# Deep Poisson Factor Modeling

Ricardo Henao, Zhe Gan, James Lu and Lawrence Carin

Department of Electrical and Computer Engineering, Duke University, Durham, NC

## Contributions

- A deep architecture for topic models based entirely on Poisson Factor Analysis (PFA) modules.
- Inherent shrinkage in all layers, thanks to the DP-like formulation of PFA.
- Block updates for binary units improve mixing.
- PFA modules can be used to easily build discriminative topic models.
- Efficient MCMC inference scales as function of the number of *non-zeros* in data and binary units.
- Scalable Bayesian inference algorithm based on Stochastic Variational Inference (SVI).

## Poisson factor analysis as a module

Assume $\mathbf{x}_n$ is an $M$-dimensional vector containing word counts for the $n$-th of $N$ documents, where $M$ is the vocabulary size. We impose the model

$$\mathbf{x}_n \sim \text{Poisson}\left(\mathbf{\Psi}(\boldsymbol{\theta}_n \circ \mathbf{h}_n)\right) , \tag{1}$$

where

- $\mathbf{\Psi} \in \mathbb{R}_+^{M\times K}$, factor loadings matrix with $K$ factors.
- $\boldsymbol{\theta}_n \in \mathbb{R}_+^{K}$, factor intensities.
- $\mathbf{h}_n \in \{0,1\}^K$, binary units indicating which factors are active for observation $n$.
- Symbol $\circ$ denotes element-wise (Hadamard) product.

Prior specification [2]:

$$\begin{aligned} x_{mn} = \textstyle\sum_{k=1}^K x_{mkn} , \quad x_{mkn} \sim \text{Poisson}(\lambda_{mkn}) , \quad \lambda_{mkn} = \psi_{mk}\theta_{kn}h_{kn} , \\ \boldsymbol{\psi}_k \sim \text{Dirichlet}(\eta \mathbf{1}_M) , \quad \theta_{kn} \sim \text{Gamma}(r_k, (1-b)b^{-1}) , \quad h_{kn} \sim \text{Bernoulli}(\pi_{kn}) . \end{aligned} \tag{2}$$

Note that $\eta$ controls for the sparsity of $\mathbf{\Psi}$, while $r_k$ accommodates for over-dispersion in $\mathbf{x}_n$ via $\boldsymbol{\theta}_n$.

**PFA module:** Conditioned on $\mathbf{h}_n$, we express

$$\mathbf{x}_n \sim \text{PFA}(\mathbf{\Psi}, \boldsymbol{\theta}_n, \mathbf{h}_n; \eta, r_k, b) . \tag{3}$$

## Deep representations with PFA modules

Develop a deep prior specification for $\mathbf{h}_n$ as

$$\begin{aligned} \mathbf{x}_n &\sim \text{PFA}\left(\mathbf{\Psi}^{(1)}, \boldsymbol{\theta}_n^{(1)}, \mathbf{h}_n^{(1)}; \eta^{(1)}, r_k^{(1)}, b^{(1)}\right) , & \mathbf{h}_n^{(1)} &= \mathbf{1}\left(\mathbf{z}_n^{(2)}\right) , \\ \mathbf{z}_n^{(2)} &\sim \text{PFA}\left(\mathbf{\Psi}^{(2)}, \boldsymbol{\theta}_n^{(2)}, \mathbf{h}_n^{(2)}; \eta^{(2)}, r_k^{(2)}, b^{(2)}\right) , & &\vdots \\ &\vdots & \mathbf{h}_n^{(L-1)} &= \mathbf{1}\left(\mathbf{z}_n^{(L)}\right) , \\ \mathbf{z}_n^{(L)} &\sim \text{PFA}\left(\mathbf{\Psi}^{(L)}, \boldsymbol{\theta}_n^{(L)}, \mathbf{h}_n^{(L)}; \eta^{(L)}, r_k^{(L)}, b^{(L)}\right) , & \mathbf{h}_n^{(L)} &= \mathbf{1}\left(\mathbf{z}_n^{(L+1)}\right) , \end{aligned} \tag{4}$$

where

- Function $\mathbf{1}(\cdot)$ is defined component-wise as

$$h_{nk}^{(\ell)} = 1 \;\text{ if }\; z_{nk}^{\ell+1} > 0 , \qquad \text{otherwise} \;\; h_{nk}^{(\ell)} = 0 . \tag{5}$$

