Small Variance Asymptotics (SVA) for Nonparametric Latent Feature Relational Model

Anupreet Porwal (12817143)
Avani Samdariya (13173)
Kanupriya Agarwal (13338)

Advisor: Dr. Piyush Rai
IIT Kanpur

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \tag{1}$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \tag{2}$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \tag{3}$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \quad (1)$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \quad (2)$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \quad (3)$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \quad (1)$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \quad (2)$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \quad (3)$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \tag{1}$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \tag{2}$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \tag{3}$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \tag{1}$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \tag{2}$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \tag{3}$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Non-parametric Latent Feature Relational Model (LFRM)[MJG09]

- Uses nonparametric Bayesian (NPB) approach for inferring latent binary features in relational entities
- Simultaneously infers the number of features as well as learns the entities which have that feature
- Generative Story for the model

$$Z \sim IBP(\alpha) \tag{1}$$

$$w_{kk'} \sim \mathcal{N}(0, \sigma^2) \ \forall \text{k, k' features that are non zero} \tag{2}$$

$$y_{ij} \sim \sigma(Z_i^T W Z_j) \text{ for each observation} \tag{3}$$

- Difficult in implementing the sampling algorithms/ variational inference techniques required for approximate inference
- Limited scalability to large-scale data

Small Variance Asymptotics (SVA) approach

- Inspired by connection between mixture of gaussians and K-means algorithm obtained as a limit of EM algorithm
- Yields conceptual link between probabilistic models and their scalable non probabilistic counterpart
- Helps devise a new, simpler and scalable K-means like objective function of the model in question
- SVA has been applied to sequential([RJK13], [HSJ14]) and vector valued i.i.d. data ([JKJ12],[BKJ13], [XMY+15],[KJ11]) but not for models pertaining to relational data

.

Small Variance Asymptotics (SVA) approach

- Inspired by connection between mixture of gaussians and K-means algorithm obtained as a limit of EM algorithm
- Yields conceptual link between probabilistic models and their scalable non probabilistic counterpart
- Helps devise a new, simpler and scalable K-means like objective function of the model in question
- SVA has been applied to sequential([RJK13], [HSJ14]) and vector valued i.i.d. data ([JKJ12],[BKJ13], [XMY+15],[KJ11]) but not for models pertaining to relational data

.

Small Variance Asymptotics (SVA) approach

- Inspired by connection between mixture of gaussians and K-means algorithm obtained as a limit of EM algorithm
- Yields conceptual link between probabilistic models and their scalable non probabilistic counterpart
- Helps devise a new, simpler and scalable K-means like objective function of the model in question
- SVA has been applied to sequential([RJK13], [HSJ14]) and vector valued i.i.d. data ([JKJ12],[BKJ13], [XMY+15],[KJ11]) but not for models pertaining to relational data

.

Small Variance Asymptotics (SVA) approach

- Inspired by connection between mixture of gaussians and K-means algorithm obtained as a limit of EM algorithm
- Yields conceptual link between probabilistic models and their scalable non probabilistic counterpart
- Helps devise a new, simpler and scalable K-means like objective function of the model in question
- SVA has been applied to sequential([RJK13], [HSJ14]) and vector valued i.i.d. data ([JKJ12],[BKJ13], [XMY^{+}15],[KJ11]) but not for models pertaining to relational data

.

Methodology for solving the problem

- Utilised connection between exponential families and Bregman divergence[JKJ12] to scale the covariance of exponential families
- SVA is directly applied to the posterior of Bayesian Non-Parametric model using MAP-based asymptotics derivation [BKJ13] to obtain a k-means like objective
- Formulation of an efficient greedy algorithm, inspired by [XMY+15], to solve the optimisation problem

Methodology for solving the problem

- Utilised connection between exponential families and Bregman divergence[JKJ12] to scale the covariance of exponential families
- SVA is directly applied to the posterior of Bayesian Non-Parametric model using MAP-based asymptotics derivation [BKJ13] to obtain a k-means like objective
- Formulation of an efficient greedy algorithm, inspired by [XMY+15], to solve the optimisation problem

