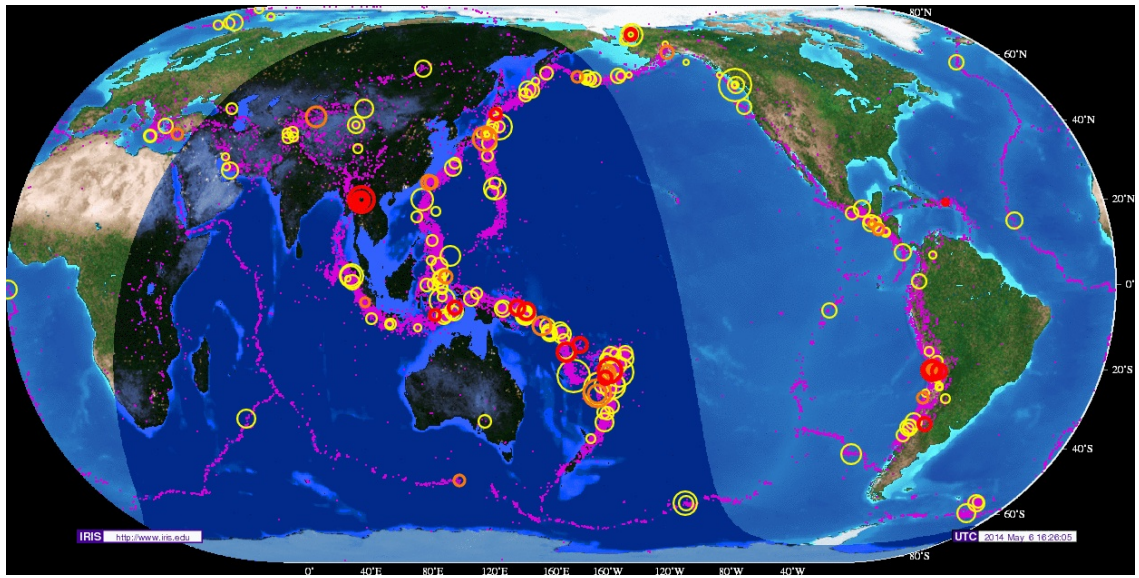
Coverage of this Outline

- Demonstrate the theory for a vector of gamma parameters used to generate a matrix of Poisson data.
- The framework presented works for all *the standard processes* we use and their normalised versions plus different *discrete data*:
  - gamma process, Dirichlet process, beta process, stable process, Pitman-Yor process, *etc.*
  - data being Bernoulli, Poisson, negative binomial, *etc.*
  - applied to LDA, Indian buffet processes, Poisson and negative gamma matrix factorisation.

Coverage of this Outline, cont.

- **This outline** gives broad overview, big picture, before we dive into detail.
- Intended to motivate the subsequent application of theory.
- Good introduction: Lijoi and Prünster, "Models beyond the Dirichlet process," 2010.

Refresh: Earthquakes in 2014



- The process part, points generated by a rate, denoted $\rho(\mathit{latitude}, \mathit{longitude})$:
  - number of earthquakes in 2014 per unit area, *i.e.* this rate is spatial
  - rate is not the same as a probability
  - can generate a countably infinite number of points too
- The marked part attaches auxiliary variables to the points in the process generated by a probability $p(\mathit{magnitude} \mid \mathit{latitude}, \mathit{longitude})$.

Our Stochastic Processes

We will build processes from the ground up using PPPs.

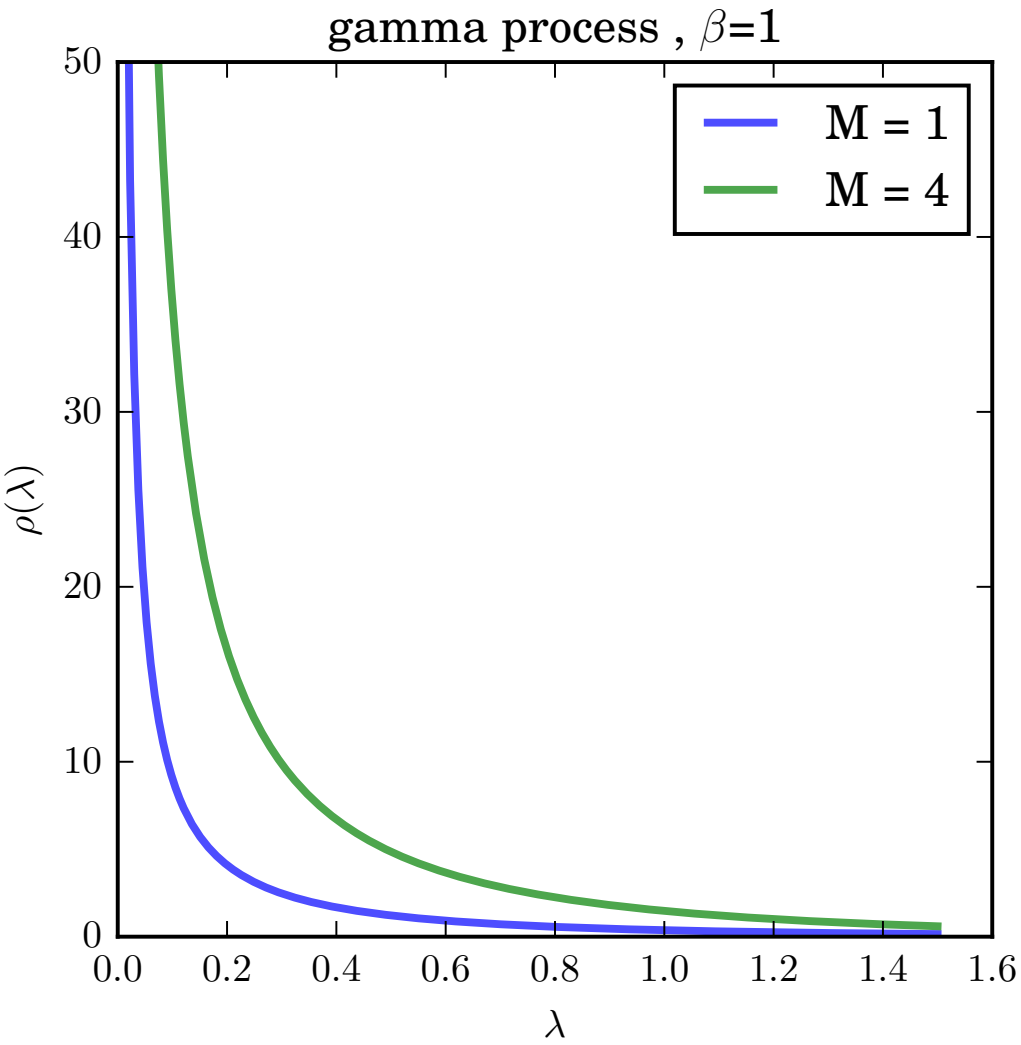
The Process Part: strengths/weights/probabilities for objects

- generated using Poisson processes
- generated in a 1-D space from $\mathcal{R}^+$
- usually generate an infinite supply, but mostly near zero and "unused"

The Marked Part: objects in the model we want to generate many of

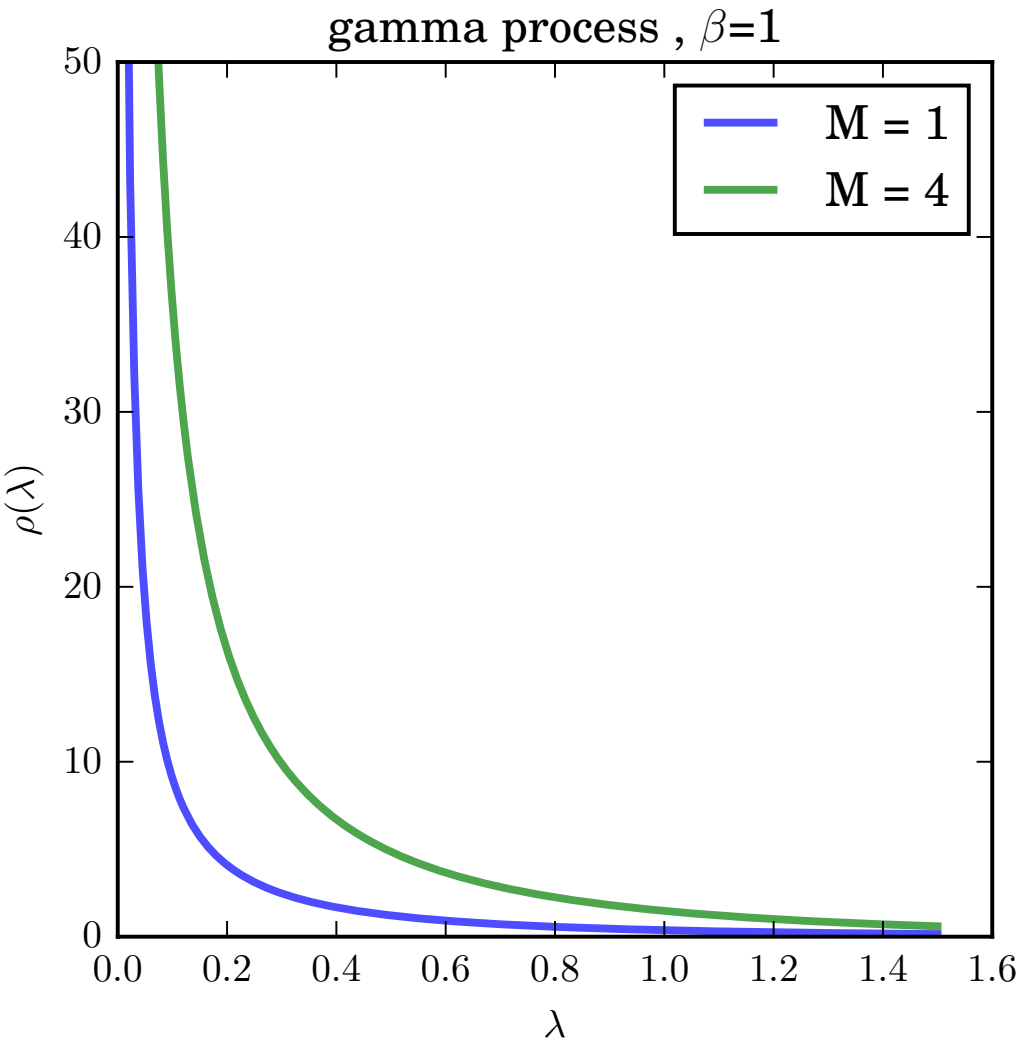
- multivariate Gaussians for a mixture model
- Dirichlet word vectors for LDA
- user rating vectors for a recommender system

Example Process Part: Gamma Process



- rate $\rho(\lambda) = M\lambda^{-1}e^{-\beta\lambda}$
- M is a general multiplier
- integral of the rate diverges to ∞ when including $\theta \to 0$

Example Process Part: Gamma Process



- rate $\rho(\lambda) = M\lambda^{-1}e^{-\beta\lambda}$
- M is a general multiplier
- integral of the rate diverges to ∞ when including $\theta \to 0$
- so the spatial process gets an infinite number of points near zero
  - but with infinitesimal λ, it most likely won't generate any data!

Alternative Versions of the Gamma Process

1. **Axiomatically** based on a (finite) measure $p(\cdot)$ on $\mathcal{X}$.
2. **Fully as a PPP** based on a rate $\rho(\lambda) = M\lambda^{-1}e^{-\beta\lambda}$ and a marking probability measure $p(\cdot)$ on $\mathcal{X}$.

Alternative Versions of the Gamma Process

1. **Axiomatically** based on a (finite) measure $p(\cdot)$ on $\mathcal{X}$.
2. **Fully as a PPP** based on a rate $\rho(\lambda) = M\lambda^{-1}e^{-\beta\lambda}$ and a marking probability measure $p(\cdot)$ on $\mathcal{X}$.

- How do we know these are the same?
  - This requires some true PPP theory to derive ... later.

Alternative Versions of the Gamma Process

1. **Axiomatically** based on a (finite) measure $p(\cdot)$ on $\mathcal{X}$.
2. **Fully as a PPP** based on a rate $\rho(\lambda) = M\lambda^{-1}e^{-\beta\lambda}$ and a marking probability measure $p(\cdot)$ on $\mathcal{X}$.

- How do we know these are the same?
  - This requires some true PPP theory to derive ... later.
- How do we do statistical/posterior inference on these, have been given some data?
  - This we will describe next.

Outline of Process Models

With respect to gamma processes, Dirichlet processes, beta processes, stable processes, Pitman-Yor processes, *etc.*:

Outline of Process Models

With respect to gamma processes, Dirichlet processes, beta processes, stable processes, Pitman-Yor processes, *etc.*:

- When using them at the root of a hierarchy, we're really using them to generate an infinite parameter vector:
  - do constructively using PPPs,
  - most are near zero and will generate zero data.

Outline of Process Models

With respect to gamma processes, Dirichlet processes, beta processes, stable processes, Pitman-Yor processes, *etc.*:

- When using them at the root of a hierarchy, we're really using them to generate an infinite parameter vector:
  - do constructively using PPPs,
  - most are near zero and will generate zero data.
- When using them below the root of a hierarchy, they are functionally different. They are really vector mapping operations using distributions:
  - using instead the associated additive distribution,

  e.g. to map $\vec{\mu}$ to $\vec{\lambda}$, for $k = 1, ..., K$ generate $\lambda_{i,k} \sim \text{gamma}(M\mu_k, \beta)$

Outline of Process Models

With respect to gamma processes, Dirichlet processes, beta processes, stable processes, Pitman-Yor processes, *etc.*:

- When using them at the root of a hierarchy, we're really using them to generate an infinite parameter vector:
  - do constructively using PPPs,
  - most are near zero and will generate zero data.
- When using them below the root of a hierarchy, they are functionally different. They are really vector mapping operations using distributions:
  - using instead the associated additive distribution,

  e.g. to map $\vec{\mu}$ to $\vec{\lambda}$, for $k = 1, ..., K$ generate $\lambda_{i,k} \sim \text{gamma}(M\mu_k, \beta)$
- The PPP interpretation and the axiomatic interpretation complement each other.

