Spatial Normalized Gamma Processes

Vinayak Rao & Yee Whye Teh

Presented by Dae Il Kim

Motivation

The problem…

1) Dirichlet Processes (DP) are commonly used in Bayesian nonparametrics, but assumes datapoints are infinitely exchangeable.

2) The Dependent Dirichlet Process (DDP) was first introduced by MacEachern (1999) to account for this by creating a dependent set of random measures that were all marginally DP.

Since then, a number of DDP related publications have appeared.

- Time-sensitive DP *(Zhu and Lafferty, 05)*
- Hierarchical DP *(Teh et. al, 06)*
- Dynamic HDP *(Ren et. al, 08)*
- Generalized Polya Urn *(Caron et. al, 07)*
- Recurrent CRP *(Ahmed and Xing, 08)*
- πDDP *(Griffin and Steel, 06)*
- Local DP *(Chung and Dunson, 09)*

Topics Covered

Contribution: Formalizes the development of DDPs by providing a simple framework based on normalized gamma processes.

This talk will focus on the following:

1) The construction of the DDP via normalized gamma processes and some of its theoretical foundations.

2) Its representation as a DP mixture model.

3) Experiments on synthetic data and NIPS corpus

Most of the discussion regarding learning and inference in this model will be skipped while interesting will be skipped over in the interest of time.

Gamma Processes

Let (Θ, Ω) be a measure space on where we define a Gamma process ΓP.

↓

Define a Poisson process over a product space with mean measure:

$$\mu(d\theta dw) = \alpha(d\theta)w^{-1}e^{-w}dw$$

↓

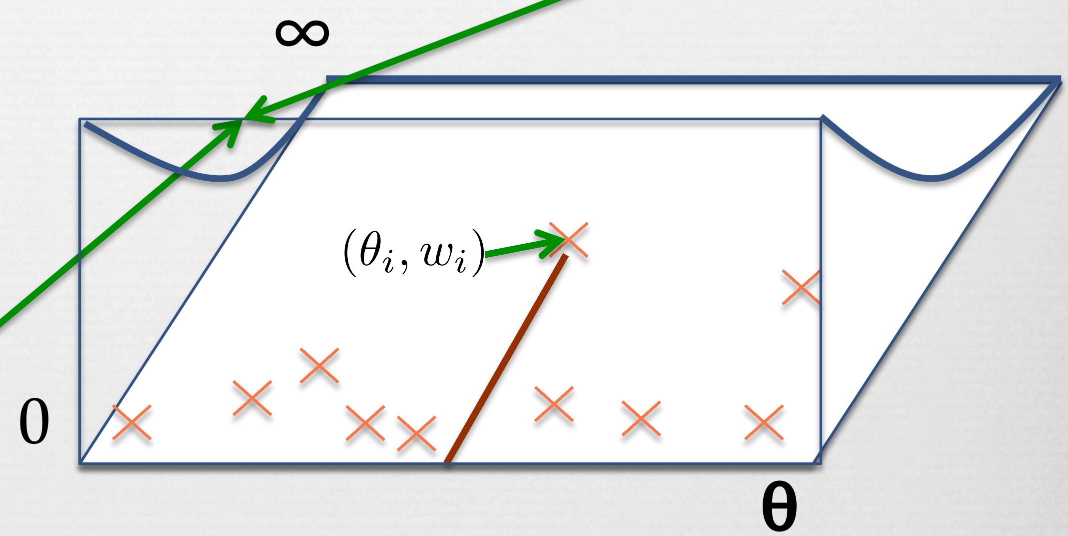
This results in an infinite set of atoms

