Spatial Normalized Gamma Process

Vinayak Rao
Yee Whye Teh

Presented at NIPS 2009

Discussion and Slides by Eric Wang
June 23, 2010
Outline

• Introduction

• Motivation

• The Gamma Process

• Spatial Normalized Gamma Process

• Inference

• Experiments

• Discussion
Introduction

• The Dirichlet Process (DP) is widely known and used in Bayesian nonparametrics.

• As originally proposed, the DP assumes the data are infinitely exchangeable (exchangeability).

• This assumption greatly simplifies model inference but is false in many situations.

• *Dependent DPs* were proposed by MacEachern with a dependent set of random measures that are all marginally DPs.

• The authors of this paper propose a general framework for constructing dependent DPs on arbitrary spaces.
Motivation

• The basic idea behind the author’s approach can be understood via an analogy: draws from a Dirichlet distribution (DD) can be constituted by drawing from a set of independent Gamma distributions and normalizing. A DP can be constructed by drawing from a Gamma process ($\Gamma P$) and normalizing.

• Conveniently, any restriction of a $\Gamma P$ is itself a $\Gamma P$. This allows the straightforward construction of dependent DPs in this paper.

• The resulting model is called the Spatial Normalized Gamma Process, parameter inference is straightforward and can be carried out with Gibbs and MH sampling.
The Gamma Process

• Consider \((\Theta, \Omega)\), a measure space on which we would like to define a \(\Gamma P\).

• Like the DP, draws from a \(\Gamma P\) consist of discrete atoms \(\theta \in \Theta\) and associated weights \(w > 0\). Here the weights do not have to sum to 1.

• Based on this, a sample from a \(\Gamma P\) is defined as

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

where \(\alpha\) is a base measure on \((\Theta, \Omega)\).

• The total mass \(G(S) = \sum_{i=1}^{\infty} w_i 1(\theta_i \in S)\) is Gamma distributed with shape parameter \(\alpha(S)\).

• A sample from a DP, \(D\), can be constituted as \(D = G/G(\Theta) \sim DP(\alpha)\).
The Gamma Process

- The base measure $\alpha$ is unusual in this case. Based on the introduced notation the usual (equivalent) parameterization of the DP are
  - Strength parameter $\alpha(\Theta)$.
  - Base distribution $\alpha/\alpha(\Theta)$.

- Two properties that will be useful:

  Firstly, if $S \in \Omega$ then the restriction $G'(d\theta) = G(d\theta \cap S)$ onto $S$ is a GP with base measure $\alpha'(d\theta) = \alpha(d\theta \cap S)$. Secondly, 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 $\Theta_1$ is also a GP with base measure $\alpha''(d\theta_1) = \int_{\Theta_2} \alpha(d\theta_1 d\theta_2)$.

- The basic idea of the above properties is that restrictions/projections of Gamma Processes are themselves Gamma Processes.

Wang 2010
Spatial Normalized Gamma Process

- Dependent DPs can be constructed as follows. Let \((\Theta, \Omega)\) define a probability space and \(\mathbb{T}\) an index space.

- The goal is construct one \(D_t\) for each \(t \in \mathbb{T}\) such that each \(D_t\) is marginally a DP.

- The authors’ approach is to define a gamma process \(G\) over an extended space and let each \(D_t\) be a normalized restriction/projection of \(G\). The previous property of Gamma Processes arising from restrictions and projections of Gamma Processes assures that each \(D_t\) will be DP distributed.
Spatial Normalized Gamma Process

• Let $\mathcal{Y}$ be an auxiliary space and for each $t \in \mathbb{T}$, let $Y_t \subset \mathcal{Y}$ be a measurable set, and let $\mu$ be an arbitrary measure over $\Theta \otimes \mathcal{Y}$. We can define a restriction projection $\mu_t$ as

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

• With this definition, let $\alpha$ be a base measure over the product space $\Theta \otimes \mathcal{Y}$ then the gamma process over $\Theta \otimes \mathcal{Y}$ is $G \sim \Gamma \text{P}(\alpha)$.

