Mathematical Derivation of Formulas for Specialized Review Recommender Systems

Luciana B. Maroun
lubm@dcc.ufmg.br

1 CAP

Context-Aware review helpfulness rating Prediction (CAP) is a method for recommending reviews through a sum of latent factors and dot product of latent vectors [3]. Each latent factor or vector has a normal distribution centered on the regression of observed features, being classified in the set of Regression Based Latent Factor Models [1]. Latent variables and regression weights are fitted using a Monte Carlo Expectation Maximization algorithm. In order to do so, it is necessary to devise the complete log-likelihood expression, the expectation of this value with respect to the posterior distribution of Ω, and derivatives of the latter, whose calculations are presented in the following subsections.

1.1 Complete log-likelihood

In this subsection, we derive the complete log-likelihood for the CAP system.

For this calculation, we use the formulas for normal distribution ($\mathcal{N}$) and multivariate normal distribution ($\mathcal{MVN}$) defined below.

$$P_{\mathcal{N}}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$P_{\mathcal{MVN}}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k det\mathbf{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

Using these probability densities, the log-likelihood is given by the joint probability of observing the data (H_{ij}) and the latent variables $\Omega_{ij} = (\alpha_i, \beta_j, \xi_k, \lambda_i^k, \gamma_i^k, \mathbf{u}_i, \mathbf{v}_j)$. Thus,

$$\begin{aligned} L(\Omega, \Theta) = &\prod_{i,j} P(H_{ij}|\Omega_{ij}) \cdot \prod_i P(\alpha_i) \cdot \prod_j P(\beta_j) \cdot \prod_k P(\xi_k) \\ &\cdot \prod_{i,k:k\in S_i} P(\gamma_i^k) \cdot \prod_{i,k:k\in T_i} P(\lambda_i^k) \\ &\cdot \prod_i P(\mathbf{u}_i) \cdot \prod_j P(\mathbf{v}_j) \end{aligned}$$

In this formula, S_i is the set of users similar to user i and T_i is the set of users trusted by user i. Also k is uniquely defined for a given review j through a function a which retrieves the author

of a review of index j, that is, $a(j) = k$. We omitted the ranges of indices for simpler exposition: i, j and k are in the range of all indices for voter, review and author; (i, j) is in the set of user-review pairs with an associated vote (from user to review); (i, k) is in the set of user-user pairs with an available vote from i to a review writer by k, restricted to a condition (belonging to S_i or T_i). For a comprehensive list of definitions of each variable, refer to original work [3].

All latent variables are considered to be independent of each other, except H_{ij} which depends on variables Ω_{ij} through mean of distribution $\hat{H}_i j$. The presented probability is the factorization over the probabilistic graphic model presented in section 1.3.

Then, we replace the probabilities using the probability density formula.

$$
\begin{aligned}
L(\Omega, \Theta) = & \prod_{i,j} \left(\frac{1}{\sigma_H \sqrt{2\pi}} e^{\frac{-(H_{ij} - \hat{H}_{ij})^2}{2\sigma_H^2}} \right) \cdot \prod_i \left(\frac{1}{\sigma_\alpha \sqrt{2\pi}} e^{\frac{-(\alpha_i - \mathbf{d}^T \mathbf{y}_i)^2}{2\sigma_\alpha^2}} \right) \\
& \cdot \prod_j \left(\frac{1}{\sigma_\beta \sqrt{2\pi}} e^{\frac{-(\beta_j - \mathbf{g}^T \mathbf{x}_j)^2}{2\sigma_\beta^2}} \right) \cdot \prod_k \left(\frac{1}{\sigma_\xi \sqrt{2\pi}} e^{\frac{-(\xi_k - \mathbf{b}^T \mathbf{z}_k)^2}{2\sigma_\xi^2}} \right) \\
& \cdot \prod_{i,k:k \in S_i} \left(\frac{1}{\sigma_\gamma \sqrt{2\pi}} e^{\frac{-(\gamma_i^k - f(\mathbf{r}^T \mathbf{p}_i^k))^2}{2\sigma_\gamma^2}} \right) \cdot \prod_{i,k:k \in T_i} \left(\frac{1}{\sigma_\lambda \sqrt{2\pi}} e^{\frac{-(\lambda_i^k - f(\mathbf{n}^T \mathbf{q}_i^k))^2}{2\sigma_\lambda^2}} \right) \\
& \cdot \prod_i \left(\frac{1}{\sqrt{(2\pi)^K det \mathbf{A}_u}} e^{-\frac{1}{2}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T \mathbf{A}_u^{-1} (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)} \right) \\
& \cdot \prod_j \left(\frac{1}{\sqrt{(2\pi)^K det \mathbf{A}_v}} e^{-\frac{1}{2}(\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T \mathbf{A}_v^{-1} (\mathbf{v}_i - \mathbf{V}\mathbf{x}_j)} \right)
\end{aligned}
$$