$$\{\theta_i, w_i\}_{i=1}^{\infty}$$

↓

$$G = \sum_{i=1}^{\infty} w_i \delta_{\theta_i} \sim \Gamma P(\alpha)$$

Levy Measure or rate function

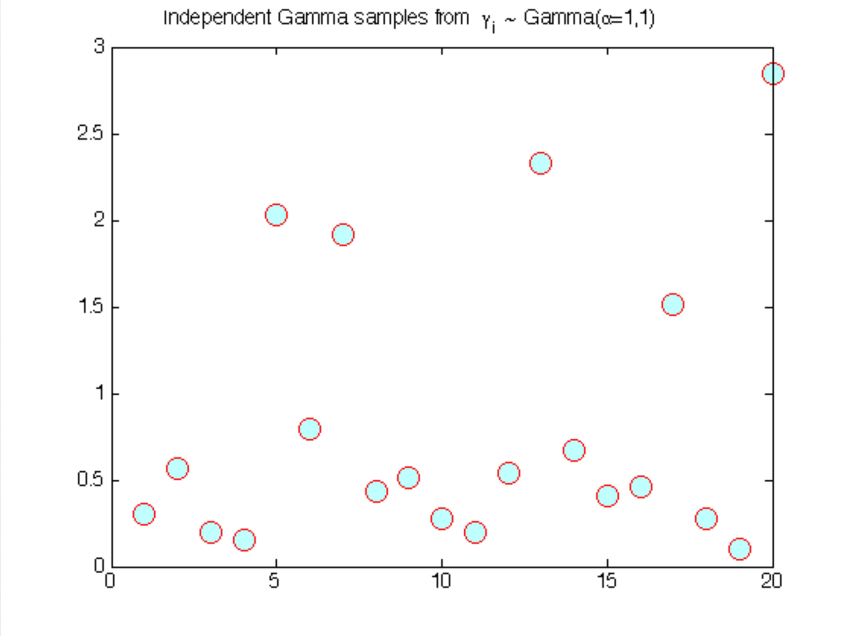


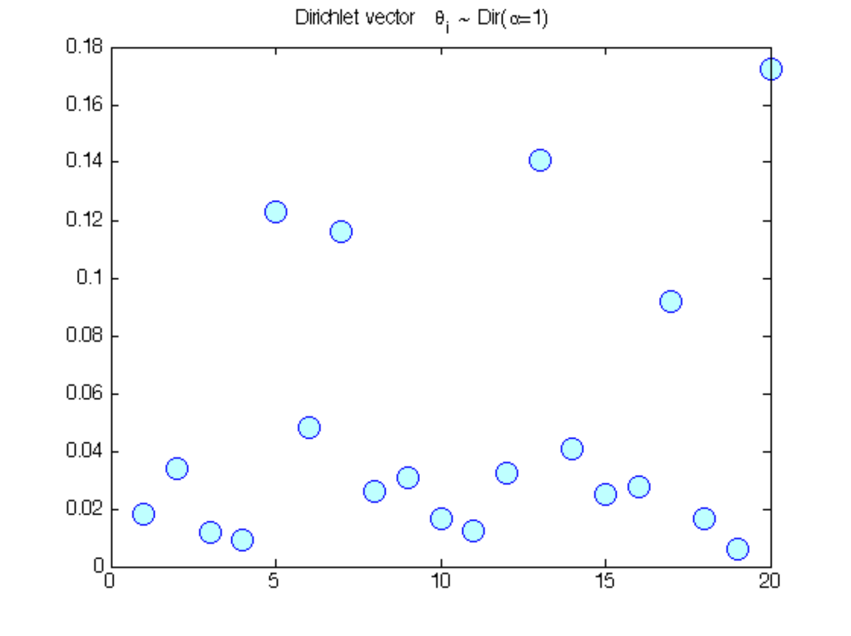
Product Space
$\Theta \otimes [0, \infty)$

A sample from a gamma process can be defined using these atoms, where α is a base measure on the (Θ, Ω) space and will be referred to as the base measure of our Gamma process.

Gamma Process → Dirichlet Process

Question: How can we sample from a Dirichlet distribution $\theta_i \sim Dir(\alpha)$?





Answer: Simply sample K independent gamma variables and normalize!

Gamma Process → Dirichlet Process

Any measurable subset S is simply gamma distributed with shape parameter $\alpha(S)$, thus the term gamma process.

$$G(S) = \sum_{i=1}^{\infty} w_i \mathbf{1}(\theta_i \in S) \; \forall \; S \subset \Theta$$

To create the Dirichlet process, we simply take G and normalize it.

$$D = \frac{G}{G(\Theta)} \sim DP(\alpha)$$

Here an atypical representation of the DP is used. The equivalent representation would be a strength parameter $\alpha(\Theta)$ and base distribution $\alpha / \alpha(\Theta)$.

Properties of the Gamma Process

The gamma process is an example of a completely random measure. (Kingman 1968)

Assigns independent mass to nonintersecting subsets.

