Bayesian Factor Analysis for Count Data

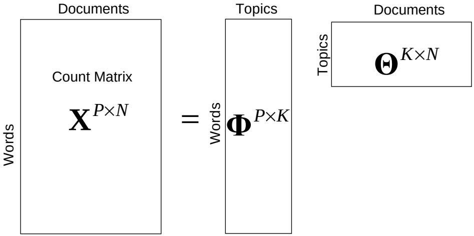
Mingyuan Zhou

IROM Department, McCombs School of Business
The University of Texas at Austin

Duke-Tshinghua Machine Learning Summer School
Duke-Kushan University, Kunshan, China
August 02, 2016

Outline

- Analysis of count data
- Latent variable models for discrete data
  - Poisson factor analysis
  - Nonnegative matrix factorization
  - Latent Dirichlet allocation



- Negative binomial processes

Count data is common

- Nonnegative and discrete:
  - Number of auto insurance claims / highway accidents / crimes
  - Consumer behavior, labor mobility, marketing, voting
  - Photon counting
  - Species sampling
  - Text analysis
  - Infectious diseases, Google Flu Trends
  - Next generation sequencing (statistical genomics)
- Mixture modeling can be viewed as a count-modeling problem
  - Number of points in a cluster (mixture model, we are modeling a count vector)
  - Number of words assigned to topic k in document j (we are modeling a $K \times J$ latent count matrix in a topic model/mixed-membership model)

Count data is common

- Nonnegative and discrete:
    - Number of auto insurance claims / highway accidents / crimes
    - Consumer behavior, labor mobility, marketing, voting
    - Photon counting
    - Species sampling
    - Text analysis
    - Infectious diseases, Google Flu Trends
    - Next generation sequencing (statistical genomics)
- Mixture modeling can be viewed as a count-modeling problem
    - Number of points in a cluster (mixture model, we are modeling a count vector)
    - Number of words assigned to topic k in document j (we are modeling a $K \times J$ latent count matrix in a topic model/mixed-membership model)

Poisson distribution

Siméon-Denis Poisson

(21 June 1781 – 25 April 1840)

"Life is good for only two things: doing mathematics and teaching it."

http://en.wikipedia.org

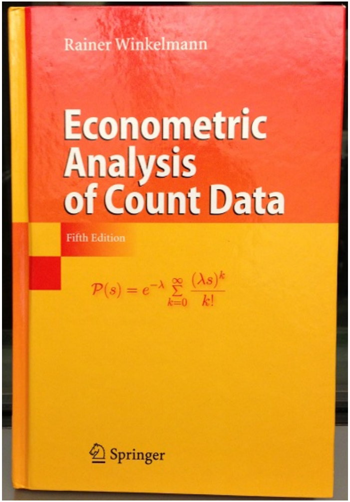
Poisson distribution

Siméon-Denis Poisson

(21 June 1781 – 25 April 1840)

"Life is good for only two things: doing mathematics and teaching it."

http://en.wikipedia.org



- Poisson distribution $x \sim \text{Pois}(\lambda)$
  - Probability mass function:

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, \dots\}$$

  - The mean and variance are the same: $\mathbb{E}[x] = \text{Var}[x] = \lambda$.
  - Restrictive to model over-dispersed (variance greater than the mean) counts that are commonly observed in practice.
  - A basic building block to construct more flexible count distributions.
- Overdispersed count data are commonly observed due to
  - Heterogeneity: difference between individuals
  - Contagion: dependence between the occurrence of events

Poisson and multinomial distributions

- Suppose that $x_1, \ldots, x_K$ are independent Poisson random variables with

$$x_k \sim \mathrm{Pois}(\lambda_k), \quad x = \sum_{k=1}^K x_k.$$

Set $\lambda = \sum_{k=1}^K \lambda_k$; let $(y, y_1, \ldots, y_K)$ be random variables such that

$$y \sim \mathrm{Pois}(\lambda), \; (y_1, \ldots, y_k) \,|\, y \sim \mathrm{Mult}\left(y; \frac{\lambda_1}{\lambda}, \ldots, \frac{\lambda_K}{\lambda}\right).$$

Then the distribution of $\boldsymbol{x} = (x, x_1, \ldots, x_K)$ is the same as the distribution of $\boldsymbol{y} = (y, y_1, \ldots, y_K)$.

Multinomial and Dirichlet distributions

- Model:

$$(x_{i1}, \ldots, x_{ik}) \sim \text{Multinomial}(n_i, p_1, \ldots, p_k),$$

$$(p_1, \ldots, p_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k p_j^{\alpha_j - 1}$$

- The conditional posterior of $(p_1, \ldots, p_k)$ is Dirichlet distributed as

$$(p_1, \ldots, p_k \,|\, -) \sim \text{Dirichlet}\left(\alpha_1 + \sum_i x_{i1}, \ldots, \alpha_k + \sum_i x_{ik}\right)$$