Methodology for solving the problem

- Utilised connection between exponential families and Bregman divergence[JKJ12] to scale the covariance of exponential families
- SVA is directly applied to the posterior of Bayesian Non-Parametric model using MAP-based asymptotics derivation [BKJ13] to obtain a k-means like objective
- Formulation of an efficient greedy algorithm, inspired by [XMY+15], to solve the optimisation problem

Methodology for solving the problem

- Utilised connection between exponential families and Bregman divergence[JKJ12] to scale the covariance of exponential families
- SVA is directly applied to the posterior of Bayesian Non-Parametric model using MAP-based asymptotics derivation [BKJ13] to obtain a k-means like objective
- Formulation of an efficient greedy algorithm, inspired by [XMY+15], to solve the optimisation problem

Bregman Divergence and Scaled Bernoulli Distribution

Express Bernoulli Distribution in its canonical form using bijective relationship between bregman divergence and exponential families [BMDG05]

$$P(X|\eta,\psi) = \exp\left[x\eta - \psi(\eta) - h_1(x)\right]$$

$$h_1(x) = 0 \qquad \eta = log\left(\frac{q}{1-q}\right) \qquad \psi(\eta) = \log\left(1+e^{\eta}\right)$$

Using lemma 3.1 of [KJ11] and bregman divergence corresponding to Bernoulli [BMDG05], we get

$$\tilde{P}(x|\tilde{\eta},\tilde{\psi}) = \tilde{P}(x|\tilde{\mu}) = \exp\left\{-d_{\tilde{\phi}}(x,\tilde{\mu})\right\} \times f_{\tilde{\phi}}(x) = \exp\left\{-d_{\tilde{\phi}}(x,\mu)\right\} \times f_{\tilde{\phi}}(x)$$

where, $f_{\tilde{\phi}}(x) = (f_{\phi}(x))^{\beta}$, $\tilde{\phi} = \beta\phi$ and

$$d_{\phi}(x,\mu) = x\log\frac{x}{q} + (1-x)\log\frac{1-x}{1-q}$$

$$f_{\phi}(x) = \exp\left(x\log x + (1-x)\log(1-x)\right)$$

Bregman Divergence and Scaled Bernoulli Distribution

Express Bernoulli Distribution in its canonical form using bijective relationship between bregman divergence and exponential families [BMDG05]

$$P(X|\eta,\psi) = \exp\left[x\eta - \psi(\eta) - h_1(x)\right]$$

$$h_1(x) = 0 \qquad \eta = log\left(\frac{q}{1-q}\right) \qquad \psi(\eta) = \log\left(1 + e^{\eta}\right)$$

Using lemma 3.1 of [KJ11] and bregman divergence corresponding to Bernoulli [BMDG05], we get

$$\tilde{P}(x|\tilde{\eta},\tilde{\psi}) = \tilde{P}(x|\tilde{\mu}) = \exp\left\{-d_{\tilde{\phi}}(x,\tilde{\mu})\right\} \times f_{\tilde{\phi}}(x) = \exp\left\{-d_{\tilde{\phi}}(x,\mu)\right\} \times f_{\tilde{\phi}}(x)$$

where, $f_{\tilde{\phi}}(x) = (f_{\phi}(x))^{\beta}$, $\tilde{\phi} = \beta\phi$ and

$$d_{\phi}(x,\mu) = x\log\frac{x}{q} + (1-x)\log\frac{1-x}{1-q}$$

$$f_{\phi}(x) = \exp\left(x\log x + (1-x)\log(1-x)\right)$$

Express Bernoulli Distribution in its canonical form using bijective relationship between bregman divergence and exponential families [BMDG05]

$$P(X|\eta, \psi) = \exp\left[x\eta - \psi(\eta) - h_1(x)\right]$$

$$h_1(x) = 0 \qquad \eta = log\left(\frac{q}{1-q}\right) \qquad \psi(\eta) = \log\left(1 + e^{\eta}\right)$$