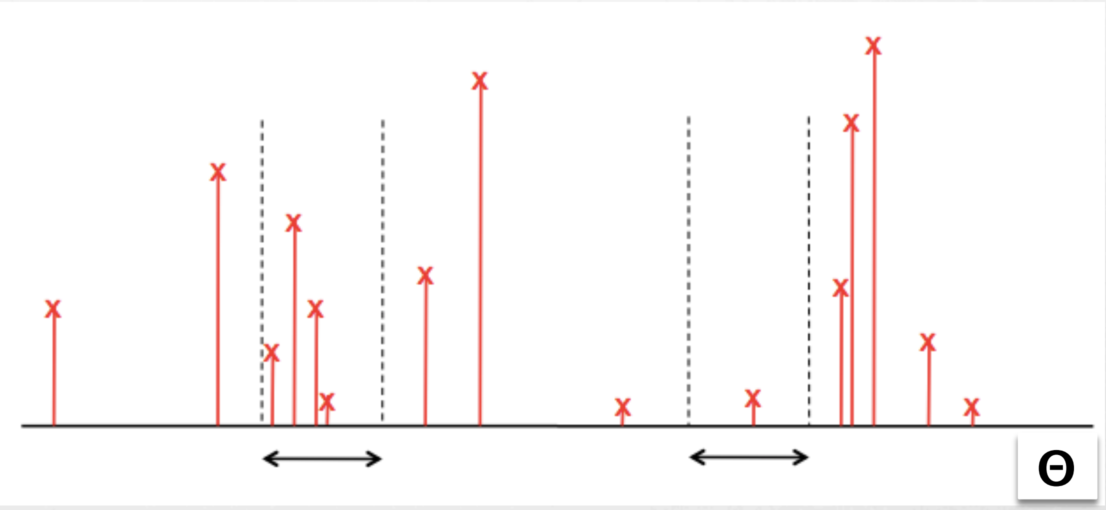


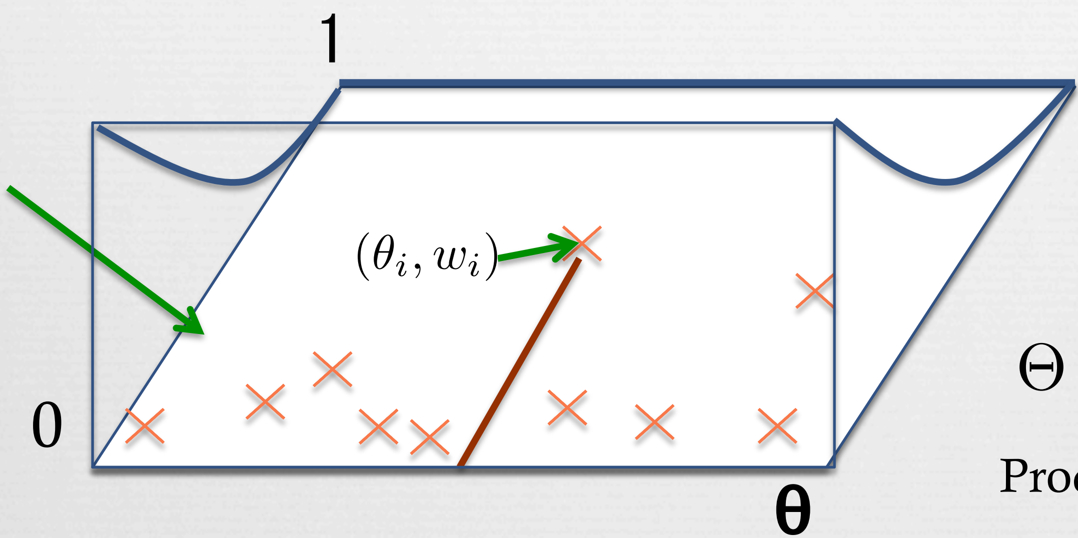
Figure from Jordan, M. (2010)

Other stochastic processes that are completely random measures are:

1) Beta Processes
2) Brownian Motion
3) Compound Poisson Processes

The Dirichlet process is not an example of a completely random measure!!

Quick Aside: Beta Processes



$$B = \sum_i w_i \delta_{\theta_i}$$

$\Theta \otimes [0, \infty)$

Product Space

The rate function is obtained from a product measure called a *Levy* measure. For the beta process, this measure lives on $\Theta \otimes (0, 1)$ and is given as follows:

$$\nu(d\theta, dw) = cw^{-1}(1-w)^{c-1} dw B_0(d\theta)$$

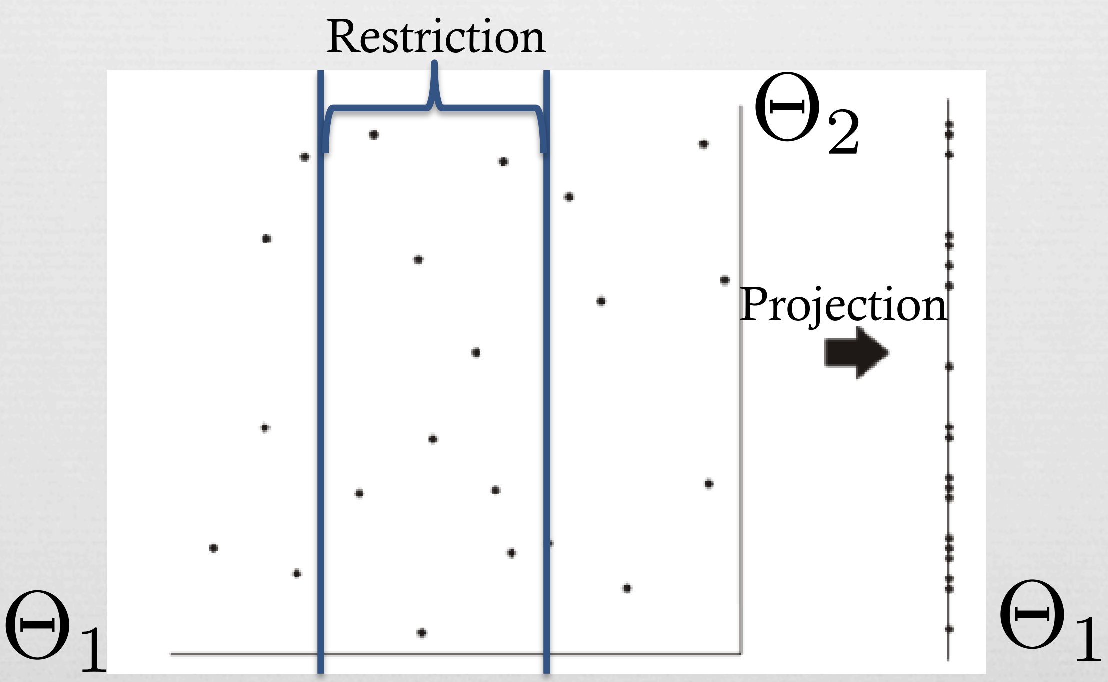
Properties of the Gamma Process

Two important consequences

1) If $S \in \Omega$ the restriction $G'(d\theta) = G(d\theta \cap S)$ onto S is a ΓP with base measure $\alpha'(d\theta) = \alpha(d\theta \cap S)$

2) If $\Theta = \Theta_1 \otimes \Theta_2$ is a two dimensional space, then the projection $G''(d\theta_1) = \int_{\theta_2} G(d\theta_1 d\theta_2)$ onto Θ_1 is also a ΓP with base measure $\alpha''(d\theta_1) = \int_{\theta_2} \alpha(d\theta_1 d\theta_2)$

In other words, projections and restrictions from gamma processes are still gamma processes.



Spatial Normalized Gamma Processes

The General Strategy

Let (Θ, Ω) be a probability space and $\mathbb{T}$ an index space.

Desiderata: We want to construct a set of dependent random measures over (Θ, Ω) with one D_t for each $t \in \mathbb{T}$ so that each D_t is marginally DP.

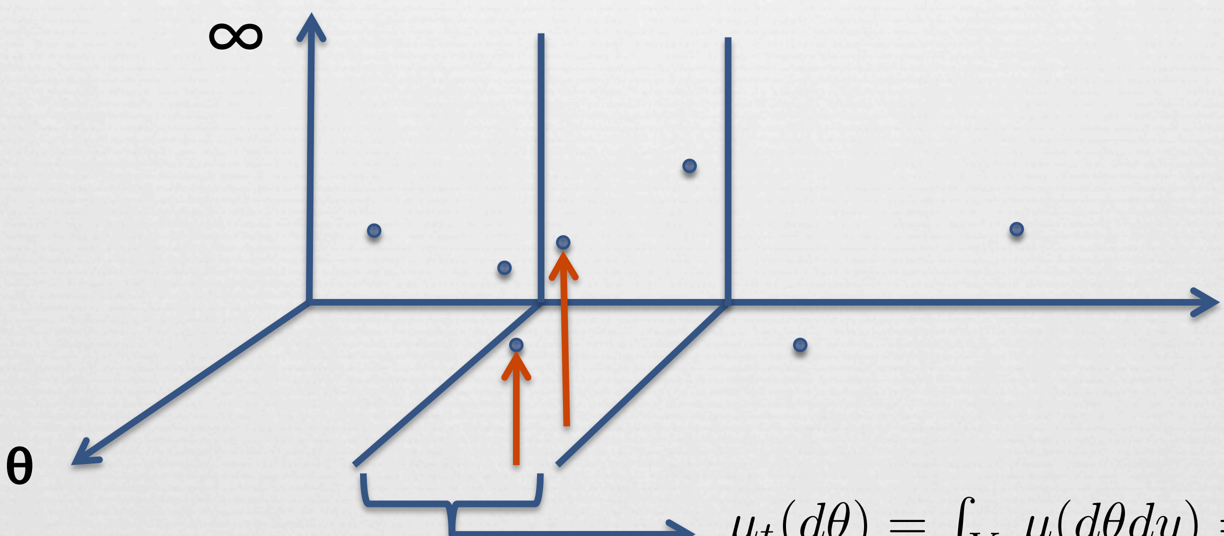
Solution

1. Define a gamma process G over an extended space.
2. Let each D_t be a normalized restriction/projection of G. 3. Since restrictions/projections of gamma processes are still gamma processes, each D_t will be DP distributed.

