Dependent Normalized Random Measures

Changyou Chen[1,2]

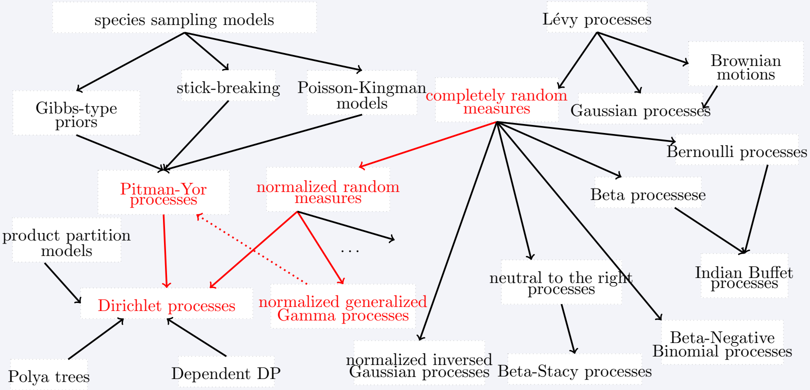
[1]ANU College of Engineering and Computer Science
The Australian National University
[2]National ICT, Australia

Joint work with Wray Buntine & Nan Ding

June 9, 2012

Outline

Nonparametric Bayesian family



Dynamic topic models

- Dynamic topic models try to model topic evolution over time.
- There are several related dynamic topic models, *e.g.*, Blei&Lafferty's DTM [BL06], Ahmed&Xing's iDTM [AX10].

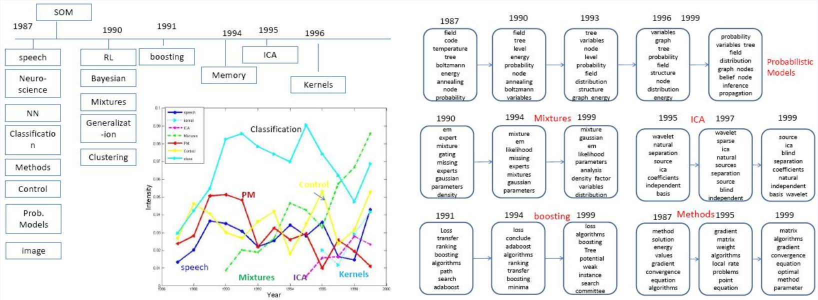


Figure: Topic evolution in NIPS, taken from [AX10].

Main contributions

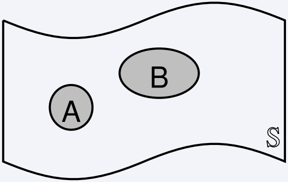
- Posterior analysis for normalized random measures.
- Develop dependent normalized random measures.
- Apply dependent normalized random measures to dynamic topic modeling to model *birth-death processes*, *dependency* and *power-law* phenomena in topic distributions over time.

Outline

Outline

Completely random measures

- Basic idea:
  - measurable space: $\mathbb{S}$.
  - disjoint subsets: $A, B \in \mathbb{S}$.
  - random function: $\Phi\colon \mathbb{S} \longmapsto \mathbb{R}^+$.



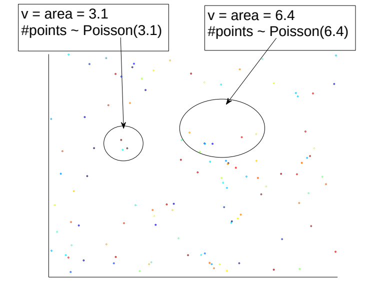
$$\Phi(A) \perp\!\!\!\perp \Phi(B)$$

- It is shown that completely random measures can be constructed from Poisson processes.

Completely random measures

Definition (Poisson Processes:)

A *Poisson process* on $\mathbb{S}$ is a random subset $\Pi \in \mathbb{S}$ such that if $N(A)$ is the number of points of Π in $A \subseteq \mathbb{S}$, then $N(A)$ is a Poisson random variable with mean $\nu(A)$, and $N(A_1), \cdots, N(A_n)$ are independent if $A_1, \cdots, A_n$ are disjoint.



