Dependent Random Probability Measures

Vinayak Rao

Gatsby Unit,
University College London

April 27, 2012

Introduction

The Dirichlet Process (DP) [Ferguson, 1973]: the cornerstone of nonparametric models of probability measures.

Introduction

The Dirichlet Process (DP) [Ferguson, 1973]: the cornerstone of nonparametric models of probability measures.

By itself, the DP assumes observations are exchangeable

Not always true: eg. time-series, spatial data, conditional density modelling.

Introduction

The Dirichlet Process (DP) [Ferguson, 1973]: the cornerstone of nonparametric models of probability measures.

By itself, the DP assumes observations are exchangeable

Not always true: eg. time-series, spatial data, conditional density modelling.

One might wish to model data *conditioned* on some covariate (time, location, features, experimental condition etc.)

Introduction

The Dirichlet Process (DP) [Ferguson, 1973]: the cornerstone of nonparametric models of probability measures.

By itself, the DP assumes observations are exchangeable

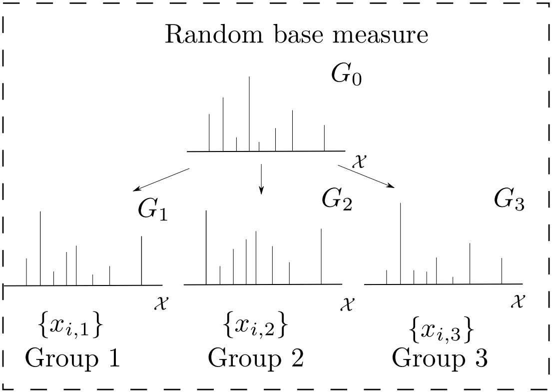
Not always true: eg. time-series, spatial data, conditional density modelling.

One might wish to model data *conditioned* on some covariate (time, location, features, experimental condition etc.)

We look at extensions of the DP (and other random probability measures) to model more structured data

dependent random probability measures (dRPMs)

- We have already seen examples of dRPMs: the HDP and its derivatives (HPY, PY-language model, the sequence memoizer).



$G_0 \sim DP(\alpha_0, H)$

$G_j \sim DP(\alpha, G_0)$

$x_{i,j} \sim G_j$

dependent random probability measures (dRPMs)

- We are interested in constructing RPMs G on some space $(\mathcal{X}, \Sigma)$.
- Consider some (usually metric) space $\mathcal{T}$, with elements t (eg. $\mathbb{R}$, $\mathbb{R}^d$).
- We want to index the RPMs by elements $t \in \mathcal{T}$.

dependent random probability measures (dRPMs)

- We are interested in constructing RPMs G on some space $(\mathcal{X}, \Sigma)$.
- Consider some (usually metric) space $\mathcal{T}$, with elements t (eg. $\mathbb{R}$, $\mathbb{R}^d$).
- We want to index the RPMs by elements $t \in \mathcal{T}$.

Desiderata:

- Similarity between G_{t_1} and G_{t_2} should decay smoothly with $\|t_1 - t_2\|$
- Ideally, we would like to decouple the marginal and correlation structures.

dependent random probability measures (dRPMs)

- We are interested in constructing RPMs G on some space $(\mathcal{X}, \Sigma)$.
- Consider some (usually metric) space $\mathcal{T}$, with elements t (eg. $\mathbb{R}$, $\mathbb{R}^d$).
- We want to index the RPMs by elements $t \in \mathcal{T}$.

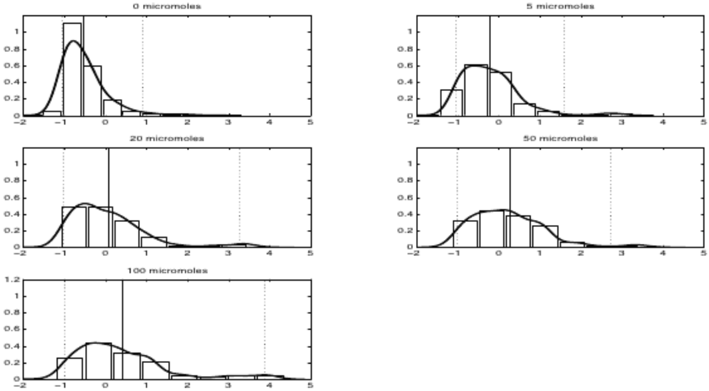
Desiderata:

- Similarity between G_{t_1} and G_{t_2} should decay smoothly with $\|t_1 - t_2\|$
- Ideally, we would like to decouple the marginal and correlation structures.

- We want to define a family of (usually uncountably infinite) dRPMs, G_t.
- G_t: a measure-valued stochastic process.

Motivating examples: genotoxicity experiments

[Dunson, 2006]



x: freq. of DNA strand breaks, t: strength of H_2O_2 dose.
Question: How does response distrbution vary with experimental conditions?

Motivating examples: Gestational age vs DDE exposure

[Dunson and Park, 2008]

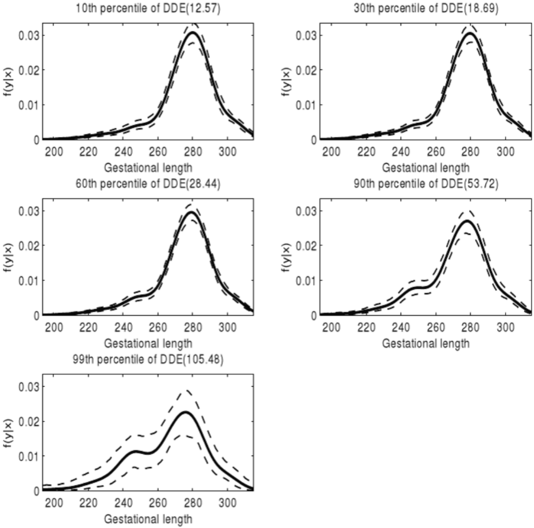
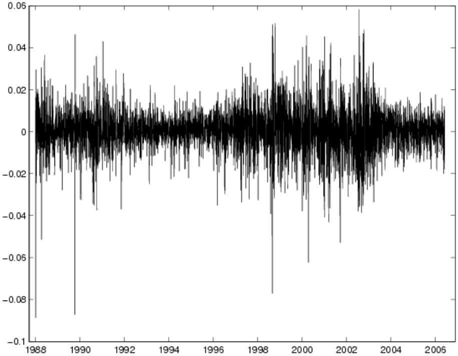


Figure 4: Estimated gestational age at delivery (in days) densities and pointwise 95% credible intervals conditional on DDE.

Motivating examples: volatility clustering

- Financial time series (http://staff.science.uva.nl/ marvisse/volatility.html)

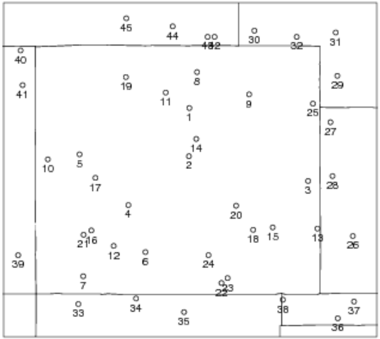


the daily percentage changes in the value of the S&P 500 index

Motivating examples: spatial data

[MacEachern et al., 2001]

Average temperature mid-July over a number of years at a number of locations



Motivating examples: spatial data

[MacEachern et al., 2001]