Spatial Normalized Gamma Processes

Motivating Example

Let $\mathbb{Y}$ be an auxiliary space for each $t \in \mathbb{T}$, let $Y_t \subset \mathbb{Y}$ be a measurable set and have μ be an arbitrary measure over the product space $\Theta \otimes \mathbb{Y}$. The following restriction projection μ_t is:



$$\mu_t(d\theta) = \int_{Y_t} \mu(d\theta dy) = \mu(d\theta \otimes Y_t)$$

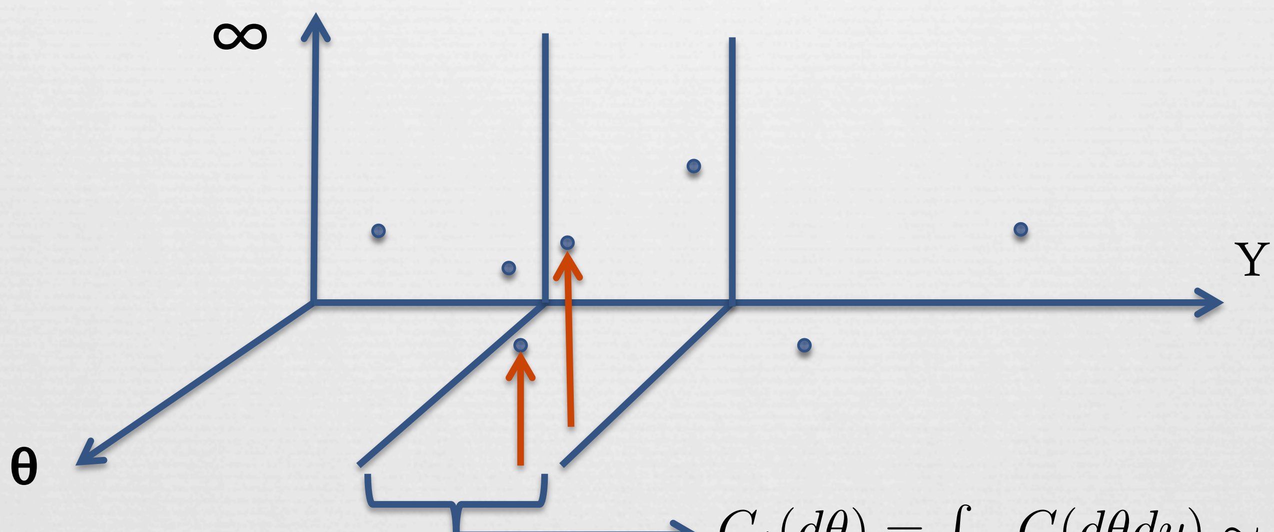
Note that μ_t is a measure over Θ for each $t \in \mathbb{T}$

Spatial Normalized Gamma Processes

We now let α represent a base measure over the same product space $\Theta \otimes \mathbb{Y}$ and consider a gamma process:

$$G \sim \Gamma P(\alpha)$$

We now perform a similar restricted projection for G so that G_t is a ΓP over Θ with base measure α_t:



$$G_t(d\theta) = \int_{Y_t} G(d\theta dy) \sim \Gamma P(\alpha_t)$$

Normalizing: $D_t = G_t / G_t(\Theta) \sim DP(\alpha_t)$

The resulting set of dependent DPs $\{D_t\}_{t \in \mathbb{T}}$ is now referred to as *spatial normalized gamma processes* ($SN\Gamma P$).

More Examples

Playing around with our base measure $\alpha(d\theta dy)$ can allow us to tune the dependency strength between D_t and D_s for $s, t \in \mathbb{T}$.

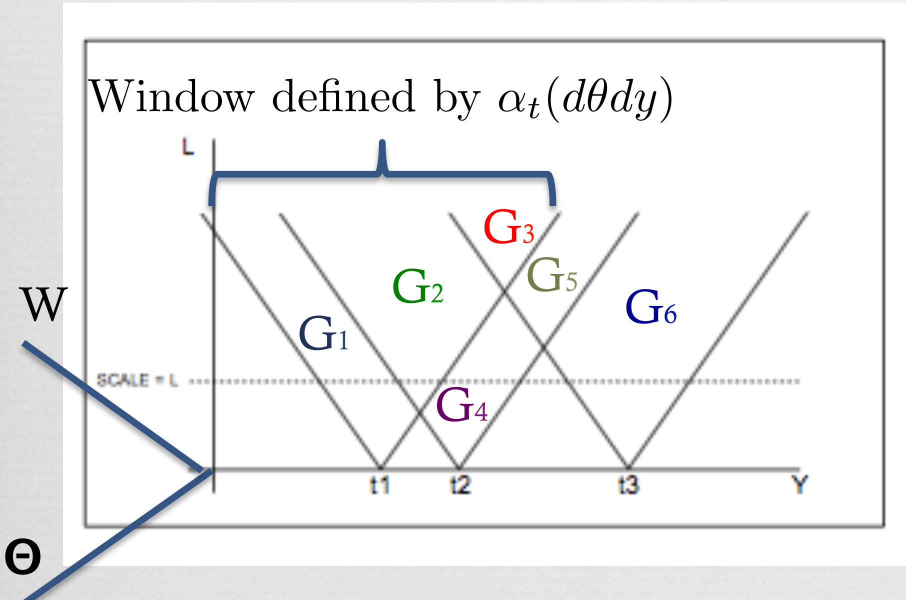
Let $\mathbb{T} = \mathbb{R}, \mathbb{Y} = \mathbb{R}, Y_t = [t - L, t + L]$ with $L > 0$. Furthermore, let H be a base distribution over Θ and $\gamma > 0$ be a concentration parameter. Defining the base measure $\alpha(d\theta dy) = \gamma H(d\theta)dy/2L$, the restricted projection of this base measure is then:

$$\alpha_t(d\theta dy) = \int_{t-L}^{t+L} \gamma H(d\theta)dy/2L = \gamma H(d\theta)$$

