Dependent Normalized Random Measures

Changyou Chen[1,2]

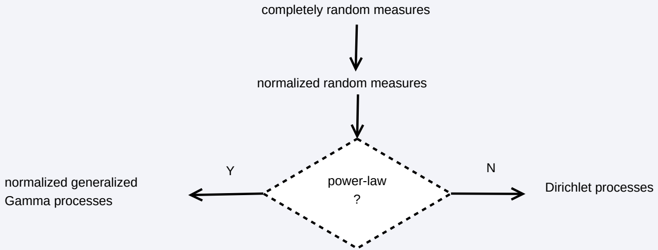
[1]ANU College of Engineering and Computer Science
The Australian National University
[2]National ICT, Australia

Joint work with Vinayak Rao, Wray Buntine & Yee Whye Teh
June 17, 2013

Outline

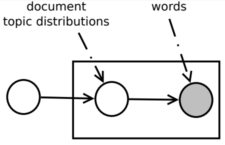
From Dirichlet processes to normalized random measures

- Why normalized random measures (NRM)?
  - More general and flexible
  - More convenient in dependency construction
  - Theoretically tractable



Motivation

- *Dependent normalize random measures* (dNRM) are useful in real applications to model dependent probability vectors:
  - Topic modeling: topic distributions of docs are dNRM.
  - Image annotation: annotation distributions and image feature distributions are dNRM.
  - ...



city,
lights
building,
night,
...

- Hierarchical Dirichlet processes (HDP):
  - Flexible, good performances.
  - Limitation: lack of some theoretical properties such as *marginal DPs*.

Related work

To name a few:

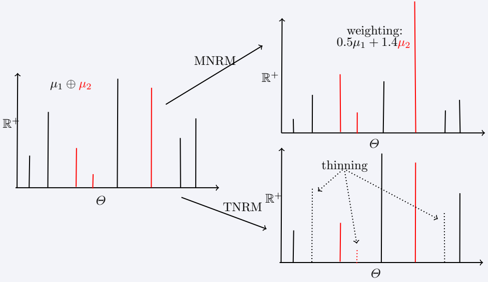
- Dependent DPs: [GriffinS06], [CaronDD07], [AX10], [LGF10, LinFisher12], [CDB12].
- Spatial DPs: [MacEachernKG01], [GelfandKM05], [RaoT09].

Limitations:

- Most limited to DP.
- Some are lacked of theoretical properties, *e.g.*, marginal DP.
- Some have theoretical flaws in model posteriors.
- See the paper for detailed analysis.

Contribution

- Propose two constructions of Dependent Normalized Random Measures
  - Mixed Normalized Random Measures (MNRM)
  - Thinned Normalized Random Measures (TNRM)
- Analyze their distributional properties
- Analyze their posterior structures
- Application in time series dynamic topic modeling



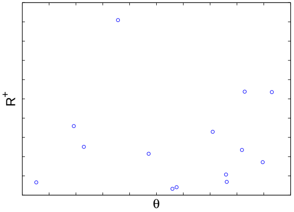
Outline

Completely random measures (CRM)

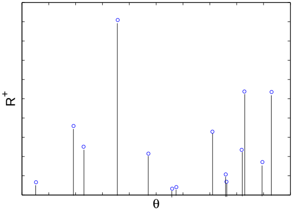
Definition (CRMs constructed from Poisson processes)

- $N(\mathrm{d}w, \mathrm{d}\theta)$: a Poisson random measure on $\mathbb{R}^+ \times \Theta$
- $\nu(\mathrm{d}w, \mathrm{d}\theta)$: intensity of the Poisson process, also the Lévy measure of the CRM

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



$\rightarrow$



$N(B) = \sum_{(w_k,\theta_k) \in \Pi \cap (\mathbb{R}^+ \times B)} \delta_{(w_k,\theta_k)}$ $\qquad$ $\tilde{\mu}(B) = \sum_{(w_k,\theta_k) \in \Pi \cap (\mathbb{R}^+ \times B)} w_k \delta_{\theta_k}$