Average temperature mid-July over a number of years at a number of locations

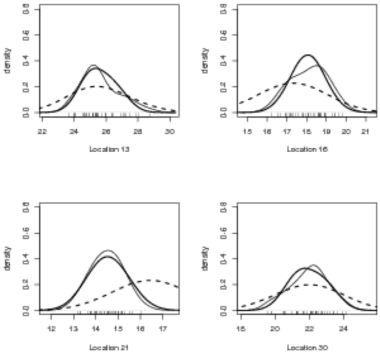
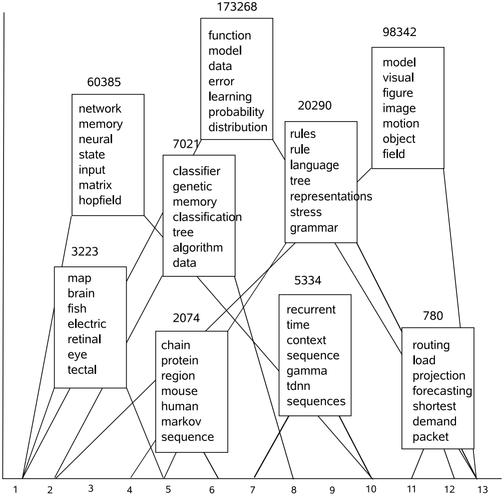


Figure 3: *Posterior predictive densities $Y_{new}(s)|data$ for the SDP(thick line $-$) and GP (thick dotted line $= \ -$). The lighter dotted line $(-)$ is the estimated density from the 40 replicates in the Colorado dataset (real data).*

Motivating examples: topic modelling

[Rao and Teh, 2009]



Simple approach: Link RPMs via the base measure

Recall:

$$G \sim DP(\alpha, H)$$

[Cifarelli and Regazzini, 1978]: introduce a regression on the base-measure H

$$G_t \sim DP(\alpha, H_t)$$

Simple approach: Link RPMs via the base measure

Recall:

$$G \sim DP(\alpha, H)$$

[Cifarelli and Regazzini, 1978]: introduce a regression on the base-measure H

$$G_t \sim DP(\alpha, H_t)$$

$$H_t = \mathcal{N}(\mu_t, \Sigma), \quad \mu_t \sim GP(0, K(\cdot,\cdot)) \implies G_t \sim DP\left(\alpha, \mathcal{N}(0, \Sigma + K(t,t))\right)$$

Simple approach: Link RPMs via the base measure

Recall:

$$G \sim DP(\alpha, H)$$

[Cifarelli and Regazzini, 1978]: introduce a regression on the base-measure H

$$G_t \sim DP(\alpha, H_t)$$

$$H_t = \mathcal{N}(\mu_t, \Sigma), \quad \mu_t \sim GP(0, K(\cdot, \cdot)) \implies G_t \sim DP\left(\alpha, \mathcal{N}(0, \Sigma + K(t,t))\right)$$

$$H_t = \mathcal{N}(\beta t, \Sigma), \quad \beta \sim \mathcal{N}(0, \sigma^2) \implies G_t \sim DP\left(\alpha, \mathcal{N}(0, \Sigma + t^2\sigma^2)\right)$$

Can also introduce dependence in α, though now we get a *mixture of DPs*

Simple approach: Link RPMs via the base measure

Recall:

$$G \sim DP(\alpha, H)$$

[Cifarelli and Regazzini, 1978]: introduce a regression on the base-measure H

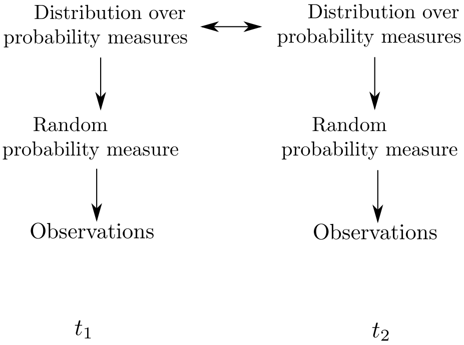
$$G_t \sim DP(\alpha, H_t)$$

$$H_t = \mathcal{N}(\mu_t, \Sigma), \quad \mu_t \sim GP(0, K(\cdot,\cdot)) \implies G_t \sim DP\left(\alpha, \mathcal{N}(0, \Sigma + K(t,t))\right)$$
$$H_t = \mathcal{N}(\beta t, \Sigma), \quad \beta \sim \mathcal{N}(0, \sigma^2) \implies G_t \sim DP\left(\alpha, \mathcal{N}(0, \Sigma + t^2\sigma^2)\right)$$

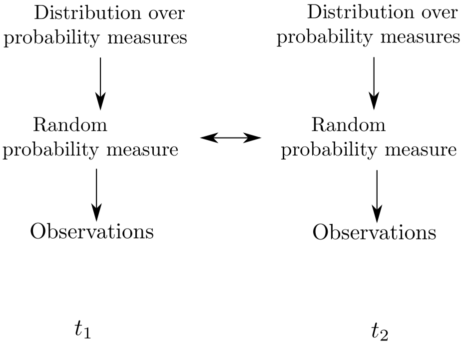
Can also introduce dependence in α, though now we get a *mixture of DPs*

For many applications, this dependence is too weak.
We want the *realizations* G_t and $G_{t+\delta}$ to be similar, not just their distributions

dependent DPs (dDP)



dependent DPs (dDP)



dependent DPs (dDP) [MacEachern, 1999]

Recall:

$$G = \sum_{i=1}^{\infty} p_i \delta_{x_i}$$

dependent DPs (dDP) [MacEachern, 1999]

Recall:

$$G = \sum_{i=1}^{\infty} p_i \delta_{x_i}$$

[MacEachern, 1999]: 'single p'-dDPs:

$$G_t = \sum_{i=1}^{\infty} p_i \delta_{x_{i,t}}$$

- A shared set of weights p drawn from eg. a stick-breaking process.
- Locations of atoms vary smoothly across measures.

dependent DPs (dDP) [MacEachern, 1999]

Recall:

$$G = \sum_{i=1}^{\infty} p_i \delta_{x_i}$$

[MacEachern, 1999]: 'single p'-dDPs:

$$G_t = \sum_{i=1}^{\infty} p_i \delta_{x_{i,t}}$$

- A shared set of weights p drawn from eg. a stick-breaking process.
- Locations of atoms vary smoothly across measures.

Eg. $x_{i,\cdot} \sim GP\,(0, K(\cdot,\cdot))$
At any t, $x_{i,t} \sim \mathcal{N}(0, K(t,t)) \equiv H_t$
Coupled with the stick breaking construction of p, we have that $G_t \sim DP(0, H_t)$

dependent DPs (dDP)

Advantages: Simple and fairly flexible

Disadvantages:

- not flexible enough?
- Global sharing of $p \implies$ lack of 'locality'

dependent DPs (dDP)

Advantages: Simple and fairly flexible

Disadvantages:

- not flexible enough?
- Global sharing of $p \implies$ lack of 'locality'
- posterior does not tend to prior as we move away from observations.

dependent DPs (dDP)

Advantages: Simple and fairly flexible

Disadvantages:

- not flexible enough?
- Global sharing of $p \implies$ lack of 'locality'
- posterior does not tend to prior as we move away from observations.

Suppose $x_{i,\cdot} \sim GP(0, K)$. single-p dDP is just a DP mixture of GPs!

general dependent DPs

How do we allow **p** vary across $\mathcal{T}$?

[MacEachern, 1999] does not provide a construction.

Remaining methods look at different approaches to this problem.

For simplicity, we shall assume the atoms locations are fixed: 'single-**x**' dRPMs.
Of course, easy to generalize.