Where K is the number of dimensions of latent vectors u and v. To obtain the log-likelihood $\ell(\Omega, \Theta)$, we take the neperian logarithm of this value using the property that the logarithm of a product of terms is the sum of the logarithm of each term.

$$
\begin{aligned}
\ell(\Omega,\Theta) = &\sum_{i,j}\left(\log\frac{1}{\sigma_H\sqrt{2\pi}} - \frac{(H_{ij}-\hat{H}_{ij})^2}{2\sigma_H^2}\right) + \sum_{i}\left(\log\frac{1}{\sigma_\alpha\sqrt{2\pi}} - \frac{(\alpha_i - \mathbf{d}^T\mathbf{y}_i)^2}{2\sigma_\alpha^2}\right) \\
&+ \sum_{j}\left(\log\frac{1}{\sigma_\beta\sqrt{2\pi}} - \frac{(\beta_j - \mathbf{g}^T\mathbf{x}_j)^2}{2\sigma_\beta^2}\right) + \sum_{k}\left(\log\frac{1}{\sigma_\xi\sqrt{2\pi}} - \frac{(\xi_k - \mathbf{b}^T\mathbf{z}_k)^2}{2\sigma_\xi^2}\right) \\
&+ \sum_{i,k:k\in S_i}\left(\log\frac{1}{\sigma_\gamma\sqrt{2\pi}} - \frac{(\gamma_i^k - f(\mathbf{r}^T\mathbf{p}_i^k))^2}{2\sigma_\gamma^2}\right) \\
&+ \sum_{i,k:k\in T_i}\left(\log\frac{1}{\sigma_\lambda\sqrt{2\pi}} - \frac{(\lambda_i^k - f(\mathbf{n}^T\mathbf{q}_i^k))^2}{2\sigma_\lambda^2}\right) \\
&+ \sum_{i}\left(\log\frac{1}{\sqrt{(2\pi)^K det\mathbf{A}_u}} - \frac{1}{2}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T\mathbf{A}_u^{-1}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)\right) \\
&+ \sum_{j}\left(\log\frac{1}{\sqrt{(2\pi)^K det\mathbf{A}_v}} - \frac{1}{2}(\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T\mathbf{A}_v^{-1}(\mathbf{v}_i - \mathbf{V}\mathbf{x}_j)\right)
\end{aligned}
$$

Simplifying further,

$$
\begin{aligned}
\ell(\Omega,\Theta) = &-\sum_{i,j}\left(\log\sigma_H + \frac{1}{2}\log 2\pi + \frac{(H_{ij}-\hat{H}_{ij})^2}{2\sigma_H^2}\right) \\
&- \sum_{i}\left(\log\sigma_\alpha + \frac{1}{2}\log 2\pi + \frac{(\alpha_i - \mathbf{d}^T\mathbf{y}_i)^2}{2\sigma_\alpha^2}\right) \\
&- \sum_{j}\left(\log\sigma_\beta + \frac{1}{2}\log 2\pi + \frac{(\beta_j - \mathbf{g}^T\mathbf{x}_j)^2}{2\sigma_\beta^2}\right) \\
&- \sum_{k}\left(\log\sigma_\xi + \frac{1}{2}\log 2\pi + \frac{(\xi_k - \mathbf{b}^T\mathbf{z}_k)^2}{2\sigma_\xi^2}\right) \\
&- \sum_{i,k:k\in S_i}\left(\log\sigma_\gamma + \frac{1}{2}\log 2\pi + \frac{(\gamma_i^k - f(\mathbf{r}^T\mathbf{p}_i^k))^2}{2\sigma_\gamma^2}\right) \\
&- \sum_{i,k:k\in T_i}\left(\log\sigma_\lambda + \frac{1}{2}\log 2\pi + \frac{(\lambda_i^k - f(\mathbf{n}^T\mathbf{q}_i^k))^2}{2\sigma_\lambda^2}\right) \\
&- \sum_{i}\left(\frac{K}{2}\log 2\pi + \frac{1}{2}\log det\mathbf{A}_u + \frac{1}{2}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T\mathbf{A}_u^{-1}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)\right) \\
&- \sum_{j}\left(\frac{K}{2}\log 2\pi + \frac{1}{2}\log det\mathbf{A}_v + \frac{1}{2}(\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T\mathbf{A}_v^{-1}(\mathbf{v}_i - \mathbf{V}\mathbf{x}_j)\right)
\end{aligned}
$$

Collapsing the terms which are constant across EM iterations, whose sum is represented by C, and replacing standard deviation by variance, we obtain the following.

