Construction of Dependent Dirichlet Processes based on Poisson Processes

Dahua Lin   Eric Grimson   John Fisher

CSAIL
MIT

NIPS 2010 Outstanding Student Paper Award
Presented by Shouyuan Chen

Outline

Mixture Models: From Static to Dynamic

- Evolutionary clustering
  - add/remove clusters
  - movement of clusters
- Document modeling
  - add/remove topics
  - evolution of topics
- Other applications
  - image modeling
  - location base services
  - financial analysis

Dynamic Mixture Models

Model the behavior of latent components *overtime*

- Creation of new components.
- Removal of existing components.
- Variation of component parameters.

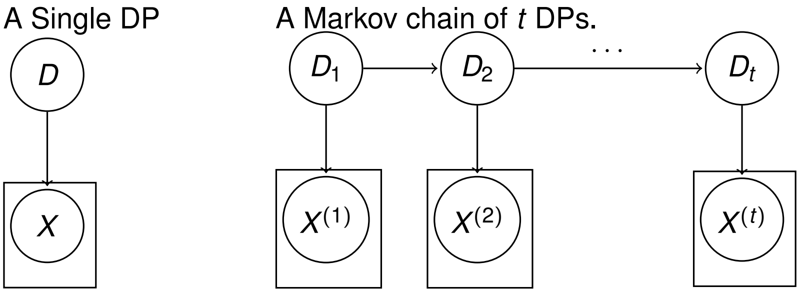
Components can be

- Clusters $\rightarrow$ Dynamic Gaussian Mixture Model
- Topics $\rightarrow$ Dynamic Topic Model

Dirichlet process (DP) $\approx$ infinite limit of Dirichlet distribution.

- Finite mixture models.
  - Prior: Dir($\vec{\alpha}$): *k*-dimensional Dirichlet distribution
  - Pre-specified number of components *k*.
- Dirichlet process mixture models (DPMM).
  - Prior: DP(α, H): "infinite dimensional Dirichlet distribution"
  - Learn hidden *k* automatically.

Extending DP to Dependent DPs



Key problem

How to design the Markov chain to support 3 key dependencies between $D_{t-1} \to D_t$:

Creation	Add a new component
Removal	Remove an existing component
Transition	Varying component parameters

Equivalent constructions for DP

Random measure Basic definition

Posterior Chinese restaurant process

Atomic construction Stick breaking process

Construct DP(μ) by ΓP and PP

- Generate compound poisson process PP($\mu \times \gamma$)
- Gamma process ΓP(μ) is transformed from compound poisson process
- Dirichlet process DP(μ) is normalized Gamma process

Given a measurable space (Ω, Σ, μ)

- **Compound Poisson Process**

$$\Pi^* \sim \text{PP}(\mu \times \gamma), \quad \gamma(dw) = w^{-1} e^{-w} dw$$

  Π is a point process (collection of infinite random points) on product space $\mu \times \gamma$

$$\Pi = \sum_{i=0}^{\infty} \delta_{(\theta, \omega_\theta)}$$

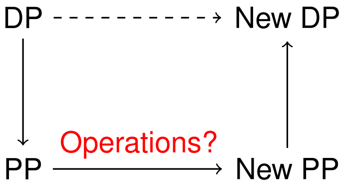
- **Gamma Process**: Transformed from compound poisson process

$$G \triangleq \sum_{(\theta, \omega_\theta) \in \Pi} \omega_\theta \delta_\theta \sim \Gamma\text{P}(\mu)$$

- **Dirichlet Process**: Normalized gamma process

$$D \triangleq G / G(\mu) \sim \text{DP}(\mu)$$

Key Idea for Transforming DPs



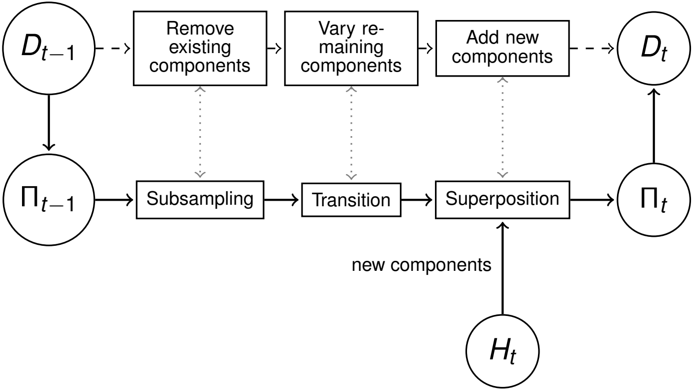
Complete randomness

A random measure of which the measure values of disjoint subsets are independent.

Complete Randomness Preserving Operations

Applying any operations that preserve complete randomness to Poisson processes results in a new Poisson process.