Order-based dependent DPs [Griffin and Steel, 2006]

- Recall that the ith atom has mass $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
- Use a common collection of stick-breaking proportions $\mathcal{V} = \{V_i\}_{i=1}^{\infty}$ for all $G_t,\ t \in \mathcal{T}$

Order-based dependent DPs [Griffin and Steel, 2006]

- Recall that the ith atom has mass $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
- Use a common collection of stick-breaking proportions $\mathcal{V} = \{V_i\}_{i=1}^{\infty}$ for all $G_t,\ t \in \mathcal{T}$
- Since V_i's are i.i.d., valid for any permutation π:

$$p_i = V_{\pi(i)} \prod_{j=1}^{i-1}(1 - V_{\pi(j)})$$

Order-based dependent DPs [Griffin and Steel, 2006]

- Recall that the ith atom has mass $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
- Use a common collection of stick-breaking proportions $\mathcal{V} = \{V_i\}_{i=1}^{\infty}$ for all $G_t,\ t \in \mathcal{T}$
- Since V_i's are i.i.d., valid for any permutation π:

$$p_i = V_{\pi(i)} \prod_{j=1}^{i-1}(1 - V_{\pi(j)})$$

- Allow the permutation π_t to vary with t.
- At any time, we have the usual stick breaking construction: marginally DP.

Order-based dependent DPs [Griffin and Steel, 2006]

- Recall that the ith atom has mass $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
- Use a common collection of stick-breaking proportions $\mathcal{V} = \{V_i\}_{i=1}^{\infty}$ for all $G_t,\ t \in \mathcal{T}$
- Since V_i's are i.i.d., valid for any permutation π:

$$p_i = V_{\pi(i)} \prod_{j=1}^{i-1}(1 - V_{\pi(j)})$$

- Allow the permutation π_t to vary with t.
- At any time, we have the usual stick breaking construction: marginally DP.
- The influence of a V_i decays as it moves down the ranking: allows us to impose 'localness'.

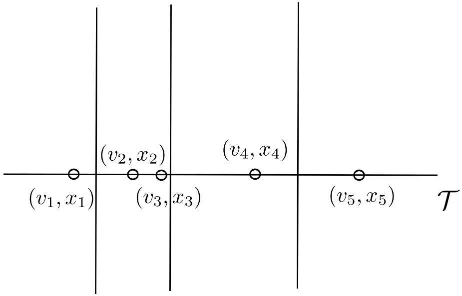
Order-based dependent DPs [Griffin and Steel, 2006]

- Recall that the ith atom has mass $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$
- Use a common collection of stick-breaking proportions $\mathcal{V} = \{V_i\}_{i=1}^{\infty}$ for all $G_t,\ t \in \mathcal{T}$
- Since V_i's are i.i.d., valid for any permutation π:

$$p_i = V_{\pi(i)} \prod_{j=1}^{i-1}(1 - V_{\pi(j)})$$

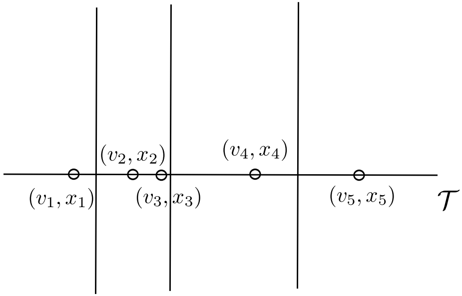
- Allow the permutation π_t to vary with t.
- At any time, we have the usual stick breaking construction: marginally DP.
- The influence of a V_i decays as it moves down the ranking: allows us to impose 'localness'.
- Challenge: construct a smoothly varying stochastic process π_t taking values in the space of all permutations.

Order-based dependent DPs [Griffin and Steel, 2006]



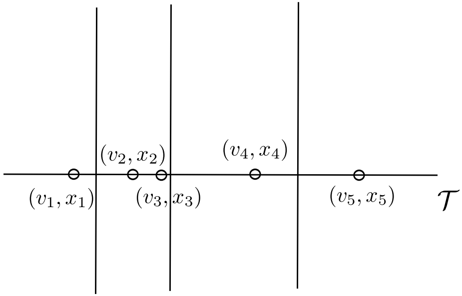
- Assign each (v, x) pair a time t.
- Permutation at t^* orders sticks by increasing distance from t^*.

Order-based dependent DPs [Griffin and Steel, 2006]



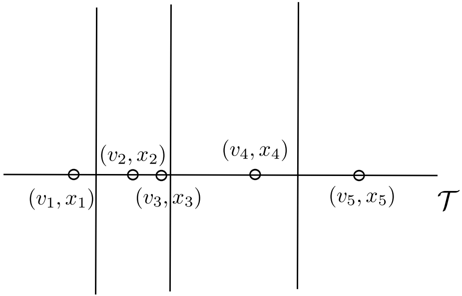
- Let π be the permutation at t, and π^* at t^*.
- Let s^* be the smallest element associated with data at t.

Order-based dependent DPs [Griffin and Steel, 2006]



- Let π be the permutation at t, and π^* at t^*.
- Let s^* be the smallest element associated with data at t.
- We want $P(\pi_{t^*}(s^*) < C) \to 0$ for any C as $d(t, t^*) \to \infty$.
- Posterior at t^* tends to prior as distance of t^* from all observations tends to 0

Order-based dependent DPs [Griffin and Steel, 2006]



- $corr(G_{t1}(B), G_{t2}(B)) = corr(G_{t1}, G_{t2}) = \left(1 + \frac{2\lambda d}{\alpha+2}\right) \exp(-\frac{2\lambda d}{\alpha+1})$
- Place priors on α and λ

Order-based dependent DPs [Griffin and Steel, 2006]

Inference:

- Truncated stick-breaking representation
- Instantiate $V, Z, x, s, \lambda, \alpha$ (Z: Poisson events, s: stick assignments)
- Instantiate Z on a bounded set containing covariates
- Update Z via birth-death and random walk processes
- Elegant model, messy inference

Kernel stick-breaking process

[Dunson and Park, 2008]

- Introduce a countable sequence of mutually independent random components,

$$\{\Gamma_h, V_h, G_h^*, h = 1, \cdots, \infty\}$$

Kernel stick-breaking process

[Dunson and Park, 2008]

- Introduce a countable sequence of mutually independent random components,

$$\{\Gamma_h, V_h, G_h^*, h = 1, \cdots, \infty\}$$

- $\Gamma_h \sim H$ is a location on $\mathcal{T}$ (can be more general).
- $V_h \sim Beta(a_h, b_h)$ is a stick-breaking proportion.
- $G_h^* \sim Q$ is a probability measure on $(\mathcal{X}, \Sigma)$.

Kernel stick-breaking process

[Dunson and Park, 2008]

- Introduce a countable sequence of mutually independent random components,

$$\{\Gamma_h, V_h, G_h^*, h = 1, \cdots, \infty\}$$

- $\Gamma_h \sim H$ is a location on $\mathcal{T}$ (can be more general).
- $V_h \sim Beta(a_h, b_h)$ is a stick-breaking proportion.
- $G_h^* \sim Q$ is a probability measure on $(\mathcal{X}, \Sigma)$.

- Consider a bounded kernel $K : \mathcal{T} \times \mathcal{T} \to [0, 1]$.
- $G_t \equiv \sum_{h=1}^{\infty} p_h(t) G_h^* \quad \forall t \in \mathcal{T}$
- $p_h(t) = \left\{ V_h K(t, \Gamma_h) \prod_{l<h} (1 - V_l K(t, \Gamma_l)) \right\}$

Kernel stick-breaking process

[Dunson and Park, 2008]

- $K = 1,\ G_h^* = \delta_{x_h},\ x_h \sim H$: A single DP
- $K = 1,\ G_h^* \sim DP(\alpha, G_0)$: A DP mixture of DPs

Typically, choose kernels like $K(x, \Gamma) = \exp(-\sigma\|x - \Gamma\|)$
If x is far from the first component, then it's breaking proportion is small.
More of the stick remains for the rest of the components

Kernel stick-breaking process

[Dunson and Park, 2008]

- $K = 1,\ G_h^* = \delta_{x_h},\ x_h \sim H$: A single DP
- $K = 1,\ G_h^* \sim DP(\alpha, G_0)$: A DP mixture of DPs

Typically, choose kernels like $K(x, \Gamma) = \exp(-\sigma\|x - \Gamma\|)$
If x is far from the first component, then it's breaking proportion is small.
More of the stick remains for the rest of the components

Not DP marginally (but can calculate marginal mean, variance, etc)
Can calculate correlation: shows localness
Inference: Block-Gibbs sampler: instantiate finite number of atoms.