- For top layer

$$z_{kn}^{(L+1)} \sim \text{Poisson}\left(\lambda_k^{(L+1)}\right) , \qquad \lambda_k^{(L+1)} \sim \text{Gamma}\left(a_0, b_0\right) . \tag{6}$$

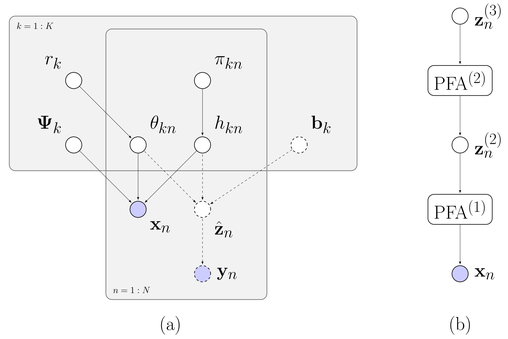
Binary units are constituted as [1]

$$h_{kn}^{(\ell-1)} = 1\left(z_{kn}^{(\ell)} \geq 1\right) , \qquad z_{kn}^{(\ell)} \sim \text{Poisson}\left(\tilde{\lambda}_{kn}^{(\ell)}\right) , \qquad \tilde{\lambda}_{kn}^{(\ell)} = \sum_{k'=1}^{K_\ell} \psi_{kk'}^{(\ell)} \theta_{k'n}^{(\ell)} h_{k'n}^{(\ell)} . \tag{7}$$

Equivalently

$$p(h_{kn}^{(\ell-1)} = 1) = \text{Bernoulli}\left(\pi_{kn}^{(\ell)}\right) , \qquad \pi_{kn}^{(\ell)} = 1 - \exp\left(-\tilde{\lambda}_{kn}^{(\ell)}\right) . \tag{8}$$

**Inference:** Analytic Gibbs updates due to local conjugacy. SVI for large datasets.




(a) (b)

FIGURE 1: Graphical models. (a) PFA module. Nodes ($\mathbf{b}_k$, $\mathbf{z}_n$ and $\mathbf{y}_n$) and edges drawn with dashed lines correspond to the discriminative PFA. (b) DPFM.

## PFA modules for discriminative tasks

Assume that there is a label $y_n \in \{1, \ldots, C\}$ associated with document $n$. We impose the model

$$\widehat{\mathbf{y}}_n \sim \text{Multinomial}\left(1, \widehat{\boldsymbol{\lambda}}_n\right) , \qquad \widehat{\lambda}_{cn} = \lambda_{cn} / \textstyle\sum_{c=1}^C \lambda_{cn} , \tag{9}$$

where

- $y_n$ is represented as a $C$-dimensional *one-hot* vector, $\widehat{\mathbf{y}}_n$.
- $\boldsymbol{\lambda}_n = \mathbf{B}(\boldsymbol{\theta}_n^{(1)} \circ \mathbf{h}_n^{(1)})$ and $\lambda_{cn}$ is element $c$ of $\boldsymbol{\lambda}_n$.
- $\mathbf{B} \in \mathbb{R}_+^{C\times K}$, matrix of nonnegative classification weights.
- $\mathbf{b}_k \sim \text{Dirichlet}(\zeta \mathbf{1}_C)$, for $\mathbf{b}_k$ column of $\mathbf{B}$.