Hierarchical Pitman-Yor Process

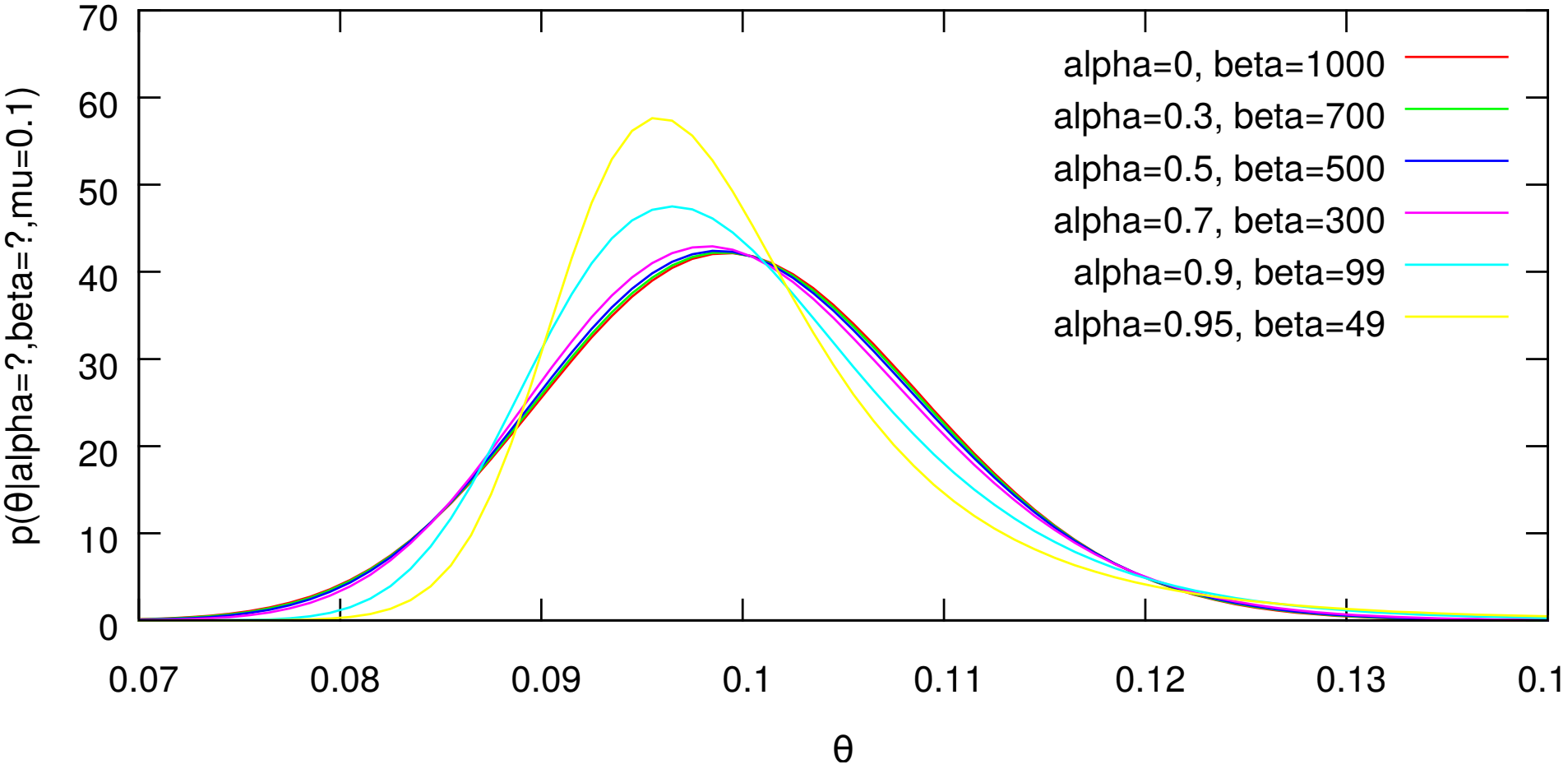
What distribution does the PYP look like when used below the root?

Integral Formula for the Hierarchical PYP

The K dimensional version on probability vector $\vec{\theta} = (\theta_1, ..., \theta_K)$ where $\vec{\theta} \sim PYP(\alpha, \beta, \vec{\mu})$ for $\alpha > 0$ and $\beta \geq 0$ and where $\dim(\vec{\mu}) = K$ introduces corresponding latent variables $\vec{\nu} = (\nu_1, ..., \nu_K)$ and has the form

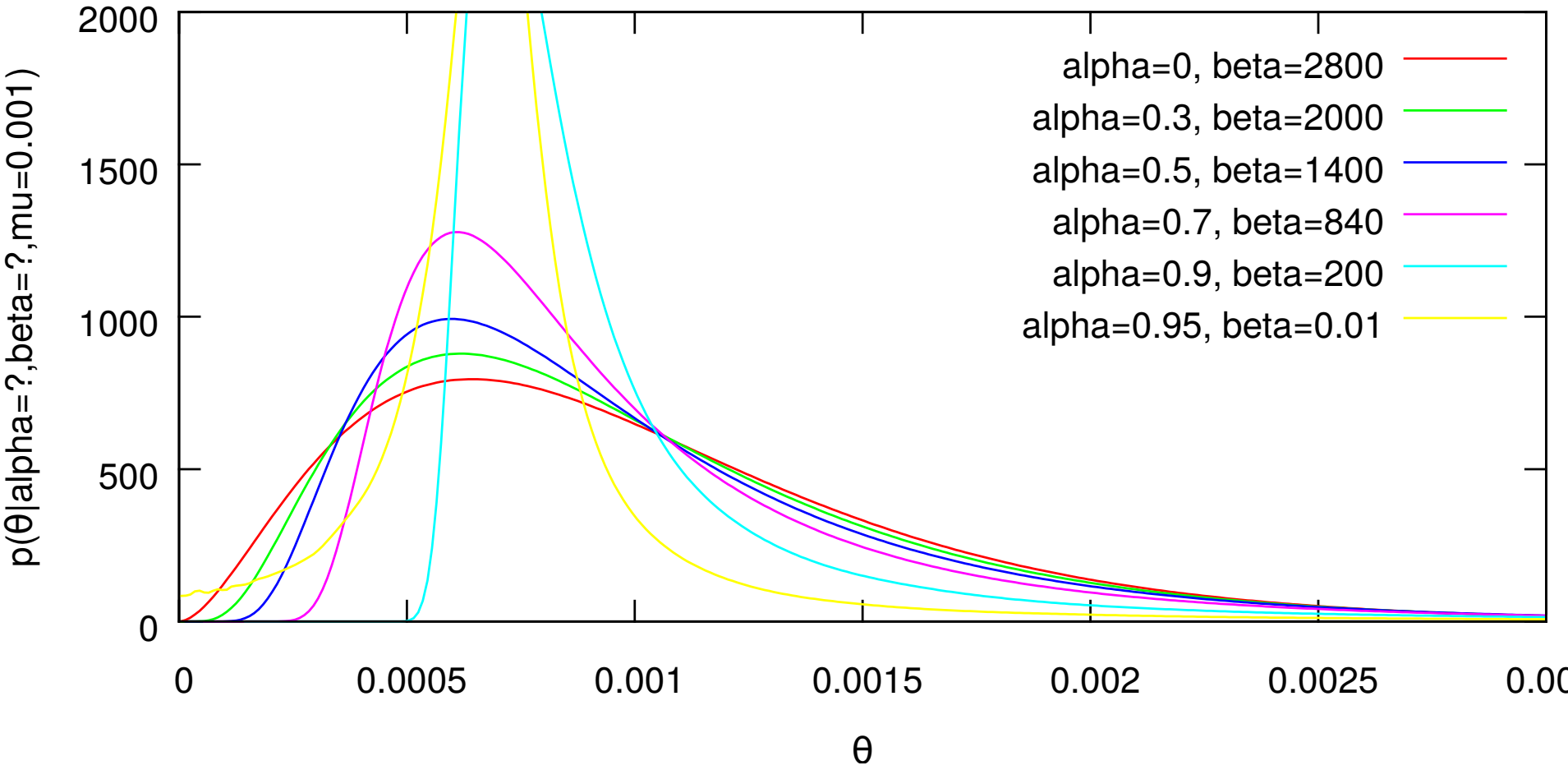
$$p(\vec{\theta}|\text{PYP}, \alpha, \beta, \vec{\mu}) = \frac{\alpha^{K-1}\Gamma(1+\beta)}{(1-\alpha)^{K-1}\pi^K\Gamma(1+\beta/\alpha)}\Gamma(K+\beta(1-\alpha)/\alpha)$$
$$\int_{\mathcal{R}^{+K}} \frac{\prod_{k=1}^K a_\alpha(\nu_k)\mu_k^{1/(1-\alpha)}\theta_k^{-1/(1-\alpha)}}{\left(\sum_{k=1}^K a_\alpha(\nu_k)\mu_k^{1/(1-\alpha)}\theta_k^{-\alpha/(1-\alpha)}\right)^{K+\beta(1-\alpha)/\alpha}} \mathrm{d}\,\vec{\nu}\ .$$

Hierarchical Pitman-Yor Process, cont.



(variations with location $\mu_1 = 0.1$ and identical variance)

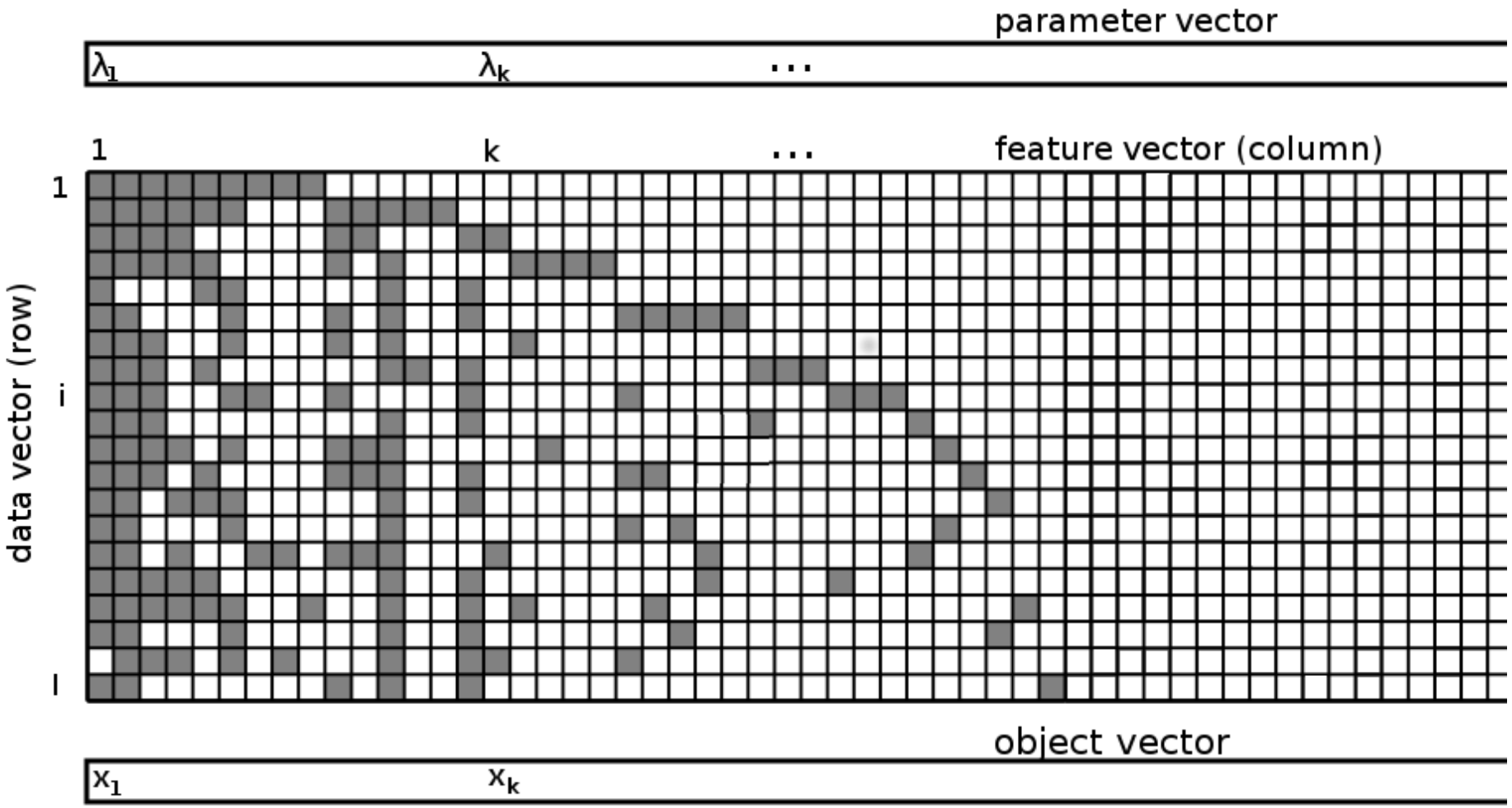
Hierarchical Pitman-Yor Process, cont.



(variations with location $\mu_1 = 0.001$ and identical variance)

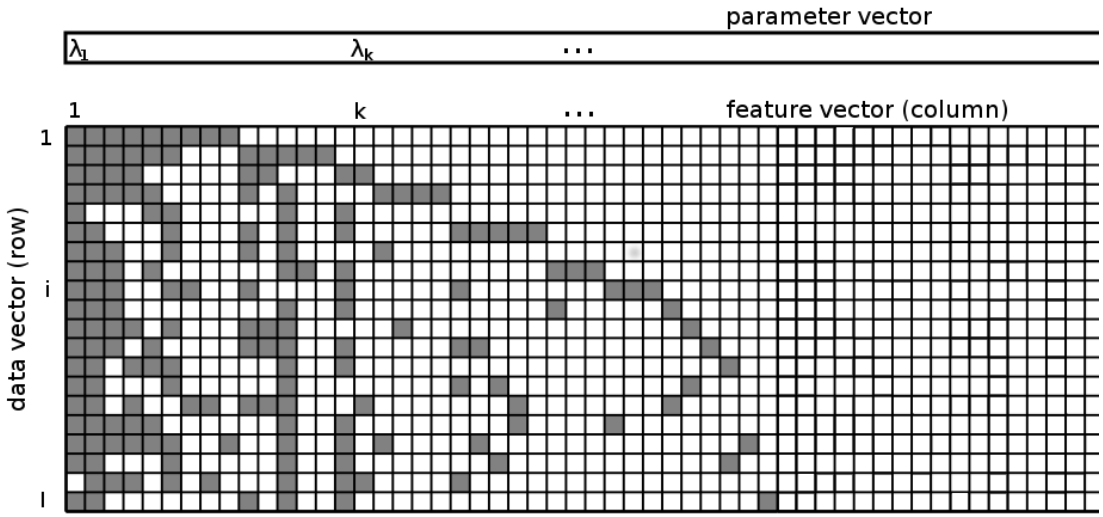
Outline

Discrete Feature Matrices



Worked Example: Gamma Poisson Model

- used widely for matrix factorisation
  - generalises LDA, originally due to Canny 2004
- easily extended to Boolean and robust versions
- fixed but unknown dimension of features K
- known dimension I of documents/images/rows



i.e. K is unobserved but must be greater or equal than that observed in the data

Gamma Poisson Model, cont.