Local Dirichlet process [Chung and Dunson, 2011]

- Three sequences of global, mutually independent components:

$$\Gamma_h, V_h, x_h, where \quad (1)$$

$$\Gamma_h \sim H, V_h \sim Beta, x_h \sim G_0 \quad (2)$$

- H is a prob measure on a space that may or may not correpond to $\mathcal{T}$.
- For some distance measure $\mathrm{d}(x, \Gamma)$ define an r-neighbourhood around x:

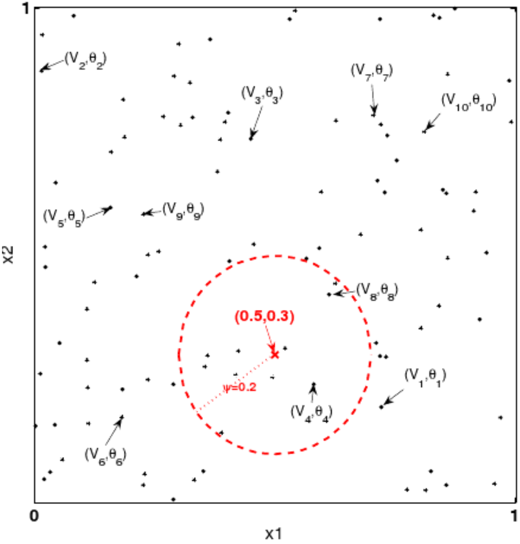
$$\mathcal{L}_x^r = \{h : d(x, \Gamma_h) < r\}$$

Now, letting π_i index the ith component in $\mathcal{L}^r$:

$$G_t = \sum_{i=1}^{\infty} p_i(t)\delta_{\theta_i}, \quad p_i(t) = V_{\pi_i(x)} \prod_{j=1}^{i-1}(1 - V_{\pi_j(x)})$$

Local Dirichlet process [Chung and Dunson, 2011]

Figure 1. Graphical Illustration for IDP formulation



Dependency via generalized Polya urn schemes

[Caron et al., 2007]

'single-p' models:

A clustering of observations at t_1	$\Longrightarrow$	a seating of customers at a restaurant
A new observation at t_2	$\Longrightarrow$	a new customer enters the *same* restaurant, even if $d(t_1, t_2)$ is large. His dish/parameter could be unrelated.

Dependency via generalized Polya urn schemes

[Caron et al., 2007]

'single-p' models:

A clustering of observations at t_1	$\Longrightarrow$	a seating of customers at a restaurant
A new observation at t_2	$\Longrightarrow$	a new customer enters the *same* restaurant, even if $d(t_1, t_2)$ is large. His dish/parameter could be unrelated.

Generalized Polya Urn for Time-varying Dirichlet Process Mixtures, [Caron et al., 2007]:

Introduce dependence across times by allowing the seating arrangement to evolve with time.

Dependency via generalized Polya urn schemes

[Caron et al., 2007] describe 3 update steps:

Dependency via generalized Polya urn schemes

[Caron et al., 2007] describe 3 update steps:

- Change parameters at all tables by some eg. Markov process

Dependency via generalized Polya urn schemes

[Caron et al., 2007] describe 3 update steps:

- Change parameters at all tables by some eg. Markov process
- Uniform deletion: Let C_t be the clustering at time t. Delete each customer with probability $\rho < 1$

Dependency via generalized Polya urn schemes

[Caron et al., 2007] describe 3 update steps:

- Change parameters at all tables by some eg. Markov process
- Uniform deletion: Let C_t be the clustering at time t. Delete each customer with probability $\rho < 1$
- Size-biased deletion: Pick a table proportinal to the number of customers seated at it. Delete all those customers.

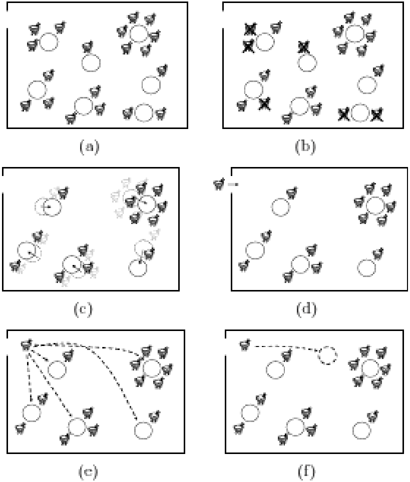
Dependency via generalized Polya urn schemes

[Caron et al., 2007] describe 3 update steps:

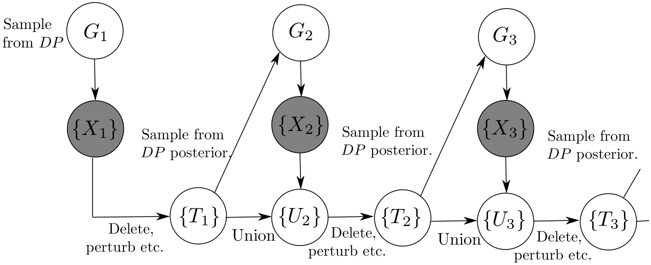
- Change parameters at all tables by some eg. Markov process
- Uniform deletion: Let C_t be the clustering at time t. Delete each customer with probability $\rho < 1$
- Size-biased deletion: Pick a table proportinal to the number of customers seated at it. Delete all those customers.

Cite [Kingman, 1975] to show that the Ewens sampling formula is still satisfied after deletion.

Dependency via generalized Polya urn schemes

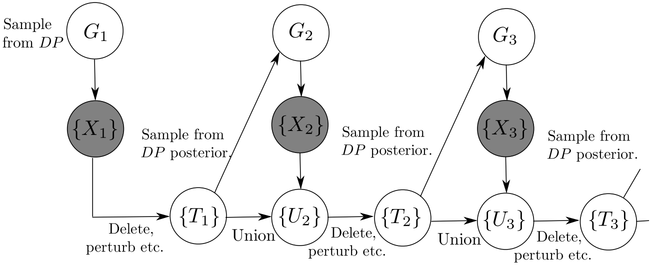


Dependency via generalized Polya urn schemes



$$
\begin{aligned}
G_i &\sim DP|\{T_{i-1}\} \\
\{X_i\} &\sim G_i \\
\{U_i\} &= T_{i-1} \cup X_i \\
\{T_i\} &= \mathrm{K}(\{U_i\})
\end{aligned}
$$

Dependency via generalized Polya urn schemes



Inference [Caron et al., 2007]:
Sequential MC and MCMC, working with the CRP representation (marginalizing out the G's).

Normalized random measures

- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.

Normalized random measures

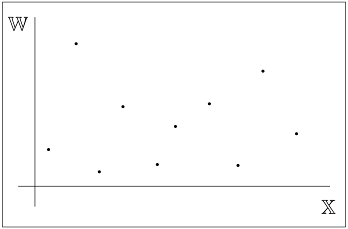
- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$

Normalized random measures

- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$
- (x_i, w_i) : events of a Poisson process on the space $\mathcal{X} \times \mathcal{W}$, where $\mathcal{W} = [0, \infty)$.

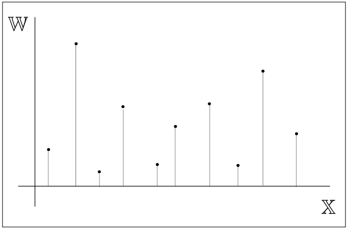
Normalized random measures

- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$
- (x_i, w_i) : events of a Poisson process on the space $\mathcal{X} \times \mathcal{W}$, where $\mathcal{W} = [0, \infty)$.



