Parametric Bayesian Models: Part II

Mingyuan Zhou and Lizhen Lin

Department of Information, Risk, and Operations Management
Department of Statistics and Data Sciences
The University of Texas at Austin

Machine Learning Summer School, Austin, TX
January 08, 2015
Outline for Part II

- Bayesian modeling of count data
  - Poisson, gamma, and negative binomial distributions
  - Bayesian inference for the negative binomial distribution
  - Regression analysis for counts

- Latent variable models for discrete data
  - Latent Dirichlet allocation
  - Poisson factor analysis

- Relational network analysis
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
Outline

Analysis of count data
  Motivations
  Count distributions
    Negative binomial distribution
    Relationships between distributions
    Count regression

Count matrix factorization and topic modeling

Relational network analysis

Main references

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:
    $$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, \ldots\}$$
  • The mean and variance is 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
Mixed Poisson distribution

\[ x \sim \text{Pois}(\lambda), \quad \lambda \sim f_\Lambda(\lambda) \]

- 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 \(\lambda\) is a constant.
- The gamma distribution is a popular choice as it is conjugate to the Poisson distribution.
• Mixing the gamma distribution with the Poisson distribution as

\[ x \sim \text{Pois}(\lambda), \quad \lambda \sim \text{Gamma} \left( r, \frac{p}{1 - p} \right), \]

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

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

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

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

\[ m \sim \text{NB}(r, p) \]

- \( r \) is the dispersion parameter
- \( p \) is the probability parameter
- Probability mass function

\[
f_M(m|r, p) = \frac{\Gamma(r + m)}{m! \Gamma(r)} p^m (1 - p)^r = (-1)^m \binom{-r}{m} p^m (1 - p)^r
\]

- 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 \( \frac{rp}{1-p} \)
- \( \text{Var}[m] = \mathbb{E}[m] + \frac{(\mathbb{E}[m])^2}{r} \)
• The conjugate prior for the negative binomial probability parameter $p$ is the beta distribution: if $m_i \sim \text{NB}(r, p)$, $p \sim \text{Beta}(a_0, b_0)$, then

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

• 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

\[
f_L(l|m,r) = \frac{\Gamma(r)}{\Gamma(m + r)}|s(m, l)|r^l, \quad l = 0, 1, \ldots, m.
\]

\( |s(m, l)| \) 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

\[
f_{M,L}(m, l|r, p) = \frac{|s(m, l)|r^l}{m!} (1 - p)^r p^m.
\]
Poisson-logarithmic bivariate count distribution

- Probability mass function:

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

- 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:

1. Draw \( \text{NegBino}(r, p) \) customers
2. Assign customers to tables using a Chinese restaurant process with concentration parameter \( r \)
3. Draw \( \text{Poisson}(-r \ln (1 - p)) \) tables
4. Draw \( \text{Logarithmic}(p) \) customers on each table
Bayesian inference for the negative binomial distribution

Negative binomial count modeling:

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

- Gibbs sampling via data augmentation:
  \[ (p|\cdot) \sim \text{Beta} \left( a_0 + \sum_{i=1}^{n} m_i, b_0 + nr \right); \]
  \[ (\ell_i|\cdot) = \sum_{t=1}^{m_i} b_t, \quad b_t \sim \text{Bernoulli} \left( \frac{r}{t+r-1} \right); \]
  \[ (r|\cdot) \sim \text{Gamma} \left( e_0 + \sum_{i=1}^{n} \ell_i, \frac{1}{f_0 - n \ln(1-p)} \right). \]

- Expectation-Maximization
- Variational Bayes
Bayesian inference for the negative binomial distribution

Negative binomial count modeling:

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

- Gibbs sampling via data augmentation:
  \[ (p|-) \sim \text{Beta} \left( a_0 + \sum_{i=1}^{n} m_i, b_0 + nr \right); \]
  \[ (\ell_i|-) = \sum_{t=1}^{m_i} b_t, \quad b_t \sim \text{Bernoulli} \left( \frac{r}{t+r-1} \right); \]
  \[ (r|-) \sim \text{Gamma} \left( e_0 + \sum_{i=1}^{n} \ell_i, \frac{1}{f_0-n\ln(1-p)} \right). \]