1. generate **number of columns** $K \sim \text{Poisson}(\lambda)$
   - some data columns $\vec{n}_{\cdot,k}$ may be all zeros
2. generate **parameter vector** (dim K)

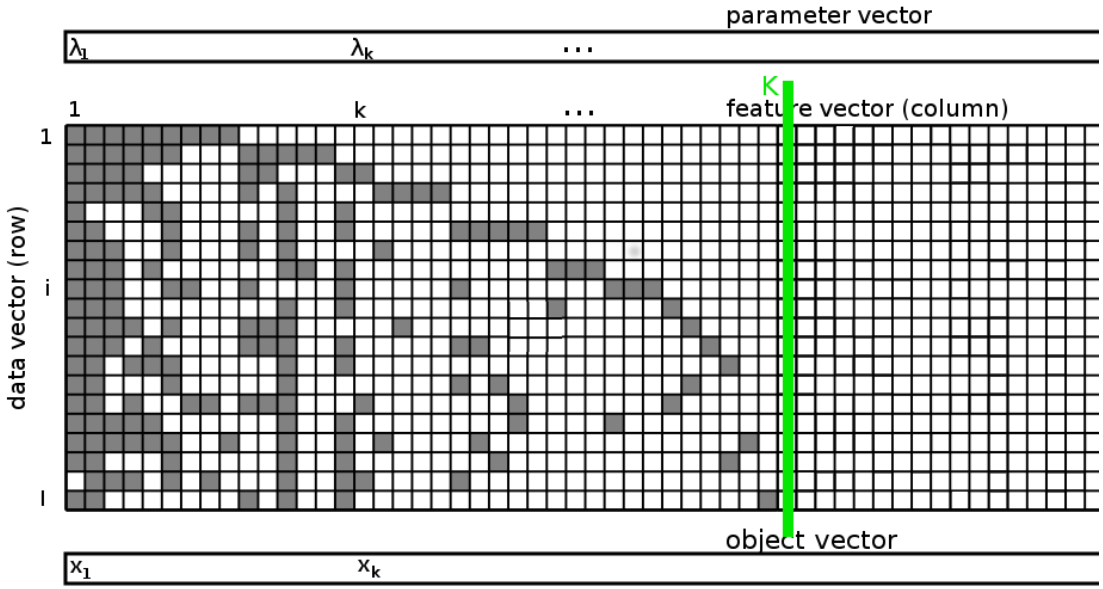
$$\lambda_k \sim \text{gamma}(\alpha, \beta) \qquad \text{for } k = 1, ..., K$$

3. generate **data matrix** (dim $I \times K$, one column per λ_k) as

$$n_{i,k} \sim \text{Poisson}(\lambda_k) \qquad \text{for } i = 1, ..., I$$

Gamma-process Poisson Model

- developed from previous Gamma-Poisson case
- natural extension, but focuses on non-zero columns (*i.e.*, those that are not all zeros)
- infinite dimension in columns but only first K columns have non-zeros in data matrix with I documents/images/rows



i.e. K is observed in the data

Gamma-process Poisson Model, cont.

1. generate **number of non-zero columns** $K \sim \text{Poisson}(\lambda_I)$
   - no data column $\vec{n}_{\cdot,k}$ is all zeros for all $k \leq K$
   - non-zero rate λ_I grows with I
2. generate **parameter vector** (dim K)

$$\lambda_k \sim \text{gamma-process}(M, \beta \,|\, \vec{n}_{\cdot,k} \neq \vec{0}) \qquad \text{for } k = 1, ..., K$$

3. generate **data matrix** (dim $I \times K$, one column per λ_k) as

$$n_{i,k} \sim \text{Poisson}(\lambda_k) \qquad \text{for } i = 1, ..., I$$

   - but constrain so $\vec{n}_{\cdot,k}$ is not all zero!

differences in orange!

Example Non-zero Rates

Name	Non-zero rate
generalised beta process with Bernoulli data	$M \sum_{i=0}^{I-1} \frac{\Gamma(\alpha+\beta+i)}{\Gamma(1+\beta+i)}$
generalised gamma process with Poisson data	$M\left((I+\beta)^{\alpha}-\beta^{\alpha}\right)$
gamma process with negative binomial data	$M\left(\log(I \log \frac{1}{1-\rho}+\beta)-\log\beta\right)$

- illustrates different processes, different kinds of data
- details later, or in Wray's report

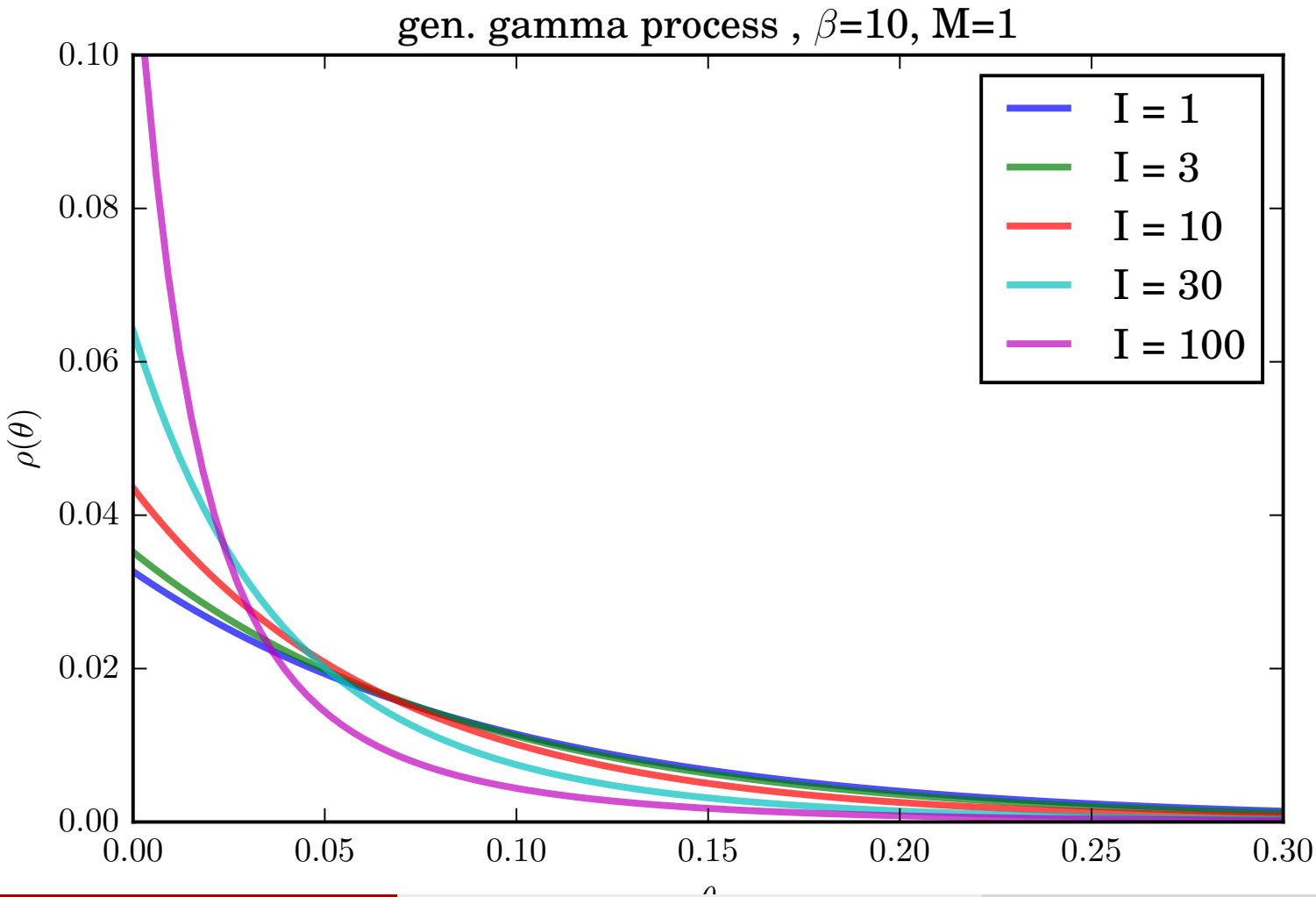
Non-zero Rates, cont.

Generally, non-zero rate is $O(\log I)$ or $O(I^{\alpha})$ (for $0 < \alpha < 1$), where I is the number of data rows.

i.e. number of non-zero features grows logarithmically or sub-linearly with number of documents/images/datum

Parameters Given Non-zero Data

want parameter vector $\vec{\lambda}$ using $\lambda_k \sim$ gamma-process$(M, \beta \mid \vec{n}_{\cdot,k} \neq \vec{0})$, where the number of data rows is I



Outline of Discrete Feature Matrices

- Generate the number of non-zero features using a "non-zero rate" dependent on the count of rows.

  e.g. with more rows, expect to see more non-zero features

- Then generate the parameters associated with non-zero features.

  e.g. given the column is non-zero, the unnormalisable rate is converted to a posterior probability.

Outline of Discrete Feature Matrices

- Generate the number of non-zero features using a "non-zero rate" dependent on the count of rows.

  e.g. with more rows, expect to see more non-zero features
- Then generate the parameters associated with non-zero features.

  e.g. given the column is non-zero, the unnormalisable rate is converted to a posterior probability.
- Uses the framework of Poisson point process:
  - theory is truly elegant and moderately simple

Outline of Discrete Feature Matrices

- Generate the number of non-zero features using a "non-zero rate" dependent on the count of rows.
  e.g. with more rows, expect to see more non-zero features
- Then generate the parameters associated with non-zero features.
  e.g. given the column is non-zero, the unnormalisable rate is converted to a posterior probability.
- Uses the framework of Poisson point process:
  - theory is truly elegant and moderately simple
- Non-zero rate λ_f and other parts of model derived from PPP theory:
  - cannot use any old formula for consistency.

Outline of Discrete Feature Matrices

- Generate the number of non-zero features using a "non-zero rate" dependent on the count of rows.
  e.g. with more rows, expect to see more non-zero features
- Then generate the parameters associated with non-zero features.
  e.g. given the column is non-zero, the unnormalisable rate is converted to a posterior probability.
- Uses the framework of Poisson point process:
  - theory is truly elegant and moderately simple
- Non-zero rate λ_l and other parts of model derived from PPP theory:
  - cannot use any old formula for consistency.
- General theory developed in James' *"Bayesian Poisson Calculus for Latent Feature Modeling* via Generalized Indian Buffet Process Priors" 2016

Outline

Theoretical Framwork

- Most of Bayesian non-parametric theory in this area is dressed up in the framework of
  - completely random measures$^{\mathcal{W}}$ (CRMs), and
  - normalised random measures with independent increments (NRMIs)
- For understanding of basic results and effects, we don't need this extra complexity.
- The ideas are not that complex, but this is a short tutorial!

Theoretical Framwork

- Most of Bayesian non-parametric theory in this area is dressed up in the framework of
  - completely random measures$^{\mathcal{W}}$ (CRMs), and
  - normalised random measures with independent increments (NRMIs)
- For understanding of basic results and effects, we don't need this extra complexity.
- The ideas are not that complex, but this is a short tutorial!
- These are reviewed next, ... but we will skip this now.

Building Infinite Vectors – Motivation

- Rather than having a vector indexed by integers $1, 2, ...$, we index it by a countable number of points from a domain $\mathcal{X}$, $\{x_1, x_2, ...\}$.
- So vector $\vec{\mu}$ corresponds to function $\mu(\cdot)$ defined as

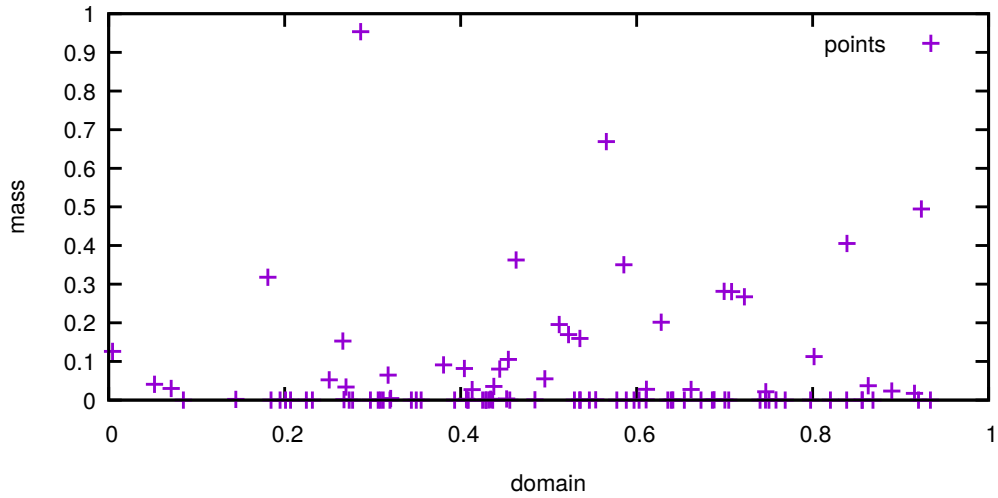
$$\mu(x) = \sum_{k=1}^{\infty} w_k \delta_{x_k}(x) \ .$$

- When all the points x_k are distinct, this behaves like an infinite vector.
- The form $\mu(\cdot)$ is a kind of completely random measure$^{\mathcal{W}}$ (CRM).
- We can sample random $\mu(\cdot)$'s using a PPP construct.
- If we also normalise $\mu(\cdot)$, it is an *infinite probability vector*:

$$\tilde{\mu}(x) = \frac{\sum_{k=1}^{\infty} w_k \delta_{x_k}(x)}{\sum_{k=1}^{\infty} w_k}$$

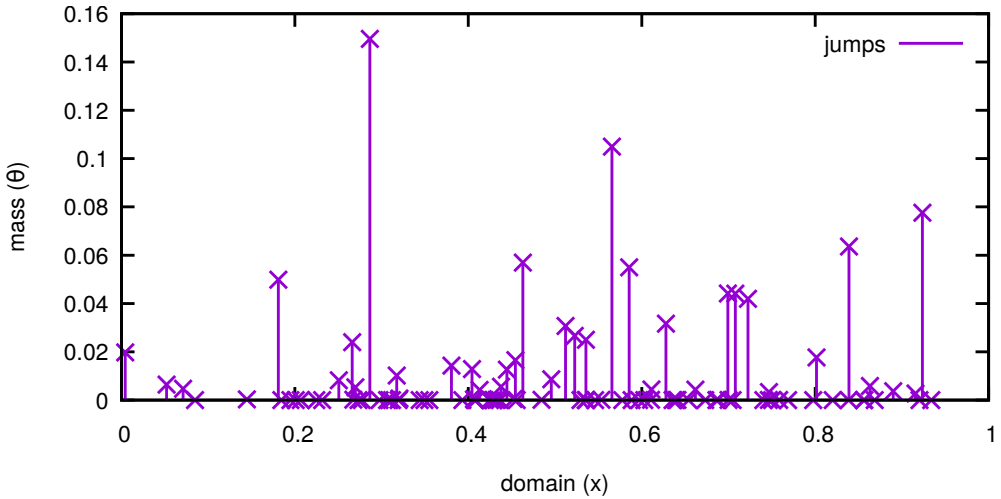
called a normalised random measure with independent increments (NRMI).