• We also have

$$
G_t(d\theta) = \int_{Y_t} G(d\theta dy) \sim \Gamma \text{P}(\alpha_t) \quad \text{and} \quad D_t = G_t / G_t(\Theta) \sim \text{D} \text{P}(\alpha_t)
$$

• The amount of dependence between $D_s$ and $D_t$ for $s, t \in \mathbb{T}$ is related to the amount of overlap between $Y_s$ and $Y_t$. 

Wang 2010
Spatial Normalized Gamma Process

• We now present two simple examples of SNGP (in time)

• Let \( T = \mathbb{R} \), \( Y = \mathbb{R} \), \( Y_t = [t - L, t + L] \), with \( L > 0 \). Defining the base measure \( \alpha(d\theta dy) = \gamma H(d\theta)dy/2L \), the restricted base measure is then

\[
\alpha_t(d\theta) = \frac{1}{2L} \int_{t-L}^{t+L} \gamma H(d\theta)dy = \gamma H(d\theta)
\]

• If the length of the time window \( L \) is not fixed then the base measure becomes

\[
\alpha_t(d\theta) = \int_{|y-t| \leq l} \gamma H(d\theta)dy \beta(dl)/2l = \int_0^{\infty} \beta(dl) \int_{t-l}^{t+l} dy/2l = \gamma H(d\theta)
\]

where \( \beta(dl) \) is a distribution over the window lengths in \([0, \infty)\).
Spatial Normalized Gamma Process

• In both cases, the resulting DP, $D_t = \frac{G_t}{G_t(\Theta)} \sim \text{DP}(\alpha_t)$ works out to be $D_t \sim \text{DP}(\gamma H)$.

• In the first example, each atom in the overall gamma process $G$ has time-stamp $y$ and time-span $[y - L, y + L]$, so it will only appear in those DPs within the time window $t \in [y - L, y + L]$.

• Two DPs $D_s$ and $D_t$ will share more atoms the closer $s$ and $t$ are to each other, and no atoms if $|s - t| > 2L$ (that is, the DPs are independent).

• In the second example, $D_s$ and $D_t$ will always be dependent, with the amount of dependence decreasing as $|s - t|$ increases.

Wang 2010
Spatial Normalized Gamma Process

- In practice, the auxiliary space is only observed as finite locations, therefore, the authors define $\mathcal{R}$ as the smallest collection of disjoint regions of $\mathcal{Y}$ s.t. each $Y_{t_j}$ is a union of subsets in $\mathcal{R}$.

- Let $\mathcal{R}_j$ be the collection of regions in $\mathcal{R}$ contained in $Y_{t_j}$ so $\bigcup_{R \in \mathcal{R}_j} = Y_{t_j}$.

- For each region $R$ in $\mathcal{R}$ define $G_R(d\theta) = G(d\theta \otimes R)$.

- Since each $G_R$ is itself a gamma process we have the following mixture-of-DPs interpretation of the model:

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

where $G_R/G_R(\Theta) \sim \text{DP}(\alpha_R)$, and base measure $\alpha_R(d\theta) = \alpha(d\theta \otimes R)$. 

Wang 2010
Spatial Normalized Gamma Process

• In this interpretation, each $D_{t_j}$ is a mixture where each component $D_R$ is drawn independently from a DP.

• The mixing proportions are Dirichlet distributed and independent.

• Based on this definition, the hierarchical construction of SNGP is

$$D_R \sim \text{DP}(\alpha_R) \quad g_R \sim \text{Gamma}(\alpha_R(\Theta))$$

$$D_{t_j} = \sum_{R \in \mathcal{R}_j} \pi_{jR} D_R$$

$$\pi_{jR} = \frac{g_R}{\sum_{R' \in \mathcal{R}_j} g_{R'}}$$

With $\alpha_R(d\theta) = \alpha(d\theta \otimes R)$.