Gamma and Dirichlet distributions

- Suppose that random variables y and $(y_1, \ldots, y_K)$ are independent with

$$y \sim \mathrm{Gamma}(\gamma, 1/c), \quad (y_1, \ldots, y_K) \sim \mathrm{Dir}(\gamma p_1, \cdots, \gamma p_K)$$

where $\sum_{k=1}^{K} p_k = 1$; Let

$$x_k = y y_k$$

then $\{x_k\}_{1,K}$ are independent gamma random variables with

$$x_k \sim \mathrm{Gamma}(\gamma p_k, 1/c).$$

- The proof can be found in arXiv:1209.3442v1

Poisson factor alaysis

- Factorize the term-document word count matrix $\mathbf{M} \in \mathbb{Z}_+^{V\times N}$ under the Poisson likelihood as

$$\mathbf{M} \sim \text{Pois}(\boldsymbol{\Phi}\boldsymbol{\Theta})$$

  where $\mathbb{Z}_+ = \{0, 1, \ldots\}$ and $\mathbb{R}_+ = \{x : x > 0\}$.
- m_{vj} is the number of times that term v appears in document j.
- Factor loading matrix: $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_K) \in \mathbb{R}_+^{V\times K}$.
- Factor score matrix: $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \in \mathbb{R}_+^{K\times N}$.
- A large number of discrete latent variable models can be united under the Poisson factor analysis framework, with the main differences on how the priors for ϕ_k and $\boldsymbol{\theta}_j$ are constructed.

Two equivalent augmentations

- Poisson factor analysis

$$m_{vj} \sim \text{Pois}\left(\sum_{k=1}^{K} \phi_{vk}\theta_{jk}\right)$$

  - Augmentation 1:

$$m_{vj} = \sum_{k=1}^{K} n_{vjk}, \;\; n_{vjk} \sim \text{Pois}(\phi_{vk}\theta_{jk})$$

  - Augmentation 2:

$$m_{vj} \sim \text{Pois}\left(\sum_{k=1}^{K} \phi_{vk}\theta_{jk}\right), \;\; \zeta_{vjk} = \frac{\phi_{vk}\theta_{jk}}{\sum_{k=1}^{K} \phi_{vk}\theta_{jk}}$$

$$[n_{vj1}, \cdots, n_{vjK}] \sim \text{Mult}\left(m_{vj}; \zeta_{vj1}, \cdots, \zeta_{vjK}\right)$$

Hierarchical model for gamma-Poisson factor analysis

- Poisson factor analysis with gamma priors on $\mathbf{\Phi}$ and $\mathbf{\Theta}$:

$$m_{vj} = \text{Pois}\left(\sum_{k=1}^{K} \phi_{vk}\theta_{jk}\right),$$
$$\phi_{vk} \sim \text{Gamma}(a_\phi, 1/b_\phi),$$
$$\theta_{jk} \sim \text{Gamma}(a_\theta, 1/b_\theta).$$

- Note here the number of factors K is a tuning parameter, and we will show later how to construct nonparametric Bayesian Poisson factor analysis.

Gibbs sampling

- Denote $n_{v\cdot k} = \sum_j n_{vjk}$, $n_{jk} = \sum_v n_{vjk}$, $n_{\cdot k} = \sum_j n_{jk}$, $\theta_{\cdot k} = \sum_j \theta_{jk}$, and $\phi_{\cdot k} = \sum_v \phi_{vk}$.
- Gibbs sampling:

$$
\begin{aligned}
([n_{vj1}, \cdots, n_{vjK}] \,|\, -) &\sim \text{Mult}\left(m_{vj}; \zeta_{vj1}, \cdots, \zeta_{vjK}\right) \\
(\phi_{vk} \,|\, -) &\sim \text{Gamma}\left[a_\phi + n_{v\cdot k}, 1/(b_\phi + \theta_{\cdot k})\right] \\
(\theta_{jk} \,|\, -) &\sim \text{Gamma}\left[a_\theta + n_{jk}, 1/(b_\theta + \phi_{\cdot k})\right]
\end{aligned}
$$

- Homework: derive these Gibbs sampling update equations

Variational Bayes

- Variational Bayes: we approximate $P(\{n_{vjk}\}, \boldsymbol{\Phi}, \boldsymbol{\Theta} \,|\, \mathbf{M})$ with

$$Q = \left[\prod_k \prod_v Q(\phi_{vk})\right] \left[\prod_k \prod_j Q(\theta_{jk})\right] \times \left[\prod_v \prod_j Q(n_{vj1}, \ldots, n_{vjK})\right]$$

- We seek the Q that minimizes $\mathrm{KL}(Q||P)$ or (equivalently) maximizes $\mathcal{L}(Q) = \mathbb{E}_Q[\ln P(\{n_{vjk}\}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \mathbf{M})] - \mathbb{E}_Q[\ln(Q)]$.