Discrete Functions from a PPP



- have a PPP on $\mathcal{X} \times \mathcal{R}^+$
- interpret each point (x, t) as (domain,mass) pair
- sample set of points $N \sim$ PPP, as shown
- usually the rate factors: $\rho(x, w) = \rho_1(w)\rho_2(x)$, said to be homogeneous
- when $\rho_2(x)$ is a PDF, call it the base distribution

Discrete Functions from a PPP, cont.



- sample $N \sim$ PPP as before
- define a function on $\mathcal{X}$ by $\mu(x) = \sum_{(w,x') \in N} w \delta_{x'}(x)$
- the function, now shown on the figure, must be discrete
- want total $\mu(\mathcal{X}) = \sum_{(w,x') \in N} w < \infty$
- if there are an infinite number of points, they must be close to the X axis (mass $w \to 0$) to get $\mu(\mathcal{X}) < \infty$

Completely Random Measure – Definitions

Definition (roughly) of a Measure

A measure[W] on domain $\mathcal{X}$ is a non-negative additive function $m(\cdot)$ on (certain well behaved) subsets of X so that $m(\emptyset) = 0$.

A measure represents the intuitive notion of size.

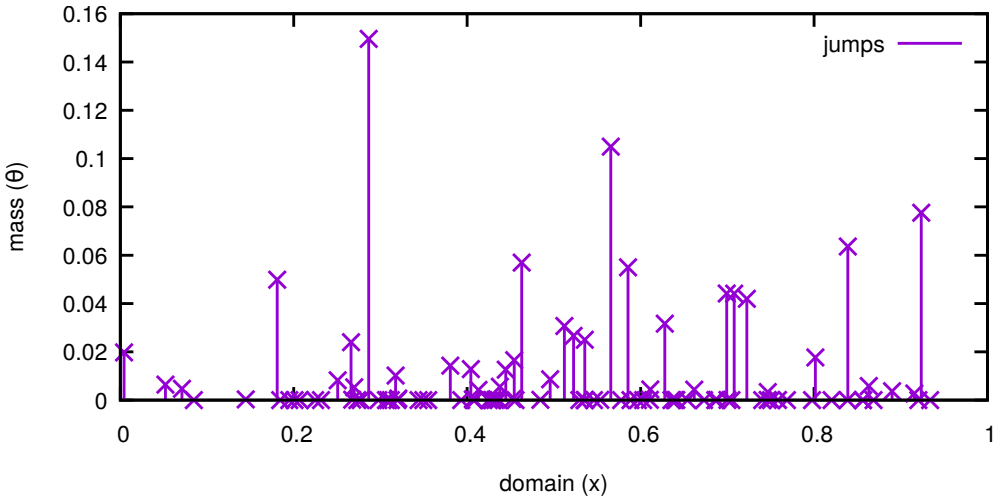
Definition (roughly) of a Random Measure

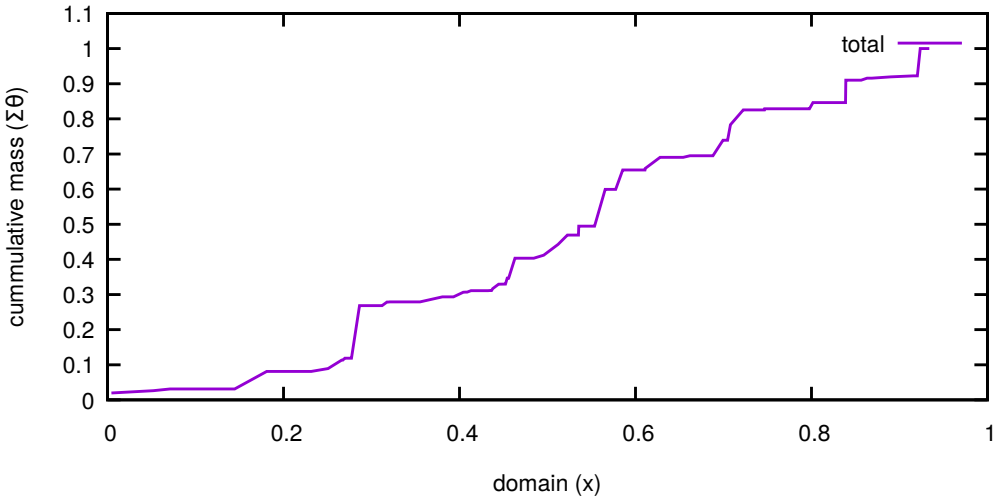
A random measure[W] is a measure that is a random variable.

Definition (roughly) of a Completely Random Measure

A completely random measure[W] (CRM) is a random measure such that the values it takes on disjoint sets are independent: if $A \cap B = \emptyset$ then $N(A) \perp\!\!\!\perp N(B)$.

Building Infinite Probability Vectors from a PPP or CRM





- form the CRM as before
- now rescale the CRM so it sums to 1
- this forms an NRMI (defined next)

$$\mu(x) = \frac{\sum_{(w,x')\in N} w\delta_{x'}(x)}{\sum_{(w,x')\in N} w}$$

- must have $\sum_{(w,x')\in N} w < \infty$

Normalised Random Measure with Independent Increments – Definition

Definition (roughly) of a Normalised Random Measure with Independent Increments

A normalised random measure with independent increments (NRMI) is formed by normalising a (pure jump) CRM that has finite total mass.

- have a PPP on $\mathcal{X} \times \mathcal{R}^+$; sample $N \sim$ PPP
- homogeneous case $\rho(x, w) = \rho(w)h(x)$ for PDF $h(x)$ called the base distribution
- let the points (w_k, x_k) for $k = 1, \ldots$ be some ordering of points in N
- define the NRMI

$$\tilde{\mu}(x) = \frac{\sum_{k=1}^{\infty} w_k \delta_{x_k}(x)}{\sum_{k=1}^{\infty} w_k}$$

- represents an infinite probability vector

Outline

Poisson Process Examples – Motivation

- Main models used are:
  - beta processes,
  - gamma processes, and
  - stable processes
- They are used because they can be used.
- We review then because they are variations of the well known beta and gamma distributions.

Beta Process Example

- The beta distribution used to generate probabilities.
- The beta distribution, $\theta \sim \text{beta}(\alpha, \beta)$ (for $\alpha, \beta > 0$)

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Beta Process Example

- The beta distribution used to generate probabilities.
- The beta distribution, $\theta \sim \text{beta}(\alpha, \beta)$ (for $\alpha, \beta > 0$)

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The generalised (or 3-parameter) beta process,
  $\vec{\theta} \sim \text{beta-proc}(M, \alpha, \beta)$
  (for $0 \leq \alpha < 1$, $\beta > 0$) and $0 \leq \theta < 1$

$$\rho(\theta) = \frac{M}{\Gamma(1-\alpha)}\theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-1}$$

- Rather like a scaled "improper" beta$(-\alpha, \alpha+\beta)$ distribution.
- $\alpha = 0$ most commonly seen, *e.g.* original Indian buffet process

Gamma Process Example

- The gamma distribution used to generate rates for a Poisson.
- The gamma distribution $\lambda \sim \text{gamma}(\alpha, \beta)$ (for $\alpha, \beta > 0$) and $\lambda > 0$

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Gamma Process Example

- The gamma distribution used to generate rates for a Poisson.
- The gamma distribution $\lambda \sim \text{gamma}(\alpha, \beta)$ (for $\alpha, \beta > 0$) and $\lambda > 0$

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- The generalised (or 3-parameter) gamma process, $\vec{\lambda} \sim \text{gen-gamma-proc}(M, \alpha, \beta)$ (for $0 \leq \alpha < 1$, $\beta > 0$)

$$\rho(\lambda) = M \frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)} \lambda^{-\alpha-1} e^{-\lambda\beta}$$

- Rather like a scaled "improper" gamma$(-\alpha, \beta)$ distribution.
- $\alpha = 0$ most commonly seen, called gamma process[W]

Example Rates for Mass

Name	Parameters	Rate
$\theta \sim$ beta-proc(M, α, β)	$0 \leq \alpha < 1$	$\frac{M}{\Gamma(1-\alpha)}\theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-}$
$\lambda \sim$ gen-gamma-proc(M, α, β)	$0 \leq \alpha < 1$	$M\frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}e^{-\lambda\beta}$
$\lambda \sim$ stable-proc(M, α)	$0 < \alpha < 1$	$\frac{M\alpha}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}$

for $M > 0$ a constant rate, $\beta > 0$ and variables $0 < \theta < 1$, $\lambda > 0$.

- the stable process is an improper gamma$(-\alpha, 0)$ distribution

Beta Process, cont.

- The beta process cannot be normalised

$$\int_0^1 \rho(\theta) \mathrm{d}\,\theta = M \int_0^1 \theta^{-1}(1-\theta)^{\alpha+\beta-1} \mathrm{d}\,\theta = \infty$$

Beta Process, cont.

- The beta process cannot be normalised

$$\int_0^1 \rho(\theta)\mathrm{d}\,\theta = M \int_0^1 \theta^{-1}(1-\theta)^{\alpha+\beta-1}\mathrm{d}\,\theta = \infty$$

- (Warning: this is an informal argument.)
  With one positive sample, likelihood θ, the joint total rate becomes

$$\begin{aligned}\int_0^1 \theta\rho(\theta)\mathrm{d}\,\theta &= M\mathrm{beta}(1,\beta)\\ &= \frac{M}{\beta}\end{aligned}$$

  so posterior inference should be possible
- Similarly for any sample with at least one positive.

Gamma Process, cont.

- The gamma process with $\alpha = 0$ cannot be normalised,

$$\int_0^\infty \rho(\theta) \mathrm{d}\,\theta = M \int_0^\infty \lambda^{-1} e^{-\lambda\beta} \mathrm{d}\,\lambda = \infty$$

Gamma Process, cont.

- The gamma process with $\alpha = 0$ cannot be normalised,

$$\int_0^\infty \rho(\theta) \mathrm{d}\,\theta = M \int_0^\infty \lambda^{-1} e^{-\lambda\beta} \mathrm{d}\,\lambda = \infty$$

- (Warning: this is an informal argument.)
  With a non-zero sample, likelihood $(1 - e^{-\lambda})$, the joint total rate becomes

$$\int_0^\infty (1 - e^{-\lambda}) \rho(\lambda) \mathrm{d}\,\lambda \quad = \quad M \left(\log(1 + \beta) - \log \beta\right)$$

  so posterior inference should be possible

Things We Do with a Process

Consider the generalised (or 3-parameter) gamma process,
$\lambda \sim$ gen-gamma-proc(M, α, β) (for $0 \leq \alpha < 1$, $\beta > 0$)

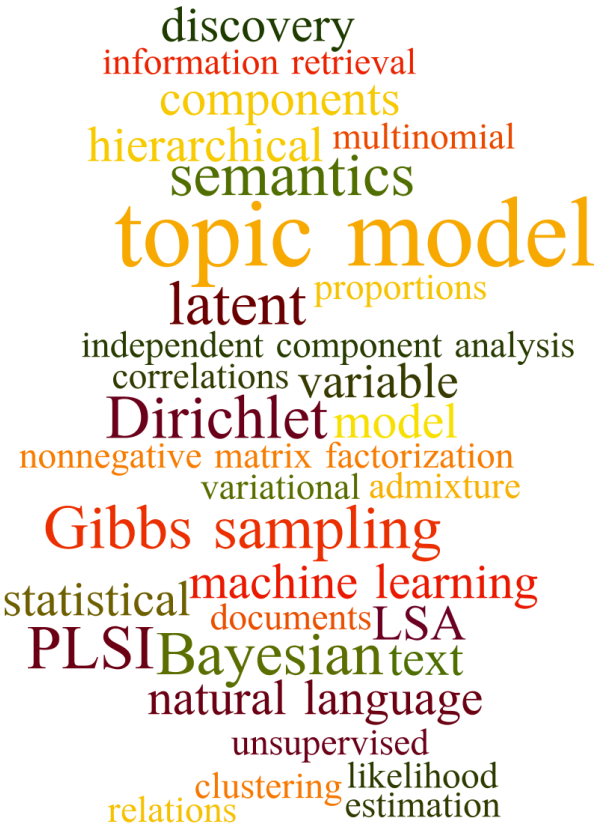
$$M \frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)} \lambda^{-\alpha-1} e^{-\lambda\beta}$$

- We can sample a typical vector. Be careful, almost surely any entry is arbitrarily close to zero!
- We can normalise it and generate an infinite probability vector:

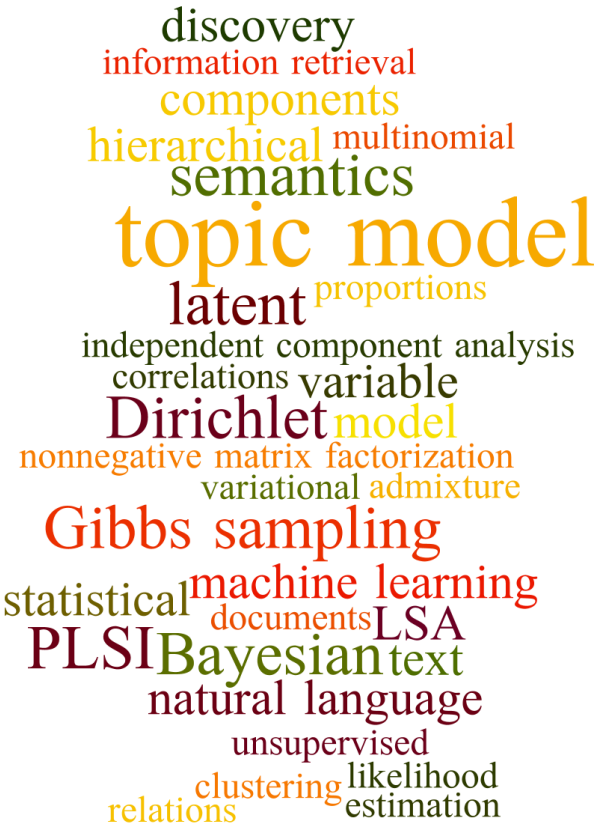
$$\frac{1}{\sum_{k=1}^{\infty} \lambda_k} \vec{\lambda}$$

- We can do inference about it based on a matrix of data.