Normalized random measures

Definition (Normalized Random Measure (NRM))

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

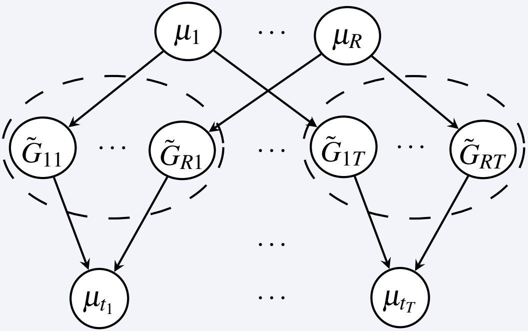
$$\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\Theta)} = \sum_k \frac{w_k}{\sum_{k'} w_{k'}} \delta_{\theta_k^*} \ .$$

- NRM is flexible in that it is the generalization of a number of well known stochastic processes, by varying its Lévy measures $\nu(\mathrm{d}w, \mathrm{d}\theta)$.
  - Dirichlet processes: $\nu(\mathrm{d}w, \mathrm{d}\theta) = \alpha w^{-1} e^{-w} \mathrm{d}w H(\mathrm{d}\theta)$.

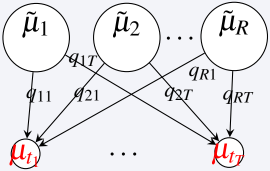
Outline

An Overview Construction

- $\mu_1, \cdots, \mu_R$: R independent NRMs, each generated from a *Region* r.
- $\tilde{G}_{1t_i}, \cdots, \tilde{G}_{Rt_i}$: intermediate results by applying weighting/thinning on μ_i's.
- μ_{t_i}: dependent NRM at time t_i.



dNRM-1: Mixed Normalized Random Measures



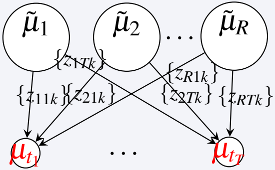
- Construction by weighting: $\tilde{\mu}_r \to q_{rt}\tilde{\mu}_r$

$$\tilde{\mu}_r(\mathrm{d}\theta) = \int_{\mathbb{R}^+} w N_r(\mathrm{d}w, \mathrm{d}\theta), \qquad \text{for each region } r$$

$$\hat{\mu}_t(\mathrm{d}\theta) = \sum_{k=1}^{\infty} q_{rt} w_{rk} \delta_{\theta_{rk}}, \qquad \text{for each time } t$$

$$\mu_t(\mathrm{d}\theta) = \frac{1}{Z_t}\hat{\mu}(\mathrm{d}\theta), \text{ where } Z_t = \tilde{\mu}_t(\Theta) \qquad \text{for each time } t$$

dNRM-2: Thinned Normalized Random Measures



- Construction by thinning $\tilde{\mu}_r$: keep an atom (w_k, θ_k) or not

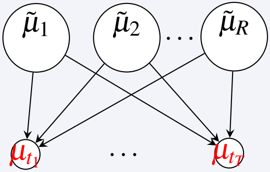
$$\tilde{\mu}_r(\mathrm{d}\theta) = \int_{\mathbb{R}^+} w N_r(\mathrm{d}w, \mathrm{d}\theta), \qquad \text{for each region } r$$

$$z_{rtk} \sim \text{Bernoulli}(q_{rt}), \qquad \text{for each atom } k$$

$$\tilde{\mu}_t(\mathrm{d}\theta) = \sum_{k=1}^{\infty} z_{rtk} w_{rk} \delta_{\theta_{rk}}, \qquad \text{for each time } t$$

$$\mu_t(\mathrm{d}\theta) = \frac{1}{Z_t} \hat{\mu}(\mathrm{d}\theta),, \text{ where } Z_t = \tilde{\mu}_t(\Theta) \qquad \text{for each time } t$$

Distributional Properties



- These constructions preserve the marginal NRM property.