$$
\begin{aligned}
\ell(\Omega, \Theta) = C &- \sum_{i,j} \left(\frac{1}{2} \log \sigma_H^2 + \frac{(H_{ij} - \hat{H}_{ij})^2}{2\sigma_H^2} \right) \\
&- \sum_{i} \left(\frac{1}{2} \log \sigma_\alpha^2 + \frac{(\alpha_i - \mathbf{d}^T \mathbf{y}_i)^2}{2\sigma_\alpha^2} \right) - \sum_{j} \left(\frac{1}{2} \log \sigma_\beta^2 + \frac{(\beta_j - \mathbf{g}^T \mathbf{x}_j)^2}{2\sigma_\beta^2} \right) \\
&- \sum_{k} \left(\frac{1}{2} \log \sigma_\xi^2 + \frac{(\xi_k - \mathbf{b}^T \mathbf{z}_k)^2}{2\sigma_\xi^2} \right) - \sum_{i,k:k \in S_i} \left(\frac{1}{2} \log \sigma_\gamma^2 + \frac{(\gamma_i^k - f(\mathbf{r}^T \mathbf{p}_i^k))^2}{2\sigma_\gamma^2} \right) \\
&- \sum_{i,k:k \in T_i} \left(\frac{1}{2} \log \sigma_\lambda^2 + \frac{(\lambda_i^k - f(\mathbf{n}^T \mathbf{q}_i^k))^2}{2\sigma_\lambda^2} \right) \\
&- \sum_{i} \left(\frac{1}{2} \log det \mathbf{A}_u + \frac{1}{2} (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T \mathbf{A}_u^{-1} (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i) \right) \\
&- \sum_{j} \left(\frac{1}{2} \log det \mathbf{A}_v + \frac{1}{2} (\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T \mathbf{A}_v^{-1} (\mathbf{v}_i - \mathbf{V}\mathbf{x}_j) \right) \\
&= C - \sum_{i,j} \frac{1}{2} \left(\log \sigma_H^2 + \frac{(H_{ij} - \hat{H}_{ij})^2}{\sigma_H^2} \right) \\
&- \sum_{i} \frac{1}{2} \left(\log \sigma_\alpha^2 + \frac{(\alpha_i - \mathbf{d}^T \mathbf{y}_i)^2}{\sigma_\alpha^2} \right) - \sum_{j} \frac{1}{2} \left(\log \sigma_\beta^2 + \frac{(\beta_j - \mathbf{g}^T \mathbf{x}_j)^2}{\sigma_\beta^2} \right) \\
&- \sum_{k} \frac{1}{2} \left(\log \sigma_\xi^2 + \frac{(\xi_k - \mathbf{b}^T \mathbf{z}_k)^2}{\sigma_\xi^2} \right) - \sum_{i,k:k \in S_i} \frac{1}{2} \left(\log \sigma_\gamma^2 + \frac{(\gamma_i^k - f(\mathbf{r}^T \mathbf{p}_i^k))^2}{\sigma_\gamma^2} \right) \\
&- \sum_{i,k:k \in T_i} \frac{1}{2} \left(\log \sigma_\lambda^2 + \frac{(\lambda_i^k - f(\mathbf{n}^T \mathbf{q}_i^k))^2}{\sigma_\lambda^2} \right) \\
&- \sum_{i} \frac{1}{2} \left(\log det \mathbf{A}_u + (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T \mathbf{A}_u^{-1} (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i) \right) \\
&- \sum_{j} \frac{1}{2} \left(\log det \mathbf{A}_v + (\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T \mathbf{A}_v^{-1} (\mathbf{v}_i - \mathbf{V}\mathbf{x}_j) \right)
\end{aligned}
$$

We can expand $\hat{H}_{ij}$ in terms of latent variables, since its probability is conditioned on the set of latent variables Ω_{ij}.

$$
\begin{aligned}
\ell(\Omega,\Theta) = C &- \sum_{i,j}\frac{1}{2}\left(\log\sigma_H^2 + \frac{(H_{ij} - \mathbf{u}_i^T\mathbf{v_j} - \alpha_i - \beta_j - \xi_k - \delta_1(i,k)\gamma_i^k - \delta_2(i,k)\lambda_i^k)^2}{\sigma_H^2}\right) \\
&- \sum_i \frac{1}{2}\left(\log\sigma_\alpha^2 + \frac{(\alpha_i - \mathbf{d}^T\mathbf{y}_i)^2}{\sigma_\alpha^2}\right) - \sum_j \frac{1}{2}\left(\log\sigma_\beta^2 + \frac{(\beta_j - \mathbf{g}^T\mathbf{x}_j)^2}{\sigma_\beta^2}\right) \\
&- \sum_k \frac{1}{2}\left(\log\sigma_\xi^2 + \frac{(\xi_k - \mathbf{b}^T\mathbf{z}_k)^2}{\sigma_\xi^2}\right) - \sum_{i,k:k\in S_i} \frac{1}{2}\left(\log\sigma_\gamma^2 + \frac{(\gamma_i^k - f(\mathbf{r}^T\mathbf{p}_i^k))^2}{\sigma_\gamma^2}\right) \\
&- \sum_{i,k:k\in T_i} \frac{1}{2}\left(\log\sigma_\lambda^2 + \frac{(\lambda_i^k - f(\mathbf{n}^T\mathbf{q}_i^k))^2}{\sigma_\lambda^2}\right) \\
&- \sum_i \frac{1}{2}\left(\log(det\mathbf{A}_u) + (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T\mathbf{A}_u^{-1}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)\right) \\
&- \sum_j \frac{1}{2}\left(\log det\mathbf{A}_v + (\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T\mathbf{A}_v^{-1}(\mathbf{v}_i - \mathbf{V}\mathbf{x}_j)\right)
\end{aligned}
$$