Variational Bayes

- We choose

$$Q(n_{vj1}, \cdots, n_{vjK}) = \text{Mult}\left(m_{vj}; \tilde{\zeta}_{vj1}, \cdots, \tilde{\zeta}_{vjK}\right)$$

$$Q(\phi_{vk}) \sim \text{Gamma}\left(\tilde{a}_{\phi_{vk}}, 1/\tilde{b}_{\phi_{vk}}\right)$$

$$Q(\theta_{jk}) \sim \text{Gamma}\left(\tilde{a}_{\theta_{jk}}, 1/\tilde{b}_{\theta_{jk}}\right)$$

- Update equations

$$\tilde{\zeta}_{vjk} \propto \exp[\langle \ln \phi_{vk} \rangle + \langle \ln \theta_{jk} \rangle]$$

$$\tilde{a}_{\phi_{vk}} = a_\phi + \langle n_{v \cdot k} \rangle, \quad \tilde{b}_{\phi_{vk}} = b_\phi + \langle \theta_{\cdot k} \rangle$$

$$\tilde{a}_{\theta_{jk}} = a_\theta + \langle n_{jk} \rangle, \quad \tilde{b}_{\theta_{jk}} = b_\theta + \langle \phi_{\cdot k} \rangle$$

- These expectations can be calculated as $\langle \ln \phi_{vk} \rangle = \psi(\tilde{a}_{\phi_{vk}}) - \ln \tilde{b}_{\phi_{vk}}$, $\langle \ln \theta_{jk} \rangle = \psi(\tilde{a}_{\theta_{jk}}) - \ln \tilde{b}_{\theta_{jk}}$, $\langle n_{vjk} \rangle = m_{vj} \tilde{\zeta}_{vjk}$, $\langle \phi_{\cdot k} \rangle = \sum_v \tilde{a}_{\phi_{vk}} / \tilde{b}_{\phi_{vk}}$, $\langle \theta_{\cdot k} \rangle = \sum_j \tilde{a}_{\theta_{jk}} / \tilde{b}_{\theta_{jk}}$
- Optional homework: derive Variational Bayes update equations

Nonnegative matrix factorization and gamma-Poisson factor analysis

- Expectation-Maximization (EM) algorithm:

$$\phi_{vk} = \phi_{vk} \frac{\frac{a_\phi - 1}{\phi_{vk}} + \sum_{i=1}^{N} \frac{m_{vj}\theta_{jk}}{\sum_{k=1}^{K} \phi_{vk}\theta_{jk}}}{b_\phi + \theta_{k\cdot}}$$

$$\theta_{jk} = \theta_{jk} \frac{\frac{a_\theta - 1}{\theta_{jk}} + \sum_{p=1}^{P} \frac{m_{vj}\phi_{vk}}{\sum_{k=1}^{K} \phi_{vk}\theta_{jk}}}{b_\theta + \phi_{\cdot k}}.$$

- If we set $b_\phi = b_\theta = 0$ and $a_\phi = a_\theta = 1$, then the EM algorithm is the same as those of non-negative matrix factorization (Lee and Seung, 2000) with an objective function of minimizing the KL divergence $D_{KL}(\mathbf{M}||\mathbf{\Phi\Theta})$.

Mixed Poisson distribution

$$x \sim \text{Pois}(\lambda), \;\; \lambda \sim f_\Lambda(\lambda)$$

- Mixing the Poisson rate parameter with a positive distribution leads to a mixed Poisson distribution.
- A mixed Poisson distribution is always over-dispersed.
  - Law of total expectation:

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x \,|\, \lambda]] = \mathbb{E}[\lambda].$$

  - Law of total variance:

$$\text{Var}[x] = \text{Var}[\mathbb{E}[x \,|\, \lambda]] + \mathbb{E}[\text{Var}[x \,|\, \lambda]] = \text{Var}[\lambda] + \mathbb{E}[\lambda].$$

  - Thus $\text{Var}[x] > \mathbb{E}[x]$ unless λ is a constant.

- Mixing the gamma distribution with the Poisson distribution as

$$x \sim \mathrm{Pois}(\lambda), \ \lambda \sim \mathrm{Gamma}\left(r, \frac{p}{1-p}\right),$$

where $p/(1-p)$ is the gamma scale parameter, leads to the negative binomial distribution $x \sim \mathrm{NB}(r, p)$ with probability mass function

$$P(x \mid r, p) = \frac{\Gamma(x+r)}{x!\Gamma(r)} p^x (1-p)^r, \quad x \in \{0, 1, \ldots\}$$

Compound Poisson distribution

- A compound Poisson distribution is the summation of a Poisson random number of *i.i.d.* random variables.
- If $x = \sum_{i=1}^{n} y_i$, where $n \sim \text{Pois}(\lambda)$ and y_i are *i.i.d.* random variable, then x is a compound Poisson random variable.
- The negative binomial random variable $x \sim \text{NB}(r, p)$ can also be generated as a compound Poisson random variable as

$$x = \sum_{i=1}^{l} u_i, \;\; l \sim \text{Pois}[-r\ln(1-p)], \;\; u_i \sim \text{Log}(p)$$

where $u \sim \text{Log}(p)$ is the logarithmic distribution with probability mass function

$$P(u \,|\, p) = \frac{-1}{\ln(1-p)} \frac{p^u}{u}, \quad u \in \{1, 2, \cdots\}.$$