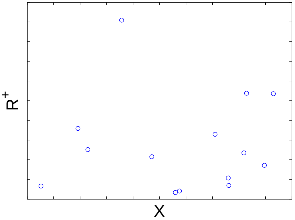
- Space: $\mathbb{S}$
- Positive measure: $\nu : \mathbb{S} \longmapsto \mathbb{R}^+$
- Poisson random measure: $N : \mathbb{S} \longmapsto$ integers
- $N(A) \sim \text{Poisson}(\nu(A))$

Completely random measures

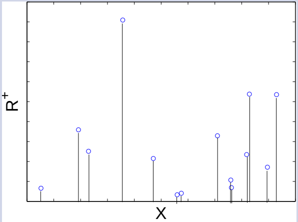
Definition (Construction from Poisson processes)

Let $N(\mathrm{d}t, \mathrm{d}x)$ being a Poisson random measure on a product space $\mathbb{S} = R^+ \times \mathbb{X}$ with mean measure $\nu(\mathrm{d}t, \mathrm{d}x)$. Construct a random measure $\tilde{\mu}$ to be a linear functional of $N(\mathrm{d}t, \mathrm{d}x)$ as

$$\tilde{\mu}(B) = \int_{\mathbb{R}^+ \times B} t N(\mathrm{d}t, \mathrm{d}x), \forall B \in \mathscr{B}(\mathbb{X}).$$



$\rightarrow$



$N(B) = \sum_{(J_k, x) \in \Pi \cap (\mathbb{R}^+ \times B)} \delta_{(J_k, x)}$ $\quad$ $\tilde{\mu}(B) = \sum_{(J_k, x) \in \Pi \cap (\mathbb{R}^+ \times B)} J_k \delta_x$

Completely random measures

$$\tilde{\mu}(B) = \int_{\mathbb{R}^+ \times B} tN(\mathrm{d}t, \mathrm{d}x) = \sum_{(J_k, x) \in \Pi \cap (\mathbb{R}^+ \times B)} J_k \delta_x$$

Proposition

$\tilde{\mu}$ is a completely random measure on $\mathbb{X}$.

- Call $\nu(\mathrm{d}t, \mathrm{d}x) = \rho(\mathrm{d}t|x)H(\mathrm{d}x)$ the *Lévy measure* of $\tilde{\mu}$.
- Taking different Lévy measures $\nu(\mathrm{d}t, \mathrm{d}x)$ we get different CRMs.

Example (Gamma CRM (Gamma processes))

A Gamma process on $\mathbb{X}$ is obtained by setting

$$\nu(\mathrm{d}t, \mathrm{d}x) = \frac{e^{-t}}{t} \mathrm{d}t H(\mathrm{d}x).$$

Sampling a CRM

$$\tilde{\mu}(B) = \sum_{(J_k, x) \in \Pi \cap (\mathbb{R}^+ \times B)} J_k \delta_x$$

- Cannot directly sample from the Lévy measure $\nu(\mathrm{d}x, \mathrm{d}t) = \rho(\mathrm{d}t|x)H(x)$ because it is improper.

Size biased sampling starting from the largest jump, then the second largest largest jump $\cdots$, given by Ferguson and Klass [FK72].

- Draw *i.i.d.* samples x_i from the base measure $H(\mathrm{d}x)$.
- The k-th largest jump has cumulative distribution function:

$$P(J_k \le j_k | J_{k-1} = j_{k-1}) = \exp\left\{ -\int_{\mathbb{X}} \int_{j_k}^{j_{k-1}} \nu(\mathrm{d}t, \mathrm{d}x) \right\} .$$

Normalized random measures

Definition (Normalized Random Measures (NRM))

An NRM is obtained by normalizing the CRM $\tilde{\mu}$ as:

$$\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})} = \sum_k \frac{J_k}{\sum_{k'} J_{k'}} \delta_{X_k^*} \ .$$

Definition (Normalized generalized Gamma processes (NGG))

A normalized generalized Gamma process is an NRM with *Lévy measure* being $\nu(\mathrm{d}t, \mathrm{d}x) = M \frac{e^{-t}}{t^{1+a}} H(\mathrm{d}x), (0 < a < 1)$[a].