The result is that each $D_t \sim DP(\gamma H)$ where each atom in the overall ΓP G has a time stamp y and a time-spam of $[y - L, y + L]$ so that it will only appear in windows $t \in [y - L, y + L]$.

Mixtures of DPs representation

Define $\mathcal{R}$ as the smallest collection of disjoint *regions* of $\mathbb{Y}$ s.t each Y_{t_j} is a union of subsets in $\mathcal{R}$. For $1 \leq j \leq m$ let $\mathcal{R}_j$ be the collection of regions in $\mathcal{R}$ contained in Y_{t_j}, so that $\cup_{R\in\mathcal{R}_j} = Y_{t_j}$



For each $R \in \mathcal{R}$ define:

$$G_R(d\theta) = G(d\theta \otimes R) \sim \Gamma P(\alpha_R)$$
$$\alpha_R(d\theta) = \alpha(d\theta \otimes R)$$

$$D_R = G_R / G_R(\Theta) \sim DP(\alpha_R)$$

Mixture of DPs $\longrightarrow$ $D_{t_j}(d\theta) = \sum_{R\in\mathcal{R}_j} \frac{G_R(\Theta)}{\sum_{R\in\mathcal{R}_j} G_{R'}(\Theta)} D_R(d\theta)$

Mixtures of DPs representation

Each D_{t_j} is a mixture where each component D_R is drawn independently from a DP. The mixing proportions are also Dirichlet distributed and independent from the components by virtue of each $G_R(\theta)$ being gamma distributed and independent from D_R.

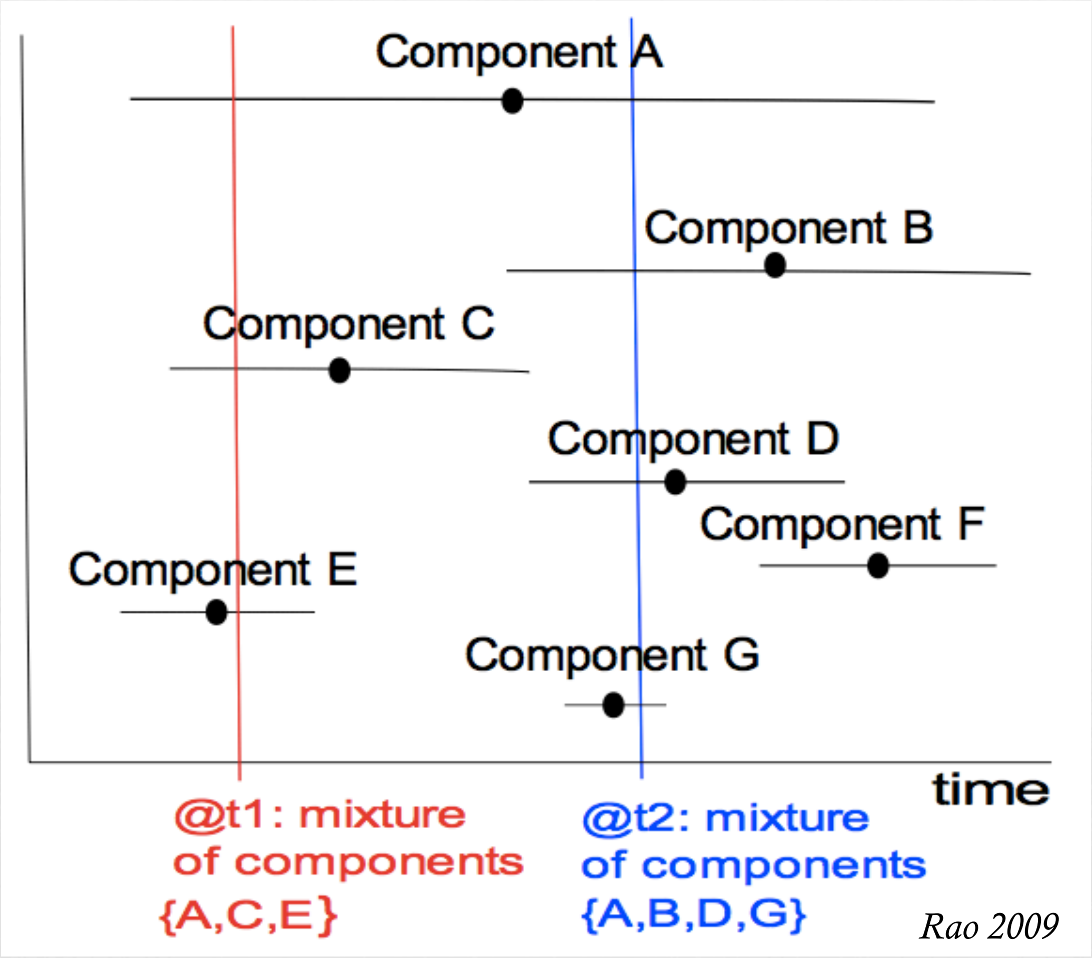
$$D_R \sim DP(\alpha_R) \qquad g_R \sim Gamma(\alpha_R(\Theta))$$