Aggregating terms from the same summation, we obtain the following.

$$
\begin{aligned}
\ell(\Omega,\Theta) = C &- \frac{1}{2}\sum_{i,j}\left(\frac{1}{\sigma_H^2}(H_{ij} - \mathbf{u}_i^T\mathbf{v_j} - \alpha_i - \beta_j - \xi_k - \delta_1(i,k)\gamma_i^k - \delta_2(i,k)\lambda_i^k)^2 + \log\sigma_H^2\right) \\
&- \frac{1}{2}\sum_i\left(\frac{(\alpha_i - \mathbf{d}^T\mathbf{y}_i)^2}{\sigma_\alpha^2} + \log\sigma_\alpha^2 + (\mathbf{u}_i - \mathbf{W}\mathbf{y}_i)^T\mathbf{A}_u^{-1}(\mathbf{u}_i - \mathbf{W}\mathbf{y}_i) + \log(det\mathbf{A}_u)\right) \\
&- \frac{1}{2}\sum_j\left(\frac{(\beta_j - \mathbf{g}^T\mathbf{x}_j)^2}{\sigma_\beta^2} + \log\sigma_\beta^2 + (\mathbf{v}_j - \mathbf{V}\mathbf{x}_j)^T\mathbf{A}_v^{-1}(\mathbf{v}_i - \mathbf{V}\mathbf{x}_j) + \log(det\mathbf{A}_v) + \right) \\
&- \frac{1}{2}\sum_k\left(\log\sigma_\xi^2 + \frac{(\xi_k - \mathbf{b}^T\mathbf{z}_k)^2}{\sigma_\xi^2}\right) - \frac{1}{2}\sum_{i,k:k\in S_i}\left(\log\sigma_\gamma^2 + \frac{(\gamma_i^k - f(\mathbf{r}^T\mathbf{p}_i^k))^2}{\sigma_\gamma^2}\right) \\
&- \frac{1}{2}\sum_{i,k:k\in T_i}\left(\log\sigma_\lambda^2 + \frac{(\lambda_i^k - f(\mathbf{n}^T\mathbf{q}_i^k))^2}{\sigma_\lambda^2}\right)
\end{aligned}
$$

We can observe that this is a modification of the formula presented for Regression Based Latent Factor Models [1] by incorporating new latent variables and dropping linear regression on dyadic features term ($x_{ij}^T b$). Using the definition of $E[\phi|Rest] = \hat{\phi}$ such that $\ell'(\hat{\phi}) = 0$ and $V[\phi|Rest] = \ell''(\hat{\phi})^{-1}$ [3], where ϕ is any variable, we have estimations of mean and variance, thus the probability function, for each variable to be used in Gibbs Sampling.

1.2 Expectation of log-likelihood with respect to posterior of Ω

We do not have a closed formula for the joint distribution of variables due to the product $\mathbf{u}_i^T\mathbf{v}_j$, that is, we do not know the distribution of a product of $\mathcal{MVN}s$. Then, this distribution is approximated

through Gibbs Sampling. An empirical expectation (E^*) can be taken from this approximated distribution, considering a sample of L elements.

$$E[X] = \sum_x x \; p(x)$$
$$E^*[X] = \frac{1}{L} \sum_{i=1}^{L} x_i$$

Where x_i is the value of random variable X in i-th sample. We want the expected value of a function. Thus,

$$E[f(X)] = \sum_x f(x) \; p(x)$$
$$E^*[f(X)] = \frac{1}{L} \sum_{i=1}^{L} f(x_i)$$

After Gibbs Sampling step, we have L samples of the whole latent vector:

$$\Omega = (\alpha_1, ..., \alpha_{N_v}, \beta_1, ..., \beta_M, \xi_1, ..., \xi_{N_a}, \mathbf{u}_1, ..., \mathbf{u}_{N_v}, \mathbf{v}_1, ..., \mathbf{v}_{N_r}, \gamma_{i_1}^{k_1}, ..., \gamma_{i_{N_s}}^{k_{N_s}}, \lambda_{i_1}^{k_1}, ..., \lambda_{i_{N_t}}^{k_{N_t}})$$