[a] The general form is $\nu(\mathrm{d}t, \mathrm{d}x) = M \frac{e^{-bt}}{t^{1+a}} H(\mathrm{d}x)(0 < a < 1, b > 0)$, but b can be absorbed into M, thus we use $b = 1$.

- We denote a NRM with parameters a, M and base measure $H(\cdot)$ as $\mathrm{NRM}(a, M, H(\cdot))$.

Outline

Posterior analysis for the NGG

- General form for the posteriors of the NRM is developed in [JLP09], here we focus on NGG[1].

Theorem (Posterior of the NGG)

Consider the NGG$(a,M,H(\cdot))$. For a data vector $\vec{X}$ of length N there are K distinct values $X_1^*,\ldots,X_K^*$ with counts $n_1,\ldots,n_K$ respectively. The posterior marginal is given by

$$p\left(\vec{X}|NGG(a,M,H(\cdot)\right) = \frac{e^M a^{K-1} T_{a,M}^{N,K}}{\Gamma(N)} \prod_{k=1}^{K} (1-a)_{n_k-1} h(X_k^*) \ . \quad (1)$$

where

$$T_{a,M}^{N,K} = a\frac{M^K}{e^M} \int_{\mathbb{R}^+} \frac{u^{N-1}}{(1+u)^{N-Ka}} e^{M-M(1+u)^a} \mathrm{d}u \ . \quad (2)$$

[1] [FT12] also derives some similar results.

Posterior analysis for the NGG

- Compare NGG with PYP (Pitman-Yor process)

$$p(\vec{X}|\text{NGG},\cdots) = \frac{e^M a^{K-1} T_{a,M}^{N,K}}{\Gamma(N)} \prod_{k=1}^{K} (1-a)_{n_k-1} h(X_k^*) \ .$$

$$p(\vec{X}|\text{PYP},\cdots) = \frac{(b|a)_K}{(b)_N} \prod_{k=1}^{K} (1-a)_{n_k-1} h(X_k^*) \ .$$

Corollary (NGG$\longleftrightarrow$PYP)

Let $\vec{\mu} \sim$ NGG$(a, M, H(\cdot))$ and suppose $M \sim \Gamma(b/a, 1)$ then it follows that $\vec{\mu} \sim$ PYP$(a, b, H(\cdot))$

- If we also sample M using the prior $\Gamma(b/a, 1)$ for NGG, then we are sampling from a PYP.

Posterior analysis for the NGG

- This relationship is different from the Poisson-Kingman construction of the PYP, where it is constructed by exponentially tilting an σ-stable process, but we believe they are closely related.
- One problem of the above posterior sampling is the evaluation of $T^{N,K}_{a,M}$, which is computationally expensive and cannot easily be tabulated.

Conditional posterior for the NGG

- Likelihood of NGG: $\frac{\prod_{k=1}^{K} J_k^{n_k}}{\left(\sum_{k'=1}^{\infty} J_{k'}\right)^{\sum_{k=1}^{K} n_k}}$.
- A well studied auxiliary variable is introduced to eliminate this power term in the denominator. We call it *latent relative mass*.

Definition (Latent relative mass)

The latent relative mass is an auxiliary variable U_N defined as

$$U_N = \Gamma_N / (\sum_{k=1}^{\infty} J_k), \text{ where } \Gamma_N \sim \gamma(1, N)$$

- After a change of variable, we then have:

$$\frac{1}{(\sum_{k=1}^{\infty} J_k)^N} p(\Gamma_N) \mathrm{d}\Gamma_N = \exp\left\{-U_N \sum_{k=1}^{\infty} J_k\right\} \mathrm{d}U_N.$$

Conditional posterior for the NGG

Theorem (Conditional posterior)

Given NGG$(a, M, H(\cdot))$ and N observed data $\vec{X}$, assume there are K jumps such that $n_k > 0$, then (marginalize out jumps)

$$p\left(\vec{X}, U_N = u, K \,|\, N, \text{NGG}(a, M, H(\cdot))\right)$$
$$= \frac{u^{N-1}}{(1+u)^{N-Ka}} (Ma)^K e^{M - M(1+u)^a} \prod_{k=1}^{K} (1-a)_{n_k - 1} h(X_k^*) \ . \quad (3)$$