Using lemma 3.1 of [KJ11] and bregman divergence corresponding to Bernoulli [BMDG05], we get

$$\tilde{P}(x|\tilde{\eta}, \tilde{\psi}) = \tilde{P}(x|\tilde{\mu}) = \exp\left\{-d_{\tilde{\phi}}(x, \tilde{\mu})\right\} \times f_{\tilde{\phi}}(x) = \exp\left\{-d_{\tilde{\phi}}(x, \mu)\right\} \times f_{\tilde{\phi}}(x)$$

where, $f_{\tilde{\phi}}(x) = (f_{\phi}(x))^{\beta}$, $\tilde{\phi} = \beta\phi$ and

$$d_{\phi}(x, \mu) = x \log \frac{x}{q} + (1-x) \log \frac{1-x}{1-q}$$

$$f_{\phi}(x) = \exp\left(x \log x + (1-x) \log (1-x)\right)$$

Express Bernoulli Distribution in its canonical form using bijective relationship between bregman divergence and exponential families [BMDG05]

$$P(X|\eta, \psi) = \exp\left[x\eta - \psi(\eta) - h_1(x)\right]$$

$$h_1(x) = 0 \qquad \eta = log\left(\frac{q}{1-q}\right) \qquad \psi(\eta) = \log\left(1 + e^{\eta}\right)$$

Using lemma 3.1 of [KJ11] and bregman divergence corresponding to Bernoulli [BMDG05], we get

$$\tilde{P}(x|\tilde{\eta}, \tilde{\psi}) = \tilde{P}(x|\tilde{\mu}) = \exp\left\{-d_{\tilde{\phi}}(x, \tilde{\mu})\right\} \times f_{\tilde{\phi}}(x) = \exp\left\{-d_{\tilde{\phi}}(x, \mu)\right\} \times f_{\tilde{\phi}}(x)$$

where, $f_{\tilde{\phi}}(x) = (f_{\phi}(x))^{\beta}$, $\tilde{\phi} = \beta\phi$ and

$$d_{\phi}(x, \mu) = x \log\frac{x}{q} + (1-x)\log\frac{1-x}{1-q}$$

$$f_{\phi}(x) = \exp\left(x \log x + (1-x)\log(1-x)\right)$$

Posterior and Prior

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \;+\; const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp\left(-\beta\lambda^2\right)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \ + \ const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp(-\beta\lambda^2)$. Chosen such so that number of features get smaller as $\beta \rightarrow \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^{+}\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Posterior and Prior

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \; + \; const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp\left(-\beta\lambda^2\right)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \; + \; const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp\left(-\beta\lambda^2\right)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \ + \ const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp(-\beta\lambda^2)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp-(\beta\lambda^2)}{n} + constant(w.r.t.\ \beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Posterior and Prior

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) + const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp(-\beta\lambda^2)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\,\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \ + \ const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp(-\beta\lambda^2)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Posterior and Prior

Joint Posterior:

$$L(W,Z) = P(Z,W|Y) \propto P(Y|Z,W)P(Z)P(W)$$

$$-\log L(W,Z) = -\log P(Y|W,Z) - \log P(Z) - \log P(W) \ + \ const.$$

Prior on Feature Matrix: $Z \sim IBP(\alpha)$

- $\alpha = \exp(-\beta\lambda^2)$. Chosen such so that number of features get smaller as $\beta \to \infty$
- Avoids over-fitting of data to features

$$-\log P(Z) = K^+\beta\lambda^2 + \sum_{n=1}^{N} \frac{\exp -(\beta\lambda^2)}{n} + constant(w.r.t.\ \beta)$$

Prior on Weights: $W \sim N(0, \sigma_w^2)$

$$-\log P(W) = \sum_{k=1}^{K^+} \sum_{k'=1}^{K^+} \frac{w_{kk'}}{2\sigma^2} + constant(\sigma)$$

Likelihood

Scaled Bernoulli Likelihood:

$$
\begin{aligned}
P(Y|Z,W) &= \prod_{i,j=1}^{N} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\
&= \prod_{i,j=1}^{N} \exp\left[-\beta\left[y_{ij}\log\frac{y_{ij}}{p_{ij}} + (1-y_{ij})\log\left(\frac{1-y_{ij}}{1-p_{ij}}\right)\right]\right] \\
&\quad \times \exp\left[\beta\left[y_{ij}\log y_{ij} + (1-y_{ij})\log(1-y_{ij})\right]\right]
\end{aligned}
$$

$$
-\log P(Y|Z,W) = \sum_{i=1}^{N}\sum_{j=1}^{N}\beta\left[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})\right]
$$