- Expectation-Maximization
- Variational Bayes

Main references

Bayesian inference for the negative binomial distribution

Negative binomial count modeling:

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

- Gibbs sampling via data augmentation:
  \[ (p|-) \sim \text{Beta} \left( a_0 + \sum_{i=1}^{n} m_i, b_0 + nr \right); \]
  \[ (\ell_i|-) = \sum_{t=1}^{m_i} b_t, \quad b_t \sim \text{Bernoulli} \left( \frac{r}{t+r-1} \right); \]
  \[ (r|-) \sim \text{Gamma} \left( e_0 + \sum_{i=1}^{n} \ell_i, \frac{1}{f_0-n\ln(1-p)} \right). \]

- Expectation-Maximization
- Variational Bayes

Main references
- 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

**Augmentation**

(CRT, NegBino)-Gamma-Gamma-...
Negative binomial gamma chain

NegBino-Gamma-Gamma-...

Augmentation

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

Equivalence

(Log, Poisson)-Gamma-Gamma-...
Negative binomial gamma chain

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

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

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

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

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

Poisson and multinomial distributions
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
  
  $$y \sim \text{Gamma}(\gamma, 1/c), \quad (y_1, \ldots, y_K) \sim \text{Dir}(\gamma p_1, \ldots, \gamma p_K)$$

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

  $$x_k = yy_k$$

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

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

- The proof can be found in arXiv:1209.3442v1
Relationships between various distributions

Outline
- Analysis of count data
  - Motivations
  - Count distributions
  - Negative binomial distribution
  - Relationships between distributions
  - Count regression
- Count matrix factorization and topic modeling
- Relational network analysis
- Main references

### Count Modeling
- Gaussian
- Logit
- Polya-Gamma
- Chinese Restaurant
- Bernoulli

### Mixture Modeling
- Logarithmic
- Poisson
- Gamma

### Latent Gaussian
- Multinomial
- Dirichlet

**Negative Binomial**

**Augmentation**

**Equivalence**

**Marginalization**

**NegBino** - Gamma - Gamma - ...
(CRT, NegBino) - Gamma - Gamma - ...
(Log, Poisson) - Gamma - Gamma - ...
NegBino - Gamma - ...

**Chinese Restaurant**

**Beta**

**Polya-Gamma**

**Logit**

**Logarithmic**

**Beta**

**Gamma**

**Multinomial**

**Dirichlet**
Poisson regression

- Model:
  \[ y_i \sim \text{Pois}(\lambda_i), \quad \lambda_i = \exp(x_i^T \beta) \]

- Model assumption:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] = \exp(x_i^T \beta). \]

- Poisson regression does not model over-dispersion.
Poisson regression with multiplicative random effects

- Model:
  \[ y_i \sim \text{Pois}(\lambda_i), \quad \lambda_i \sim \exp(x_i^T \beta) \epsilon_i \]

- Model property:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \frac{\text{Var}[\epsilon_i]}{\mathbb{E}^2[\epsilon_i]} \mathbb{E}^2[y_i|x_i]. \]

- Negative binomial regression (gamma random effect):
  \[ \epsilon_i \sim \text{Gamma}(r, 1/r) = \frac{r^r \epsilon_i^{r-1} e^{-r \epsilon_i}}{\Gamma(r)} \]

- Lognormal-Poisson regression (lognormal random effect):
  \[ \epsilon_i \sim \text{ln}\mathcal{N}(0, \sigma^2) \]
Poisson regression with multiplicative random effects

- **Model:**

\[ y_i \sim \text{Pois}(\lambda_i), \quad \lambda_i \sim \exp(x_i^T \beta) \epsilon_i \]

- **Model property:**

\[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \frac{\text{Var}[\epsilon_i]}{\mathbb{E}^2[\epsilon_i]} \mathbb{E}^2[y_i|x_i]. \]

- **Negative binomial regression (gamma random effect):**

\[ \epsilon_i \sim \text{Gamma}(r, 1/r) = \frac{r^r}{\Gamma(r)} \epsilon_i^{r-1} e^{-r\epsilon_i}. \]

- **Lognormal-Poisson regression (lognormal random effect):**

\[ \epsilon_i \sim \text{lnN}(0, \sigma^2). \]
Poisson regression with multiplicative random effects