Moreover (retain jumps),

$$p\left(\vec{X}, U_N = u, K, J_1, \ldots, J_K \,|\, N, \text{NGG}(a, M, H(\cdot))\right)$$
$$= u^{N-1} \left(\frac{Ma}{\Gamma(1-a)}\right)^K e^{M - M(1+u)^a} \prod_{k=1}^{K} J_k^{n_k - a - 1} e^{-(1+u)J_k} h(X_k^*) \quad (4)$$

Posterior sampling for the NGG

$$Q = \sum_k J_k \delta_{X_k^*}$$

- Conditional posterior sampling for the NGG:
  - Sample the auxiliary variable:

$$p(U_N|\cdot) \propto \frac{U_N^{N-1}}{(1+U_N)^{N-Ka}} e^{-M(1+U_N)^a}.$$

  - For the jumps J_k with data attached:

$$J_k \sim \text{Gamma}(U_N + 1, n_k - \frac{1}{2}).$$

  - The rest of jumps form another NGG with an updated Lévy measure

$$e^{-U_N t} \nu(\mathrm{d}t, \mathrm{d}x),$$

    which is essentially $\sum_{k:n_k=0} J_k \delta_{X_k^*}$

Outline

Dependent NRMs

- There are several ways to construct dependent nonparametric Bayesian models.
- We use the standard dependency operations on Poisson processes to construct dependent NRMs, *e.g.*, *superposition*, *subsampling* and *point transition*.
- This has been used in dependent Dirichlet processes by Lin, Grimson and Fisher [LGF10].

Dependency operations

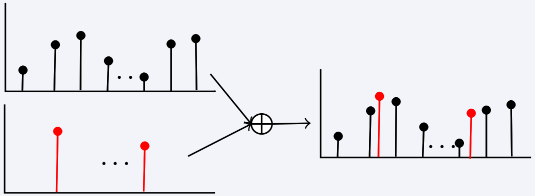
Definition (Superposition of NRMs)

Given n independent NRMs $\mu_1, \cdots, \mu_n$ on $\mathbb{X}$, the superposition ($\oplus$) is:

$$\mu_1 \oplus \mu_2 \oplus \cdots \oplus \mu_n := c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n \ .$$

where the weights[a] $c_m = \frac{\tilde{\mu}_m(\mathbb{X})}{\sum_j \tilde{\mu}_j(\mathbb{X})}$ and $\tilde{\mu}_m$ is the unnormalized random measures corresponding to μ_m.

[a] This is different from Lin *et al.*'s [LGF10]

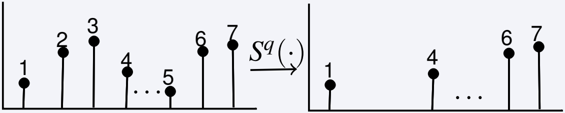


Dependency operations

Definition (Subsampling of NRMs)

Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, and a measurable function $q : \mathbb{X} \to [0,1]$. If we independently draw $z(\theta) \in \{0,1\}$ for each $\theta \in \mathbb{X}$ with $p(z(\theta)=1) = q(\theta)$, the subsampling of μ, is defined as

$$S^q(\mu) := \sum_{k:z(\theta_k)=1} \frac{r_k}{\sum_j z(\theta_j) r_j} \delta_{\theta_k}, \tag{5}$$

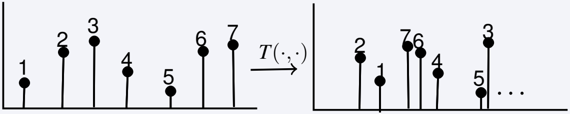


Dependency operations

Definition (Point transition of NRMs)

Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, the point transition of μ, is to draw atoms θ'_k from a transformed base measure to yield a new NRM as

$$T(\mu) := \sum_{k=1}^{\infty} r_k \delta_{\theta'_k} .$$



Properties of operations

Theorem (Posterior under superposition)