In this definition, N_v is the number of voters, N_r is the number of reviews, N_a is the number of author, N_s is the number of pairs (i, k) where user i evaluated a review written by k and i has k in its list of similar users, and N_t is the number of pairs (i, k) where user i evaluated a review written by k and i trusts k. We can, then, compute the probabilities for each occurred value or simply sum over the function f evaluated over each sample to obtain the empirical distribution of ℓ.

$$E^* = \sum_{\{\Omega\}} p(\Omega | y_{ij}, \Theta) \; \ell(\Omega, \Theta)$$
$$= \frac{1}{L} \sum_{l=1}^{L} \; \ell(\Omega^l, \Theta)$$

Where $\{\Omega\}$ is the set of all possible full latent vectors, Ω_{ij}^l is the value of latent factor vector in l-th sample. In this case, the log-likelihood is a function of the latent variables, but, after the sample, we marginalize out of these values through the summation. We know how to calculate $\ell(\Omega, \Theta)$ because we have the values of the latent variables on each sample as well as the true value of the helpfulness (H_{ij}). Then, the result E^* is a function of only the parameters Θ, after the sample.

Since the value of E* is known, it is possible to derive it. As it is a summation, we can use the property of the derivative of a sum being the sum of derivatives. The formulas for all parameters can be easily found and set to zero in order to find an optimal value, which falls in OLS linear regression on empiric mean of each variable. However, for parameters $\mathbf{r}$, $\mathbf{h}$ and their variances, we have a non-standard linear regression to solve, because of the sigmoid function f. Instead, Newton-Raphson method can be used, which turns out to be much simpler in this situation. Newton-Raphson is a method for finding a root of a function; since we want to maximize E^*, we find the root of its derivative through this method. Consequently, we have to calculate the first and second partial derivatives of E^* with respect to $\mathbf{r}$ and $\mathbf{h}$, the first is the function whose root is going to be estimated and the latter is its derivative. Only terms in $\ell(\Theta, \Omega)$ containing r are relevant and, first, we consider a single dimension r_j.

$$
\begin{aligned}
\frac{\partial E^*}{\partial \mathbf{r}_j} &= \frac{\partial}{\partial \mathbf{r}_j} \frac{1}{L} \sum_{l=1}^{L} -\frac{1}{2} \sum_{i,k:k\in S_i} \left(\frac{(\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))^2}{\sigma_\gamma^2} \right) \\
&= \frac{1}{L} \sum_{l=1}^{L} -\frac{1}{2} \frac{2}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))(-1) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} \frac{1}{L} \sum_{l=1}^{L} \sum_{i,k:k\in S_i} (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k)) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \frac{1}{L} \sum_{l=1}^{L} (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k)) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \sum_{i,k:k\in S_i} \frac{1}{L} \sum_{l=1}^{L} \gamma_i^{kl} - \frac{1}{l} \sum_{\{\delta_{ij}^l\}} f(\mathbf{r}^T \mathbf{p}_i^k) \\
&= \frac{1}{\sigma_\gamma^2} f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \sum_{i,k:k\in S_i} \left(E^*(\gamma_i^k) - \frac{1}{l} l f(\mathbf{r}^T \mathbf{p}_i^k) \right) \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left(E^*(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k) \right) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k
\end{aligned}
$$

We consider $f'(\mathbf{r}^T \mathbf{p}_i^k)$ to be $\frac{\partial f(\mathbf{r}^T \mathbf{p}_i^k)}{\partial (\mathbf{r}^T \mathbf{p}_i^k)}$. $E^*(\gamma_i^k)$ is the empiric mean of γ_i^k variable considering L samples. This solution is no different than a non-linear regression on the empiric mean of a variable by minimizing sum of squares. In this calculation, we use the Chain Rule of a composite function. In this scenario,

$$
\frac{\partial (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))^2)}{\partial \mathbf{r}_j} = \frac{\partial (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))^2}{\partial (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))} \cdot \frac{\partial (\gamma_i^{kl} - f(\mathbf{r}^T \mathbf{p}_i^k))}{\partial (f(\mathbf{r}^T \mathbf{p}_i^k))} \cdot \frac{\partial (f(\mathbf{r}^T \mathbf{p}_i^k))}{\partial (\mathbf{r}^T \mathbf{p}_i^k)} \cdot \frac{\partial (\mathbf{r}^T \mathbf{p}_i^k)}{\partial \mathbf{r}_j}
$$

The vector derivative is just the extrapolation of this using the definition below.

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}_1} \; \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}_2} \; \cdots \; \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}_n} \right]$$

Then,

$$\begin{aligned}
\frac{\partial E^*}{\partial \mathbf{r}} &= \left[\frac{\partial E^*}{\partial \mathbf{r}_1} \; \frac{\partial E^*}{\partial \mathbf{r}_2} \; \cdots \; \frac{\partial E^*}{\partial \mathbf{r}_n} \right] \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left(E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k) \right) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_i^k
\end{aligned}$$

The term $\mathbf{p}_i^k$ is a vector, which turns this result into a vector. The term left are the same for all $\mathbf{r}_j$, which makes this formula easily extensible to vector case. The formula for $\mathbf{h}$ is similar.