- Model:
  \[ y_i \sim \text{Pois}(\lambda_i), \quad \lambda_i \sim \exp(x_i^T \beta) \epsilon_i \]

- Model property:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \frac{\text{Var}[\epsilon_i]}{\mathbb{E}^2[\epsilon_i]} \mathbb{E}^2[y_i|x_i]. \]

- Negative binomial regression (gamma random effect):
  \[ \epsilon_i \sim \text{Gamma}(r, 1/r) = \frac{r^r}{\Gamma(r)} \epsilon_i^{r-1} e^{-r\epsilon_i}. \]

- Lognormal-Poisson regression (lognormal random effect):
  \[ \epsilon_i \sim \text{ln}\mathcal{N}(0, \sigma^2). \]
Lognormal and gamma mixed negative binomial regression

- Model (Zhou et al., ICML2012):
  \[ y_i \sim \text{NegBino}(r, p_i), \quad r \sim \text{Gamma}(a_0, 1/h) \]

- Bayesian inference with the Polya-Gamma distribution.

- Model properties:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \left( e^{\sigma^2} (1 + r^{-1}) - 1 \right) \mathbb{E}^2[y_i|x_i]. \]

- Special cases:
  - Negative binomial regression: \( \sigma^2 = 0; \)
  - Lognormal-Poisson regression: \( r \to \infty; \)
  - Poisson regression: \( \sigma^2 = 0 \) and \( r \to \infty. \)
Lognormal and gamma mixed negative binomial regression

- Model (Zhou et al., ICML2012):
  
  \[ y_i \sim \text{NegBino}(r, p_i), \quad r \sim \text{Gamma}(a_0, 1/h) \]
  
  \[ \psi_i = \log \left( \frac{p_i}{1 - p_i} \right) = x_i^T \beta + \ln \epsilon_i, \quad \epsilon_i \sim \ln \mathcal{N}(0, \sigma^2) \]

- Bayesian inference with the Polya-Gamma distribution.

- Model properties:
  
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \left( e^{\sigma^2 (1 + r^{-1})} - 1 \right) \mathbb{E}^2[y_i|x_i]. \]

- Special cases:
  
  - Negative binomial regression: \( \sigma^2 = 0; \)
  - Lognormal-Poisson regression: \( r \to \infty; \)
  - Poisson regression: \( \sigma^2 = 0 \) and \( r \to \infty. \)
Lognormal and gamma mixed negative binomial regression

- Model (Zhou et al., ICML2012):
  \[ y_i \sim \text{NegBino}(r, p_i), \quad r \sim \text{Gamma}(a_0, 1/h) \]
  \[ \psi_i = \log \left( \frac{p_i}{1 - p_i} \right) = x_i^T \beta + \ln \epsilon_i, \quad \epsilon_i \sim \ln \mathcal{N}(0, \sigma^2) \]

- Bayesian inference with the Polya-Gamma distribution.

- Model properties:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \left( e^{\sigma^2 (1 + r^{-1}) - 1} \right) \mathbb{E}^2[y_i|x_i]. \]

- Special cases:
  - Negative binomial regression: \( \sigma^2 = 0 \);
  - Lognormal-Poisson regression: \( r \to \infty \);
  - Poisson regression: \( \sigma^2 = 0 \) and \( r \to \infty \).
Lognormal and gamma mixed negative binomial regression

- Model (Zhou et al., ICML2012):
  \[ y_i \sim \text{NegBino} (r, p_i), \quad r \sim \text{Gamma}(a_0, 1/h) \]
  \[ \psi_i = \log \left( \frac{p_i}{1 - p_i} \right) = x_i^T \beta + \ln \epsilon_i, \quad \epsilon_i \sim \ln \mathcal{N}(0, \sigma^2) \]

- Bayesian inference with the Polya-Gamma distribution.
- Model properties:
  \[ \text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i] + \left( e^{\sigma^2} (1 + r^{-1}) - 1 \right) \mathbb{E}^2[y_i|x_i]. \]