$$D_{t_j} = \textstyle\sum_{R \in \mathcal{R}_j} \pi_{jR} D_R \qquad \pi_{jR} = \frac{g_R}{\sum_{R' \in \mathcal{R}_j} g_R}$$

DPs in this construction are all defined only over Θ

References to the auxiliary space $\mathbb{Y}$ and base measure α are only used to define the individual base measures α_R and shape parameters for g_R

Mixtures of DPs representation



A visual representation of $SN\Gamma P$ as a Dirichlet Process Mixture Model

Inference in SN Γ P

They derive both a Gibbs Sampling scheme and a Metropolis Hastings Update to improve the convergence of the sampler.

Gibbs Sampling

1) Integrate out region DPs and replace them with chinese restaurants. Iteratively resample latent variables associated with the model and use conjugate base distribution H to the mixture distributions.

MH – Proposals

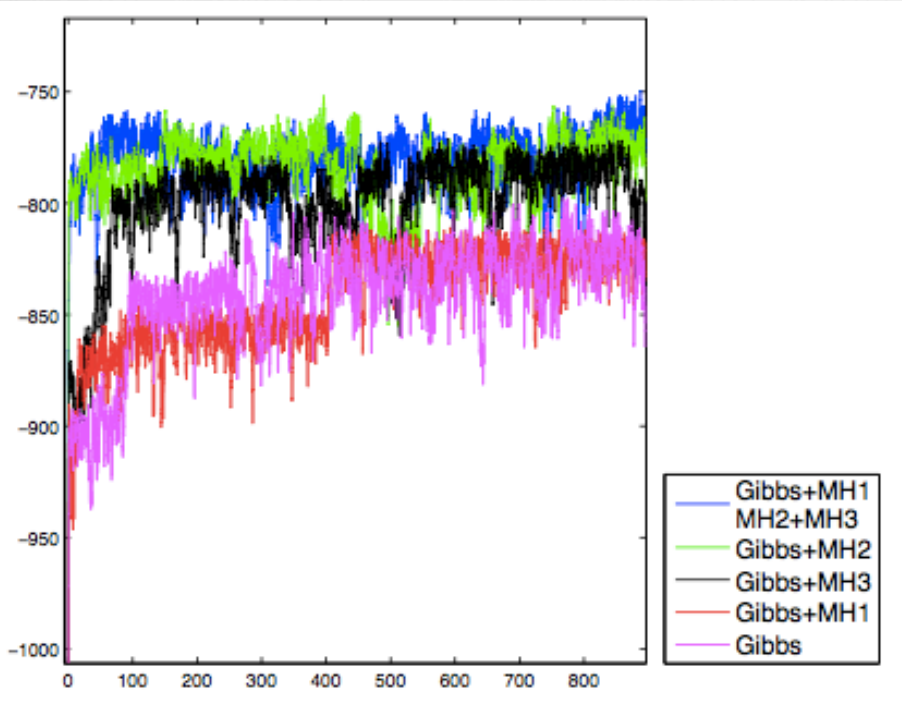
1) Utilize a split-merge move to split an existing cluster in a region into two new clusters in the *same* region or merge two existing clusters.

2) Move a picked cluster from one region to another. The new region is chosen from a region neighboring the current one.

