Deep Poisson Factor Analysis

Lawrence Carin

Duke University

24 March 2017

Agenda

- Basic Poisson factor analysis (PFA) for vector count data
- Extension of PFA modules to a generative deep architecture
- Bernoulli-Poisson link function
- Experiments
- Summary and future directions

Poisson Factor Model for Count Data

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

- $\boldsymbol{\Psi} \in \mathbb{R}_+^{M \times K}$ is 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

Poisson Factor Analysis (PFA)

$$\mathbf{x}_n \sim \mathrm{Poisson}\left(\boldsymbol{\Psi}(\boldsymbol{\theta}_n \circ \mathbf{h}_n)\right)$$

- $\boldsymbol{\psi}_k$ column k of $\boldsymbol{\Psi}$, and typical prior specification:

$$\boldsymbol{\psi}_k \sim \ \mathrm{Dirichlet}(\eta \mathbf{1}_M)$$

- Typical prior specification for $\boldsymbol{\theta}_n$: $\theta_{kn} \sim \ \mathrm{Gamma}(r_k, (1-b)/b)$

- Conditioned on $\mathbf{h}_n$, we express

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

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

Agenda

- Basic Poisson factor analysis (PFA) for vector count data
- Extension of PFA modules to a generative deep architecture
- Bernoulli-Poisson link function
- Experiments
- Summary and future directions

Two-Layer Model

$$
\begin{aligned}
\mathbf{x}_n &\sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\theta}_n^{(1)}, \mathbf{h}_n^{(1)}; \eta^{(1)}, r_k^{(1)}, b^{(1)}\right) \\
\mathbf{h}_n^{(1)} &= \mathbf{1}\left(\mathbf{z}_n^{(2)}\right) \\
\mathbf{z}_n^{(2)} &\sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(2)}, \boldsymbol{\theta}^{(2)}, \mathbf{h}_n^{(2)}; \eta^{(2)}, r_k^{(2)}, b^{(2)}\right)
\end{aligned}
$$

- where $\mathbf{1}\left(\mathbf{z}_n^{(2)}\right)$ defined component-wise as

$$
h_{nk} = 1 \text{ if } z_{nk}^{(2)} > 0, \qquad \text{otherwise } h_{nk} = 0
$$

- May be repeated for more than two layers, constituting a deep architecture
- Repeated use of PFA modules

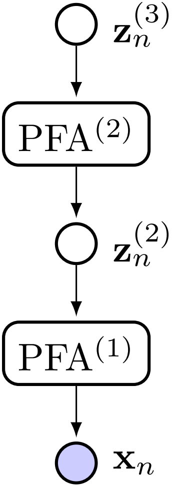
L-Layer Model

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

- For top layer:

$$
z_{kn}^{(L+1)} \sim \mathrm{Poisson}(\lambda_k^{(L+1)}), \quad \lambda_k^{(L+1)} \sim \mathrm{Gamma}(a_0, b_0)
$$

Graphical Model



Agenda

- Basic Poisson factor analysis (PFA) for vector count data
- Extension of PFA modules to a generative deep architecture
- Bernoulli-Poisson link function
- Experiments
- Summary and future directions

More on Model of Binary Hidden Units

- Binary units constituted as

$$\begin{aligned}\mathbf{h}_n^{(l-1)} &= \mathbf{1}\left(\mathbf{z}_n^{(l)}\right)\\ \mathbf{z}_n^{(l)} &\sim \mathrm{PFA}\left(\mathbf{\Psi}^{(l)}, \boldsymbol{\theta}^{(l)}, \mathbf{h}_n^{(l)}; \eta^{(l)}, r_k^{(l)}, b^{(l)}\right)\end{aligned}$$

- Draw for $h_{nk}^{(l-1)}$, component k of $\mathbf{h}_n^{(l-1)}$, may be expressed as

$$\begin{aligned}h_{nk}^{(l-1)} &= 1(z_{nk}^{(l)})\\ z_{nk}^{(l)} &\sim \text{Poisson}(\lambda_{nk}^{(l)})\\ \lambda_{nk}^{(l)} &= \sum_{k'=1}^{K_l} \psi_{kk'}^{(l)} \theta_{k'}^{(l)} h_{nk'}^{(l)}\end{aligned}$$