Outline



Outline



Distribution of the Total Mass

- An important quantity for a Poisson process with rate $\rho(\lambda)$ is the total mass $T = \sum_{k=1}^{\infty} \lambda_k$.
- This completes the link between axiomatic definitions of processes from additive distributions to the PPP definition of processes.

Distribution of the Total Mass

- An important quantity for a Poisson process with rate $\rho(\lambda)$ is the total mass $T = \sum_{k=1}^{\infty} \lambda_k$.
- This completes the link between axiomatic definitions of processes from additive distributions to the PPP definition of processes.
- Poisson process theory shows how to derive this.
- We cannot derive the total mass distribution directly, but we can get its moment-generating function$^{\mathcal{W}}$, sometimes re-expressed as a cumulant-generating function by $\mathrm{CGF}(t) = \log \mathrm{MGF}(t)$.
- For examples of distributions (not proocesses) see Moment-generating function$^{\mathcal{W}}$ (*MGF*).

Theory – Total Mass Distribution

- In Lévy process[W] theory, the MGF is developed using the Laplace functional[W] got via the Lévy-Khintchine formula.
  - quite simple to argue if you don't need to be formal
- Express as by $\text{CGF}(t) = \log \text{MGF}(t)$.

CGF for the total mass

For the PPP with rate $\rho(\lambda)$ for $\lambda \in \mathcal{R}^+$, let $T = \sum_{k=1}^{\infty} \lambda_k$ be the total mass for a sample $\vec{\lambda}$. Then the CGF of T, if it exists, is given by

$$\text{CGF}(t) = -\int_{\mathcal{R}^+} \left(1 - e^{t\lambda}\right) \rho(\lambda) \mathrm{d}\,\lambda \ .$$

- Requires some tricky integration to evaluate the integral.

Example Total Mass Distributions[1]

rate for $\vec{\lambda}$	distribution of total $\sum_k \lambda_k$
beta-proc$(M, 0, 1)$	Dickman(M)
gamma-proc$(M, 0, \beta)$	gamma(M, β)
gen-gamma-proc(M, α, β)	Tweedie$(\alpha, M^{1/\alpha}, \beta)$
stable-proc(M, α)	positive-stable$(\alpha, M^{1/\alpha})$

- all these distributions are additive (or infinitely divisible) in the parameter M
- moments (mathematics)$^{\mathcal{W}}$ of the Poisson rate turn out to be cumulants$^{\mathcal{W}}$ of the total distribution

[1] Lots of unexplained definitions!

Normalised Processes

- A useful variant of a Poisson process with rate $\rho(\lambda)$ is the normalised vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k} \vec{\lambda}$.
- Not all variants are meaningful:
  - the beta process already generates probabilities (that do not together sum to one)
  - no literature on normalising it.

Example Normalised Processes

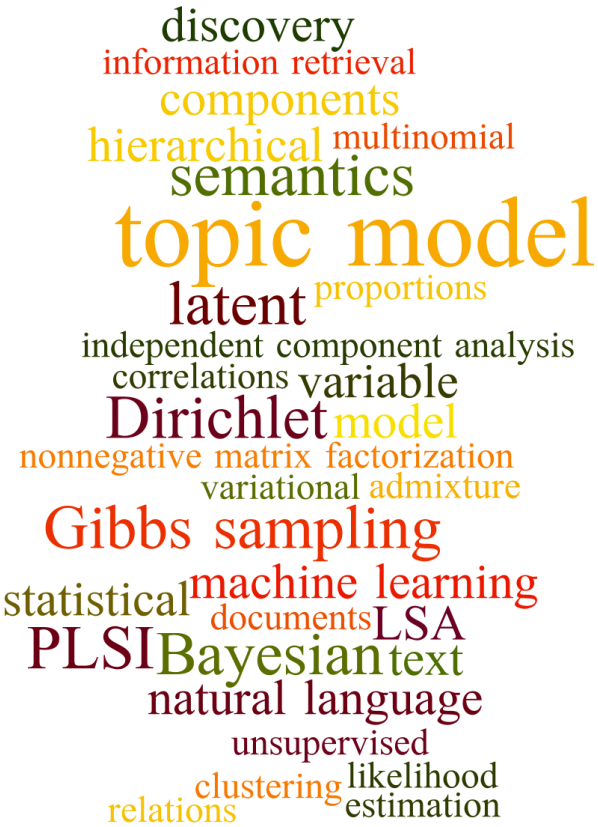
Mass Rate	Normalised version
beta-proc(M, α, β)	not in general use
gamma-proc$(M, 0, \beta)$	Dirichlet process with concentration M, DP(M)
gen-gamma-proc(M, α, β)	normalised generalised gamma process with concentration M and discount α, NGG(α, M)
stable-proc(M, α)	Pitman-Yor process with discount α and concentration 0

Example Normalised Processes

Mass Rate	Normalised version
beta-proc(M, α, β)	not in general use
gamma-proc($M, 0, \beta$)	Dirichlet process with concentration M, DP(M)
gen-gamma-proc(M, α, β)	normalised generalised gamma process with concentration M and discount α, NGG(α, M)
stable-proc(M, α)	Pitman-Yor process with discount α and concentration 0

What about the Pitman-Yor process?

Outline



Theory — Model Equivalences

Let $\vec{\lambda}$ denote an infinite vector in $\mathcal{R}^\infty$, and Λ be its total, $\Lambda = \sum_{k=1}^{\infty} \lambda_k$.

$$\vec{\lambda} \sim \text{gamma-process}(M, \beta)$$
is equivalent to
$$\Lambda \sim \text{gamma}(M, \beta) \text{ and independently } \frac{1}{\Lambda}\vec{\lambda} \sim \text{Dirichlet-process}(M)$$

$$\vec{\lambda} \sim \text{gen-gamma-process}(M, \alpha, \delta) \text{ and } M \sim \text{gamma}(\beta/\alpha, \delta^\alpha)$$
is equivalent to
$$\Lambda \sim \text{gamma}(\beta, \delta) \text{ and independently } \frac{1}{\Lambda}\vec{\lambda} \sim \text{Pitman-Yor-process}(\alpha, \beta)$$

$$\vec{\lambda} \sim \text{stable-process}(M, \alpha)$$
is equivalent to
$$\Lambda \sim \text{pstable}(\alpha, M^{1/\alpha}) \text{ and independently } \vec{\lambda}/\Lambda \sim \text{PYP}(\alpha, 0)$$

Model Equivalences, cont.

- Hierarchical Poisson matrix factorisation (HPF)
  - Gopalan, Ruiz, Ranganath and Blei's *"Bayesian Nonpara. Poisson Factorization for Rec. Sys."* 2014 uses a gamma process (though they don't say it)

is equivalent to
HDP-LDA (2006, 2011, ...) plus an independent gamma intensity component

- Robust (negative binomial) Poisson factorisation
  - Zhou, Hannah, Dunson and Carin *"Beta-Negative Binomial Process and Poisson Factor Analysis"*, 2012

is equivalent to
Bursty HDP-LDA
  - Buntine and Mishra *"Experiments with non-parametric topic models"* 2014

plus an independent gamma intensity component

Pitman-Yor Process

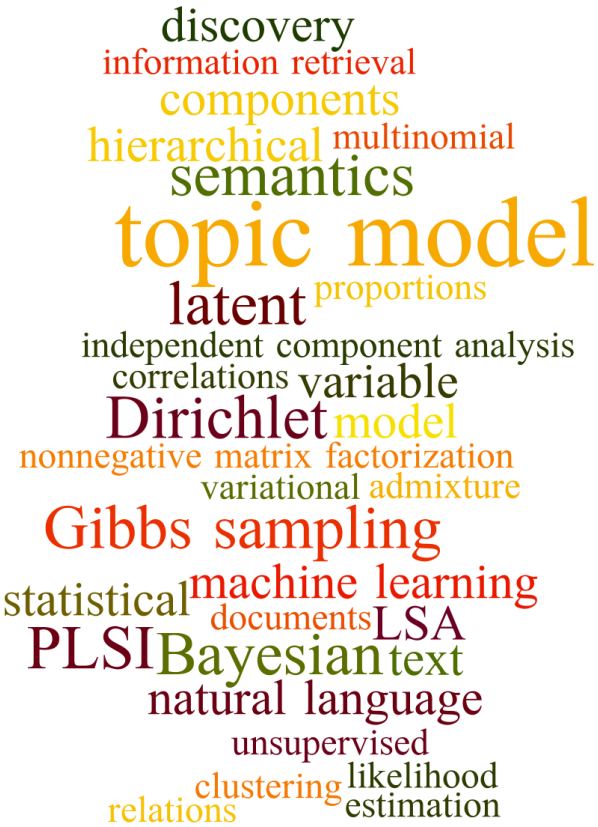
$\vec{\lambda} \sim$ gen-gamma-process(M, α, δ) and $M \sim$ gamma$(\beta/\alpha, \delta^\alpha)$

is equivalent to

$\Lambda \sim$ gamma(β, δ) and independently $\frac{1}{\Lambda}\vec{\lambda} \sim$ Pitman-Yor-process(α, β)

- We see it can be derived by normalising a generalised gamma process and marginalising out the concentration.
- It also shares many convenient properties with the Dirichlet process: stick breaking definition, *etc.*
- Pitman developed an alternative formulation that is related, the Poisson-Kingman model.
  - in Pitman, "Poisson-Kingman partitions," 2003,
  - some complexity for the BNP Initiates!

Outline



Sampling Vectors

- Assume we have an infinite vector from a PPP with rate $\rho(\lambda)$.
- We cannot arbitrarily pick a single element from the vector because almost surely the element $\lambda_k < \epsilon$ for any $\epsilon > 0$.
- Instead, pick the element λ_k proportionally to $\frac{\lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$.
  - alternatively, sample some data and note the indices!
- We note that PPP theory solves this problem:
  - in Pitman, "Poisson-Kingman partitions," 2003,
  - derivation uses the famous Campbell's Theorem of PPP theory,
  - complex .... we wont look at the details
- Modern tricks like slice sampling make it easier to implement in the general case.

Theory — Size-Biased Sampling

Definition of Size-biased Sampling

Assume we have an infinite vector from a PPP with rate $\rho(\lambda)$, and the total is $\Lambda = \sum_{k=1}^{\infty} \lambda_k$. Given a sample $\vec{\lambda}$ generated by the PPP, sampling λ_k proportionally to $\frac{\lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$ is called size-biased sampling.

1. Sample j_1 according to probability vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k}\vec{\lambda}$, and use $\lambda_1^* = \lambda_{j_1}$.
2. Sample j_2 according to probability vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k}\vec{\lambda}$, but don't allow j_1, and use $\lambda_2^* = \lambda_{j_2}$.
3. Sample j_3 according to probability vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k}\vec{\lambda}$, but don't allow j_1, j_2, and use $\lambda_3^* = \lambda_{j_3}$.
4. *etc.*

Theory — Size-Biased Sampling

Definition of Size-biased Sampling

Assume we have an infinite vector from a PPP with rate $\rho(\lambda)$, and the total is $\Lambda = \sum_{k=1}^{\infty} \lambda_k$. Given a sample $\vec{\lambda}$ generated by the PPP, sampling λ_k proportionally to $\frac{\lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$ is called size-biased sampling.

Size-biased Sampling Lemma

Assume we have an infinite vector from a PPP with rate $\rho(\lambda)$, and the PDF for the total $\Lambda = \sum_{k=1}^{\infty} \lambda_k$ is known and given by $f(\Lambda)$. Then

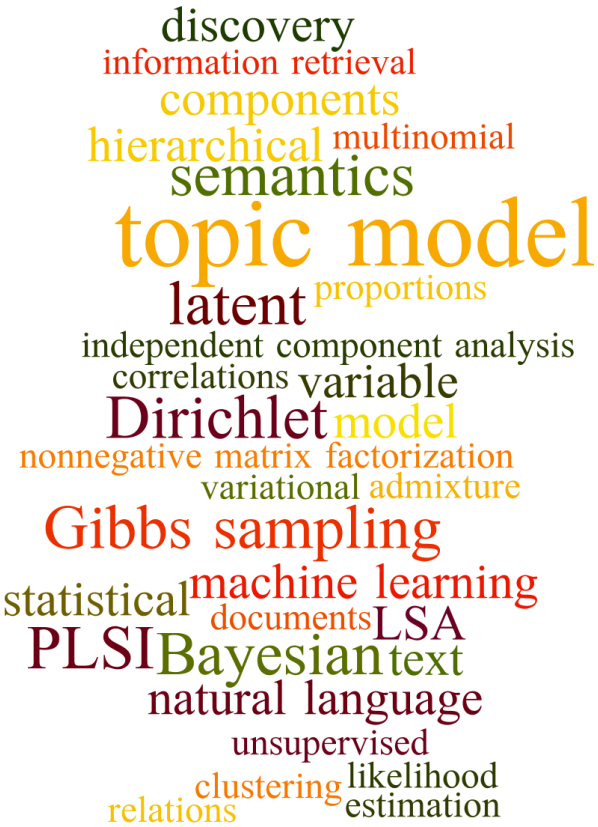
$$p(\lambda \mid \Lambda, \rho(\lambda)) = \frac{\lambda}{\Lambda} \rho(\lambda) \frac{f(\Lambda - \lambda)}{f(\Lambda)}$$

samples λ_k by size-biased sampling.