Scaled Bernoulli Likelihood:

$$
\begin{aligned}
P(Y|Z,W) &= \prod_{i,j=1}^{N} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\
&= \prod_{i,j=1}^{N} \exp\left[-\beta[y_{ij}\log\frac{y_{ij}}{p_{ij}} + (1-y_{ij})\log(\frac{1-y_{ij}}{1-p_{ij}})]\right] \\
&\quad \times \exp\left[\beta[y_{ij}\log y_{ij} + (1-y_{ij})\log(1-y_{ij})]\right]
\end{aligned}
$$

$$
-\log P(Y|Z,W) = \sum_{i=1}^{N}\sum_{j=1}^{N} \beta[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})]
$$

Likelihood

Scaled Bernoulli Likelihood:

$$
\begin{aligned}
P(Y|Z,W) &= \prod_{i,j=1}^{N} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \\
&= \prod_{i,j=1}^{N} \exp\left[-\beta\left[y_{ij}\log\frac{y_{ij}}{p_{ij}} + (1-y_{ij})\log\left(\frac{1-y_{ij}}{1-p_{ij}}\right)\right]\right] \\
&\quad \times \exp\left[\beta\left[y_{ij}\log y_{ij} + (1-y_{ij})\log(1-y_{ij})\right]\right]
\end{aligned}
$$

$$
-\log P(Y|Z,W) = \sum_{i=1}^{N}\sum_{j=1}^{N}\beta\left[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})\right]
$$

After substituting the priors for W and Z and also, Bernoulli likelihood:

$$-\frac{\log L(W,Z)}{\beta} = K^{+}\lambda^{2} + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp-(\beta\lambda^{2})}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \to \infty$, Objective function, $Q(W,Z)$ is obtained

$$Q(W,Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^{T}Wz_j) + \log(1+\exp(z_i^{T}Wz_j))] + K^{+}\lambda^{2}$$

$$where, p_{ij} = \sigma(z_i^{T}Wz_j)$$

It has to be minimized w.r.t. W and Z

After substituting the priors for W and Z and also, Bernoulli likelihood:

$$-\frac{\log L(W,Z)}{\beta} = K^{+}\lambda^{2} + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp-(\beta\lambda^{2})}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \rightarrow \infty$, Objective function, $Q(W,Z)$ is obtained

$$Q(W,Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^{T}Wz_j) + \log(1+\exp(z_i^{T}Wz_j))] + K^{+}\lambda^{2}$$

$$where, p_{ij} = \sigma(z_i^{T}Wz_j)$$

It has to be minimized w.r.t. W and Z

After substituting the priors for W and Z and also, Bernoulli likelihood:

$$-\frac{\log L(W,Z)}{\beta} = K^{+}\lambda^{2} + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp -(\beta\lambda^{2})}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \to \infty$, Objective function, $Q(W,Z)$ is obtained

$$Q(W,Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^T W z_j) + \log(1+\exp(z_i^T W z_j))] + K^{+}\lambda^{2}$$

$$where, p_{ij} = \sigma(z_i^T W z_j)$$

It has to be minimized w.r.t. W and Z

After substituting the priors for W and Z and also, Bernoulli likelihood:

$$-\frac{\log L(W,Z)}{\beta} = K^{+}\lambda^{2} + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp -(\beta\lambda^{2})}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \to \infty$, Objective function, $Q(W,Z)$ is obtained

$$Q(W,Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^T W z_j) + \log(1+\exp(z_i^T W z_j))] + K^{+}\lambda^{2}$$

$$where, p_{ij} = \sigma(z_i^T W z_j)$$

It has to be minimized w.r.t. W and Z

After substituting the priors for W and Z and also, Bernoulli likelihood:

$$-\frac{\log L(W,Z)}{\beta} = K^{+}\lambda^{2} + -\sum_{i=1}^{N}\sum_{j=1}^{N}[y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})] + \frac{\exp-(\beta\lambda^{2})}{\beta}\sum_{n=1}^{N}\frac{1}{n} + O(\frac{1}{\beta})$$