Bernoulli-Poisson Link (BPL)

$$\begin{aligned} h_{nk}^{(l-1)} &= 1(z_{nk}^{(l)}) \\ z_{nk}^{(l)} &\sim \text{Poisson}(\lambda_{nk}^{(l)}) \end{aligned}$$

- Expressed concisely as

$$h_{nk}^{(l-1)} \sim \text{BPL}(\lambda_{nk}^{(l)})$$

- Equivalent to

$$h_{nk}^{(l-1)} \sim \text{Bernoulli}(\pi_{nk}^{(l)}) \ , \quad \pi_{nk}^{(l)} = 1 - \exp(-\lambda_{nk}^{(l)})$$

- Generalization of the logistic link, for which

$$\pi_{nk}^{(l)} = 1/[1 + \exp(-\lambda_{nk}^{(l)})]$$

Advantages of the Bernoulli-Poisson Link

- Update equations only depend on the non-zero components of $\mathbf{x}_n$ and $\mathbf{h}_n^{(l)}$
- Yields significant acceleration for typically sparse data
- Entire deep architecture built upon clean repeated use of Poisson elements

Meaning of Deep Architecture for Count Data

$$\mathbf{x}_n \sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\theta}_n^{(1)}, \mathbf{h}_n^{(1)}; \eta^{(1)}, r_k^{(1)}, b^{(1)}\right), \qquad \mathbf{h}_n^{(1)} = \mathbf{1}\left(\mathbf{z}_n^{(2)}\right),$$

$$\mathbf{z}_n^{(2)} \sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(2)}, \boldsymbol{\theta}^{(2)}, \mathbf{h}_n^{(2)}; \eta^{(2)}, r_k^{(2)}, b^{(2)}\right), \qquad \vdots$$

$$\vdots \qquad \mathbf{h}_n^{(L-1)} = \mathbf{1}\left(\mathbf{z}_n^{(L)}\right),$$

$$\mathbf{z}_n^{(L)} \sim \mathrm{PFA}\left(\boldsymbol{\Psi}^{(L)}, \boldsymbol{\theta}^{(L)}, \mathbf{h}_n^{(L)}; \eta^{(L)}, r_k^{(L)}, b^{(L)}\right), \qquad \mathbf{h}_n^{(L)} = \mathbf{1}\left(\mathbf{z}_n^{(L+1)}\right),$$

- Columns of $\boldsymbol{\Psi}^{(1)}$ correspond to topics for representation of $\mathbf{x}_n$
- Columns of $\boldsymbol{\Psi}^{(2)}$ corresponds to meta-topics, indicating which layer-one topics are likely to be turned on (imposes correlation across layer-1 topics)
- Further sophisticated correlation structure imposed as we go deeper in the model

Learning and Inference

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

- Global parameters to be learned: $\boldsymbol{\Psi}^{(1:L)}, \boldsymbol{\theta}^{(2:L)}$
- Local parameters to be inferred for each $\mathbf{x}_n$: $\boldsymbol{\theta}_n^{(1)}, \mathbf{h}_n^{(1:L)}$
- As consequence of local conjugacy, analytic Gibbs updates throughout

Agenda

- Basic Poisson factor analysis (PFA) for vector count data
- Extension of PFA modules to a generative deep architecture
- Bernoulli-Poisson link function
- Experiments
- Summary and future directions