• Notice here that all DPs are defined over $\Theta$. 
Inference

• The hierarchical model with auxiliary random variables is

\[
D_R \sim \text{DP}(\alpha_R) \quad g_R \sim \text{Gamma}(\alpha_R(\Theta)) \quad \text{for } R \in \mathcal{R}
\]

\[
D_{t_j} = \sum_{R \in \mathcal{R}_j} \pi_{jR} D_R \quad \pi_{jR} = \frac{g_R}{\sum_{R' \in \mathcal{R}_j} g_{R'}} \quad \text{for } R \in \mathcal{R}_j
\]

\[
r_{ji} \sim \text{Mult}([\pi_{jR} : R \in \mathcal{R}_j]) \quad \theta_{ji} \sim D_{r_{ji}} \quad x_{ji} \sim f_{\theta_{ji}}
\]

• Let \( c_{ji} \) indicate the cluster assignment of \( x_{ji} \) in \( D_{r_{ji}} \), then we sample \( r_{ji} \) and \( c_{ji} \) together
  – Existing cluster

\[
p(r_{ji} = R, c_{ji} = c | \text{others}) \propto \left( \frac{g_R}{\sum_{r \in \mathcal{R}_j} g_r} \right) \left( \frac{m_{Rc}^{-ji}}{m_{Rc}^{-ji} + \alpha_R(\Theta)} \right) f_{Rc}^{-ji}(x_{ji})
\]

  – New cluster

\[
p(r_{ji} = R, c_{ji} = c_{\text{new}} | \text{others}) \propto \left( \frac{g_R}{\sum_{r \in \mathcal{R}_j} g_r} \right) \left( \frac{\alpha_R(\Theta)}{m_{Rc}^{-ji} + \alpha_R(\Theta)} \right) f_{Rc_{\text{new}}}(x_{ji})
\]
Inference

• Notice that the mixing weights have been marginalized out. To sample the \( g_R \)'s takes is a bit more complicated. An additional random variable \( A_j \) is introduced and the sampling is

\[
g_R | \text{others} \sim \text{Gamma}(\alpha_R(\Theta) + m_{R_i}, 1 + \sum_{j \in J_R} A_j) \\
A_j | \text{others} \sim \text{Gamma}(m_{j..}, \sum_{R \in R_j} g_R)
\]

• To speed convergence and improve mixing, the authors discuss three different Metropolis-Hastings proposals. While not necessary to the implementation of the model, the MH steps offer empirical improvements over the standard Gibbs sampler inference.

• The three approaches operate on the clusters, proposing split/merges within a region, across regions, or both.
Experiments

- Synthetic Data: 60 data points where gathered at each of 5 times by sampling from a mixture of 10 Gaussians.

- The authors note that Gibbs sampling alone generally did not yield good mixing.

Figure 2: log-likelihoods (the coloured lines are ordered at iteration 80 like the legend).

Figure 4: Acceptance rates of the MH proposals for Gibbs+MH1+MH2+MH3 after burn-in (percentages).

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Synthetic</th>
<th>NIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH-Proposal 1</td>
<td>0.51</td>
<td>0.6621</td>
</tr>
<tr>
<td>MH-Proposal 2</td>
<td>11.7</td>
<td>0.6548</td>
</tr>
<tr>
<td>MH-Proposal 3</td>
<td>0.22</td>
<td>0.0249</td>
</tr>
</tbody>
</table>
Experiments

• NIPS dataset: 13 years of NIPS with 1740 documents, and 13000 unique tokens.

• Each document is modeled with a DP with base measure drawn from a single DP for that year (HDP). A SNGP prior is placed on the top level DPs.

• Each topic is a distribution over the words.

• Concentration parameters were set \textit{a priori} and not inferred.

• The average perplexity of the SNGP was 3023.4, while a standard HDP model achieved an average perplexity of 3046.5.
Experiments

- The inferred topics and timespans of the topics. The MH acceptance rates are generally between 1 and 5 percent.

- 100 topics were generated over the entire corpus.

- The model was initialized randomly.

- **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, idm, sequences
- **Topic H**: chain, protein, region, mouse, human, markov, sequence
- **Topic I**: routing, load, projection, forecasting, shortest, demand, packet
Discussion

• A simple framework for constructing dependent DPs based on normalized gamma processes.

• A contribution of this paper is that the proposed framework can be generalized to arbitrary spaces, whereas previous work constructed dependent DPs only on the real line.

• The interpretation of the proposed model as mixtures of DPs allows for straightforward parameter inference. The mixing weights of the DPs were marginalized out, and a Chinese Restaurant construction was used.

• Various MH proposals were tested to improve convergence and mixing.