Scalable Deep Poisson Factor Analysis for Topic Modeling

Zhe Gan, Changyou Chen,
Ricardo Henao, David Carlson, Lawrence Carin

Duke University

July 9th, 2015

Outline

Introduction

Problem of interest: How to develop deep generative models for documents that are represented in bag-of-words form?

- **Directed Graphical Models:**
  - Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
  - Focused Topic Model (FTM) (Williamson et al., 2010)
  - Poisson Factor Analysis (PFA) (Zhou et al., 2012)
- **Going "Deep"?**
  - Hierarchical tree-structured topic models
  - nested Chinese Restaurant Process (nCRP) (Blei et al., 2004)
  - Hierarchical Dirichlet Process (HDP) (Teh et al., 2006)
  - nested Hierarchical Dirichlet Process (nHDP) (Paisley et al., 2015)
- How about we want to model general topic correlations?

- **Undirected Graphical Models:**
  - Replicated Softmax Model (RSM) (Salakhutdinov and Hinton, 2009b)
  - One generalization of the Restricted Boltzmann Machine (RBM) (Hinton, 2002)
- **Going Deep?**
  - Deep Belief Networks (DBN) (Hinton et al., 2006; Hinton and Salakhutdinov, 2011)
  - Deep Boltzmann Machines (DBM) (Salakhutdinov and Hinton, 2009a; Srivastava et al., 2013)
- Topics are not defined “properly”.

Main idea:

- Poisson Factor Analysis (PFA) + Deep Sigmoid Belief Network (SBN) or Restricted Boltzmann Machine (RBM).
- PFA is employed to interact with data at the bottom layer.
- Deep SBN or RBM serve as a flexible prior for revealing topic structure.

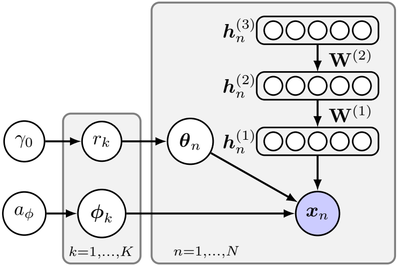


Figure: Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a *deep Boltzmann machine*.

Model Formulation

Poisson Factor Analysis: (Zhou et al., 2012)

- We represent a discrete matrix $\mathbf{X} \in \mathbb{Z}_+^{P \times N}$ containing counts from N documents and P words as

$$\mathbf{X} = \text{Pois}(\boldsymbol{\Phi}(\boldsymbol{\Theta} \circ \mathbf{H}^{(1)})) \,. \tag{1}$$

- Each column of $\boldsymbol{\Phi}$, ϕ_k, encodes the relative importance of each word in topic k.
- Each column of $\boldsymbol{\Theta}$, θ_n, contains relative topic intensities specific to document n.
- Each column of $\mathbf{H}^{(1)}$, $\boldsymbol{h}_n^{(1)}$, defines a sparse set of topics associated with each document.

Model Formulation

Poisson Factor Analysis: (Zhou et al., 2012)

- We construct PFAs by placing Dirichlet priors on ϕ_k and gamma priors on $\boldsymbol{\theta}_n$.

$$x_{pn} = \sum_{k=1}^{K} x_{pnk} \,, \quad x_{pnk} \sim \text{Pois}(\phi_{pk}\theta_{kn}h_{kn}^{(1)}) \,, \qquad (2)$$

  with priors specified as $\phi_k \sim \text{Dir}(a_\phi, \ldots, a_\phi)$, $\theta_{kn} \sim \text{Gamma}(r_k, p_n/(1-p_n))$, $r_k \sim \text{Gamma}(\gamma_0, 1/c_0)$, and $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$.

- Previously, a beta-Bernoulli process prior is defined on $\boldsymbol{h}_n^{(1)}$, assuming topic independence (Zhou and Carin, 2015).
- The **novelty** in our models comes from the prior for $\boldsymbol{h}_n^{(1)}$.