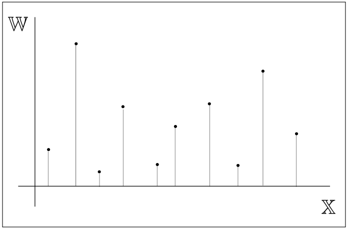
Normalized random measures

- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$
- (x_i, w_i) : events of a Poisson process on the space $\mathcal{X} \times \mathcal{W}$, where $\mathcal{W} = [0, \infty)$.



Normalized random measures

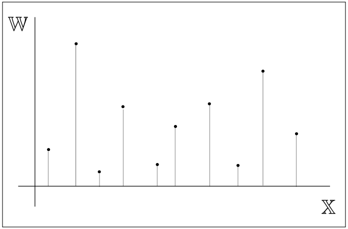
- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$
- (x_i, w_i) : events of a Poisson process on the space $\mathcal{X} \times \mathcal{W}$, where $\mathcal{W} = [0, \infty)$.



- Normalize to construct a random probability measure G: $G(\cdot) = \frac{\mu(\cdot)}{\mu(\Omega)}$

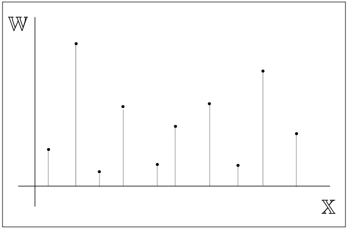
Normalized random measures

- Sample a *random measure* μ on some space $(\mathcal{X}, \Sigma)$.
- $\mu = \sum_i w_i \delta_{x_i}$
- (x_i, w_i) : events of a Poisson process on the space $\mathcal{X} \times \mathcal{W}$, where $\mathcal{W} = [0, \infty)$.



- Normalize to construct a random probability measure G: $G(\cdot) = \frac{\mu(\cdot)}{\mu(\Omega)}$
- The Levy intensity λ needs to ensure that the normalization constant $Z = \mu(\Omega)$ is strictly positive and finite a.s. Let $f(Z)$ be its distribution.

Normalized random measures



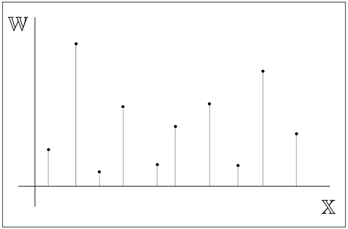
When $\lambda(dx, dw) =$

$$\alpha w^{-1} e^{-w} dw \ H(dx) : \text{Gamma process}$$

$$\alpha w^{-3/2} e^{-\tau w} dw \ H(dx) : \text{Inverse Gaussian process}$$

$$w^{-1-\beta} e^{-\tau w} dw \ H(dx) : \text{Stable process}$$

Completely Random Measures

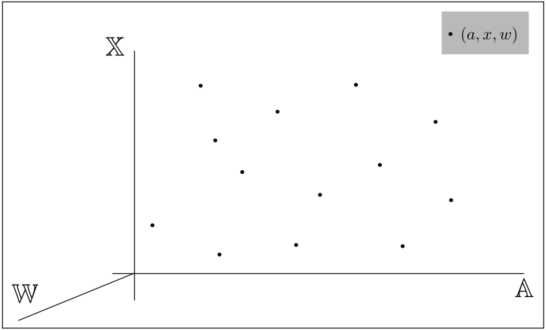


- From its Poisson construction, μ is a *completely random measure* [Kingman, 1993] :
  $\mu(A) \perp\!\!\!\perp \mu(B)$ if A and B are disjoint
- Similarly, the projection of a Poisson process is a Poisson process, resulting in μ being closed under projection in location space.

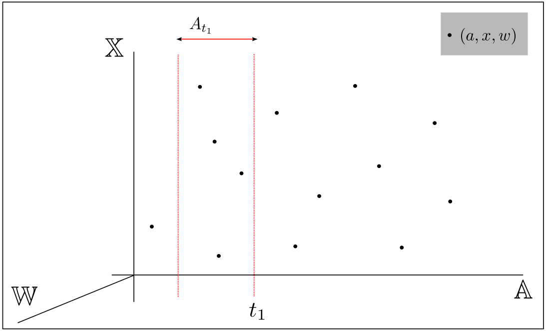
Spatial Normalized Random Measures [Rao and Teh, 2009]

- Instantiate a Poisson process on some augmented space
- Consider restrictions whose projections onto the original space define normalized random measures
- Dependency is controlled by controlling the amount of overlap of the restrictions

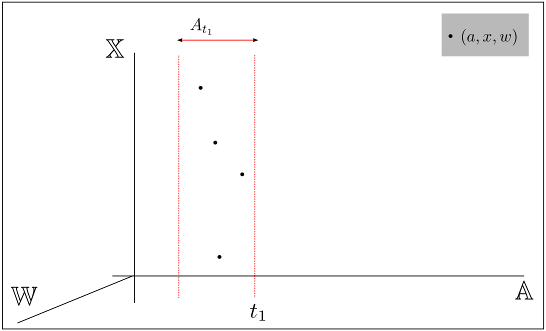
An example



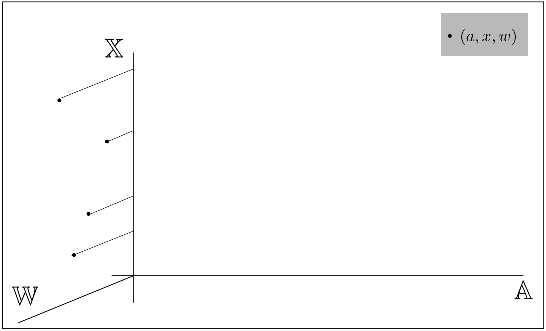
An example



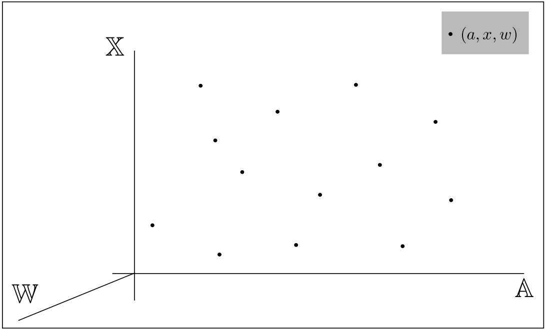
An example



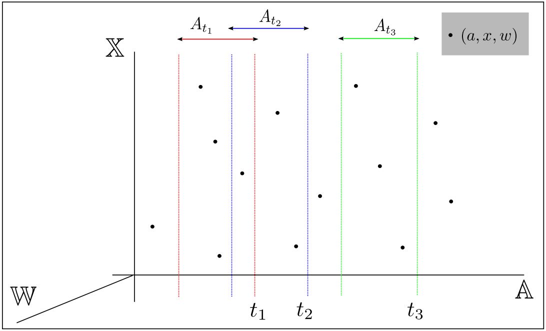
An example



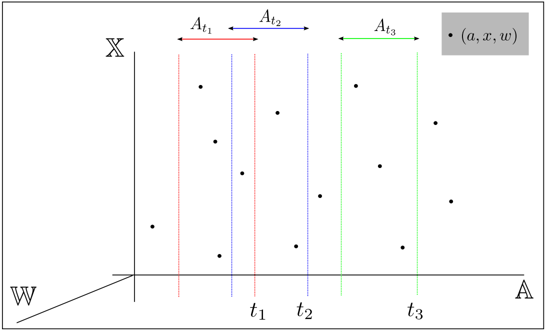
An example



An example



An example



- NRMs whose windows overlap share atoms
- NRMs that are 'closer' share more atoms
- NRMs separated by more than t_0 are independent