Negative binomial distribution

$$m \sim \mathrm{NB}(r, p)$$

- r is the dispersion parameter
- p is the probability parameter
- Probability mass function

$$f_M(m \mid r, p) = \frac{\Gamma(r+m)}{m!\Gamma(r)} p^m (1-p)^r = (-1)^m \binom{-r}{m} p^m (1-p)^r$$

- It is a gamma-Poisson mixture distribution
- It is a compound Poisson distribution
- Its variance $\frac{rp}{(1-p)^2}$ is greater that its mean $\frac{rp}{1-p}$
- $\mathrm{Var}[m] = \mathbb{E}[m] + \frac{(\mathbb{E}[m])^2}{r}$

- The conjugate prior for the negative binomial probability parameter p is the beta distribution: if $m_i \sim \mathrm{NB}(r, p),\ p \sim \mathrm{Beta}(a_0, b_0)$, then

$$(p \,|\, -) = \mathrm{Beta}\left(a_0 + \sum_{i=1}^{n} m_i, b_0 + nr\right)$$

- The conjugate prior for the negative binomial dispersion parameter r is unknown, but we have a simple data augmentation technique to derive closed-form Gibbs sampling update equations for r.

- If we assign m customers to tables using a Chinese restaurant process with concentration parameter r, then the random number of occupied tables l follows the Chinese Restaurant Table (CRT) distribution

$$f_L(l \mid m, r) = \frac{\Gamma(r)}{\Gamma(m+r)} |s(m,l)| r^l, \quad l = 0, 1, \cdots, m.$$

  $|s(m,l)|$ are unsigned Stirling numbers of the first kind.

- The joint distribution of the customer count $m \sim \text{NB}(r,p)$ and table count is the Poisson-logarithmic bivariate count distribution

$$f_{M,L}(m, l \mid r, p) = \frac{|s(m,l)| r^l}{m!} (1-p)^r p^m.$$

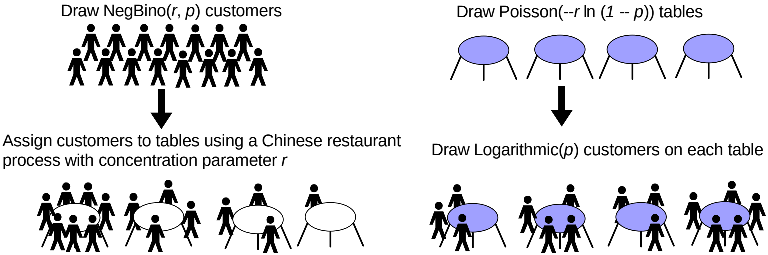
Poisson-logarithmic bivariate count distribution

- Probability mass function:

$$f_{M,L}(m, l; r, p) = \frac{|s(m, l)| r^l}{m!} (1 - p)^r p^m.$$

- It is clear that the gamma distribution is a conjugate prior for r to this bivariate count distribution.

The joint distribution of the customer count and table count are equivalent:



Bayesian inference for the negative binomial distribution

Negative binomial count modeling:

$$m_i \sim \text{NegBino}(r, p),\ \ p \sim \text{Beta}(a_0, b_0),\ \ r \sim \text{Gamma}(e_0, 1/f_0).$$

- Gibbs sampling via data augmetantion:

$$(p \,|\, -) \sim \text{Beta}\left(a_0 + \textstyle\sum_{i=1}^n m_i, b_0 + nr\right);$$

$$(\ell_i \,|\, -) = \textstyle\sum_{t=1}^{m_i} b_t,\ \ b_t \sim \text{Bernoulli}\left(\frac{r}{t+r-1}\right);$$

$$(r \,|\, -) \sim \text{Gamma}\left(e_0 + \textstyle\sum_{i=1}^n \ell_i, \frac{1}{f_0 - n\ln(1-p)}\right).$$

- Expectation-Maximization
- Variational Bayes

Bayesian inference for the negative binomial distribution

Negative binomial count modeling:

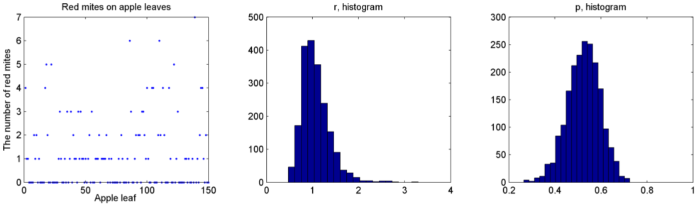
$$m_i \sim \text{NegBino}(r, p), \;\; p \sim \text{Beta}(a_0, b_0), \;\; r \sim \text{Gamma}(e_0, 1/f_0).$$

- Gibbs sampling via data augmetantion:

$$(p \,|\, -) \sim \text{Beta}\left(a_0 + \textstyle\sum_{i=1}^n m_i, b_0 + nr\right);$$
$$(\ell_i \,|\, -) = \textstyle\sum_{t=1}^{m_i} b_t, \;\; b_t \sim \text{Bernoulli}\left(\frac{r}{t+r-1}\right);$$
$$(r \,|\, -) \sim \text{Gamma}\left(e_0 + \textstyle\sum_{i=1}^n \ell_i, \frac{1}{f_0 - n\ln(1-p)}\right).$$

- Expectation-Maximization
- Variational Bayes

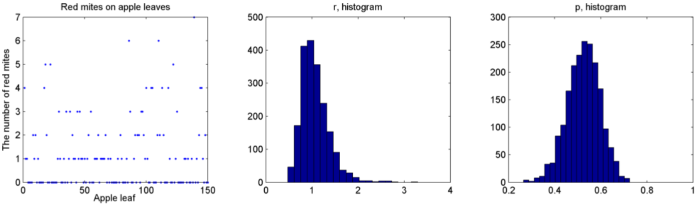
- Gibbs sampling: $\mathbb{E}[r] = 1.076$, $\mathbb{E}[p] = 0.525$.