Model Formulation

Structured Priors on the Latent Binary matrix:

- Assume $\boldsymbol{h}_n^{(1)} \in \{0,1\}^{K_1}$, we define another hidden set of units $\boldsymbol{h}_n^{(2)} \in \{0,1\}^{K_2}$ placed at a layer "above" $\boldsymbol{h}_n^{(1)}$.
- **Modeling with the RBM:** (Undirected)

$$-E(\boldsymbol{h}_n^{(1)}, \boldsymbol{h}_n^{(2)}) = (\boldsymbol{h}_n^{(1)})^\top \boldsymbol{c}^{(1)} + (\boldsymbol{h}_n^{(1)})^\top \mathbf{W}^{(1)} \boldsymbol{h}_n^{(2)} + (\boldsymbol{h}_n^{(2)})^\top \boldsymbol{c}^{(2)}. \quad (3)$$

- **Modeling with the SBN** (Neal, 1992): (Directed)

$$p(h_{k_2 n}^{(2)} = 1) = \sigma(c_{k_2}^{(2)}), \quad (4)$$

$$p(h_{k_1 n}^{(1)} = 1 | \boldsymbol{h}_n^{(2)}) = \sigma\left((\boldsymbol{w}_{k_1}^{(1)})^\top \boldsymbol{h}_n^{(2)} + c_{k_1}^{(1)}\right). \quad (5)$$

Going Deep?

- Add multiple layers of SBNs or RBMs.

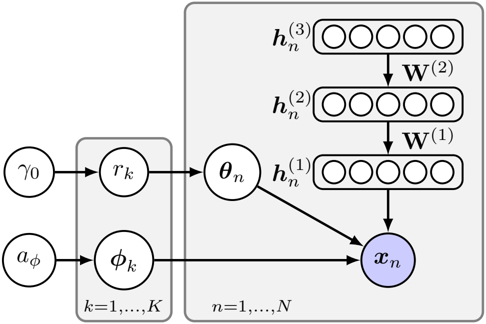


Figure: Graphical model for the Deep Poisson Factor Analysis with three layers of hidden binary hierarchies. The directed binary hierarchy may be replaced by a *deep Boltzmann machine*.

Scalable Posterior Inference

Challenge: Designing scalable Bayesian inference algorithms.
Solutions: Scaling up inference by stochastic algorithms.

- Applying Bayesian conditional density filtering algorithm (Guhaniyogi et al., 2014).
- Extending recently proposed work on stochastic gradient thermostats (Ding et al., 2014).

Scalable Posterior Inference

Bayesian conditional density filtering (BCDF):

- Repeatedly updating the surrogate conditional sufficient statistics (SCSS) using the current mini-batch.
- Drawing samples from the conditional posterior distributions of model parameters, based on SCSS.
- "stochastic Gibbs-style" updates.

Input: text documents, *i.e.*, a count matrix $\mathbf{X}$.
Initialize $\boldsymbol{\Psi}_g^{(0)}$ randomly and set $\mathbf{S}_g^{(0)}$ all to zero.
for $t = 1$ **to** ∞ **do**
  Get one mini-batch $\mathbf{X}^{(t)}$.
  Initialize $\boldsymbol{\Psi}_g^{(t)} = \boldsymbol{\Psi}_g^{(t-1)}$, and $\mathbf{S}_g^{(t)} = \mathbf{S}_g^{(t-1)}$.
  Initialize $\boldsymbol{\Psi}_l^{(t)}$ randomly.
  for $s = 1$ **to** S **do**
    Gibbs sampling for DPFA on $\mathbf{X}^{(t)}$.
    Collect samples $\boldsymbol{\Psi}_g^{1:S}$,$\boldsymbol{\Psi}_l^{1:S}$ and $\mathbf{S}_g^{1:S}$.
  end for
  Set $\boldsymbol{\Psi}_g^{(t)} = \mathrm{mean}(\boldsymbol{\Psi}_g^{1:S})$, and $\mathbf{S}_g^{(t)} = \mathrm{mean}(\mathbf{S}_g^{1:S})$.
end for