As $\beta \to \infty$, Objective function, $Q(W,Z)$ is obtained

$$Q(W,Z) = \sum_{i=1}^{N}\sum_{j=1}^{N}[-y_{ij}(z_i^T W z_j) + \log(1+\exp(z_i^T W z_j))] + K^{+}\lambda^{2}$$

$$where, p_{ij} = \sigma(z_i^T W z_j)$$

It has to be minimized w.r.t. W and Z

K-LAFTER I(Latent feature learning on relational data) algorithm

1: Set C=1. Initialise Z as a N x C matrix by setting $Z_{n1} = 1$ with probability 0.5 $\forall n = 1 \dots N$
2: Initialise W as C x C matrix with entries drawn from $\mathcal{N}(0, \sigma^2)$
Iterate until no convergence
3: $\forall n$, optimise Q with respect to Z_n over all $2^C - 1$ possibilities
4: Optimise Q(W,Z) with respect to W for current Z and C values
5: Construct Z' from Z by adding a new feature as (C+1) column with one randomly initialised n having that feature
6: Augment W in a similar manner by drawing entries from $\mathcal{N}(0, \sigma^2)$ to form a C+1 dimensional square matrix W'
7: Optimise Q with respect to W' for Z' and C+1 features
8: Optimise Q with respect to Z' for current W' and C+1 features
9: If (C+1, W', Z') lowers Q from (C,W,Z), replace latter with former

Empirical Results

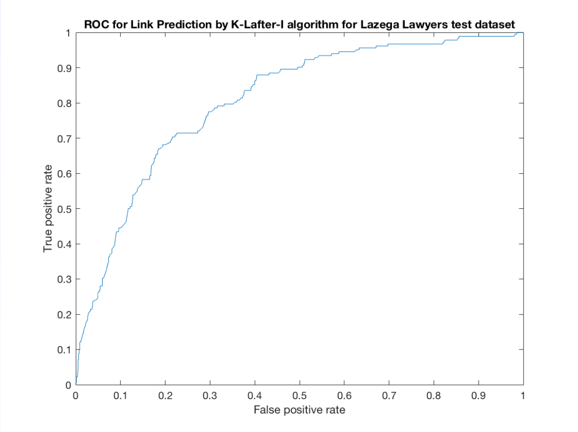

Table: AUC score and Time Complexity

AUC score	0.8067
Time Taken	5.9 Hrs

Lets get more Greedy!

K-LAFTER II(Latent feature learning on relational data) algorithm

1: Set C=1. Initialise Z as a N x C matrix by setting $Z_{n1} = 1$ with probability 0.5 $\forall n = 1 \dots N$
2: Initialise W as C x C matrix with entries drawn from $\mathcal{N}(0, \sigma^2)$
Iterate until no convergence
3: $\forall n, c$, Choose the optimal value(0 or 1) of each Z_{nc}
4: Optimise Q(W,Z) with respect to W for current Z and C values
5: Construct Z' from Z by adding a new feature as (C+1) column with one randomly initialised n having that feature
6: Augment W in a similar manner by drawing entries from $\mathcal{N}(0, \sigma^2)$ to form a C+1 dimensional square matrix W'
7: Optimise Q with respect to W' for Z' and C+1 features
8: Optimise Q with respect to Z' for current W' and C+1 features
9: If (C+1, W', Z') lowers Q from (C,W,Z), replace latter with former

Reduced time with comparable accuracy :)