- Special cases:
  - Negative binomial regression: \( \sigma^2 = 0; \)
  - Lognormal-Poisson regression: \( r \to \infty; \)
  - Poisson regression: \( \sigma^2 = 0 \) and \( r \to \infty. \)
Lognormal and gamma mixed negative binomial regression

- Model (Zhou et al., ICML2012):

\[ y_i \sim \text{NegBino}(r, p_i), \quad r \sim \text{Gamma}(a_0, 1/h) \]

\[ \psi_i = \log \left( \frac{p_i}{1 - p_i} \right) = x_i^T \beta + \ln \epsilon_i, \quad \epsilon_i \sim \ln \mathcal{N}(0, \sigma^2) \]

- Bayesian inference with the Polya-Gamma distribution.
- Model properties:

\[ \text{Var}[y_i | x_i] = \mathbb{E}[y_i | x_i] + \left( e^{\sigma^2} (1 + r^{-1}) - 1 \right) \mathbb{E}^2[y_i | x_i]. \]

- Special cases:
  - Negative binomial regression: \( \sigma^2 = 0; \)
  - Lognormal-Poisson regression: \( r \to \infty; \)
  - Poisson regression: \( \sigma^2 = 0 \) and \( r \to \infty. \)
Count regression example

- Count regression on the NASCAR dataset:

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Poisson (MLE)</th>
<th>NB (MLE)</th>
<th>LGNB (VB)</th>
<th>LGNB (Gibbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
<td>N/A</td>
<td>N/A</td>
<td>0.1396</td>
<td>0.0289</td>
</tr>
<tr>
<td>$r$</td>
<td>N/A</td>
<td>5.2484</td>
<td>18.5825</td>
<td>6.0420</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>-0.4903</td>
<td>-0.5038</td>
<td>-3.5271</td>
<td>-2.1680</td>
</tr>
<tr>
<td>$\beta_1$ (Laps)</td>
<td>0.0021</td>
<td>0.0017</td>
<td>0.0015</td>
<td>0.0013</td>
</tr>
<tr>
<td>$\beta_2$ (Drivers)</td>
<td>0.0516</td>
<td>0.0597</td>
<td>0.0674</td>
<td>0.0643</td>
</tr>
<tr>
<td>$\beta_3$ (TrkLen)</td>
<td>0.6104</td>
<td>0.5153</td>
<td>0.4192</td>
<td>0.4200</td>
</tr>
</tbody>
</table>

- Using Variational Bayes inference, we can calculate the correlation matrix for $(\beta_1, \beta_2, \beta_3)^T$ as

$$
\begin{pmatrix}
1.0000 & -0.4824 & 0.8933 \\
-0.4824 & 1.0000 & -0.7171 \\
0.8933 & -0.7171 & 1.0000
\end{pmatrix}
$$
Latent Dirichlet allocation (Blei et al., 2003)

- Hierarchical model:

\[ x_{ji} \sim \text{Mult}(\phi_{z_{ji}}) \]
\[ z_{ji} \sim \text{Mult}(\theta_j) \]
\[ \phi_k \sim \text{Dir}(\eta, \ldots, \eta) \]
\[ \theta_j \sim \text{Dir} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right) \]

- 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 \( \theta_j \) represents the proportion of the \( K \) topics in the \( j \)th document.
- \( x_{ji} \) is the \( i \)th word in the \( j \)th document.
- \( z_{ji} \) is the index of the topic selected by \( x_{ji} \).
• Denote \( n_{vjk} = \sum_i \delta(x_{ji} = v)\delta(z_{ji} = k) \), \( n_{v.k} = \sum_j n_{vjk} \), \( n_{jk} = \sum_v n_{vjk} \), and \( n_.k = \sum_j n_{jk} \).