Let $\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_n$ be n independent CRMs on $\mathbb{X}$ with Lévy measures $\nu_i(\mathrm{d}t, \mathrm{d}x)$, $\tilde{\mu} = \oplus_{i=1}^n \tilde{\mu}_i$. The posterior of $\tilde{\mu}$ given observed data $\{(X_k^*, n_k)\}$ is given by

$$\tilde{\mu}_n + \sum_{k=1}^{K} J_k \delta_{X_k^*},$$

1. $\tilde{\mu}_n$: *a CRM with* $\nu(\mathrm{d}t, \mathrm{d}x) = e^{-ut}\left(\sum_{i=1}^n \nu_i(\mathrm{d}t, \mathrm{d}x)\right)$;
2. X_k^* :*fixed points;*
3. J_k : *jumps with densities proportional to* $t^{n_k} e^{-ut}\left(\sum_{i=1}^n \nu_i(\mathrm{d}t, \mathrm{d}x)\right)$.

Properties of operations

Theorem (Lévy measure under subsampling)

Let $\tilde{\mu} = \sum_{k=1}^{\infty} J_k \delta_{X_k^*}$ *be a CRM on* $\mathbb{X}$ *with Lévy measure* $\nu(\mathrm{d}t, \mathrm{d}x)$*,* $S^q(\tilde{\mu})$ *be its subsampling version with acceptance rate* $q(\cdot)$*, then* $S^q(\tilde{\mu})$ *has the Lévy measure of*

$$q(\mathrm{d}x)\nu(\mathrm{d}t, \mathrm{d}x) \ .$$

Properties of operations

Lemma (Applied on graphs)

- *Subsampling is commutative:*

$$S^{q'}(S^q(\tilde{\mu})) = S^q(S^{q'}(\tilde{\mu})) = S^{q'q}(\tilde{\mu}))$$

- *Transitions commute under constant subsampling rates:*

$$S^q(T(\tilde{\mu})) = T(S^q(\tilde{\mu}))$$

- *Subsampling and transition distribute over superposition:*

$$S^q(\tilde{\mu} \oplus \tilde{\mu}') = S^q(\tilde{\mu}) \oplus S^q(\tilde{\mu}'), \qquad T(\tilde{\mu} \oplus \tilde{\mu}') = T(\tilde{\mu}) \oplus T(\tilde{\mu}').$$

- *Superposition is commutative and associative:*

$$\tilde{\mu} \oplus \tilde{\mu}' = \tilde{\mu}' \oplus \tilde{\mu}, \qquad (\tilde{\mu} \oplus \tilde{\mu}') \oplus \tilde{\mu}'' = \tilde{\mu} \oplus (\tilde{\mu}' \oplus \tilde{\mu}'')$$

Outline

Outline

Applications: Dynamic topic models

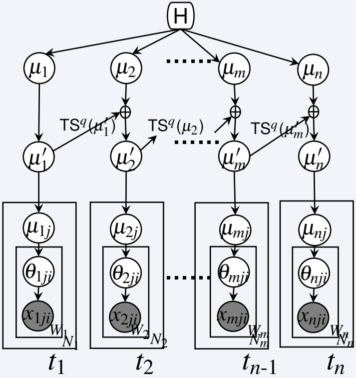
We want to model the following phenomenas in dynamic topic models:

- Birth-death processes.
- Dependency of topics between time frames (partially exchangeable).
- Power-law phenomena.

These objectives are well tackled by the dependent NRMs framework.

Model construction

$$\mu'_m = T(S^q(\mu'_{m-1})) \oplus \mu_m$$

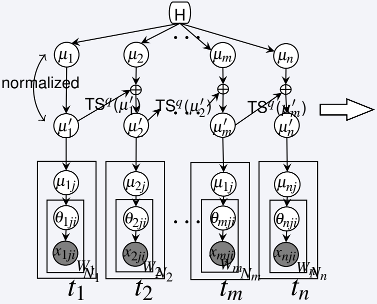


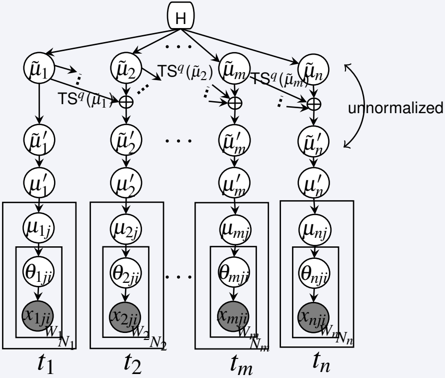
1. t_i: epochs.
2. μ_i: new topics at epoch i.
3. μ'_i: topic distribution for epoch i.
4. Each epoch has a hierarchical NRM structure.