- Expectation-Maximization: $r : 1.025$, $p : 0.528$.
- Variational Bayes: $\mathbb{E}[r] = 0.999$, $\mathbb{E}[p] = 0.534$.

- For this example, variational Bayes inference correctly identifies the modes but underestimates the posterior variances of model parameters.

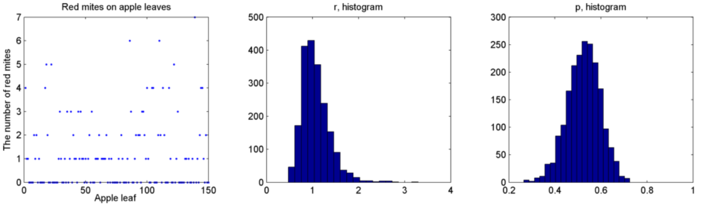
- Gibbs sampling: $\mathbb{E}[r] = 1.076$, $\mathbb{E}[p] = 0.525$.



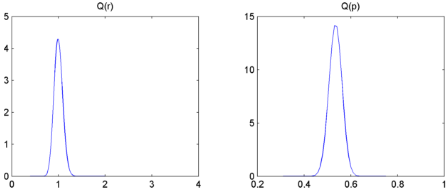
- Expectation-Maximization: $r : 1.025$, $p : 0.528$.
- Variational Bayes: $\mathbb{E}[r] = 0.999$, $\mathbb{E}[p] = 0.534$.

- For this example, variational Bayes inference correctly identifies the modes but underestimates the posterior variances of model parameters.

- Gibbs sampling: $\mathbb{E}[r] = 1.076$, $\mathbb{E}[p] = 0.525$.



- Expectation-Maximization: $r : 1.025$, $p : 0.528$.
- Variational Bayes: $\mathbb{E}[r] = 0.999$, $\mathbb{E}[p] = 0.534$.



- For this example, variational Bayes inference correctly identifies the modes but underestimates the posterior variances of model parameters.

Negative binomial gamma chain

NegBino-Gamma-Gamma-...

Negative binomial gamma chain

NegBino-Gamma-Gamma-...

Augmentation

(CRT, NegBino)-Gamma-Gamma-...

Negative binomial gamma chain

$$
\begin{array}{c}
\text{NegBino-Gamma-Gamma-...} \\
\Big\downarrow \textit{ Augmentation} \\
\text{(CRT, NegBino)-Gamma-Gamma-...} \\
\Big\downarrow \textit{ Equivalence} \\
\text{(Log, Poisson)-Gamma-Gamma-...}
\end{array}
$$

Negative binomial gamma chain

NegBino-Gamma-Gamma-...

↓ *Augmentation*

(CRT, NegBino)-Gamma-Gamma-...

↓ *Equivalence*

(Log, Poisson)-Gamma-Gamma-...

↓ *Marginalization*

NegBino-Gamma-...

Negative binomial gamma chain

NegBino-Gamma-Gamma-...

↙ *Augmentation*

(CRT, NegBino)-Gamma-Gamma-...

↓ *Equivalence*

(Log, Poisson)-Gamma-Gamma-...

↘ *Marginalization*

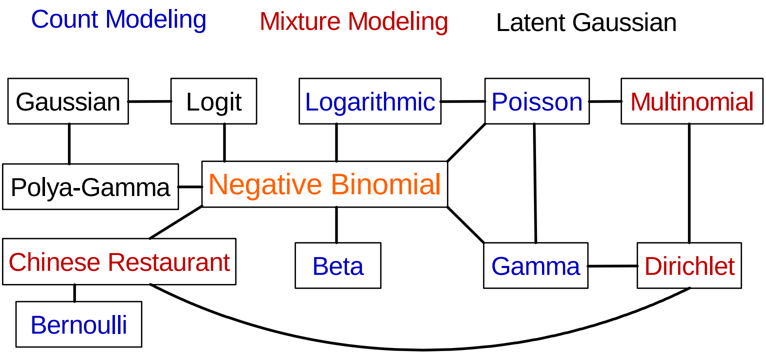
NegBino-Gamma-...