Spatial Normalized Random Measures [Rao and Teh, 2009]

- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs

Spatial Normalized Random Measures [Rao and Teh, 2009]

- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs
- Define a Poisson process N on an augmented space $\mathcal{A} \times \mathcal{X} \times \mathcal{W}$
  Let its intensity be $l(da) \times \lambda(dx, dw)$
- Associate with each index t a subset $S_t = A_t \times \mathcal{X} \times \mathcal{W}$

Spatial Normalized Random Measures [Rao and Teh, 2009]

- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs
- Define a Poisson process N on an augmented space $\mathcal{A} \times \mathcal{X} \times \mathcal{W}$
  Let its intensity be $l(da) \times \lambda(dx, dw)$
- Associate with each index t a subset $S_t = A_t \times \mathcal{X} \times \mathcal{W}$
- Define N_t as the projection of N restricted to S_t

$$N_t(dx, dw) = \int_{A_t} N(da, dx, dw)$$

Spatial Normalized Random Measures [Rao and Teh, 2009]

- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs
- Define a Poisson process N on an augmented space $\mathcal{A} \times \mathcal{X} \times \mathcal{W}$
  Let its intensity be $l(da) \times \lambda(dx, dw)$
- Associate with each index t a subset $S_t = A_t \times \mathcal{X} \times \mathcal{W}$
- Define N_t as the projection of N restricted to S_t

$$N_t(dx, dw) = \int_{A_t} N(da, dx, dw)$$

- $N_t(dx, dw)$ is a Poisson process with intensity $l(A_t)\lambda(dx, dw)$

Spatial Normalized Random Measures [Rao and Teh, 2009]

- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs
- Define a Poisson process N on an augmented space $\mathcal{A} \times \mathcal{X} \times \mathcal{W}$
  Let its intensity be $l(da) \times \lambda(dx, dw)$
- Associate with each index t a subset $S_t = A_t \times \mathcal{X} \times \mathcal{W}$
- Define N_t as the projection of N restricted to S_t

$$N_t(dx, dw) = \int_{A_t} N(da, dx, dw)$$

- $N_t(dx, dw)$ is a Poisson process with intensity $l(A_t)\lambda(dx, dw)$
- If $l(A_t)$ is finite, N_t specifies an NRM defined by $l(A_t)\lambda(dx, dw)$

Spatial Normalized Random Measures [Rao and Teh, 2009]

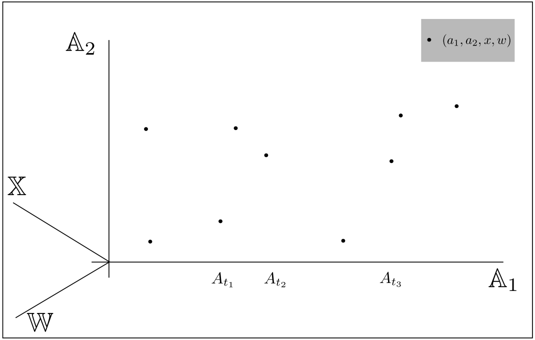
- We want:
  - a set of random probability measures G_t, $t \in \mathcal{T}$
  - for all t, G_t belongs to a class of NRMs
- Define a Poisson process N on an augmented space $\mathcal{A} \times \mathcal{X} \times \mathcal{W}$
  Let its intensity be $l(da) \times \lambda(dx, dw)$
- Associate with each index t a subset $S_t = A_t \times \mathcal{X} \times \mathcal{W}$
- Define N_t as the projection of N restricted to S_t

$$N_t(dx, dw) = \int_{A_t} N(da, dx, dw)$$

- $N_t(dx, dw)$ is a Poisson process with intensity $l(A_t)\lambda(dx, dw)$
- If $l(A_t)$ is finite, N_t specifies an NRM defined by $l(A_t)\lambda(dx, dw)$
- For two indices t_1 and t_2, if A_{t_1} and A_{t_2} overlap, the resulting NRMs share atoms and are correlated

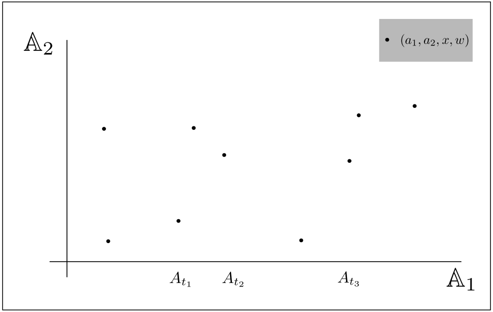
Multiple time-scales

- Allow different atoms have different scales
- Add an auxillary 'scale'-axis to the augmented space



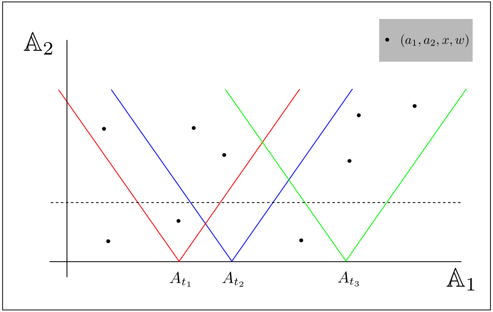
Multiple time-scales

- Allow different atoms have different scales
- Add an auxillary 'scale'-axis to the augmented space



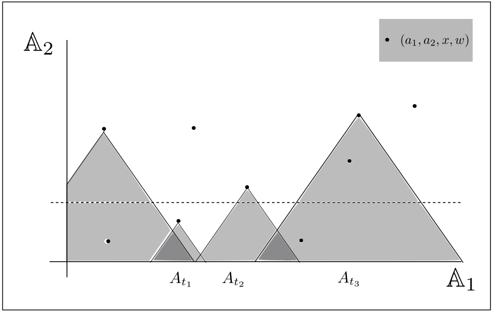
Multiple time-scales

- Allow different atoms have different scales
- Add an auxillary ‘scale’-axis to the augmented space

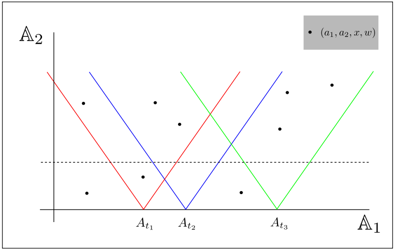


Multiple time-scales

- Allow different atoms have different scales
- Add an auxillary 'scale'-axis to the augmented space

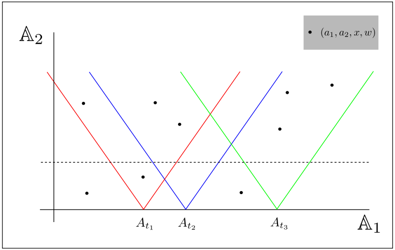


Spatial normalized Gamma processes [Rao and Teh, 2009]



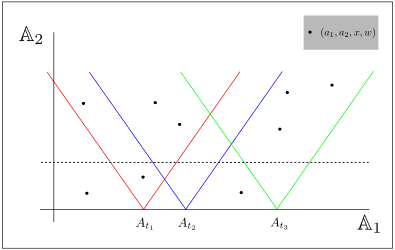
- The above procedure defines a DP for each element of $\mathcal{T}$
- In practice, we are given observations at a finite set of times

Spatial normalized Gamma processes [Rao and Teh, 2009]



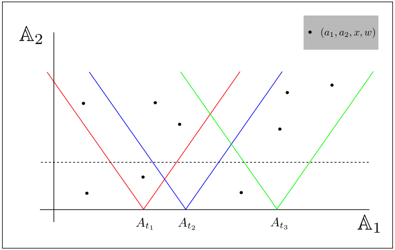
- The above procedure defines a DP for each element of $\mathcal{T}$
- In practice, we are given observations at a finite set of times
- We need only consider Poisson atoms relevant to these times
- Location of these atoms in $\mathcal{A}$ not important beyond which elements of $\mathcal{T}$ it is relevant to