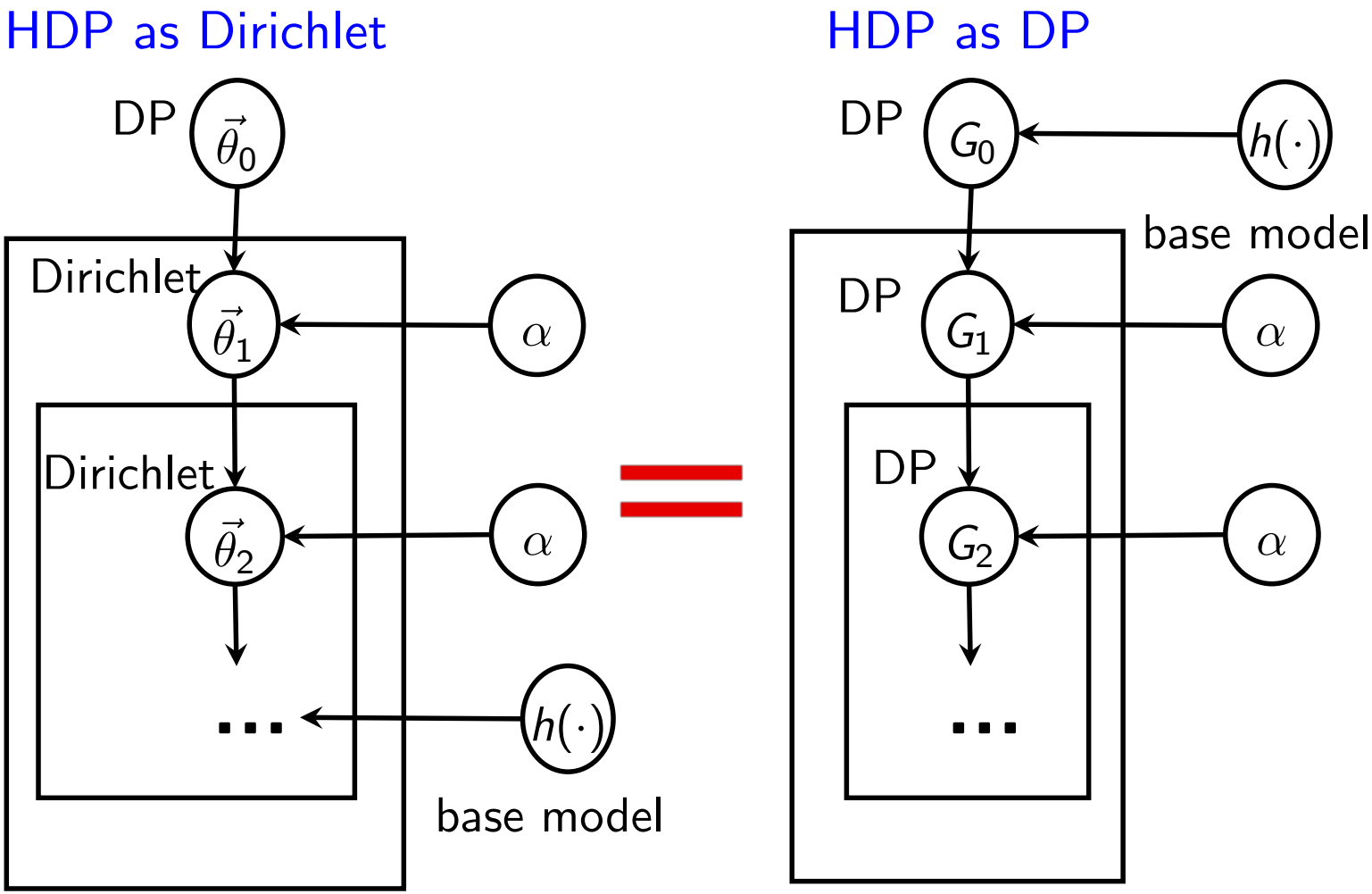
Sampling Vectors,, Note

For Dirichlet process, gamma Process, stable process and Pitman-Yor process only, equivalences mean sampling is independent of Λ.

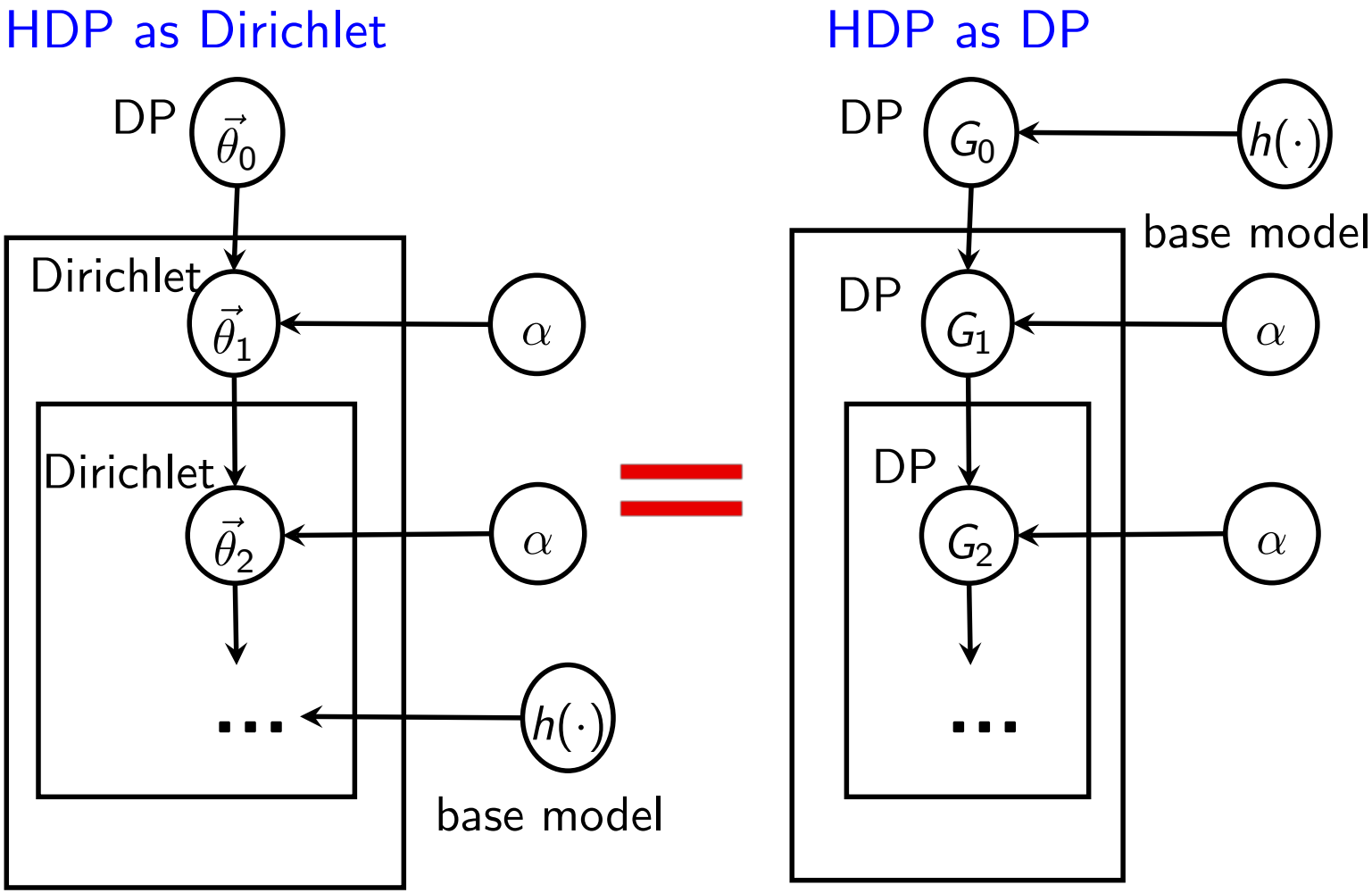
Outline



Hierarchical Processes



Hierarchical Processes



- hierarchical part of a HDP *is just* a Dirichlet

Hierarchical Processes, cont.

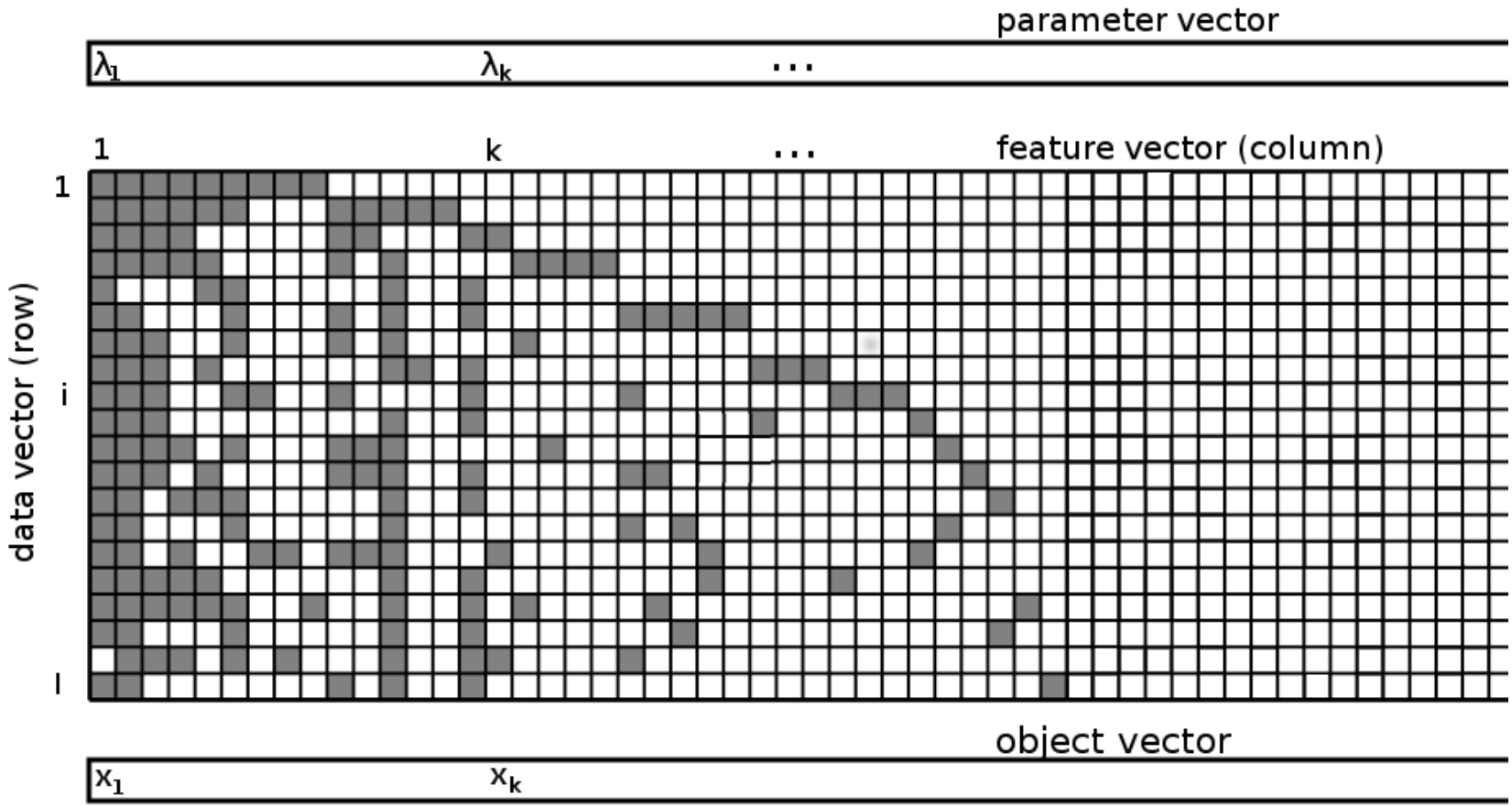
Likewise,

- hierarchical part of a gamma process *is just* an element-wise mapping of parameters using a gamma distribution
- hierarchical part of a generalised gamma process *is just* an element-wise mapping of parameters using a positive stable distribution$^{\mathcal{W}}$
- *etc.*

Now look at this in more detail.

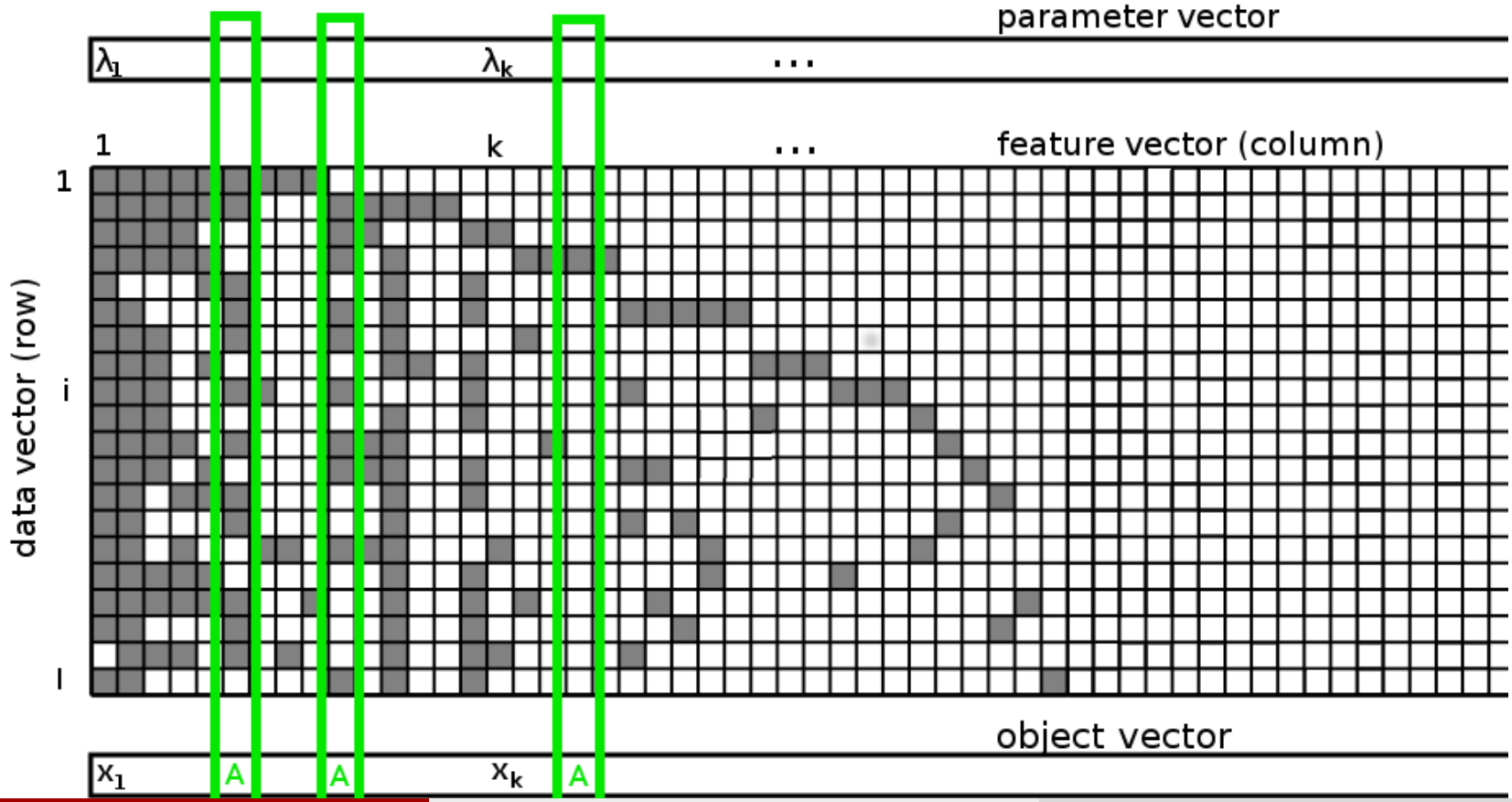
Hierarchical Modelling

- In regular use, at root of hierarchy, the objects (the marked part) with the stochastic sample are drawn from a continuous domain *e.g.*, multivariate Gaussian
  - all sampled objects are distinct



Hierarchical Modelling, cont.