Model construction

$$\mu'_m = T(S^q(\mu'_{m-1})) \oplus \mu_m$$

$$\tilde{\mu}'_m \sim \tilde{T}(\tilde{S}^q(\tilde{\mu}'_{m-1})) \oplus \tilde{\mu}_m,$$
$$\mu'_m = \frac{\tilde{\mu}'_m}{\tilde{\mu}'_m(\mathbb{X})}$$





Relation between NRMs and CRMs

Theorem

The following two generative processes are equivalent:

- *Manipulate the normalized random measures:*

$$\mu'_m \sim T(S^q(\mu'_{m-1})) \oplus \mu_m, \qquad \text{for } m > 1.$$

- *Manipulate the completely random measures:*

$$\tilde{\mu}'_m \sim \tilde{T}(\tilde{S}^q(\tilde{\mu}'_{m-1})) \oplus \tilde{\mu}_m, \quad \mu'_m = \frac{\tilde{\mu}'_m}{\tilde{\mu}'_m(\mathbb{X})} \quad \text{for } m > 1.$$

The resultant NRMs μ'_m's correspond to:

$$\mu'_m = \sum_{j=1}^{m} \frac{\left(q^{m-j}\tilde{\mu}_j\right)(\mathbb{X})}{\sum_{j'=1}^{m}\left(q^{m-j'}\tilde{\mu}_{j'}\right)(\mathbb{X})} T^{m-j}(\mu_j), \qquad \text{for } m > 1$$

Outline

Sampling

At each time frame m, we do:

- **Top level**: slice sample J_{mk} (ends up finite jumps).
- **Second Level**: subsample J'_{mk} by inheriting from $J_{m'k}(\text{top level}), m' \leq m$ with Bernoulli trials.
- **Third level**: construct μ'_m by normalizing J'_{mk}.
- **Hierarchical NRMs**:
  - sample topic assignments s_{mji} using a generalized Blackwell-MacQueen sampling scheme for the hierarchical NRM.
  - Sample #customers for the parent restaurant n'_{mk} by simulating a generalized Chinese restaurant process for the NRM.

Outline

Datasets

- Academic, news, Twitter, blog datasets.

dataset	vocab	docs	words	epochs
ICML	2k	765	44k	2007–2011
JMLR	2.4k	818	60k	12 vols
TPAMI	3k	1108	91k	2006–2011
NIPS	14k	2483	3.28M	1987-2003
Person	60k	8616	1.55M	08/96–08/97
Twitter_1	6k	3200	16k	14 months
Twitter_2	6k	3200	31k	16 months
Twitter_3	6k	3200	25k	29 months
BDT	8k	2649	234k	11/07–04/08

Table: Data statistics

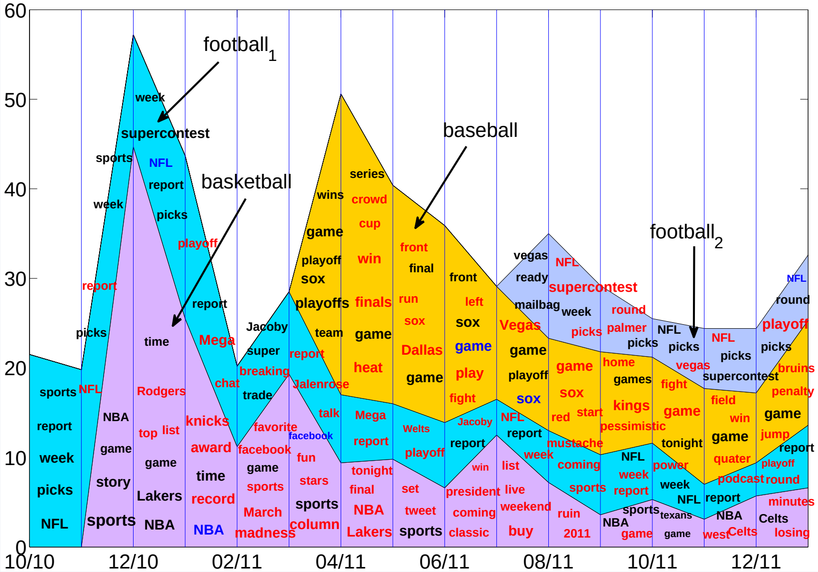
Experiments: Quantitative evaluation

- *DHNGG*: dependent hierarchical normalized generalized Gamma processes
- *DHDP*: dependent hierarchical Dirichlet processes
- *HDP*: hierarchical Dirichlet processes
- *DTM:* dynamic topic model[2]