In order to calculate the second order derivative by an index j' of $\mathbf{r}$, we have now a derivative of a vector by a scalar. This is defined as deriving each dimension j, which is $\frac{\partial E^*}{\mathbf{r}_j}$. Considering the dimensions j and j',

$$\begin{aligned}
\frac{\partial^2 E^*}{\partial \mathbf{r}_j \partial \mathbf{r}_{j'}} &= \frac{\partial}{\partial \mathbf{r}_{j'}} \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left(E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k) \right) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left(\frac{\partial (E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k))}{\partial \mathbf{r}_{j'}} f'(\mathbf{r}^T \mathbf{p}_i^k) + (E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k)) \frac{\partial f'(\mathbf{r}^T \mathbf{p}_i^k)}{\partial \mathbf{r}_{j'}} \right) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left((-1) f'(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij'}^k f'(\mathbf{r}^T \mathbf{p}_i^k) + (E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k)) f''(\mathbf{r}^T \mathbf{p}_i^k) \mathbf{p}_{ij'}^k \right) \mathbf{p}_{ij}^k \\
&= \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left((E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k)) f''(\mathbf{r}^T \mathbf{p}_i^k) - (f'(\mathbf{r}^T \mathbf{p}_i^k))^2 \right) \mathbf{p}_{ij}^k \mathbf{p}_{ij'}^k
\end{aligned}$$

We used the Product Rule in this differentiation. The first generalization consists of the second derivative in one dimension of the full vector.

$$\frac{\partial^2 E^*}{\partial \mathbf{r} \partial \mathbf{r}_{j'}} = \frac{1}{\sigma_\gamma^2} \sum_{i,k:k\in S_i} \left((E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k)) f''(\mathbf{r}^T \mathbf{p}_i^k) - (f'(\mathbf{r}^T \mathbf{p}_i^k))^2 \right) \mathbf{p}_{ij'}^k \mathbf{p}_i^k$$

Considering p_i^k to be a column vector, we have the first derivative as a column vector and the combination of columns compose a matrix. Each column of the matrix is multiplied by the corresponding $\mathbf{p}_{ij'}^k$.

$$\frac{\partial^2 E^*}{\partial \mathbf{r} \partial \mathbf{r}^T} = \frac{1}{\sigma_\gamma^2} \sum_{i,k:k \in S_i} \left((E^*_{\{\delta_{ij}^l\}}(\gamma_i^k) - f(\mathbf{r}^T \mathbf{p}_i^k)) f''(\mathbf{r}^T \mathbf{p}_i^k) - (f'(\mathbf{r}^T \mathbf{p}_i^k))^2 \right) \mathbf{p}_i^k \mathbf{p}_i^k$$

Since we consider $\mathbf{p}_i^k$ to be a column vector, the multiplication of a column vector by a row vector results in the desired multiplication $\mathbf{p}_{ij}^k \mathbf{p}_{ij'}^k$ in position (j, j') of the second derivative matrix.

We just have to define, now, the values of the derivatives of the sigmoid function (f) and the first and second order derivatives are completely defined.

$$f'(x) = \frac{e^x}{(1+e^x)^2}$$
$$f''(x) = \frac{e^x(e^x - 1)}{(1+e^x)^3}$$

1.3 Probabilistic graphical model

We may define a probabilistic graphical model for CAP method, presented in figure 1. In this figure, we use the same definition of set sizes as in Subsection 1.2.