- Below root of the hierarchical, objects with the sample are drawn from the discrete (countable number of) items of the parent.
- So sampled objects can be repeated: we get total data and cannot discern separate columns with identical objects.



Hierarchical Modelling, cont.

- Essential observation of Teh *et al.*'s 2006 HDP paper was this.
- However, same theory applies to all hierarchical processes.
- General proof is very simple: collapse all columns with identical objects, essentially means computing a total;
  - hence why we just looked at total mass distributions.

Hierarchical Modelling: Main Lemma

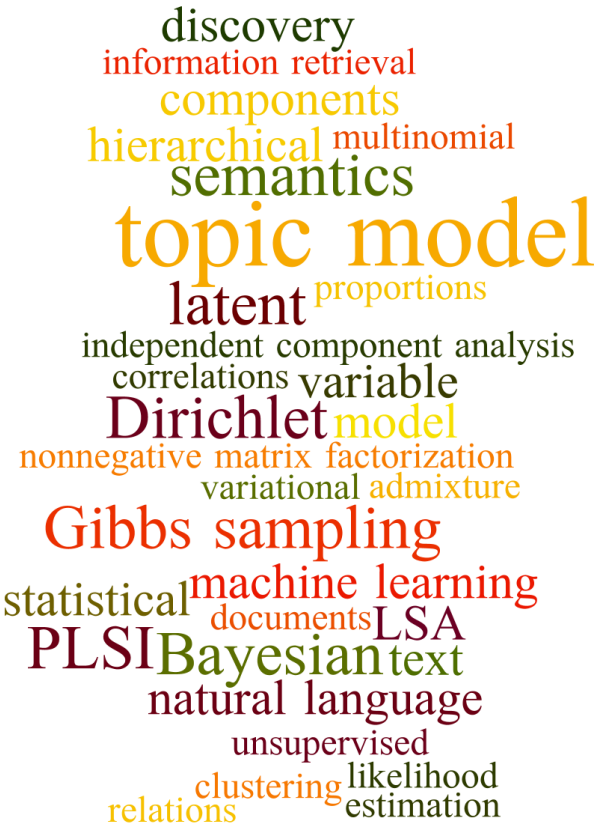
- Assume the parent process generates parameter vector $\vec{\mu}$ and object vector $\vec{x}$,
- then the child process will generate parameter vector $\vec{\lambda}$ and object vector $\vec{x}$ using an element-wise mapping of μ_k to λ_k:

rate for process	distribution of λ_k
beta-proc$(M, 0, 1)$	Dickman$(M\mu_k)$
gamma-proc$(M, 0, \beta)$	gamma$(M\mu_k, \beta)$
gen-gamma-proc(M, α, β)	Tweedie$(\alpha, (M\mu_k)^{1/\alpha}, \beta)$
stable-proc(M, α)	positive-stable$(\alpha, (M\mu_k)^{1/\alpha})$

Further Details for the BNP Initiates

- All hierarchical processes correspond to element-wise mappings of the parent vectors.
  - General proof is very simple: collapse all columns with identical objects, essentially means computing a total.
  - Alternatively, follows directly from tghe axiomatic definition of the process.
- There is also a generalised Chinese restaurant process$^{\mathcal{W}}$ for all hierarchical processes.
  - General proof involves noticing that the unfolding of a recursion linking cumulants to moments mimics counting in a Chinese restaurant process.
  - Useful as an alternative computational strategy for additive distributions.

Outline



Bayes Theorem for Gen. Beta Process

- we generate a countably infinite number of θ values using a beta process

$$\rho(\theta) = M \frac{1}{\Gamma(1-\alpha)} \theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-1}$$

  $\theta_1, \theta_2, \ldots$

- **thought experiment:** for each θ_k, sample a single Bernoulli and filter out those that return 0 in all I samples

Bayes Theorem for Gen. Beta Process

- we generate a countably infinite number of θ values using a beta process

$$\rho(\theta) = M\frac{1}{\Gamma(1-\alpha)}\theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-1}$$

  $\theta_1, \theta_2, \ldots$

- **thought experiment:** for each θ_k, sample a single Bernoulli and filter out those that return 0 in all I samples
- this means we change the rate to $\left(1-(1-\theta)^I\right)\rho(\theta)$

Bayes Theorem for Gen. Beta Process

- we generate a countably infinite number of θ values using a beta process

$$\rho(\theta) = M \frac{1}{\Gamma(1-\alpha)} \theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-1}$$

  $\theta_1, \theta_2, \ldots$
- **thought experiment:** for each θ_k, sample a single Bernoulli and filter out those that return 0 in all I samples
- this means we change the rate to $\left(1-(1-\theta)^I\right)\rho(\theta)$
- how does this new rate look?

Bayes Theorem for Gen. Beta Process

- we generate a countably infinite number of θ values using a beta process

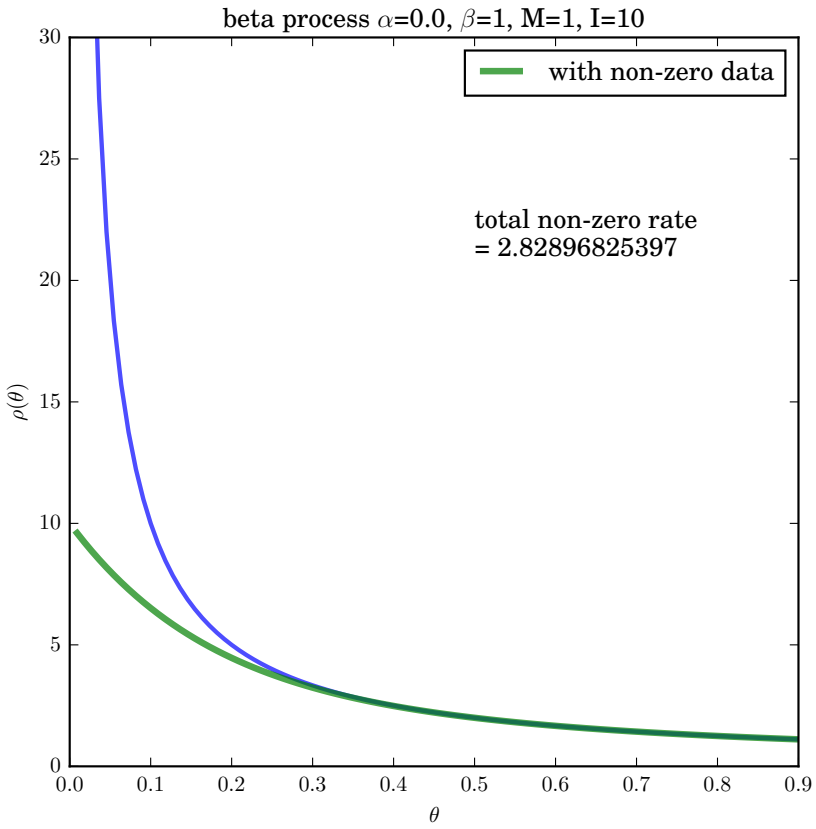
$$\rho(\theta) = M\frac{1}{\Gamma(1-\alpha)}\theta^{-\alpha-1}(1-\theta)^{\alpha+\beta-1}$$

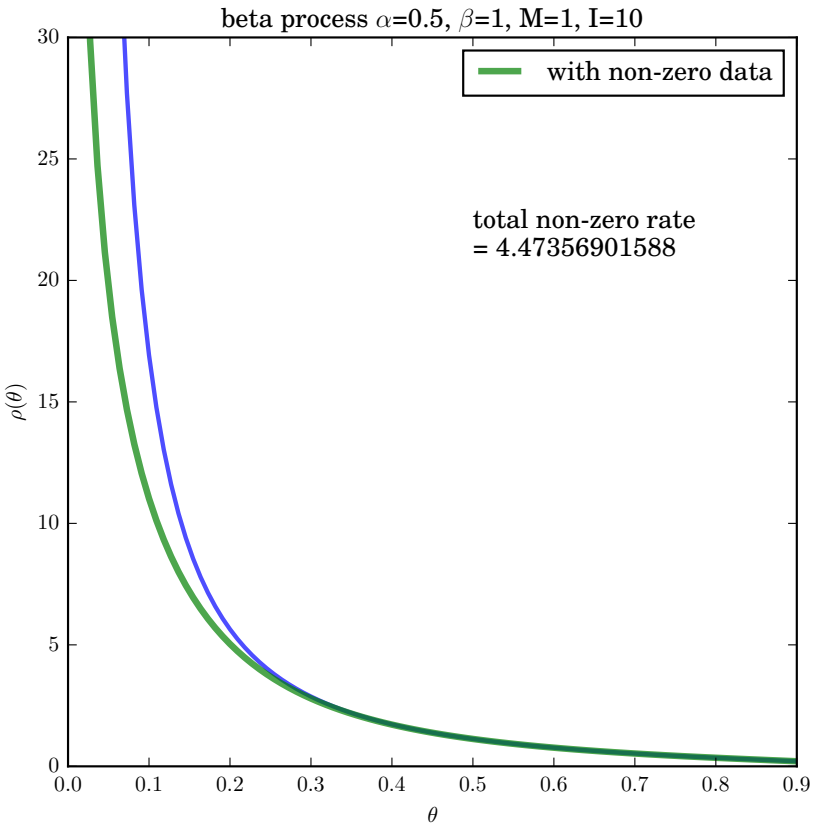
  $\theta_1, \theta_2, \ldots$
- **thought experiment:** for each θ_k, sample a single Bernoulli and filter out those that return 0 in all I samples
- this means we change the rate to $\left(1-(1-\theta)^I\right)\rho(\theta)$
- how does this new rate look? well its total is finite:

$$\Psi = \int_0^1 \left(1-(1-\theta)^I\right)\rho(\theta)\mathrm{d}\,\theta = M\sum_{i=0}^{I-1}\frac{\Gamma(\alpha+\beta+i)}{\Gamma(1+\beta+i)}$$

- call this Ψ the non-zero rate for the process

Bayes Theorem for Gen. Beta Process, cont.





Bayes Theorem for Gen. Gamma Process

- we generate a countably infinite number of λ values using a gamma process

$$\rho(\lambda) = M \frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)} \lambda^{-\alpha-1} e^{-\beta\lambda}$$

  $\lambda_1, \lambda_2, \ldots$
- **thought experiment:** for each λ_k, sample a single Poisson variable and filter out those that return 0 count all I times

Bayes Theorem for Gen. Gamma Process

- we generate a countably infinite number of λ values using a gamma process

$$\rho(\lambda) = M \frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)} \lambda^{-\alpha-1} e^{-\beta\lambda}$$

  $\lambda_1, \lambda_2, \ldots$
- **thought experiment:** for each λ_k, sample a single Poisson variable and filter out those that return 0 count all I times
- this means we change the rate to $(1 - e^{-I\lambda})\rho(\lambda)$

Bayes Theorem for Gen. Gamma Process

- we generate a countably infinite number of λ values using a gamma process

$$\rho(\lambda) = M \frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)} \lambda^{-\alpha-1} e^{-\beta\lambda}$$

  $\lambda_1, \lambda_2, \ldots$
- **thought experiment:** for each λ_k, sample a single Poisson variable and filter out those that return 0 count all I times
- this means we change the rate to $(1 - e^{-I\lambda})\rho(\lambda)$
- how does this new rate look?