Datasets	ICML	JMLR	TPAMI	NIPS	Person
DHNGG	**-5.3123e+04**	**-7.3318e+04**	**-1.1841e+05**	**-4.1866e+06**	**-2.4718e+06**
DHDP	-5.3366e+04	-7.3661e+04	-1.2006e+05	-4.4055e+06	-2.4763e+06
HDP	-5.4793e+04	-7.7442e+04	-1.2363e+05	-4.4122e+06	-2.6125e+06
DTM	-6.2982e+04	-8.7226e+04	-1.4021e+05	-5.1590e+06	-2.9023e+06
Datasets	Twitter_1	Twitter_2	Twitter_3	BDT	
DHNGG	**-1.0391e+05**	**-2.1777e+05**	**-1.5694e+05**	**-3.3909e+05**	
DHDP	-1.0711e+05	-2.2090e+05	-1.5847e+05	-3.4048e+05	
HDP	-1.0752e+05	-2.1903e+05	-1.6016e+05	-3.4833e+05	
DTM	-1.2130e+05	-2.6264e+05	-1.9929e+05	-3.9316e+05	

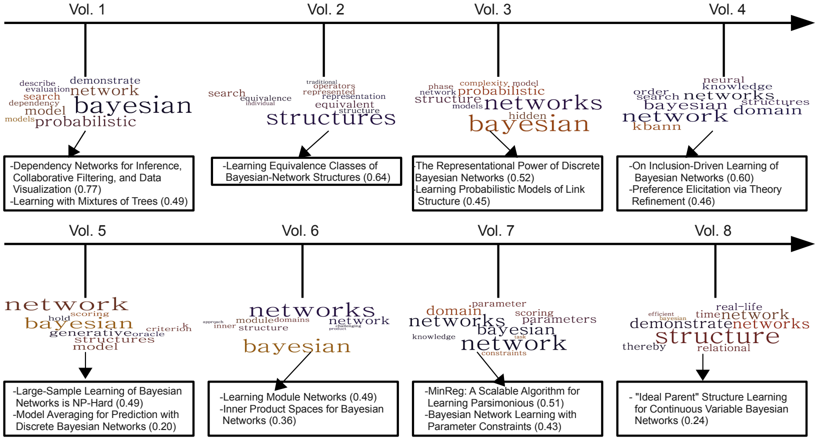
[2] Did not compare with Ahmed& Xing's iDTM [AX10], but ours is expected to be better since iDTM is comparable to HDP.

Experiments: Topic evolutions on Twitter



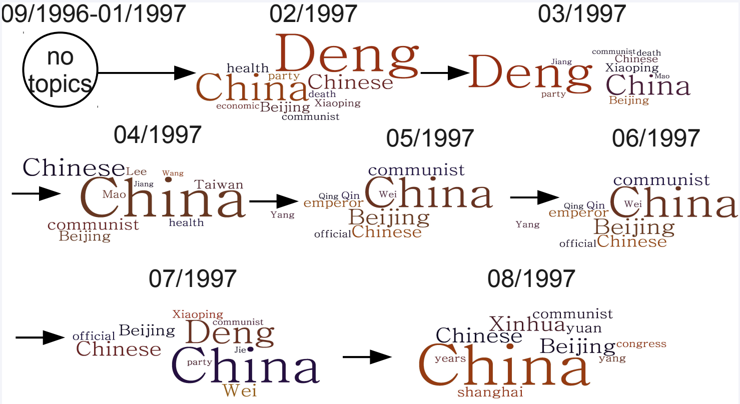
Experiments: Topic evolutions on JMLR

- 12 vol. from JMLR.