↙

↓

↘

Relationships between various distributions



Latent Dirichlet allocation (Blei et al., 2003)

- Hierarchical model:

$$\begin{aligned} x_{ji} &\sim \mathrm{Mult}(\phi_{z_{ji}}) \\ z_{ji} &\sim \mathrm{Mult}(\boldsymbol{\theta}_j) \\ \phi_k &\sim \mathrm{Dir}(\eta, \ldots, \eta) \\ \boldsymbol{\theta}_j &\sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \end{aligned}$$

- There are K topics $\{\phi_k\}_{1,K}$, each of which is a distribution over the V words in the vocabulary.
- There are N documents in the corpus and $\boldsymbol{\theta}_j$ represents the proportion of the K topics in the jth document.
- x_{ji} is the ith word in the jth document.
- z_{ji} is the index of the topic selected by x_{ji}.

- Denote $n_{vjk} = \sum_i \delta(x_{ji} = v)\delta(z_{ji} = k)$, $n_{v\cdot k} = \sum_j n_{vjk}$, $n_{jk} = \sum_v n_{vjk}$, and $n_{\cdot k} = \sum_j n_{jk}$.
- Blocked Gibbs sampling:

$$P(z_{ji} = k|-) \propto \phi_{x_{ji}k}\theta_{jk}, \quad k \in \{1, \ldots, K\}$$
$$(\phi_k|-) \sim \mathrm{Dir}(\eta + n_{1\cdot k}, \ldots, \eta + n_{V\cdot k})$$
$$(\boldsymbol{\theta}_j|-) \sim \mathrm{Dir}\left(\frac{\alpha}{K} + n_{j1}, \ldots, \frac{\alpha}{K} + n_{jK}\right)$$

- Variational Bayes inference (Blei et al., 2003).

- Collapsed Gibbs sampling (Griffiths and Steyvers, 2004):
  - Marginalizing out both the topics $\{\phi_k\}_{1,K}$ and the topic proportions $\{\theta_j\}_{1,N}$.
  - Sample z_{ji} conditioning on all the other topic assignment indices $\boldsymbol{z}^{-ji}$:

$$P(z_{ji} = k|\boldsymbol{z}^{-ji}) \propto \frac{\eta + n_{x_{ji}\cdot k}^{-ji}}{V\eta + n_{\cdot k}^{-ji}} \left(n_{jk}^{-ji} + \frac{\alpha}{K}\right), \quad k \in \{1, \ldots, K\}$$

- This is easy to understand as

$$P(z_{ji} = k|\boldsymbol{\phi}_k, \boldsymbol{\theta}_j) \propto \phi_{x_{ji}k}\theta_{jk}$$

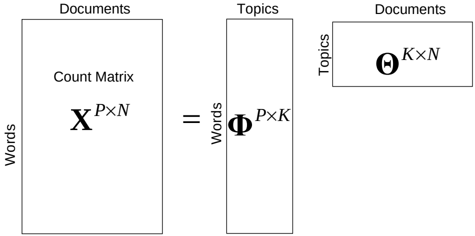
$$P(z_{ji} = k|\boldsymbol{z}^{-ji}) = \iint P(z_{ji} = k|\boldsymbol{\phi}_k, \boldsymbol{\theta}_j)P(\boldsymbol{\phi}_k, \boldsymbol{\theta}_j|\boldsymbol{z}^{-ji})d\boldsymbol{\phi}_k d\boldsymbol{\theta}_j$$

$$P(\boldsymbol{\phi}_k|\boldsymbol{z}^{-ji}) = \mathrm{Dir}(\eta + n_{1\cdot k}^{-ji}, \ldots, \eta + n_{V\cdot k}^{-ji})$$

$$P(\boldsymbol{\theta}_j|\boldsymbol{z}^{-ji}) = \mathrm{Dir}\left(\frac{\alpha}{K} + n_{j1}^{-ji}, \ldots, \frac{\alpha}{K} + n_{jK}^{-ji}\right)$$

$$P(\boldsymbol{\phi}_k, \boldsymbol{\theta}_j|\boldsymbol{z}^{-ji}) = P(\boldsymbol{\phi}_k|\boldsymbol{z}^{-ji})P(\boldsymbol{\theta}_j|\boldsymbol{z}^{-ji})$$

- In latent Dirichlet allocation, the words in a document are assumed to be exchangeable (bag-of-words assumption).
- Below we will relate latent Dirichlet allocation to Poisson factor analysis and show it essentially tries to factorize the term-document word count matrix under the Poisson likelihood:



Latent Dirichlet allocation and Dirichlet-Poisson factor analysis

- Dirichlet priors on $\mathbf{\Phi}$ and $\mathbf{\Theta}$:

$$m_{vj} = \text{Pois}\left(\sum_{k=1}^{K} \phi_{vk}\theta_{jk}\right)$$

$$\boldsymbol{\phi}_k \sim \text{Dir}(\eta, \ldots, \eta), \quad \boldsymbol{\theta}_j \sim \text{Dir}(\alpha/K, \ldots, \alpha/K).$$

- One may show that both the block Gibbs sampling inference and variational Bayes inference of the Dirichlet-Poisson factor analysis model are the same as that of the Latent Dirichlet allocation.