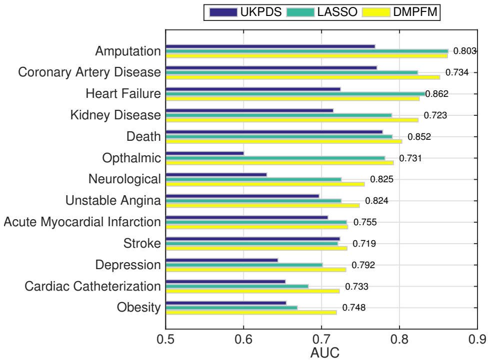
Electronic Medical Record Data

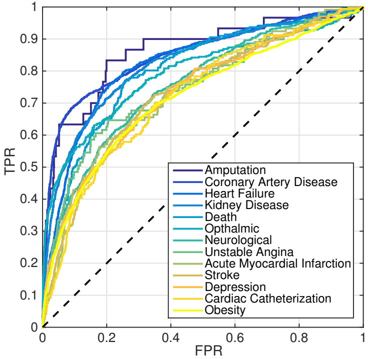
Data:

- Duke University 5-year dataset (2007-2011):
  240K patients 4.4M patient visits
- Mapping of data elements:
  - 39,429 medication names to 1,694 active ingredients using RxNorm
  - 4,391 laboratory tests to 1,869 unique tests using RELMA tool
  - 21,013 unique ICD-9 and CPT codes
- For each patient: total number of occurrences of a data element over a 6 months window defines an observation entry
- 16,756 type-2 diabetes mellitus patients (7,892/8,864 training/test)
- Modalities (filter out variables with less than 10 occurrences):
  - 253 medications (Meds)
  - 606 laboratory tests (Labs)
  - 4,222 diagnosis and procedure codes (Codes)

Electronic Medical Record Data - Classification

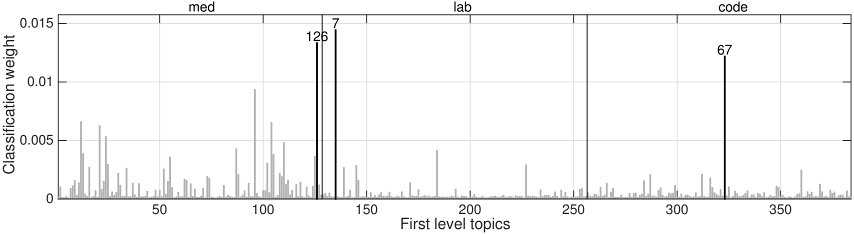
AUC and ROC curves for each outcome:





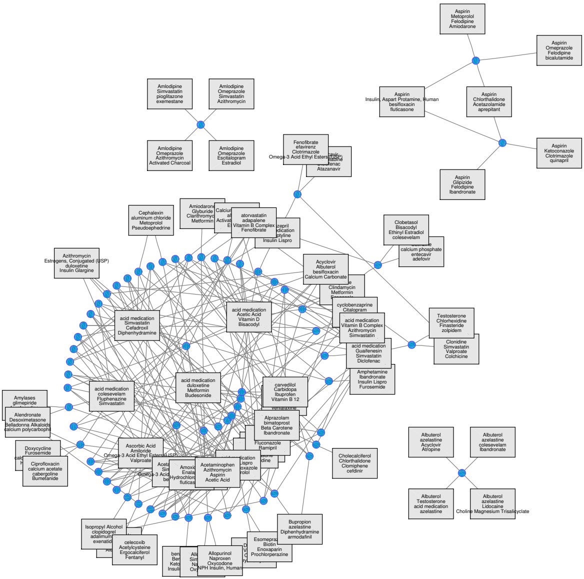
Electronic Medical Record Data - Classification

Amputation classifier



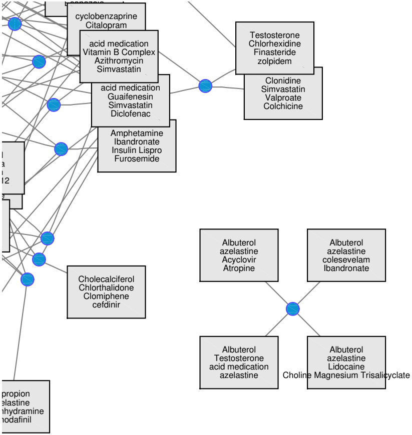
#7	#126	#67
streptococcus pyogenes ag	magnesium salt	ulcer of other part of foot
hemoglobin.gastrointestinal	aspirin	diabetes with other specifi...
glucose	lisinopril	peripheral vascular disease...
thiamine	metformin	atherosclerosis of native a...
estimated average glucose	acetic acid	rad exam foot comp min 3 vws
ferritin	acetazolamide	interview and evaluation, d...
hemoglobin a1c/hemoglobin.t...	hydromorphone	atherosclerosis of native a...
glucose	calcitriol	ulcer of heel and midfoot
potassium	amitriptyline	gangrene
hemoglobin a2/hemoglobin.total	simvastatin	cellulitis and abscess of f...