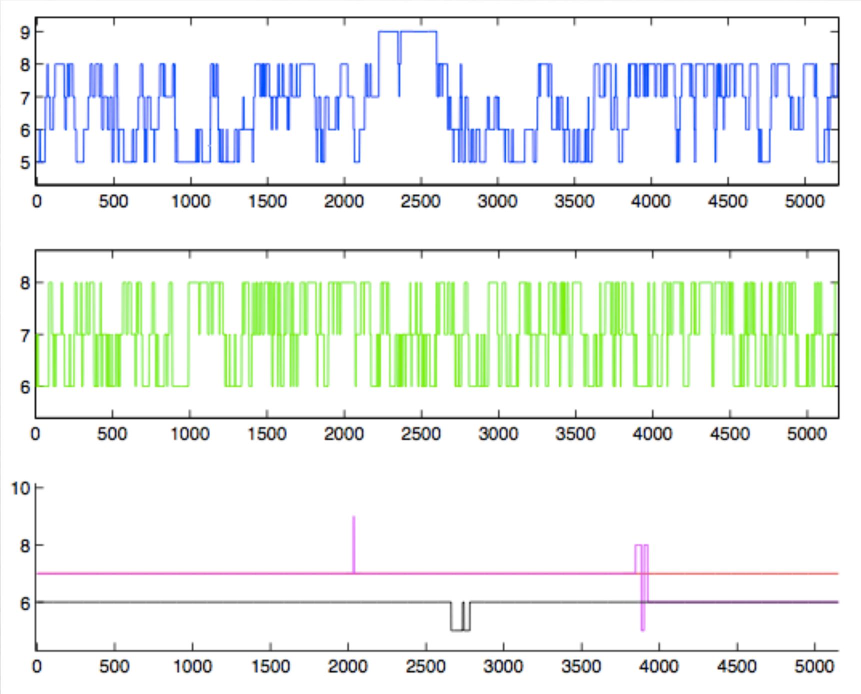
3) Make larger moves by combining the first two steps.

Experiments

Synthetic Data & Performance: Dataset generated by sampling from a mixture of 10 Gaussians with 60 generated datapoints across 5 time points. The data was modeled as a 5 DP mixture of Gaussians with a SN Γ P prior over the five dependent DPs.



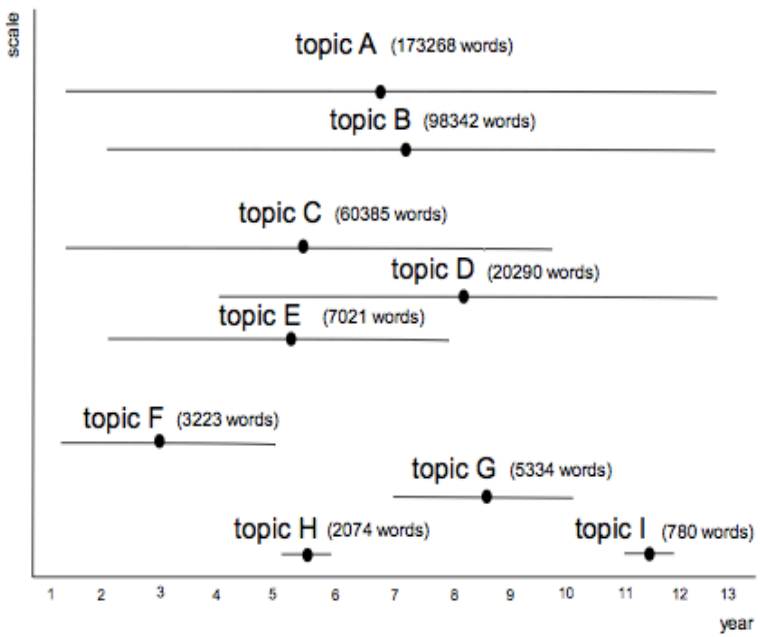
Log-likelihoods (the colored lines are ordered at iteration 80 like the legend)



Evolution of the timespan of a cluster. From top to bottom: Gibbs+MH1+MH2+MH3, Gibbs+MH2 and Gibbs+MH1 (pink), Gibbs+MH3 (black) and Gibbs (magenta).

Experiments

NIPS Dataset: The model utilized a combination of a SNΓP prior on top of an HDP model where each document was a DP sharing the same base distribution for a given year. Thus, there were a total of 13 base distributions associated with years 1987-1999 which were shared using a SNΓP.



Topic A	function, model, data, error, learning, probability, distribution
Topic B	model, visual, figure, image, motion, object, field
Topic C	network, memory, neural, state, input, matrix, hopfield
Topic D	rules, rule, language, tree, representations, stress, grammar
Topic E	classifier, genetic, memory, classification, tree, algorithm, data
Topic F	map, brain, fish, electric, retinal, eye, tectal
Topic G	recurrent, time, context, sequence, gamma, tdnn, sequences
Topic H	chain, protein, region, mouse, human, markov, sequence
Topic I	routing, load, projection, forecasting, shortest, demand, packet

Inferred topics with their timespans (the horizontal lines). In parentheses are the number of words assigned to each topic. On the right are the top ten most probable words in the topics.

Summary

- We discussed a new representation of a dependent Dirichlet Process as a spatially normalized gamma process.
- This process starts as points from a Possion process on an outer product space, which also represents a realization of a Gamma Process. Normalizing this we have a Dirichlet Process.
- Dependent DPs can be obtained by specifying a restriction/projection of a gamma process from an extended space.
- Experiments with synthetic data and NIPS corpus shows that inference needs some work.
- Lots of possibilities via normalized random measures. The Dirichlet process is just one example of this with nice computational properties.