Beta-gamma-Poisson factor analysis

- Hierachical model (Zhou et al., 2012, Zhou and Carin, 2014):

$$
\begin{aligned}
m_{vj} &= \sum_{k=1}^{K} n_{vjk}, \;\; n_{vjk} \sim \mathrm{Pois}(\phi_{vk}\theta_{jk}) \\
\phi_k &\sim \mathrm{Dir}\left(\eta, \cdots, \eta\right), \\
\theta_{jk} &\sim \mathrm{Gamma}\left[r_j, p_k/(1-p_k)\right], \\
r_j &\sim \mathrm{Gamma}(e_0, 1/f_0), \\
p_k &\sim \mathrm{Beta}[c/K, c(1-1/K)].
\end{aligned}
$$

- $n_{jk} = \sum_{v=1}^{V} n_{vjk} \sim \mathrm{NB}(r_j, p_k)$
- This parametric model becomes a nonparametric Bayesian model governed by the beta-negative binomial process as $K \to \infty$.

Gamma-gamma-Poisson factor analysis

- Hierachical model (Zhou and Carin, 2014):

$$
\begin{aligned}
m_{vj} &= \sum_{k=1}^{K} n_{vjk}, \;\; n_{vjk} \sim \mathrm{Pois}(\phi_{vk}\theta_{jk}) \\
\phi_k &\sim \mathrm{Dir}\left(\eta, \cdots, \eta\right), \\
\theta_{jk} &\sim \mathrm{Gamma}\left[r_k, p_j/(1-p_j)\right], \\
p_j &\sim \mathrm{Beta}(a_0, b_0), \\
r_k &\sim \mathrm{Gamma}(\gamma_0/K, 1/c).
\end{aligned}
$$

- $n_{jk} \sim \mathrm{NB}(r_k, p_j)$
- This parametric model becomes a nonparametric Bayesian model governed by the gamma-negative binomial process as $K \to \infty$.

Poisson factor analysis and mixed-membership modeling

- We may represent the Poisson factor analysis

$$m_{vj} = \sum_{k=1}^{K} n_{vjk}, \;\; n_{vjk} \sim \mathrm{Pois}(\phi_{vk}\theta_{jk})$$

  in terms of a mixed-membership model, whose group sizes are randomized, as

$$x_{ji} \sim \mathrm{Mult}(\phi_{z_{ji}}), \; z_{ji} \sim \sum_{k=1}^{K} \frac{\theta_{jk}}{\sum_k \theta_{jk}} \delta_k, \; m_j \sim \mathrm{Pois}\left(\sum_k \theta_{jk}\right),$$

  where $i = 1, \ldots, m_j$ in the jth document, and $n_{vjk} = \sum_{i=1}^{m_j} \delta(x_{ji} = v)\delta(z_{ji} = k)$.
- The likelihoods of the two representations are different update to a multinomial coefficient (Zhou, 2014).

Connections to previous approaches

- Nonnegative matrix factorization (K-L divergence) (NMF)
- Latent Dirichlet allocation (LDA)
- GaP: gamma-Poisson factor model (GaP) (Canny, 2004)
- Hierarchical Dirichlet process LDA (HDP-LDA) (Teh et al., 2006)

Poisson factor analysis priors on θ_{jk}	Infer (p_k, r_j)	Infer (p_j, r_k)	Support $K \to \infty$	Related algorithms
gamma	$\times$	$\times$	$\times$	NMF
Dirichlet	$\times$	$\times$	$\times$	LDA
beta-gamma	$\checkmark$	$\times$	$\checkmark$	GaP
gamma-gamma	$\times$	$\checkmark$	$\checkmark$	HDP-LDA

Blocked Gibbs sampling

- Sample z_{ji} from multinomial;
  $n_{vjk} = \sum_{i=1}^{m_j} \delta(x_{ji} = v)\delta(z_{ji} = k)$.
- Sample ϕ_k from Dirichlet
- For the beta-negative binomial model (beta-gamma-Poisson factor analysis)
  - Sample l_{jk} from $\mathrm{CRT}(n_{jk}, r_j)$
  - Sample r_j from gamma
  - Sample p_k from beta
  - Sample θ_{jk} from $\mathrm{Gamma}(r_j + n_{jk}, p_k)$
- For the gamma-negative binomial model (gamma-gamma-Poisson factor analysis)
  - Sample l_{jk} from $\mathrm{CRT}(n_{jk}, r_k)$
  - Sample r_k from gamma
  - Sample p_j from beta
  - Sample θ_{jk} from $\mathrm{Gamma}(r_k + n_{jk}, p_j)$
- Collapsed Gibbs sampling for the beta-negative binomial model can be found in (Zhou, 2014).

Example application

- Example Topics of United Nation General Assembly Resolutions inferred by the gamma-gamma-Poisson factor analysis:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
trade	rights	environment	women	economic
world	human	management	gender	summits
conference	united	protection	equality	outcomes
organization	nations	affairs	including	conferences
negotiations	commission	appropriate	system	major

- The gamma-negative binomial and beta-negative binomial models have distinct mechanisms on controlling the number of inferred factors.
- They produce state-of-the-art perplexity results when used for topic modeling of a document corpus (Zhou et al, 2012, Zhou and Carin 2014, Zhou 2014).