• Blocked Gibbs sampling:

\[
P(z_{ji} = k|\cdot) \propto \phi_{x_{ji}k}\theta_{jk}, \quad k \in \{1, \ldots, K\}
\]

\[
(\phi_k|\cdot) \sim \text{Dir}(\eta + n_{1.k}, \ldots, \eta + n_{V.k})
\]

\[
(\theta_j|\cdot) \sim \text{Dir}\left(\frac{\alpha}{K} + n_{j1}, \ldots, \frac{\alpha}{K} + n_{jK}\right)
\]

• 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 \( z^{-ji} \):

\[
P(z_{ji} = k | z^{-ji}) \propto \frac{\eta + n_{xji,k}}{V \eta + n_{-ji,k}} \left( n_{-ji}^{ji} + \frac{\alpha}{K} \right), \quad k \in \{1, \ldots, K\}
\]

- This is easy to understand as

\[
P(z_{ji} = k | \phi_k, \theta_j) \propto \phi_{xji,k} \theta_{jk}
\]

\[
P(z_{ji} = k | z^{-ji}) = \int \int P(z_{ji} = k | \phi_k, \theta_j) P(\phi_k, \theta_j | z^{-ji}) d\phi_k d\theta_j
\]

\[
P(\phi_k | z^{-ji}) = \text{Dir}(\eta + n_{-ji,1.k}, \ldots, \eta + n_{-ji,V.k})
\]

\[
P(\theta_j | z^{-ji}) = \text{Dir}\left(\frac{\alpha}{K} + n_{-ji,j1}, \ldots, \frac{\alpha}{K} + n_{-ji,jK}\right)
\]

\[
P(\phi_k, \theta_j | z^{-ji}) = P(\phi_k | z^{-ji}) P(\theta_j | z^{-ji})
\]
• 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:
Poisson factor analysis

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

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

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: $\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}_+^{V \times K}$.

- Factor score matrix: $\Theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}_+^{K \times N}$.

- 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 $\phi_k$ and $\theta_j$ 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:

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

- Augmentation 2:

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

\[ [n_{vj1}, \cdots, n_{vjk}] \sim \text{Mult} \left( m_{vj}; \zeta_{vj1}, \cdots, \zeta_{vjK} \right) \]
Nonnegative matrix factorization and gamma-Poisson factor analysis

- Gamma priors on $\Phi$ and $\Theta$:
  \[ m_{vj} = \text{Pois} \left( \sum_{k=1}^{K} \phi_{vk} \theta_{jk} \right) \]
  \[ \phi_{vk} \sim \text{Gamma}(a_\phi, 1/b_\phi), \quad \theta_{jk} \sim \text{Gamma}(a_\theta, g_k/a_\theta). \]

- Expectation-Maximization (EM) algorithm:
  \[ \phi_{vk} = \frac{\phi_{vk}^{a_\phi-1}}{b_\phi + \theta_k} \sum_{i=1}^{N} \frac{m_{vj} \theta_{jk}}{\sum_{k=1}^{K} \phi_{vk} \theta_{jk}} \]
  \[ \theta_{jk} = \frac{\theta_{jk}^{a_\theta-1}}{a_\theta/g_k + \phi_k} \sum_{p=1}^{P} \frac{m_{vj} \phi_{vk}}{\sum_{k=1}^{K} \phi_{vk} \theta_{jk}}. \]

- If we set $b_\phi = 0$, $a_\phi = a_\theta = 1$ and $g_k = \infty$, 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 $D_{KL}(M \parallel \Phi \Theta)$. 
Latent Dirichlet allocation and Dirichlet-Poisson factor analysis

- Dirichlet priors on $\Phi$ and $\Theta$:
  \[ m_{vj} = \text{Pois} \left( \sum_{k=1}^{K} \phi_{vk} \theta_{jk} \right) \]
  \[ \phi_k \sim \text{Dir}(\eta, \ldots, \eta), \quad \theta_j \sim \text{Dir}(\alpha/K, \ldots, \alpha/K). \]

- 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


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

\[ \phi_k \sim \text{Dir} (\eta, \cdots, \eta), \]

\[ \theta_{jk} \sim \text{Gamma} [r_j, p_k/(1 - p_k)], \]

\[ r_j \sim \text{Gamma}(e_0, 1/f_0), \]

\[ p_k \sim \text{Beta}[c/K, c(1 - 1/K)]. \]

- \[ n_{jk} = \sum_{v=1}^{V} n_{vjk} \sim \text{NB}(r_j, p_k) \]

- This parametric model becomes a nonparametric Bayesian model governed by the beta-negative binomial process as \( K \to \infty \).
Gamma-gamma-Poisson factor analysis