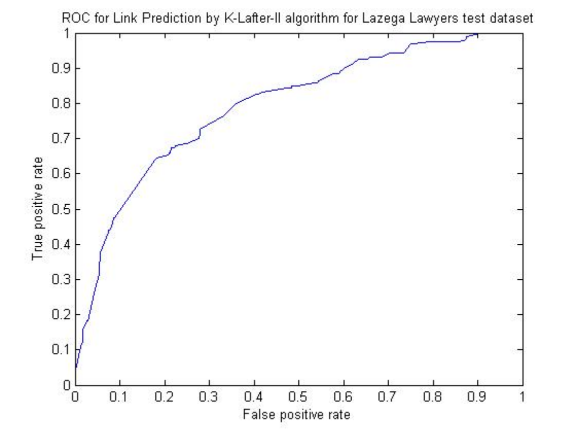

Table: AUC score and Time Complexity

AUC score	0.7932
Time Taken	0.9 Hrs

Results of K-LAFTER II for NIPS234

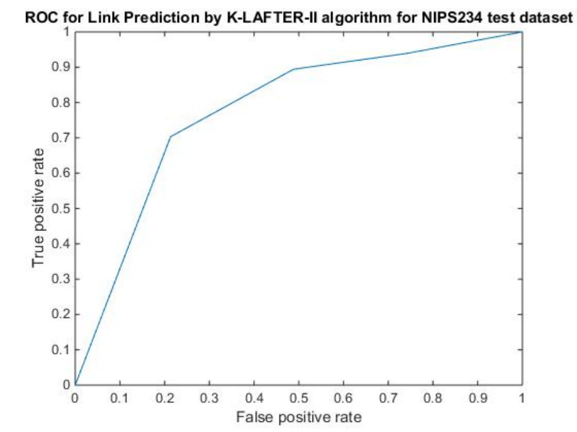

Table: AUC score and Time Complexity

AUC score	0.7732
Time Taken	3 Hrs

Conclusion

- Developed a connection between Non- Parametric LFRM and its non-probabilistic counterpart using SVA
- Obtained a scalable K-means style objective function with flexibility of NPB techniques through penalty term on number of features using MAD-Bayes approach [BKJ13]
- Propose a greedy algorithm to optimise the objective function over feature matrix Z and weight matrix W

Conclusion

- Developed a connection between Non- Parametric LFRM and its non-probabilistic counterpart using SVA
- Obtained a scalable K-means style objective function with flexibility of NPB techniques through penalty term on number of features using MAD-Bayes approach [BKJ13]
- Propose a greedy algorithm to optimise the objective function over feature matrix Z and weight matrix W

Conclusion

- Developed a connection between Non- Parametric LFRM and its non-probabilistic counterpart using SVA
- Obtained a scalable K-means style objective function with flexibility of NPB techniques through penalty term on number of features using MAD-Bayes approach [BKJ13]
- Propose a greedy algorithm to optimise the objective function over feature matrix Z and weight matrix W

Future Extensions

- Use Bernoulli-Poisson Link function[MSH11] with IBP prior on Z to devise an algorithm that scales up computationally only with number of links present

- Apply SVA to Infinite Edge Partition Model (IFPM)[Zho15], to obtain a more scalable and fast algorithm for identifying latent feature structure relational data

Future Extensions

- Use Bernoulli-Poisson Link function[MSH11] with IBP prior on Z to devise an algorithm that scales up computationally only with number of links present
- Apply SVA to Infinite Edge Partition Model (IFPM)[Zho15], to obtain a more scalable and fast algorithm for identifying latent feature structure relational data

References I

- Tamara Broderick, Brian Kulis, and Michael I Jordan. Mad-bayes: Map-based asymptotic derivations from bayes. In *ICML (3)*, pages 226–234, 2013.
- Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.
- Jonathan H Huggins, Ardavan Saeedi, and Matthew J Johnson. Detailed derivations of small-variance asymptotics for some hierarchical bayesian nonparametric models. *arXiv preprint arXiv:1501.00052*, 2014.
- Ke Jiang, Brian Kulis, and Michael I Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3158–3166, 2012.

References II

Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284, 2009.

Morten Mørup, Mikkel N Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011.

Anirban Roychowdhury, Ke Jiang, and Brian Kulis. Small-variance asymptotics for hidden markov models. In *Advances in Neural Information Processing Systems*, pages 2103–2111, 2013.

References III

Yanxun Xu, Peter Müller, Yuan Yuan, Kamalakar Gulukota, and Yuan Ji. Mad bayes for tumor heterogeneity—feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514, 2015.

Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.