3.1 Probabilistic Matrix Factorization

PMF [22] is one of the most famous matrix factorization models in collaborative filtering, which decomposes the partially observed data matrix \( X \) into the product of two low-rank latent feature matrices, \( U \) and \( V \), where \( U \in \mathbb{R}^{K \times N} \), \( V \in \mathbb{R}^{K \times M} \), and \( K \ll \min(N,M) \).

By assuming Gaussian distribution on the residual noise of observed data and placing Gaussian priors on the latent feature matrices, PMF tries to maximize the log-likelihood of the posterior distribution on the user and item features as follows:

\[
\mathcal{L}_{PMF} = - \sum_{(i,j,x) \in \mathcal{Q}} \frac{(x - U_i^T V_j)^2}{2\sigma^2} - \frac{\|U\|_F^2}{2\sigma_U^2} - \frac{\|V\|_F^2}{2\sigma_V^2}.
\]

(6)

This is equivalent to minimizing a squared loss with regularization defined as follows:

\[
\mathcal{E} = \frac{1}{2} \sum_{(i,j,x) \in \mathcal{Q}} (x - U_i^T V_j)^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2,
\]

(7)

where \( \lambda_U = \sigma^2/\sigma_U^2 \) and \( \lambda_V = \sigma^2/\sigma_V^2 \) are positive constants to control the trade-off between the loss and the regularization terms. \( \| \cdot \|_F \) denotes the Frobenious norm.

After training the PMF model via gradient descent or stochastic gradient algorithms [22], the predicted rating that user \( i \) would assign to item \( j \) can be computed as the expected mean of the Gaussian distribution \( \hat{x}_{ij} = U_i^T V_j \).

3.2 Response Aware PMF

In Sec. 2, we have demonstrated that by neglecting response patterns, not only do we lose the potential information that might boost the model performance, but also can it lead to incorrect or biased parameter estimation. Due to the effectiveness and interpretability of PMF, we will unify it with explicit response models, which we refer to as Response Aware PMF (RAPMF).

Replacing \( \theta \) in Eq. (2) by the low-rank latent feature matrices in PMF, we have

\[
P(R|X,U,V,\mu,\sigma^2) = P(R|U,V,\mu,\sigma^2)P(X|U,V,\sigma^2).
\]

(8)

The probability of full model, \( P(R|X,U,V,\mu,\sigma^2) \), is decomposed into data model \( P(X|U,V,\sigma^2) \) and the missing data model \( P(R|X,U,V,\mu,\sigma^2) \).

3.3 Response Model

Modeling the missing data successfully requires a correct and tractable distribution on the response patterns. Bernoulli distribution is an intuitive distribution to explain data missing phenomena [16]. Depending on whether users’ and items’ features are incorporated, we propose two response models, rating dominant response model and context-aware response model.

3.4 Rating Dominant Response Model

For the sake of simplification, we assume the probability that a user chooses to rate an item follows a Bernoulli distribution given the rating assigned is \( k \). Hence, for a scale of 1 to \( D \), the rating dominant response model has \( D \) parameters \( \mu_1, \mu_2, \cdots, \mu_D \).

If \( X \) is fully observed, then the response mechanism can be modeled as [16]:

\[
P(R|X,U,V,\mu,\sigma^2) = P(R|X,\mu)
\]

\[
= \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^D (\mu_k^{r_{ij}=1} (1 - \mu_k)^{r_{ij}=0}) P(x_{ij} = k|U,V,\sigma^2),
\]

(9)

where \( [r = 0] \) is an indicator variable that outputs 1 if the expression is valid and 0 otherwise. It is noted that Eq. (9) adopts the “winner-take-all” scheme, i.e., a hard assignment scheme, to model users’ response on a particular rating.

However, in real-world recommender system, the data is not fully observed. The “winner-take-all” scheme brings the risk of deteriorating assignment probability when the data is recovered based on the learned model. Hence, we adopt a soft assignment using probability of the possible rating values in the response model as follows:

\[
P(R|X,U,V,\mu,\sigma^2) = P(R|U,V,\mu,\sigma^2)
\]

\[
= \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^D (\mu_k^{r_{ij}=1} (1 - \mu_k)^{r_{ij}=0}) P(x_{ij} = k|U,V,\sigma^2),
\]

(10)

where \( P(X|U,V,\sigma^2) \), the probability of \( X \) being assigned to \( k \), can be set to \( \mathcal{N}(k|U^T V, \sigma^2) \) as is in [22].

To relieve the inaccuracy issue when recovering the original model, we further introduce a discount parameter \( \beta \) on the assignment probability

\[
P(R|X,U,V,\mu,\sigma^2) = P(R|U,V,\mu,\sigma^2)
\]

\[
\propto \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^D (\mu_k^{r_{ij}=1} (1 - \mu_k)^{r_{ij}=0}) \mathcal{N}(k|U^T V, \sigma^2)^\beta,
\]

(12)