Spatial normalized Gamma processes [Rao and Teh, 2009]



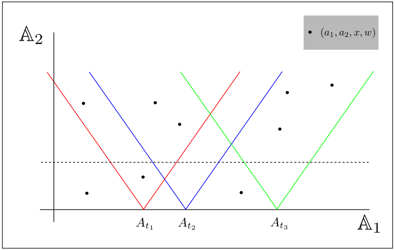
- Define *regions*
- We don't care about $\mathbb{A}$-coordinates of atoms in each region
- Associate a Gamma process with each region $\mu_r = Z_r G_r$
- ΓP at index t is the sum of the relevant ΓPs $\mu_t = \sum_{r \in R_t} \mu_r$
- DP at index t, $G_t = \sum_{r \in R_t} \frac{Z_r}{Z} G_r$

Spatial normalized Gamma processes [Rao and Teh, 2009]



Results in the following generative process:

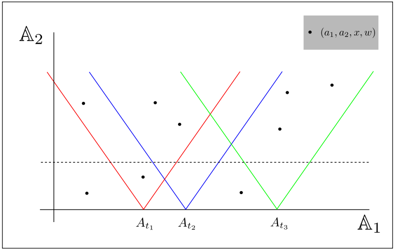
Spatial normalized Gamma processes [Rao and Teh, 2009]



Results in the following generative process:

- Assign each region $Z_r \sim Gamma(\alpha(A_r))$

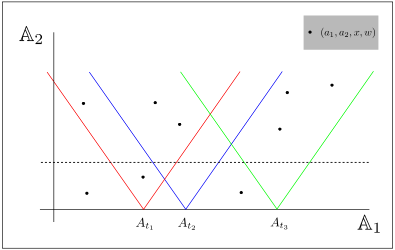
Spatial normalized Gamma processes [Rao and Teh, 2009]



Results in the following generative process:

- Assign each region $Z_r \sim Gamma(\alpha(A_r))$
- Assign an observation to a region r with probability $\propto Z_r$

Spatial normalized Gamma processes [Rao and Teh, 2009]

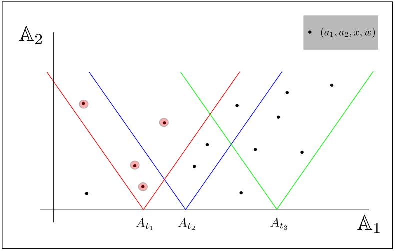


Results in the following generative process:

- Assign each region $Z_r \sim Gamma(\alpha(A_r))$
- Assign an observation to a region r with probability $\propto Z_r$
- Assign the observation to a cluster in that region according to the CRP

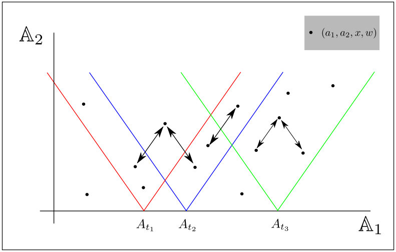
Inference

- The previous section suggests a Gibbs sampler where one conditionally updates the Z_r's, region and cluster assignments of observations and cluster parameters. The Gamma process is integrated out



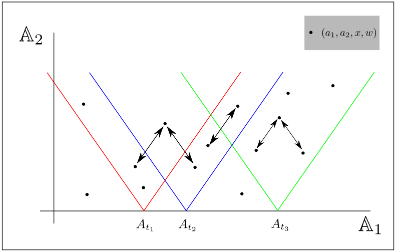
Inference

We also considered Metropolis-Hastings proposals



Inference

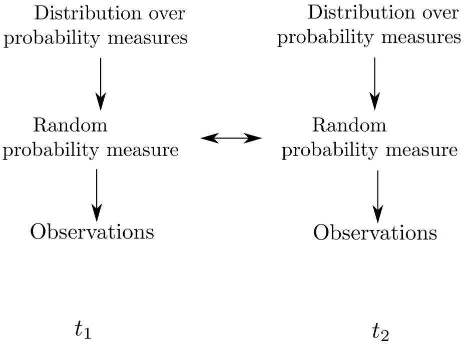
We also considered Metropolis-Hastings proposals



Also: slice sampling [Griffin and Walker, 2011]: associate with each observation a 'slice' variable $u_i \in [0, 1]$. We need instantiate only those weights greater than $\min_i u_i$

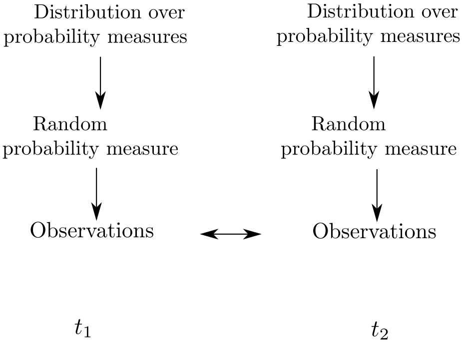
We saw two levels at which we can introduce dependence:

- At the level of the base-measure: RPMs at nearby points are similar on average
- One level below: the RPM realization itself is 'smooth'



We saw two levels at which we can introduce dependence:

- At the level of the base-measure: RPMs at nearby points are similar on average
- One level below: the RPM realization itself is 'smooth'

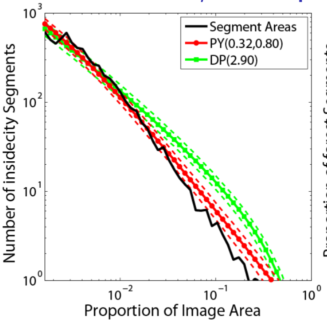


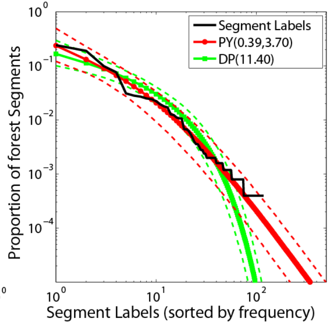
We can impose an even stronger dependence, at the level of the observations.

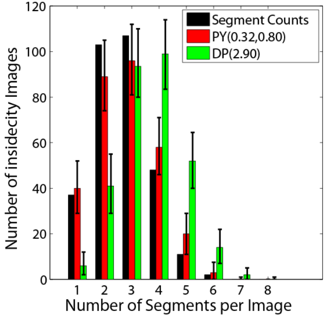
Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes.

[Sudderth and Jordan, 2008] (Generalizes [Duan et al., 2007] to the PY-process)

[Sudderth and Jordan, 2008]







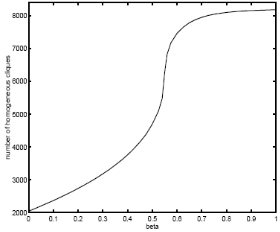


Figure 1. $< N(\mathbf{x}) >$ vs β for $64 \times 64 \times 4$-state Potts model

[Sudderth and Jordan, 2008]

Suppose $g \sim \mathcal{N}(0, 1)$

[Sudderth and Jordan, 2008]

Suppose $\quad g \sim \mathcal{N}(0,1)$

Let $\Phi(x)$ denote the standard normal cdf: $\quad u = \Phi(g) \sim \text{Unif}(0,1)$

$$P(u < V) = V, \quad P(u > V) = (1 - V) \quad \forall V \in [0,1]$$

[Sudderth and Jordan, 2008]

Suppose $\quad g \sim \mathcal{N}(0,1)$

Let $\Phi(x)$ denote the standard normal cdf: $\quad u = \Phi(g) \sim \text{Unif}(0,1)$

$$P(u < V) = V, \quad P(u > V) = (1 - V) \quad \forall V \in [0,1]$$

Now, let $\mathbf{p}$ be a stick-breaking RPM with $\quad p_i = V_i \prod_{j<i}(1 - V_j)$