- $\boldsymbol{\Psi}_g$: global parameters
- $\boldsymbol{\Psi}_l$: local hidden variables
- $\mathbf{S}_g$: SCSS for $\boldsymbol{\Psi}_g$

Stochastic Gradient Nóse-Hoover Thermostats (SGNHT):

- Extending *Hamiltonian Monte Carlo* using stochastic gradient.
- Introducing thermostat to maintain system temperature.
- Adaptively absorbing stochastic gradient noise.
- The motion of the particles in the system are defined by the stochastic differential equations (SDE)

$$\mathrm{d}\boldsymbol{\Psi}_g = \boldsymbol{v}\mathrm{d}t\,, \qquad \mathrm{d}\boldsymbol{v} = \tilde{f}(\boldsymbol{\Psi}_g)\mathrm{d}t - \xi\boldsymbol{v}\mathrm{d}t + \sqrt{D}\mathrm{d}\mathcal{W}\,,$$
$$\mathrm{d}\xi = \left(\frac{1}{M}\boldsymbol{v}^T\boldsymbol{v} - 1\right)\mathrm{d}t\,, \tag{6}$$

where $\boldsymbol{\Psi}_g \in \mathbb{R}^M$ are model parameters, $\boldsymbol{v} \in \mathbb{R}^M$ are the momentum variables, $\tilde{f}(\boldsymbol{\Psi}_g) \triangleq -\nabla_{\boldsymbol{\Psi}_g}\tilde{U}(\boldsymbol{\Psi}_g)$, and $\tilde{U}(\boldsymbol{\Psi}_g)$ is the negative log-posterior.

Extension:

- Extending the SGNHT by introducing multiple thermostat variables $(\xi_1, \cdots, \xi_M)$ into the system such that each ξ_i controls one degree of the particle momentum.
- The proposed SGNHT is defined by the following SDEs

$$\begin{aligned} \mathrm{d}\boldsymbol{\Psi}_g &= \boldsymbol{v}\mathrm{d}t\,, \qquad \mathrm{d}\boldsymbol{v} = \tilde{f}(\boldsymbol{\Psi}_g)\mathrm{d}t - \boldsymbol{\Xi}\boldsymbol{v}\mathrm{d}t + \sqrt{D}\mathrm{d}\mathcal{W}\,, \\ \mathrm{d}\boldsymbol{\Xi} &= (\mathbf{q} - \mathbf{I})\,\mathrm{d}t\,, \end{aligned} \tag{7}$$

where $\boldsymbol{\Xi} = \text{diag}(\xi_1, \xi_2, \cdots, \xi_M)$, $\mathbf{q} = \text{diag}(v_1^2, \cdots, v_M^2)$

Theorem

The equilibrium distribution of the SDE system in (7) is

$$p(\boldsymbol{\Psi}_g, \boldsymbol{v}, \boldsymbol{\Xi}) \propto \exp\left(-\frac{1}{2}\boldsymbol{v}^\top \boldsymbol{v} - U(\boldsymbol{\Psi}_g) - \frac{1}{2}tr\left\{(\boldsymbol{\Xi} - D)^\top (\boldsymbol{\Xi} - D)\right\}\right).$$

Stochastic Gradient Nóse-Hoover Thermostats (SGNHT):

Input: text documents, *i.e.*, a count matrix $\mathbf{X}$.
Random Initialization.
for $t = 1$ **to** ∞ **do**
$\boldsymbol{\Psi}_g^{(t+1)} = \boldsymbol{\Psi}_g^{(t)} + \boldsymbol{v}^{(t)} h$.
$\boldsymbol{v}^{(t+1)} = \tilde{f}(\boldsymbol{\Psi}_g^{(t+1)})h - \boldsymbol{\Xi}^{(t)} \boldsymbol{v}^{(t)} h + \sqrt{2Dh}\mathcal{N}(0, \mathbf{I})$.
$\boldsymbol{\Xi}^{(t+1)} = \boldsymbol{\Xi}^{(t)} + (\boldsymbol{q}^{(t+1)} - \mathbf{I})h$, where $\boldsymbol{q} = \mathrm{diag}(v_1^2, \ldots, v_M^2)$.
end for

