Deep Poisson Factor Modeling
Ricardo Henao, Zhe Gan, James Lu and Lawrence Carin
Department of Electrical and Computer Engineering, Duke University, Durham, NC

Poison factor analysis as a module

Assume \( x_n \) 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

\[
x_n \sim \text{Poisson}(\theta_0 + h_n),
\]

where

- \( \theta \in \mathbb{R}^K \): factor loadings matrix with \( K \) factors.
- \( h_n \in \mathbb{R}^K \): factor intensities.
- \( \beta_n \in \{0,1\}^K \), binary units indicating which factors are active for observation \( n \).
- Symbol \( \tilde{v} \) denotes element-wise (Hamadad) product.

Prior specification [2]

\[
x_{mn} = \sum_{k=1}^{K} x_{mkn} = \text{Poisson}(\lambda_{0kn}), \quad \lambda_{0kn} = \tilde{v}_{k} \theta_{k0} h_{nk}, \quad \lambda_{in} = \text{Gamma}(\kappa, \eta), \quad \beta_n \sim \text{Bernoulli}(\eta). \tag{2}
\]

Note that \( \eta \) controls the sparsity of \( \beta \), while \( \kappa \) accommodates for over-dispersion in \( x_{mn} \) via \( \lambda_{0kn} \).

PFA module: Condensed on \( h_n \), we express

\[
x_n \sim \text{Poisson}(\theta_0 + h_n). \tag{3}
\]

Deep representations with PFA modules

Develop a deep prior specification for \( h_n \) as

\[
x_n \sim \text{Poisson}(\theta_0, h_n), \quad h_n \sim \text{Multinomial}(1, \hat{x}_n), \quad \hat{x}_n = \frac{1}{\sum_{i=1}^{K} x_{mkn}}. \tag{4}
\]

Binary units are constituted as \( \beta \):

\[
\beta_n \sim \text{Bernoulli}(\eta), \quad \hat{x}_n \sim \text{Gamma}(\kappa, \eta), \quad \lambda_{0kn} \sim \text{Poisson}(\tilde{v}_k \theta_{k0} h_{nk}). \tag{5}
\]

where

- Function \( 1(\cdot) \) is defined component-wise as

\[
1(\cdot) = 1 \text{ if } \cdot > 0, \quad \text{otherwise } 1(\cdot) = 0. \tag{6}
\]

For top layer

\[
x_{mn} = \sum_{k=1}^{K} x_{mkn} = \text{Poisson}(\lambda_{mn}), \quad \lambda_{mn} = \text{Gamma}(\kappa, \eta), \quad \beta_n \sim \text{Bernoulli}(\eta). \tag{7}
\]

Equivalently

\[
p(\beta_n) = 1 = \text{Bernoulli}(\eta) \tag{8}
\]

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

Inference:

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

Benchmark corpora

- Date: 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, 10/1,000 training/test documents.

Performance: Held-out perplexity on 20% of test set.

Model: LDA, FTN, BSM, unDP, 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 on 20 News, 20 Newsgroup and Wiki. Size number of topics and/or binary units.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method-Size</th>
<th>20 News</th>
<th>20 Newsgroup</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPFM</td>
<td>SVI</td>
<td>64-32</td>
<td>381</td>
<td>954</td>
</tr>
<tr>
<td>DPFM</td>
<td>MCMC</td>
<td>64-32</td>
<td>780</td>
<td>908</td>
</tr>
<tr>
<td>DPFM</td>
<td>SVI</td>
<td>128-64-32</td>
<td>827</td>
<td>1141</td>
</tr>
<tr>
<td>DPFM</td>
<td>MCMC</td>
<td>128-64-32</td>
<td>806</td>
<td>920</td>
</tr>
<tr>
<td>DPFM</td>
<td>SVI</td>
<td>1024-512-256</td>
<td>909</td>
<td>1041</td>
</tr>
<tr>
<td>BSM</td>
<td>CDS</td>
<td>128</td>
<td>887</td>
<td>1151</td>
</tr>
<tr>
<td>FTN</td>
<td>Gibbs</td>
<td>128</td>
<td>877</td>
<td>1171</td>
</tr>
</tbody>
</table>

Classification

- Date: 20 News for document classification.
- Performance: test accuracy.
- Model: LDA, Doc-NADE, BSM, OSM and DPFM.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Method-Sizes</th>
<th>Accuracy (%)</th>
<th>20 News</th>
<th>20 Newsgroup</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA_S</td>
<td>Doc-NADE</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>LDA_S</td>
<td>BSM</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>LDA_S</td>
<td>DPFM</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
<td>66.4</td>
</tr>
</tbody>
</table>

DFPM 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

- 34,000 medication mapped to 1,691 pharmaceutical active ingredients (AI).
- Dataset: 1,039x133,264 counts matrix of AIs vs. patients.
- MCMC-based DPFM of size 64-12.

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

References