[Sudderth and Jordan, 2008]

Suppose $g \sim \mathcal{N}(0,1)$

Let $\Phi(x)$ denote the standard normal cdf: $u = \Phi(g) \sim \text{Unif}(0,1)$

$$P(u < V) = V, \quad P(u > V) = (1 - V) \quad \forall V \in [0,1]$$

Now, let $\mathbf{p}$ be a stick-breaking RPM with $p_i = V_i \prod_{j<i}(1 - V_j)$

$$z \sim \mathbf{p} \implies P(z = i) = p_i = V_i \prod_{j<i}(1 - V_j)$$

[Sudderth and Jordan, 2008]

Suppose $\quad g \sim \mathcal{N}(0,1)$

Let $\Phi(x)$ denote the standard normal cdf: $\quad u = \Phi(g) \sim \text{Unif}(0,1)$

$$P(u < V) = V, \quad P(u > V) = (1 - V) \quad \forall V \in [0,1]$$

Now, let $\mathbf{p}$ be a stick-breaking RPM with $\quad p_i = V_i \prod_{j<i}(1 - V_j)$

$$\begin{aligned} z \sim \mathbf{p} \implies P(z = i) = p_i &= V_i \prod_{j<i}(1 - V_j) \\ &= P(u_i < V_i) \prod_{j<i} P(u_j > V_j) \end{aligned}$$

[Sudderth and Jordan, 2008]

Suppose $\quad g \sim \mathcal{N}(0,1)$

Let $\Phi(x)$ denote the standard normal cdf: $\quad u = \Phi(g) \sim \text{Unif}(0,1)$

$$P(u < V) = V, \quad P(u > V) = (1 - V) \quad \forall V \in [0,1]$$

Now, let $\mathbf{p}$ be a stick-breaking RPM with $\quad p_i = V_i \prod_{j<i}(1 - V_j)$

$$\begin{aligned} z \sim \mathbf{p} \implies P(z = i) = p_i &= V_i \prod_{j<i}(1 - V_j) \\ &= P(u_i < V_i) \prod_{j<i} P(u_j > V_j) \end{aligned}$$

$g_i \sim \mathcal{N}(0,1), \quad u_i = \Phi(g_i), \quad V_i \sim Beta(a_i, b_i), \quad$ (for appropriate (a_i, b_i))

$\implies \qquad z$ is a sample from the DP/PYP.

[Sudderth and Jordan, 2008]

$g_i \sim \mathcal{N}(0,1), \quad u_i = \Phi(g_i), \quad V_i \sim Beta(a_i, b_i)$
Thus, we need an infinite number of normals, one for each stick-break V_i

[Sudderth and Jordan, 2008]

$g_i \sim \mathcal{N}(0,1), \quad u_i = \Phi(g_i), \quad V_i \sim Beta(a_i, b_i)$
Thus, we need an infinite number of normals, one for each stick-break V_i

Associate with each image a PY processes i.e. a set $\{V_i, \theta_i^*\}$.
θ^*: Value of feature vector (colour and texture histograms) for a superpixel
Place a Dirichlet distribution prior on θ^*

[Sudderth and Jordan, 2008]

$g_i \sim \mathcal{N}(0,1), \quad u_i = \Phi(g_i), \quad V_i \sim Beta(a_i, b_i)$
Thus, we need an infinite number of normals, one for each stick-break V_i

Associate with each image a PY processes i.e. a set $\{V_i, \theta_i^*\}$.
θ^*: Value of feature vector (colour and texture histograms) for a superpixel
Place a Dirichlet distribution prior on θ^*

Define an infinite collection of GPs for each image. Let the marginal at each superpixel by $\mathcal{N}(0,1)$

[Sudderth and Jordan, 2008]

$g_i \sim \mathcal{N}(0, 1), \quad u_i = \Phi(g_i), \quad V_i \sim Beta(a_i, b_i)$
Thus, we need an infinite number of normals, one for each stick-break V_i

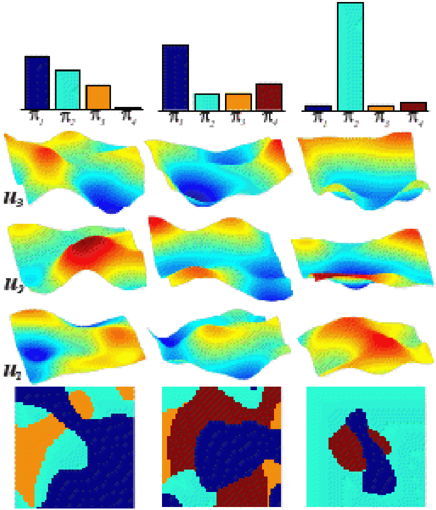
Associate with each image a PY processes i.e. a set $\{V_i, \theta_i^*\}$.
θ^*: Value of feature vector (colour and texture histograms) for a superpixel
Place a Dirichlet distribution prior on θ^*

Define an infinite collection of GPs for each image. Let the marginal at each superpixel by $\mathcal{N}(0, 1)$

Each superpixel s had an associated PY-distributed parameter $\theta(s)$.
Nearby GP values are similar $\implies$ Nearby θ are similar.

[Sudderth and Jordan, 2008]

$$P(z = i) = p_i = V_i \prod_{j<i}(1 - V_j) = P(u_i < V_i) \prod_{j<i} P(u_j > V_j)$$



Inference

Variational Bayes on a truncated stick-breaking representation
Expectation propagation

Bibliography I

Caron, F., Davy, M., and Doucet, A. (2007).
Generalized Polya urn for time-varying Dirichlet process mixtures.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 23.

Chung, Y. and Dunson, D. B. (2011).
The local Dirichlet process.
Annals of The Institute of Statistical Mathematics, 63:59–80.

Cifarelli, D. and Regazzini, E. (1978).
Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative.
Technical report, Quaderni Istituto Matematica Finanziaria dell'Università di Torino.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007).
Generalized spatial dirichlet process models.
Biometrika, 94(4):809–825.

Dunson, D. B. (2006).
Bayesian dynamic modeling of latent trait distributions.
Biostatistics, 7(4):551–68.

Bibliography II

Dunson, D. B. and Park, J.-H. (2008).
Kernel stick-breaking processes.
Biometrika, 95(2):307–323.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
Annals of Statistics, 1(2):209–230.

Griffin, J. E. and Steel, M. F. J. (2006).
Order-based dependent Dirichlet processes.
Journal of the American Statistical Association, Theory and Methods, 101:179–194.

Griffin, J. E. and Walker, S. G. (2011).
Posterior Simulation of Normalized Random Measure Mixtures.
Journal of Computational and Graphical Statistics, 20(1):241–259.

Kingman, J. F. C. (1975).
Random discrete distributions.
Journal of the Royal Statistical Society, 37:1–22.

Bibliography III

Kingman, J. F. C. (1993).
Poisson processes, volume 3 of *Oxford Studies in Probability*.
The Clarendon Press Oxford University Press, New York.
Oxford Science Publications.

MacEachern, S. (1999).
Dependent nonparametric processes.
In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association.

MacEachern, S., Kottas, A., and Gelfand, A. (2001).
Spatial nonparametric Bayesian models.
In *Proceedings of the 2001 Joint Statistical Meetings*.

Rao, V. and Teh, Y. W. (2009).
Spatial normalized gamma processes.
In *Advances in Neural Information Processing Systems*.

Bibliography IV

Sudderth, E. B. and Jordan, M. I. (2008).
Shared segmentation of natural scenes using dependent pitman-yor processes.
In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS*, pages 1585–1592. Curran Associates, Inc.