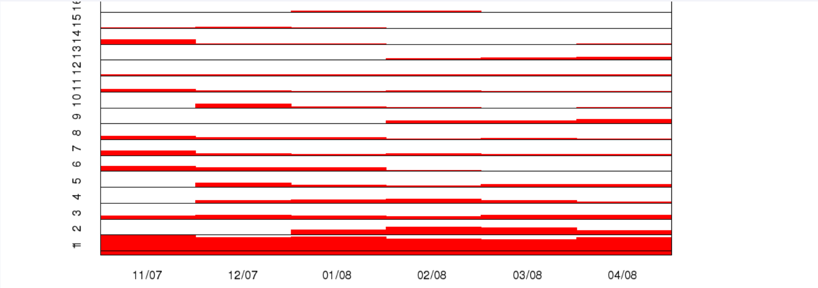
Experiments: Topic evolutions on Reuters

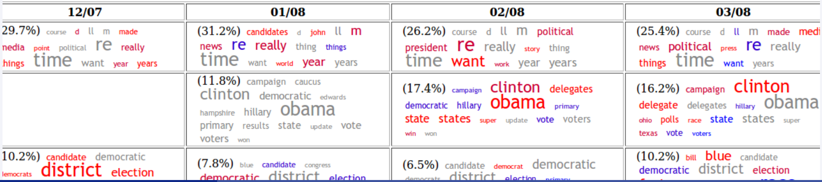
- 1 year (1996) news from Reuters.



Experiments: Topic evolutions on blogs

- 6 month blog data from Daily Kos blogs.




l have increased)

12/07	01/08	02/08	03/08
29.7%) course d ll m made nedia point political re really hings time want year years	(31.2%) candidates d john ll m news re really thing things time want world year years	(26.2%) course d ll m political president re really story thing time want work year years	(25.4%) course d ll m made medi news political press re really things time want years
	(11.8%) campaign caucus clinton democratic edwards hampshire hillary obama primary results state update vote voters won	(17.4%) campaign clinton delegates democratic hillary obama primary state states super update vote voters win won	(16.2%) campaign clinton delegate delegates hillary obama ohio polls race state states super texas vote voters
10.2%) candidate democratic lemocrats district election	(7.8%) blue candidate congress democratic district election	(6.5%) candidate democrat democratic democrats district election primary	(10.2%) bill blue candidate democratic district election

Outline

Conclusion

- Reviewed and developed theory of normalized random measures (NRM).
- Extended the Poisson based dependency operations to NRMs.
- Application on dynamic topic models.
- Developed a sampler for the proposed model.
- Future work include:
  - Develop more efficient sampler for NGG, specifically, for the dependent NGG.
  - Explore other ways of constructing dependent NRMs, *e.g.*, Lijoi, Nipoti and Prüster's dependent NRMs [LNP12], which is related but somehow different to ours.

References I

Ferguson, T. S., Klass, M. J.:
A representation of independent increment processes without Gaussian component.
The Annals of Mathematical Statistics (1972)

James, L. F., Lijoi, A., Prünster, I.:
Posterior analysis for normalized random measures with independent increments.
Scandinavian Journal of Statistics (2009)

Lin, D., Grimson, E., Fisher, J.:
Construction of Dependent Dirichlet Processes based on Poisson Processes.
Annual Conference on Neural Information Processing Systems (2010)

Ahmed, A., Xing, E.:
Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream.
Uncertainty in Artificial Intelligence (2010)

Blei, D., Lafferty, J.:
Dynamic topic models.
International Conference on Machine Learning (2006)

References II

Chen, C., Ding, N., Buntine, W.:
Dependent hierarchical normalized random measures for dynamic topic modeling.
International Conference on Machine Learning (2012)

Chen, C., Buntine, W., Ding, N.:
Theory of dependent hierarchical normalized random measures.
Technical Report arXiv:1205.4159, NICTA and ANU, Australia (2012)

Favaro, S., Teh, Y. W.:
MCMC for normalized random measure mixture models.
Submitted to the Statistical Science (2012)

Lijoi, A., Nipoti, A., Prüster, I.:
Bayesian inference with dependent normalized completely random measures.
Working paper (2012)

Thanks for your attention!!!