- Hierarchical model (Zhou and Carin, 2014):

\[ m_{vj} = \sum_{k=1}^{K} n_{vjk}, \quad n_{vjk} \sim \text{Pois}(\phi_v \theta_{jk}) \]

\[ \phi_k \sim \text{Dir}(\eta, \cdots, \eta), \]

\[ \theta_{jk} \sim \text{Gamma}[r_k, p_j/(1 - p_j)], \]

\[ p_j \sim \text{Beta}(a_0, b_0), \]

\[ r_k \sim \text{Gamma}(\gamma_0/K, 1/c). \]

- \( n_{jk} \sim \text{NB}(r_k, p_j) \)

- 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

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

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

\[ x_{ji} \sim \text{Mult}(\phi_{zji}), \quad z_{ji} \sim \sum_{k=1}^{K} \frac{\theta_{jk}}{\sum_{k} \theta_{jk}} \delta_{k}, \quad m_{j} \sim \text{Pois} \left( \sum_{k} \theta_{jk} \right), \]

where \( i = 1, \ldots, m_{j} \) in the \( j \)th 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)

<table>
<thead>
<tr>
<th>Poisson factor analysis priors on $\theta_{jk}$</th>
<th>Infer $(p_k, r_j)$</th>
<th>Infer $(p_j, r_k)$</th>
<th>Support $K \to \infty$</th>
<th>Related algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>NMF</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>LDA</td>
</tr>
<tr>
<td>beta-gamma</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>GaP</td>
</tr>
<tr>
<td>gamma-gamma</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>HDP-LDA</td>
</tr>
</tbody>
</table>
Blocked Gibbs sampling

- Sample $z_{ji}$ from multinomial;
  \[ n_{vjk} = \sum_{i=1}^{m_j} \delta(x_{ji} = v)\delta(z_{ji} = k). \]
- Sample $\phi_k$ from Dirichlet
- For the beta-negative binomial model
  (beta-gamma-Poisson factor analysis)
  - Sample $l_{jk}$ from CRT($n_{jk}, r_j$)
  - Sample $r_j$ from gamma
  - Sample $p_k$ from beta
  - Sample $\theta_{jk}$ from Gamma($r_j + n_{jk}, p_k$)
- For the gamma-negative binomial model
  (gamma-gamma-Poisson factor analysis)
  - Sample $l_{jk}$ from CRT($n_{jk}, r_k$)
  - Sample $r_k$ from gamma
  - Sample $p_j$ from beta
  - Sample $\theta_{jk}$ from 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:

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>trade</td>
<td>rights</td>
<td>environment</td>
<td>women</td>
<td>economic</td>
</tr>
<tr>
<td>world</td>
<td>human</td>
<td>management</td>
<td>gender</td>
<td>summits</td>
</tr>
<tr>
<td>conference</td>
<td>united</td>
<td>protection</td>
<td>equality</td>
<td>outcomes</td>
</tr>
<tr>
<td>organization</td>
<td>nations</td>
<td>affairs</td>
<td>including</td>
<td>conferences</td>
</tr>
<tr>
<td>negotiations</td>
<td>commission</td>
<td>appropriate</td>
<td>system</td>
<td>major</td>
</tr>
</tbody>
</table>

- The gamma-negative binomial and beta-negative binomial models have distinct mechanisms on controlling the number of inferred factors.

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 symmetric 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.
- Hierarchical model:

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

- 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 $\alpha$:

  $$b_{ij} \sim \text{Bernoulli}(p_{z_i z_j}), \text{ for } i > j$$
  $$p_{k_1 k_2} \sim \text{Beta}(a_0, b_0),$$
  $$\mathbf{z}_1, \ldots, \mathbf{z}_N \sim \text{CRP}(\alpha)$$

- 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.
Outline

Analysis of count data
Count matrix factorization and topic modeling
Relational network analysis
Main references

Main references

D. Blei, A. Ng, and M. Jordan.
Latent Dirichlet allocation.

T. L. Griffiths and M. Steyvers.
Finding scientific topics.

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.

Hierarchical Dirichlet processes.
*JASA*, 2006.

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.
Outline

Analysis of count data

Count matrix factorization and topic modeling

Relational network analysis

Main references

M. Zhou and L. Carin.

M. Zhou and L. Carin.

M. Zhou.