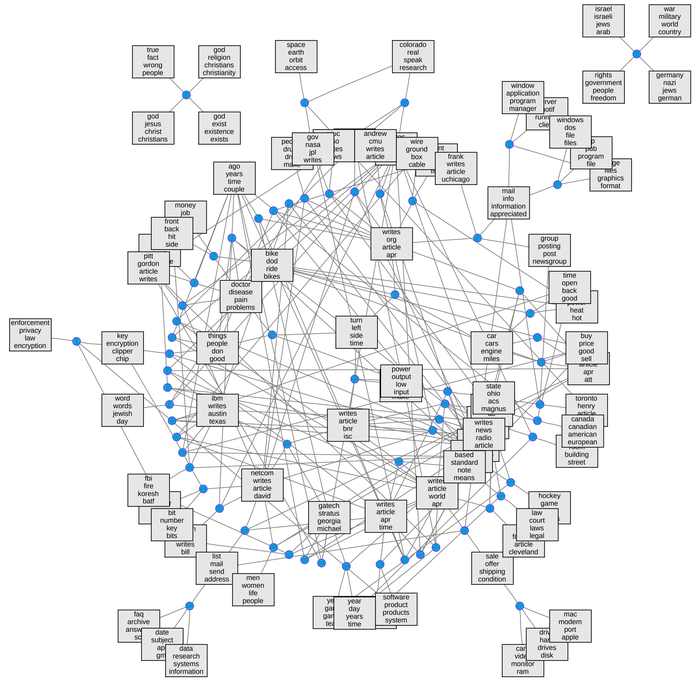



TABLE 1: Graph representation obtained from 20 News. Meta-topics are denoted by circles and layer-1 topics as boxes, with word lists corresponding to the top four words in layer-1 topics, $\boldsymbol{\psi}_k^{(1)}$. We only show the top four connections between meta-topics and their associated topics

## Experiments

### Benchmark corpora

- Data:
  - 20 Newsgroups (20 News): 2,000 words, 11,315/7,531 training/test documents.
  - Reuters corpus volume I (RCV1): 10,000 words, 794,414/10,000 training/test documents.
  - Wikipedia (Wiki): 7,702 words, $10^7$/1,000 training/test documents.
- Performance: held-out perplexity on 20% of test set.
- Models: LDA, FTM, RSM, nHDP, DPFA-SBN, DPFA-RBM and DPFM.
- Runtime: one iteration of the two-layer DPFM on 20 News takes approx. 3/2 secs, for MCMC/SVI.

TABLE 2: Held-out perplexities for 20 News, RCV1 and Wiki. Size: number of topics and/or binary units.

| Model | Method | Size | 20 News | RCV1 | Wiki |
|---|---|---|---|---|---|
| DPFM | SVI | 128-64 | 818 | 961 | 791 |
| DPFM | MCMC | 128-64 | **780** | **908** | 783 |
| DPFA-SBN | SGNHT | 1024-512-256 | — | 942 | **770** |
| DPFA-SBN | SGNHT | 128-64-32 | 827 | 1143 | 876 |
| DPFA-RBM | SGNHT | 128-64-32 | 896 | 920 | 942 |
| nHDP | SVI | (10,10,5) | 889 | 1041 | 932 |
| LDA | Gibbs | 128 | 893 | 1179 | 1059 |
| FTM | Gibbs | 128 | 887 | 1155 | 991 |
| RSM | CD5 | 128 | 877 | 1171 | 1001 |

### Classification

- Data: 20 News for document classification.
- Performance: test accuracy.
- Models: LDA, DocNADE, RSM, OSM and DPFM.

TABLE 3: Test accuracy on 20 News. Subscript accompanying model names indicate their size.

| Model | $\text{LDA}_{128}$ | $\text{DocNADE}_{512}$ | $\text{RSM}_{512}$ | $\text{OSM}_{512}$ | $\text{DPFM}_{128}$ | $\text{DPFM}_{128-64}$ |
|---|---|---|---|---|---|---|
| Accuracy (%) | 65.7 | 68.4 | 67.7 | 69.1 | 72.11 | **72.67** |

DPFM also outperforms multinomial logistic regression, SVM, supervised LDA and two-layer feed-forward neural networks, for which test accuracies ranged from 67% to 72.14%, using term frequency-inverse document frequency features.

### Medical records

- Duke University 5-year dataset (2007-2011): 240,000 patients, 4.4M visits.
- 34,000 medication mapped to 1,691 pharmaceutical active ingredients (AI).
- Dataset: 1,019×131,264 counts matrix of AIs vs. patients.
- MCMC-based DPFM of size 64-32.

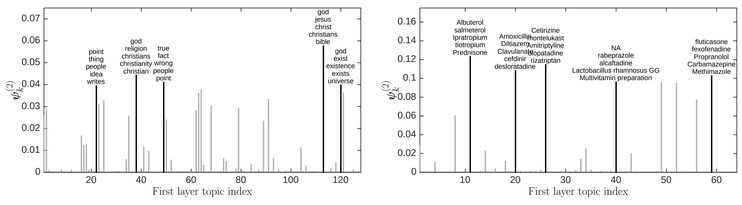



TABLE 4: Representative meta-topics obtained from (left) 20 News and (right) medical records. Meta-topic weights $\boldsymbol{\psi}_k^{(2)}$ vs. layer-1 topics indices, with word lists corresponding to the top five words in layer-1 topics, $\boldsymbol{\psi}_k^{(1)}$.

## References

[1] M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.

[2] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.