- Superposition two PP
- Subsampling a PP
- Mapping a PP point by point


D_{t-1}
Remove existing components
Vary remaining components
Add new components
D_t
Π_{t-1}
Subsampling
Transition
Superposition
Π_t
new components
H_t

Subsampling via Independent Bernoulli Trail

$\forall \eta = (\theta, p_\theta), \quad z_\eta \sim \text{Bernoulli}(q), \quad D = \sum_\eta p_\theta \delta_\theta \sim \text{DP}(\mu)$

$$S_q(D) \triangleq \frac{1}{\sum_{z_\eta = 1} p_\theta} \sum_{z_\eta = 1} p_\theta \delta_\theta$$

Theorem (Subsampling)

$$S_q(D) \sim \text{DP}(q\mu)$$

Proof sketch:

- DP $\to$ PP: $D \to \Pi \sim \text{PP}(\mu\gamma)$.
- **Subsampling PP**: $S_q(\Pi) = \{\eta \in \Pi : z_\eta = 1\} \sim \text{PP}(q\mu\gamma)$.
- PP $\to$ DP: $S_q(\Pi) \to S_q(D) \sim \text{DP}(q\mu)$

Independent movement of each point

$T(\cdot,\cdot)$: probabilistic transition kernel $D = \sum_\eta p_\theta \delta_\theta \sim \mathrm{DP}(\mu)$

$$T(D) \triangleq \sum p_\theta \delta_{T(\theta)}$$

Theorem (Transition)

$$T(D) \sim \mathrm{DP}(T\mu)$$

Proof sketch:

- DP $\rightarrow$ PP: $D \rightarrow \Pi \sim \mathrm{PP}(\mu \times \gamma)$.
- **Mapping PP**:
  $T(\Pi) = \{(T(\theta), \omega_\theta) : (\theta, \omega_\theta) \in \Pi\} \sim \mathrm{PP}(T\mu \times \gamma)$.
- PP $\rightarrow$ DP: $T(\Pi) \rightarrow T(D) \sim \mathrm{DP}(T\mu)$

Sum of independent DPs

$D_k \sim DP(\mu_k)$, $k = 1, \ldots, m$ be independent,
$(c_1, \ldots, c_m) \sim \mathrm{Dir}(\mu_1(\Omega), \ldots, \mu_m(\Omega))$

Theorem (Superposition)

$$\sum_k c_k D_k \sim \mathrm{DP}(\mu_1 + \ldots + \mu_m)$$

Proof sketch:

- DP $\to$ PP: $D_k \to \Pi_k \sim \mathrm{PP}(\mu_k \times \gamma)$.
- **Mapping PP**: $\sum_k g_k \Pi_k \sim \mathrm{PP}(\sum_k g_k \mu_k \times \gamma)$.
- PP $\to$ DP: $\frac{1}{\sum_k g_k} \sum g_k D_k = \sum c_k D_k \sim \mathrm{DP}(\sum_k \mu_k)$

Are All Poisson Things Necessary?

Basic definition of DP

$D \sim \mathrm{DP}(\mu)$ is a DP if for any partition $A_1, \ldots, A_n$ of space Ω

$$(D(A_1), \ldots, D(A_n)) \sim \mathrm{Dir}(\mu(A_1), \ldots, \mu(A_n))$$

Alternate proof of superposition theorem

Let $D = \sum_k c_k D_k$, consider any partition $A_1, \ldots, A_n$ of space Ω,

$$\begin{aligned}(D(A_1), \ldots, D(A_n)) &= \left(\sum_k c_k D_k(A_1), \ldots, \sum_k c_k D_k(A_n)\right) \\ &\sim \mathrm{Dir}(\sum_k \mu_k(A_1), \ldots, \sum_k \mu_k(A_n))\end{aligned}$$

The second step is from the property of Dirichlet distribution, and it concludes that $D \sim \mathrm{DP}(\sum_k \mu_k)$.

Proof by basic definition

- By defining DP on an extended space over functions, we can directly model all three operations: subsampling, transition and superposition without appealing to Poisson process.
- Such construction also allows DDP to be constructed over any measurable space. This paper is exactly a special case if the space is fixed to be a discrete Markov chain.

Outline

- Gibbs sampling. Sample one latent variable from posterior at each step. Consider time $1, \ldots, t$ sequentially.
- Update labels. Samples survived components (with probability q) and component assignments
- Update parameters. Samples component parameter from $T(.)$
- Iterates between step 2 and 3. Then move on to next time $t + 1$, and **never** estimate earlier distributions.

Sequential sampling

This paper doesn't derive a batch sampling algorithm. Earlier samples would likely be less accurate.

For simplicity of notation, and without loss of generality, assume the expectation of new components equals with removed components.

- Given a set of samples $\Phi \sim D_t$: ϕ_i appears c_i times
- (By DP posterior) $D_t | \Phi \sim \mathrm{DP}(\mu + \sum_k c_k \delta_{\phi_k})$
- (This paper) $D_{t+1} | \Phi \sim \mathrm{DP}(\mu + \sum_k q c_k \delta_{T(\phi_k)})$

Is that true?