Relational network

- A relational network (graph) is commonly used to describe the relationship between nodes, where a node could represent a person, a movie, a protein, etc.
- Two nodes are connected if there is an edge (link) between them.
- An undirected unweighted relational network with N nodes can be equivalently represented with a sysmetric binary affinity matrix $B \in \{0,1\}^{N \times N}$, where $b_{ij} = b_{ji} = 1$ if an edge exists between nodes i and j and $b_{ij} = b_{ji} = 0$ otherwise.

Stochastic blockmodel

- Each node is assigned to a cluster.
- The probability for an edge to exist between two nodes is solely decided by the clusters that they are assigned to.
- Hierachical model:

$$\begin{aligned} b_{ij} &\sim \text{Bernoulli}(p_{z_i z_j}), \quad \text{for } j > i \\ p_{k_1 k_2} &\sim \text{Beta}(a_0, b_0), \\ z_i &\sim \text{Mult}(\pi_1, \ldots, \pi_K), \\ (\pi_1, \ldots, \pi_K) &\sim \text{Dir}(\alpha/K, \ldots, \alpha/K) \end{aligned}$$

- Blocked Gibbs sampling:

$$P(z_i = k|-) = \pi_k \left\{ \prod_{j \neq i} p_{kz_j}^{b_{ij}} (1 - p_{kz_j})^{1-b_{ij}} \right\}$$

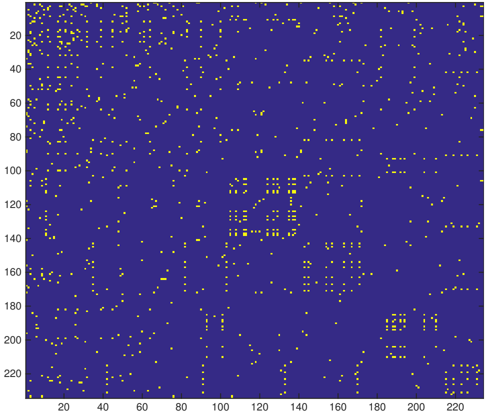
Infinite relational model (Kemp et al., 2006)

- As $K \to \infty$, the stochastic block model becomes a nonparametric Bayesian model governed by the Chinese restaurant process (CRP) with concentration parameter α:

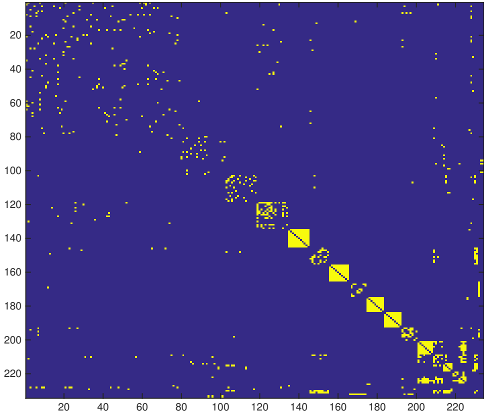
$$
\begin{aligned}
b_{ij} &\sim \mathrm{Bernoulli}(p_{z_i z_j}), \quad \text{for } i > j \\
p_{k_1 k_2} &\sim \mathrm{Beta}(a_0, b_0), \\
(z_1, \ldots, z_N) &\sim \mathrm{CRP}(\alpha)
\end{aligned}
$$

- Collapsed Gibbs sampling can be derived by marginalizing out $p_{k_1 k_2}$ and using the prediction rule of the Chinese restaurant process.

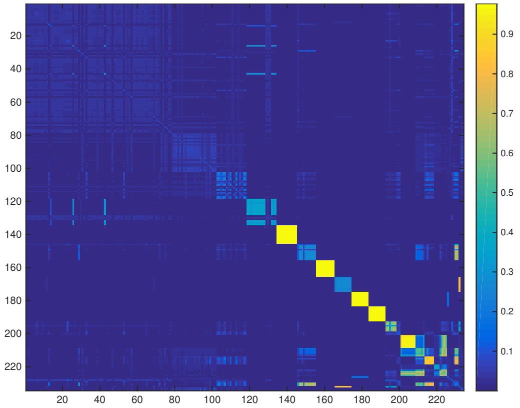
The coauthor network of the top 234 NIPS authors.



The reordered network using the stochastic blockmodel.



The estimated link probabilities within and between blocks.



D. Blei, A. Ng, and M. Jordan.
Latent Dirichlet allocation.
J. Mach. Learn. Res., 2003.

T. L. Griffiths and M. Steyvers.
Finding scientific topics.
PNAS, 2004.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda.
Learning systems of concepts with an infinite relational model.
In *AAAI*, 2006.

D. D. Lee and H. S. Seung.
Algorithms for non-negative matrix factorization.
In *NIPS*, 2000.

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

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

M. Zhou, L. Li, D. Dunson, and L. Carin.
Lognormal and gamma mixed negative binomial regression.
In *ICML*, 2012.

M. Zhou and L. Carin.
Augment-and-conquer negative binomial processes.
In *NIPS*, 2012.

M. Zhou and L. Carin.
Negative binomial process count and mixture modeling.
IEEE TPAMI, 2014.

M. Zhou.
Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling.
In *NIPS*, 2014.