Bayes Theorem for Gen. Gamma Process

- we generate a countably infinite number of λ values using a gamma process

$$\rho(\lambda) = M\frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}e^{-\beta\lambda}$$

  $\lambda_1, \lambda_2, ...$
- **thought experiment:** for each λ_k, sample a single Poisson variable and filter out those that return 0 count all I times
- this means we change the rate to $(1 - e^{-I\lambda})\rho(\lambda)$
- how does this new rate look? again its total is finite:

$$\Psi = \int_0^\infty (1 - e^{-I\lambda})\rho(\lambda)\lambda = \begin{cases} M\left(\log(I+\beta) - \log\beta\right) & \text{when } \alpha = 0 \\ M\left((I+\beta)^\alpha - \beta^\alpha\right) & \text{when } \alpha > 0 \end{cases}$$

Bayes Theorem for Gen. Gamma Process

- we generate a countably infinite number of λ values using a gamma process

$$\rho(\lambda) = M\frac{\alpha^{1_{\alpha>0}}}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}e^{-\beta\lambda}$$

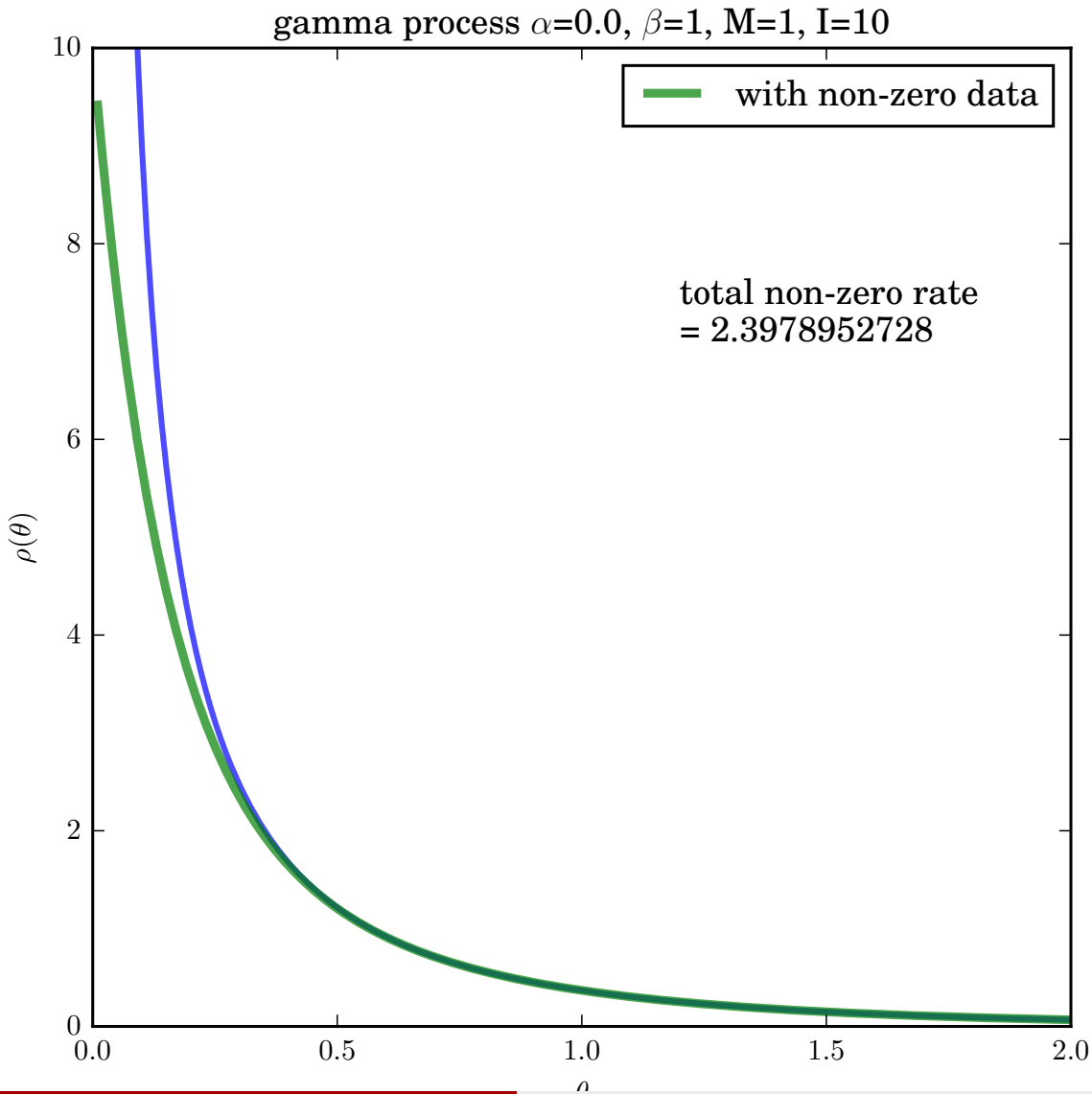
  $\lambda_1, \lambda_2, ...$
- **thought experiment:** for each λ_k, sample a single Poisson variable and filter out those that return 0 count all I times
- this means we change the rate to $(1 - e^{-I\lambda})\rho(\lambda)$
- how does this new rate look? again its total is finite:

$$\Psi = \int_0^\infty (1 - e^{-I\lambda})\rho(\lambda)\lambda = \begin{cases} M\left(\log(I+\beta) - \log\beta\right) & \text{when } \alpha = 0 \\ M\left((I+\beta)^\alpha - \beta^\alpha\right) & \text{when } \alpha > 0 \end{cases}$$

- moreover, posterior of λ_k given non-zero data in I rows

$$p(\lambda \,|\, \text{non-zero data}, I, \rho(\cdot)) = \frac{1}{\Psi}(1 - e^{-I\lambda})\rho(\lambda)$$

Bayes Theorem for Gen. Gamma Process, cont.



similar shaped plots for other settings of α and β

note λ takes value on full $\mathcal{R}^+$

Bayes Theorem for PPP

(Rough) Bayes theorem for Poisson processes

Given a Poisson process on space $\mathcal{A}$ with rate $\rho(w)$. Let A be discrete data generated using a discrete distribution $p(A|w)$, so $p(A|w) > 0$ for at least some values of w. Then if $\int_{\mathcal{A}} p(A|w)\rho(w)\mathrm{d}\, w < \infty$, it follows that:

- the *posterior distribution* for w given A is given by

$$p(w|A) = \frac{p(A|w)\rho(w)}{\int_{\mathcal{A}} p(A|w)\rho(w)\mathrm{d}\, w} ,$$

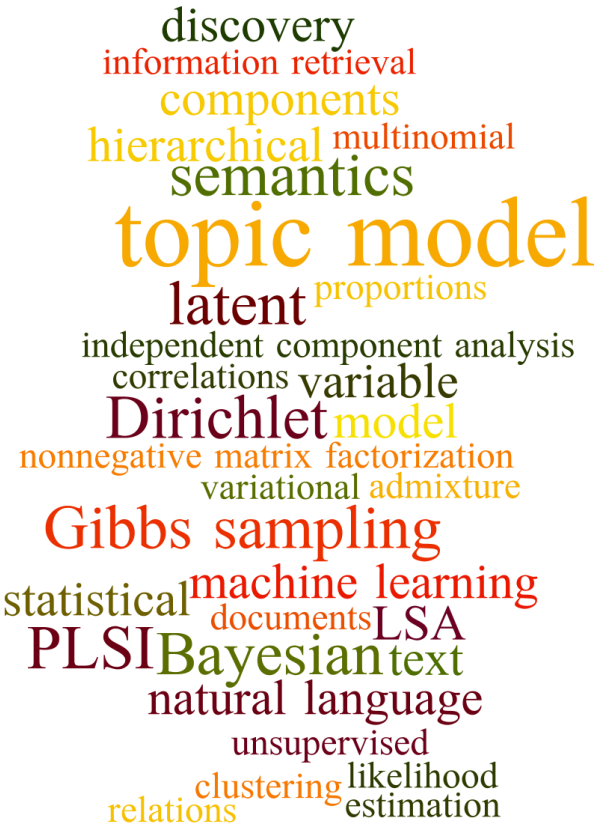
- and, the (finite) *joint rate* for (w, A) is given by

$$\rho(w, A) = p(A|w)\rho(w) .$$

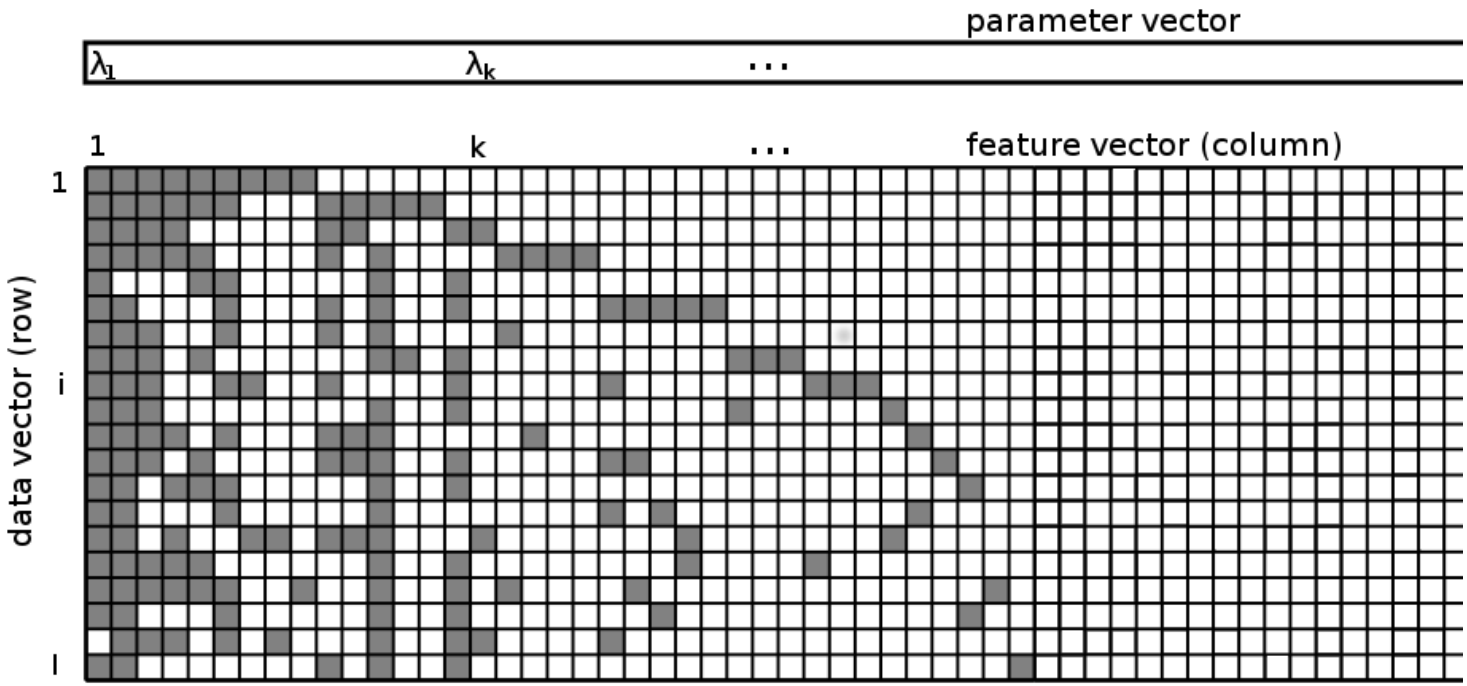
Bayes Theorem for PPP, cont.

- intuitive justification easily done.
- the joint rate has a finite total so posterior and marginal inference is feasible,
- ... even if $\int_{\mathcal{A}} \rho(\lambda) \mathrm{d}\, \lambda = \infty$.
- Correct proofs of related result can be done with James' Poisson process partition calculus (*Annals of Statistics*, 2005).
- See James' *Bayesian Poisson Calculus for Latent Feature Modeling via Generalized Indian Buffet Process Priors*, Annals of Statistics, 2016

Outline



Matrices with Unbounded Features



- **Data** = matrix of counts; each row $(1, ..., I)$ is a document/image; each column $(1, ..., \infty)$ is a feature/word
- potentially infinite features but ordered so non-zero features in first K
- model each column k as Poisson data with unknown rate λ_k
- model rate vector $\vec{\lambda}$ with gamma process

Likelihood Model for Unbounded Features

1. first, generate number of non-zero features (columns) using non-zero rate (which is a function of number of rows)

$$p(K \mid \text{non-zero data}, I, \rho(\cdot))$$

2. then for each feature (column) $k = 1, ..., K$
   1. generate λ_k given there is non-zero data

$$p(\lambda_k \mid \text{non-zero data}, I, \rho(\cdot))$$

   2. given λ_k and assuming non-zero data, generate the column of data

$$p(\vec{n}_{\cdot,k} \mid \lambda_k, \text{non-zero data}, I, \rho(\cdot))$$

Matrices with Unbounded Features, cont.

- non-zero rate $\Psi = M\left((I+\beta)^{\alpha} - \beta^{\alpha}\right)$ so $K \sim \text{Poisson}(\Psi)$
- for data vector $\vec{n}_{\cdot,k}$ in non-zero column k,

$$
\begin{aligned}
& p(\vec{n}_{\cdot,k}, \lambda_k \,|\, \text{non-zero data}) \\
&= p(\vec{n}_{\cdot,k} \,|\, \lambda_k, \text{non-zero data})\, p(\lambda_k \,|\, \text{non-zero data}) \\
&= \frac{1}{(1-e^{-I\lambda_k})} \left(\prod_{i=1}^{I} \frac{1}{n_{i,k}!} \lambda_k^{n_{i,k}} e^{-\lambda_k} \right) \frac{1}{\Psi} (1-e^{-I\lambda_k}) \rho(\lambda_k)
\end{aligned}
$$

- data marginal, $p(\vec{n}_{1,\cdot}, ..., \vec{n}_{I,\cdot} \,|\, \text{gamma process}, \alpha, \beta, M)$

$$
\begin{aligned}
&= \frac{\Psi^K e^{-\Psi}}{K!} \prod_{k=1}^{K} \int_{\mathcal{R}^+} p(\vec{n}_{\cdot,k}, \lambda_k \,|\, \text{non-zero data}) \mathrm{d}\, \lambda_k \\
&= \frac{e^{-\Psi}}{K!} \prod_{i,k=1}^{I,K} \frac{1}{n_{i,k}!} \prod_{k=1}^{K} \int_{\mathcal{R}^+} \lambda_k^{n_{\cdot,k}} e^{-I\lambda_k} \rho(\lambda_k) \mathrm{d}\, \lambda_k
\end{aligned}
$$

Matrices with Unbounded Features

- Framework works for all standard cases:
  - various versions Indian buffet process
  - gamma process or stable process models with Poisson or negative binomial data.
- Extension deals with Dirichlet process, normalised gamma process.

Non-zero Rates

Name	Non-zero rate
beta proc. Bernoulli	$M \sum_{i=0}^{I-1} \frac{1}{\beta + i}$
gen. beta proc. Bernoulli	$M \sum_{i=0}^{I-1} \frac{\Gamma(\alpha + \beta + i)}{\Gamma(1 + \beta + i)}$
gamma proc. Poisson	$M \left(\log(I + \beta) - \log \beta\right)$
gen. gamma proc. Poisson	$M \left((I + \beta)^{\alpha} - \beta^{\alpha}\right)$
gamma proc. negative binomial	$M \left(\log(I \log \frac{1}{1-\rho} + \beta) - \log \beta\right)$
gen. gamma proc. negative binomial	$M \left((I \log \frac{1}{1-\rho} + \beta)^{\alpha} - \beta^{\alpha}\right)$
stable proc. Poisson	MI^{α}