Theorem (Marginal NRMs)

μ_t's in both MNRM and TNRM are marginally Normalized Random Measures[a]*.*

[a] By marginalizing out the independent μ_i's.

- One difference: if μ_i's are DP distributed, μ_t's in TNRM would follow DP distributions, but μ_t's in MNRM would not.

Conditional Posterior of MNRM

- With appropriate auxiliary variables[1], the posterior of MNRM is simple:

Theorem (Generalized CRP for MNRM)

Conditional on some auxiliary variables, the marginal posterior of μ_t's can be seen as generalizations of the Chinese restaurant process (CRP).

- Feasible marginal and slice samplers available for MNRM.

[1] see the paper for details.

Conditional Posterior Lévy Measure of TNRM

- Marginal posterior of TNRM is complex.
- No simple result as the MNRM.

Theorem (Generalized CRP for TNRM)

Conditional on some auxiliary variables, the marginal posterior of the μ_t is a mixture of 2^T generalized Chinese restaurant processes (CRP),where T is #times.

- Marginal sampler is infeasible.
- Thus only slice sampler is practical for posterior inference.

Outline

Datasets and settings

- Use a particular class of the NRM: normalized generalized Gamma process (NGG), inducing *power-law*.
- The dNRMs (MNGG, TNGG) are use to model topic distributions for each time.
- Academic, news datasets.

dataset	vocab	docs	words	epochs
ICML	2k	765	44k	2007–2011
TPAMI	3k	1108	91k	2006–2011
NIPS	14k	2483	3.28M	1987-2003
Person	60k	8616	1.55M	08/96–08/97

Table: Data statistics

Mixing behaviors of slice and marginal samplers

- Evaluated in terms of:
  - Effective Sample Size (ESS)[2] (the larger, the better)
  - Running times

		ICML		Person		NIPS	
		ESS	Time	ESS	Time	ESS	Time
Mixed	Marginal	57.4	66s	119.4	1.0h	111.1	1.5h
	Slice	**125.4**	69s	**212.9**	1.1h	**205.2**	1.9h
Thinned	Marginal	50.3	71s	144.8	1.3h	119.1	2.3h
	Slice	**94.9**	76s	**153.2**	1.1h	**176.1**	1.9h

- The slice sampler mixes better than the marginal sampler.
- The running times are comparable.

[2] used to evaluate the mixing behavior of the MCMC

Training and testing perplexity

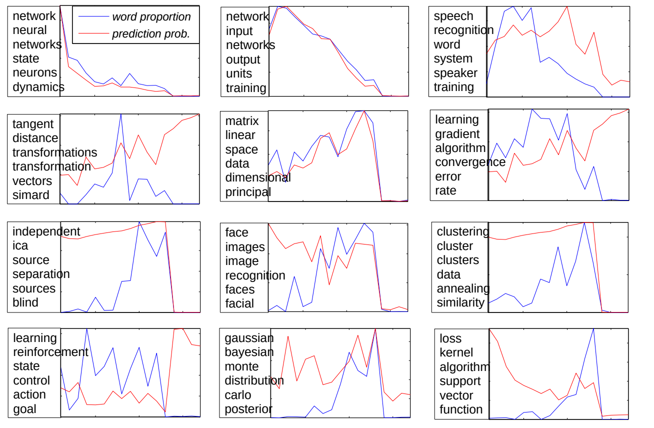
Datasets	ICML		Person		NIPS	
Models	train	test	train	test	train	test
HDP	580 ± 6	1017 ± 8	4541 ± 33	5962 ± 43	1813 ± 27	1956 ± 18
SNΓP[RaoT09]	550 ± 5	1007 ± 8	4324 ± 77	5733 ± 66	1406 ± 5	1679 ± 8
Thinned	572 ± 7	$\mathbf{945 \pm 7}$	4196 ± 29	5527 ± 47	1377 ± 5	1635 ± 3
Mixed	$\mathbf{535 \pm 6}$	1001 ± 10	$\mathbf{4083 \pm 36}$	$\mathbf{5488 \pm 44}$	$\mathbf{1366 \pm 8}$	$\mathbf{1618 \pm 5}$
MNGP	561 ± 10	995 ± 14	4118 ± 45	5519 ± 41	1370 ± 3	1634 ± 4