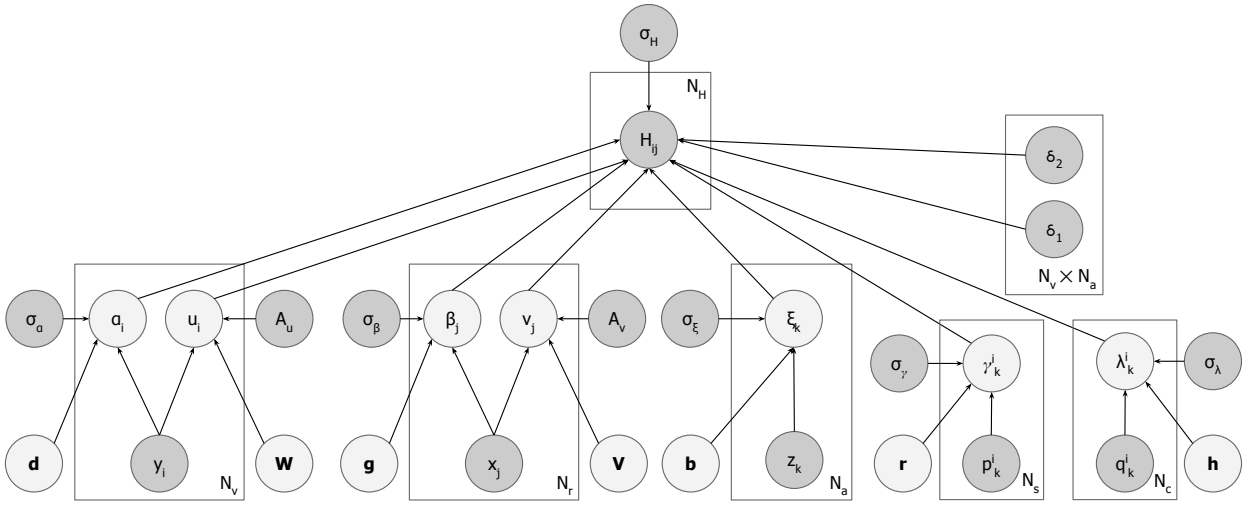

Figure 1: Probabilistic Graphical Model represented as a Bayesian Network for CAP.

2 BETF

The unBiased Extended Tensor Factorization is a predictor of review helpfulness which uses extensions over tensor factorization. The tensor factorization models three latent vectors of dimension K: related to voter, related to product and related to review's author [2]. The extension included

in this model is represented by the addition of a new factorization to the objective function related to the matrix of product ratings. Bias are adjusted to account for part of the vote not explained by interaction.

This method uses a stochastic gradient descent to fit latent factors. Thus, the derivative of the objective function has to be calculated regarding each latent factor and bias variable.

The goal is to minimize the following expression, where N_v is the number of voters (users who evaluated reviews) and N_r is the number of reviews.

$$E_{BETF} = \frac{1}{2}\sum_{i}^{N_u}\sum_{j}^{N_v}\sum_{k}^{N_p} I_{ij}^{q}(q_{ijk} - g(\hat{q_{ijk}}))^2 + \frac{1}{2}\sum_{j}^{N_v}\sum_{k}^{N_p} I_{ij}^{o}(o_{jk} - g(\hat{o_{jk}}))^2 + \frac{\lambda_u}{2}\sum_{i}^{N_u}\|\mathbf{u}_i\|_{Fro}^2 + \frac{\lambda_v}{2}\sum_{j}^{N_v}\|\mathbf{v}_j\|_{Fro}^2 + \frac{\lambda_p}{2}\sum_{k}^{N_p}\|\mathbf{p}_k\|_{Fro}^2 + \frac{\lambda_s}{2}\sum_{ijk}\|\mathbf{s}_{ijk}\|_{Fro}^2$$

Where g is the logistic function, λ_v and λ_r are $\frac{\sigma_q^2}{\sigma_v^2}$ and $\frac{\sigma_q^2}{\sigma_r^2}$, respectively. For a complete description of variables, refer to original work [2].

Stochastic gradient descent, though, only regards the derivative of a single example. So, we may consider the optimization formula only regarding a single instance (i, j, k) presence in the training set (no need of indicator function).

$$E_{BETF}^{ijk} = \frac{1}{2}(q_{ijk} - g(\hat{q_{ijk}}))^2 + \frac{1}{2}(o_{jk} - g(\hat{o_{jk}}))^2 + \frac{\lambda_u}{2}\|\mathbf{u}_i\|_{Fro}^2 + \frac{\lambda_v}{2}\|\mathbf{v}_j\|_{Fro}^2 + \frac{\lambda_p}{2}\|\mathbf{p}_k\|_{Fro}^2 + \frac{\lambda_s}{2}\|\mathbf{s}_{ijk}\|_{Fro}^2$$

In order to apply stochastic gradient descent, we have to find the derivative of this expression related to $\mathbf{u_i}$, $\mathbf{v_j}$, $\mathbf{p_k}$, $\mathbf{s}_{ijk}$, $\hat{q_{\mathbf{u}_i}}$, $\hat{q_{r_jk}}$, $\hat{o_{\mathbf{v}_j}}$, and $\hat{o_{\mathbf{p}_k}}$. The calculation of all of them are presented below.

$$\begin{aligned}\frac{\partial E_{BETF}^{ijk}}{\partial \mathbf{u}_i} &= \frac{1}{2}2(q_{ij} - g(\hat{q_{ijk}}))g'(\hat{q_{ijk}})\frac{\partial h_{ijk}}{\partial \mathbf{u}_i} + \frac{\lambda_u}{2}2\mathbf{u}_i \\ &= (q_{ij} - g(\hat{q_{ijk}}))g'(\hat{q_{ijk}})(\mathbf{s} \times_v \mathbf{v}_j \times_p \mathbf{p}_k) + \lambda_u\mathbf{u}_i\end{aligned}$$

Where we calculated $\frac{\partial h_{ijk}}{\mathbf{u}_i}$ as follows.