Let $D_t = \sum_k p_k \delta_{\theta_k}$, $D_{t+1} = \frac{1}{\sum_{z_k=1} p_k} \sum_{z_k=1} p_k \delta_{\theta_k}$. $\phi \sim D_t$ is one sample from D_t.
Fact: $D_{t+1}|\phi$ is **not** DP.

Mixture of DP is not DP

Consider $z_\phi \sim$ Bernoulli(q). There are two different cases for $D_{t+1}|\phi$:

- $z_\phi = 1$. Thus ϕ is not removed. Thus ϕ is equivalently observed in D_{t+1}. $D_{t+1}|\phi, z_\phi = 1 \sim \text{DP}(\mu + \delta_\phi)$
- $z_\phi = 0$. In this case ϕ is removed. $D_{t+1}|\phi, z_\phi = 0 \sim \text{DP}(\mu)$

Hence $D_{t+1}|\phi$ is a mixture of DPs:

$$D_{t+1}|\phi = q\text{DP}(\mu + \delta_\phi) + (1 - q)\text{DP}(\mu)$$

It is proved **NOT** a DP. [1]

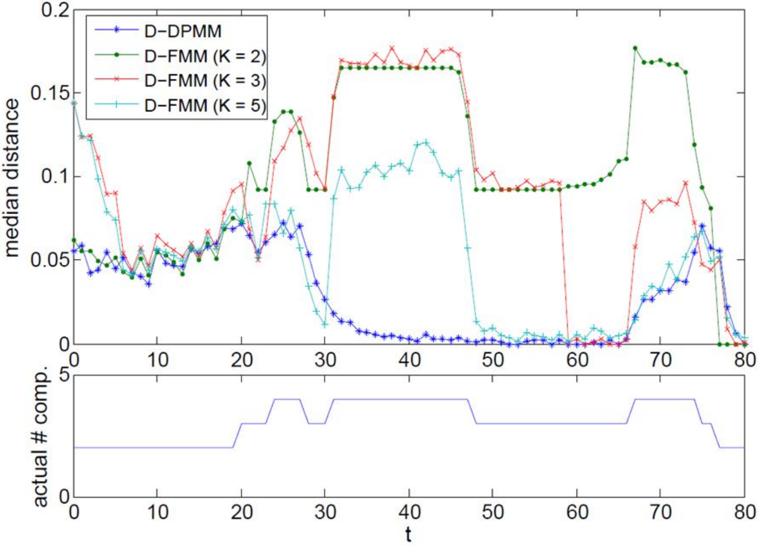
- The observation is censored.
- Only knows ϕ is not removed at now.
- The complete lifespan of a component ϕ is not observed.
- Posterior of DP under censored observations is a mixture of DP.

Setup

- Simulated over 80 phases.
- Gaussian mixture models with 2 components initially.
- The speed of introducing new components (one new component per 20 phases in average) and removing existing components is equal.
- Mean of component has a Brownian motion.
- 1000 samples per components at each phase.

Baselines

Finite mixture models with $K = 3, 5, 10$. DPM is not compared with.


D−DPMM
D−FMM (K = 2)
D−FMM (K = 3)
D−FMM (K = 5)
median distance
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80
actual # comp.
5
0
t

Evolutionary Topic Model

- Model topic evolution of research paper
- Data: all NIPS papers over years
- Method: feature extraction to generate 12 dimensions feature per document. Then use Gaussian mixture model.

People Flow

- The motion of people in New York Grand Central station.
- Data: 90,000 frames in one hour, divided into 60 phases.
- Try to group people tracks into flows depending on their motion patterns

Summary

- Propose a principled methodology to construct dependent Dirichlet processes based on the theoretical connections between Poisson, Gamma and Dirichlet processes.
- Develop a framework of evolving mixture model, which allows creation and removal of mixture components, as well as variation of parameters.
- Derive a Gibbs sampling algorithm for inferring mixture model parameters from observations.
- Test the approach on both synthetic data and real applications.

My Summary

- Poisson process is not essential for constructing DDP.
- Sequential sampling may damage the performance.
- Posterior of this model should be MDP rather than DP.

C. E. Antoniak
Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems
Annals of Statistics, 2(6):1152-1174, 1974.

A sample from DP is almost surely *discrete*.

Stick-breaking representation

Let $D \sim DP(\alpha, H)$ is a Dirichlet process. Then, almost surely

$$D = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$

$$p_{1\ldots\infty} \sim \text{GEM}(\alpha)$$

$$\forall i, \quad \theta_i \sim H, \quad \text{i.i.d}$$

The GEM distribution is called "stick-breaking" distribution.

Setup

- Simulated over 30 phases.
- Gaussian mixture models with 2 components initially.
- The speed of introducing new components (0.4 new component per phase in average) and removing existing components is equal.
- Mean of component has a Brownian motion.
- 200 samples per components at each phase.
- **Bias in posterior is fixed**

Baselines

DPM, Sequential sampling (Markov-DPM), Batch algorithm (F-DPM)

My Experiment Results