Scalable Posterior Inference

Stochastic Gradient Nóse-Hoover Thermostats (SGNHT):

Input: text documents, *i.e.*, a count matrix $\mathbf{X}$.
Random Initialization.
for $t = 1$ **to** ∞ **do**
$\boldsymbol{\Psi}_g^{(t+1)} = \boldsymbol{\Psi}_g^{(t)} + \boldsymbol{v}^{(t)} h.$
$\boldsymbol{v}^{(t+1)} = \tilde{f}(\boldsymbol{\Psi}_g^{(t+1)})h - \boldsymbol{\Xi}^{(t)} \boldsymbol{v}^{(t)} h + \sqrt{2Dh}\mathcal{N}(0, \mathbf{I}).$
$\boldsymbol{\Xi}^{(t+1)} = \boldsymbol{\Xi}^{(t)} + (\boldsymbol{q}^{(t+1)} - \mathbf{I})h$, where $\boldsymbol{q} = \mathrm{diag}(v_1^2, \ldots, v_M^2).$
end for

Discussion:

- **BCDF:** ease of implementation, but prefers the conditional densities for all the parameters.
- **SGNHT**: more general and robust, fast convergence.

Experiments

Datasets:

- **20 Newsgroups:** 20K documents with a vocabulary size of 2K.
- **RCV1-v2:** 800K documents with a vocabulary size of 10K.
- **Wikipedia:** 10M documents with a vocabulary size of 8K.

Experiments

Quantitative Evaluation:

Table: 20 Newsgroups.

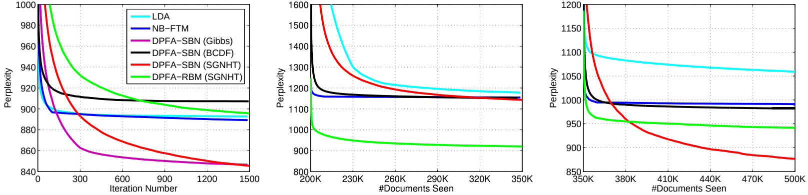
Model	Method	Dim	Perp.
DPFA-SBN-*t*	Gibbs	128-64-32	**827**
DPFA-SBN	Gibbs	128-64-32	**846**
DPFA-SBN	SGNHT	128-64-32	**846**
DPFA-RBM	SGNHT	128-64-32	896
DPFA-SBN	BCDF	128-64-32	905
DPFA-SBN	Gibbs	128-64	851
DPFA-SBN	SGNHT	128-64	850
DPFA-RBM	SGNHT	128-64	893
DPFA-SBN	BCDF	128-64	896
LDA	Gibbs	128	893
NB-FTM	Gibbs	128	887
RSM	CD5	128	877
nHDP	sVB	$(10,10,5)^{\diamond}$	889

Table: RCV1-v2 & Wikipedia.

Model	Method	Dim	RCV	Wiki
DPFA-SBN	SGNHT	1024-512-256	964	**770**
DPFA-SBN	SGNHT	512-256-128	1073	799
DPFA-SBN	SGNHT	128-64-32	1143	876
DPFA-RBM	SGNHT	128-64-32	**920**	942
DPFA-SBN	BCDF	128-64-32	1149	986
LDA	BCDF	128	1179	1059
NB-FTM	BCDF	128	1155	991
RSM	CD5	128	1171	1001
nHDP	sVB	$(10,5,5)^{\diamond}$	1041	932

Experiments

Quantitative Evaluation:



Sensitivity Analysis:

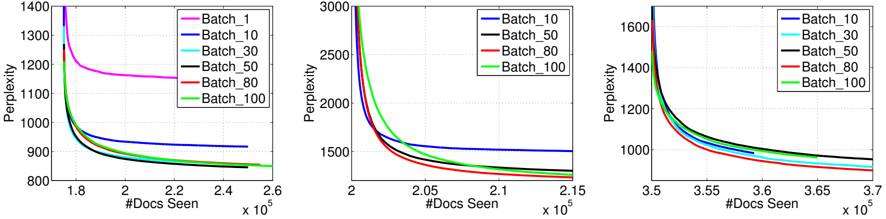


Figure: Perplexities. (Left) *20 News*. (Middle) *RCV1-v2*. (Right) *Wikipedia*.

Topics we learned on 20 Newsgroups:

T1	T3	T8	T9	T10	T14	T15	T19	T21	T24
year	people	group	world	evidence	game	israel	software	files	team
hit	real	groups	country	claim	games	israeli	modem	file	players
runs	simply	reading	countries	people	win	jews	port	ftp	player
good	world	newsgroup	germany	argument	cup	arab	mac	program	play
season	things	pro	nazi	agree	hockey	jewish	serial	format	teams
T25	T26	T29	T40	T41	T43	T50	T54	T55	T64
god	fire	people	wrong	image	boston	problem	card	windows	turkish
existence	fbi	life	doesn	program	toronto	work	video	dos	armenian
exist	koresh	death	jim	application	montreal	problems	memory	file	armenians
human	children	kill	agree	widget	chicago	system	mhz	win	turks
atheism	batf	killing	quote	color	pittsburgh	fine	bit	ms	armenia
T65	T69	T78	T81	T91	T94	T112	T118	T120	T126
truth	window	drive	makes	question	code	children	people	men	sex
true	server	disk	power	answer	mit	father	make	women	sexual
point	display	scsi	make	means	comp	child	person	man	cramer
fact	manager	hard	doesn	true	unix	mother	things	hand	gay
body	client	drives	part	people	source	son	feel	world	homosexual

Visualization:

Sports, Computers, and Poltics/Law.

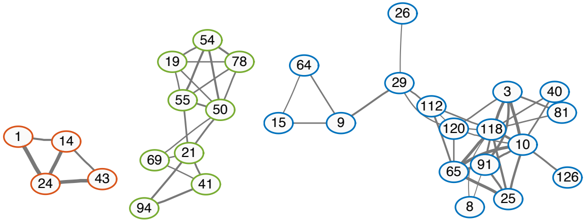


Figure: Graphs induced by the correlation structure learned by DPFA-SBN for the *20 Newsgroups*.

Summary

- **Model:** Deep Poisson Factor Analysis
  - PFA is employed to interact with data at the bottom layer.
  - Deep SBN or RBM serve as a flexible prior for revealing topic structure.
- **Scalable Inference:**
  - Bayesian conditional density filtering.
  - Stochastic gradient thermostats.

https://github.com/zhegan27/dpfa_icml2015

Questions?

References I

Blei, D. M., Griffiths, T., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *NIPS*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR*.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. (2014). Bayesian sampling using stochastic gradient thermostats. *NIPS*.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2014). Bayesian conditional density filtering. *arXiv:1401.3632*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*.

Hinton, G. E. and Salakhutdinov, R. (2011). Discovering binary codes for documents by learning deep generative models. *Topics in Cognitive Science*.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*.

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. (2015). Nested hierarchical Dirichlet processes. *PAMI*.

References II

Salakhutdinov, R. and Hinton, G. E. (2009a). Deep Boltzmann machines. *AISTATS*.

Salakhutdinov, R. and Hinton, G. E. (2009b). Replicated softmax: an undirected topic model. *NIPS*.

Srivastava, N., Salakhutdinov, R., and Hinton, G. E. (2013). Modeling documents with deep Boltzmann machines. *UAI*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *JASA*.

Williamson, S., Wang, C., Heller, K., and Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. *ICML*.

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *PAMI*.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. *AISTATS*.