$$
\begin{aligned}
\frac{\partial h_{ijk}}{\mathbf{u}_{id}} &= \frac{\partial(\sum_{x=1}^{K} u_{ix} \sum_{y=1}^{K} \sum_{z=1}^{K} \mathbf{s}_{xyz}\mathbf{v}_{jy}\mathbf{p}_{kz})}{\partial \mathbf{u}_{id}} \\
&= \frac{\partial(u_{id} \sum_{y=1}^{K} \sum_{z=1}^{K} \mathbf{s}_{dyz}\mathbf{v}_{jy}\mathbf{p}_{kz})}{\partial \mathbf{u}_{id}} \\
&= \sum_{y=1}^{K}\sum_{z=1}^{K} \mathbf{s}_{dyz}\mathbf{v}_{jy}\mathbf{p}_{kz} \\
&= \sum_{z=1}^{K}\sum_{y=1}^{K} \mathbf{s}_{dyz}\mathbf{v}_{jy}\mathbf{p}_{kz} \\
&= \sum_{z=1}^{K} \mathbf{p}_{kz} \sum_{y=1}^{K} \mathbf{s}_{dyz}\mathbf{v}_{jy} \\
&= \sum_{z=1}^{K} \mathbf{p}_{kz} (\mathbf{s} \times_v \mathbf{v}_j)_{dz} \\
&= (\mathbf{s} \times_v \mathbf{v}_j \times_p \mathbf{p}_k)_d \\
\frac{\partial h_{ijk}}{\mathbf{u}_i} &= \mathbf{s} \times_v \mathbf{v}_j \times_p \mathbf{p}_k
\end{aligned}
$$

Which is a (K, 1) vector. For $\mathbf{v}_j$ and $\mathbf{p}_k$, we also have terms related to extended matrix of ratings.

$$
\begin{aligned}
\frac{\partial E^{ijk}_{BETF}}{\partial \mathbf{v}_j} &= (q_{ij} - g(\hat{q_{ijk}}))g'(\hat{q_{ijk}})(\mathbf{s} \times_u \mathbf{u}_i \times_p \mathbf{p}_k) + (o_{jk} - g(\hat{o_{jk}}))g'(\hat{o_{jk}})\mathbf{p}_k + \lambda_v \mathbf{v}_j \\
\frac{\partial E^{ijk}_{BETF}}{\partial \mathbf{p}_k} &= (q_{ij} - g(\hat{q_{ijk}}))g'(\hat{q_{ijk}})(\mathbf{s} \times_u \mathbf{u}_i \times_v \mathbf{v}_j) + (o_{jk} - g(\hat{o_{jk}}))g'(\hat{o_{jk}})\mathbf{v}_j + \lambda_p \mathbf{p}_k
\end{aligned}
$$

In the algorithm proposed, differently from typical stochastic gradient descent, each derivative term is updated separately, avoiding multiple updates regarding the same related entities. For example, the second term of the last derivative is related to author and product pair, which repeats for all votes given to the review written by the author to the product. The same happens with regularization terms, which are updated once for each latent variable in each iteration instead of for every helpfulness vote. The proposed algorithm update terms from different sources separately: first, regarding voter error; second, regarding rating error; and finally, regarding regularization. This way, vote error is prioritized over rating error and regularization. Although we present derivative terms together, they are updated separately.

Given the definition of h_{ijk}, we have the following.

$$
\frac{\partial h_{ijk}}{\mathbf{s}_{xyz}} = \mathbf{u}_{ix}\mathbf{v}_{jy}\mathbf{p}_{kz}
$$

So, each dimension is multiplied by the corresponding dimension in $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{p}$. Thus, we can extended to the whole tensor case.

$$\frac{\partial h_{ijk}}{\mathbf{s}} = \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{p}_k$$

Where $\otimes$ is the outer product of tensors, which produces a new one with the sum of both dimensions. Consequently, we have the following derivative regarding the objective function.

$$\frac{\partial E^{ijk}_{BETF}}{\partial \mathbf{s}} = (q_{ij} - g(\hat{q_{ijk}}))g'(\hat{q_{ijk}})(\mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{p}_k) + \lambda_s \mathbf{s}$$

The fitting algorithms is a variation of stochastic gradient descent. For each seen example (i, j, k) (or (j, k) for terms related to ratings, or i or j or k for regression terms), we perform the following update for each variable ϕ.

$$\phi = \phi - \alpha \frac{\partial E^{ijk}_{BETF}}{\partial \phi}$$

In this last formula, α is the learning rate.

References

[1] Deepak Agarwal and Bee-Chung Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 19–28, New York, NY, USA, 2009. ACM.

[2] Samaneh Moghaddam, Mohsen Jamali, and Martin Ester. Etf: Extended tensor factorization model for personalizing prediction of review helpfulness. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 163–172, Seattle, Washington, USA, 2012.

[3] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 1–8, Hong Kong, China, 2013.