Electronic Medical Record Data - Analysis



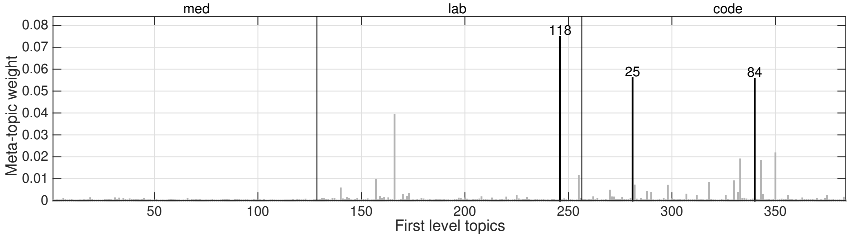
Electronic Medical Record Data - Analysis

Asthma group



Electronic Medical Record Data - Analysis

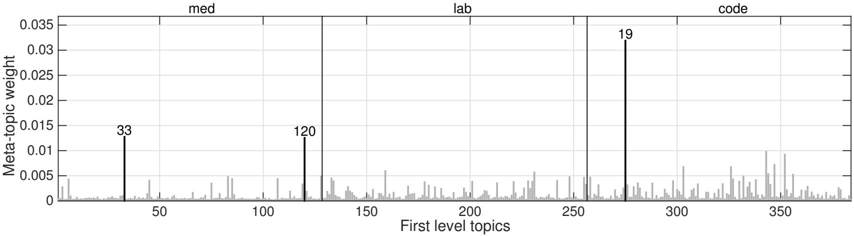
Prostate cancer meta-topic



#118	#25	#84
prostate specific ag	routine venipunture for speci	malignant neoplasm of prostate
prostate specific ag.free	routine general medical exa...	impotence of organic origin
prostate specific ag.free/p...	lipid panel	office/outpt visit est pt
testosterone	hemoglobin glycosylated	personal history of maligna...
hemoglobin a1c/hemoglobin.t...	prostate specific antgn totl	elevated prostate specific ...
estimated average glucose	office/outpt visit est pt	interview and evaluation, d...
testosterone.free	reeval healthy est 40-64yrs	other puncture of vein
calcitriol	special screening for malig...	follow-up examination follo...
glomerular filtration rate/...	postop f-up vst incl glob s...	office consult new or est pt
alanine aminotransferase	basic metabolic panel	nocturia

Electronic Medical Record Data - Analysis

Depressive disorder and chronic pain meta-topic



#33	#120	#19
calcium carbonate	alprazolam	depressive disorder, not el...
alfuzosin	bimatoprost	office/outpt visit est pt
activated charcoal	beta carotene	anxiety state, unspecified
etodolac	ibandronate	postop f-up vst incl glob s...
donepezil	amiodarone	myalgia and myositis, unspe...
adefovir	azelastine	dysthymic disorder
benazepril	clotrimazole	psych insight off/op 45-50 m
aspirin	cephalexin	insomnia, unspecified
vitamin b 12	irbesartan	chronic pain syndrome
amantadine	baclofen	major depressive disorder, ...

Agenda

- Basic Poisson factor analysis (PFA) for vector count data
- Extension of PFA modules to a generative deep architecture
- Bernoulli-Poisson link function
- Experiments
    - Comparison with state-of-the-art topic models for document analysis
    - Demonstration with multi-view electronic-medical record (EMR) data
- Summary and future directions