- The proposed models outperform related works, *e.g.*, HDP, SNΓP[3].
- Small datasets: Thinned $>$ Mixed.
- Large datasets: Mixed $>$ Thinned.
- Power-law distributions more flexible than non power-law counterparts (MNGP), but not obvious when modeling topic distributions.

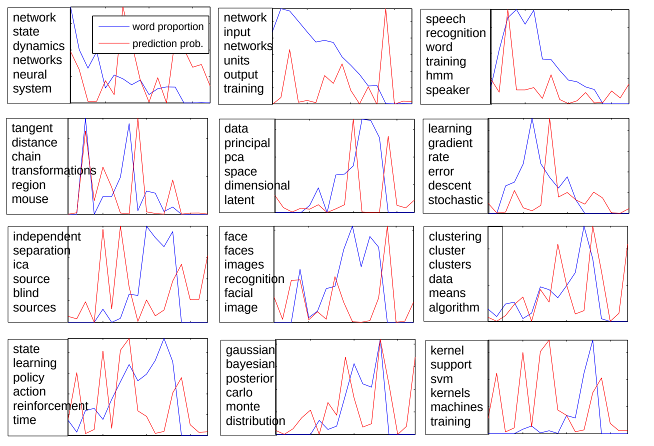
[3] Please refer to the paper for more comparisons.

Topic Evolutions on NIPS with MNRM

- word proportion: proportions of words allocated on a topic along *time*.
- prediction prob.: topic proportions evolving over time.



Topic Evolutions on NIPS with TNRM



- Mixed NRM produces smoother topic evolutions over time than Thinned NRM.

Conclusion

- Propose alternative ways to construction dNRMs, *e.g.*, by weighting and thinning.
- They are flexible, have nice theoretical properties.
- Posterior inference via slice sampler preferable.
- Many other application potentials, *e.g.*, modeling sparse distributional vectors, generalized IBPs.

References I

- 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)
- Chen, C., Ding, N., Buntine, W.:
  Dependent hierarchical normalized random measures for dynamic topic modeling.
  International Conference on Machine Learning (2012)
- Lin, D., Fisher, J.:
  Coupling Nonparametric Mixtures via Latent Dirichlet Processes.
  Annual Conference on Neural Information Processing Systems (2012)
- Rao, V., Teh, Y. W.:
  Spatial normalized Gamma processes.
  Annual Conference on Neural Information Processing Systems (2009)

References II

- Srebro, N., Roweis, S.:
  Time-varying topic models using dependent Dirichlet processes.
  Technical Report (2005)
- Griffin, J. E., Steel, M. F. J.:
  Order-based dependent Dirichlet processes.
  Journal of American Statistical Association (2006)
- Caron, F., Davy, M., Doucet, A:
  Generalized Polya urn for time-varying Dirichlet process mixtures.
  Uncertainty in Artificial Intelligence (2007)
- Gelfand, A. E., Kottas, A., MacEachern, S. N.:
  Bayesian nonparametric spatial modeling with Dirichlet process mixing.
  Journal of American Statistical Association (2005)
- MacEachern, S. N., Kottas, A., Gelfand, A. E.:
  Spatial nonparametric Bayesian models.
  Proceeding of the 2001 Joint Statistical Meetings (2001)

Thanks for your attention!!!