Bias Aware Probabilistic Boolean Matrix Factorization Supplementary Material

Changlin Wan[1,2] **Pengtao Dang**[1] **Tong Zhao**[3] **Yong Zang**[1] **Chi Zhang**[1] **Sha Cao**[1]

[1]Indiana University, Indianapolis, Indiana, United States
[2]Purdue University, West Lafayette, Indiana, United States
[3]Amazon, Seattle, Washington, United States

Appendix

A MESSAGE UPDATE

Despite the dense connections in the factor graph, max-sum belief propagation achieved admirable performance in the case of approximate the MAP of Boolean matrix factorization [Ravanbakhsh et al., 2016]. Here we also utilize this strategy that not only derive the MAP of matrix decomposition X and Y, but also infer the background row- and column-wise bias $\boldsymbol{\mu}, \boldsymbol{\nu}$. Though the information of $\boldsymbol{\mu}, \boldsymbol{\nu}$ and X, Y communicates through likelihood factor g and auxiliary variable W, their independence of each other resulted in disconnected message update between $\boldsymbol{\mu}, \boldsymbol{\nu}$ and X, Y. Conveniently, $\{X,\ Y,\ W\}$ and $\boldsymbol{\mu}, \boldsymbol{\nu},\ W\}$ can be considered as two separate systems. In this paper we focus on the message update of $\{\boldsymbol{\mu}, \boldsymbol{\nu},\ W\}$, and adopt the algorithm in Ravanbakhsh et al. [2016] for $\{X,\ Y,\ W\}$.

A.0.1 update X,Y,W

Variables to factor message.

Conveniently, all the variables in $\{X, Y, W\}$ are binary variables (X_{il}, Y_{lj}, $W_{ijl} \in \{0, 1\}$). Following the notation in Ravanbakhsh et al. [2016], we denote the message between factors and variables as $\mathbf{m}$ (e.g., $\mathbf{m}_{X_{il}\to f_{ijl}}(X_{ij}) : \{0, 1\} \to \mathcal{R}$). Max-sum BP is utilized to calculate the outgoing message, while consideration all incoming messages from neighbor factors, despite the receiving one, e.g.,

$$\mathbf{m}_{X_{il}\to f_{ijl}}(X_{ij})^{t+1} = \mathbf{m}_{h_{il}\to X_{ill}}(X_{il})^t + \sum_{j'\neq j} \mathbf{m}_{f_{ij'l}\to X_{il}}(X_{il})^t$$

Our objective is to achieve the maximum likelihood, which align with the difference between the message of $\mathbf{m}_{X_{il}\to f_{ijl}}(X_{ij} = 1)$ and $\mathbf{m}_{X_{il}\to f_{ijl}}(X_{ij} = 0)$, i.e.,

$$\hat{\Phi} = \mathbf{m}_{X_{il}\to f_{ijl}}(1) - \mathbf{m}_{X_{il}\to f_{ijl}}(0)$$

In the case of individual variable X_{il} to the factor f_{ijl}

$$\begin{aligned}\hat{\Phi}^{t+1}_{ijl} &= (\mathbf{m}_{h_{il}\to X_{ill}}(1)^t + \sum_{j'\neq j} \mathbf{m}_{f_{ij'l}\to X_{il}}(1)^t) \\ &- (\mathbf{m}_{h_{il}\to X_{ill}}(0)^t + \sum_{j'\neq j} \mathbf{m}_{f_{ij'l}\to X_{il}}(0)^t) \\ &= log(\frac{p(X_{il}=1)}{p(X_{il}=0)}) + \sum_{j'\neq j} \Phi^t_{ij'l}\end{aligned}$$

Similarly, the message $\hat{\Psi}$ can be derived as

$$\hat{\Psi}_{ijl} = log(\frac{p(Y_{lj}=1)}{p(Y_{lj}=1)}) + \sum_{i'\neq i} \Psi^t_{i'jl}$$

For W, since each variable W_{ijl} has exact two factor neighbors $g_{ij},\ f_{ijl}$, the message from W_{ijk} to either factors is the message from the other factor, i.e.,

$$\mathbf{m}_{W_{ijl}\to g_{ij}}(W_{ijl}) = \mathbf{m}_{f_{ijl}\to W_{ijl}}(W_{ijl})$$

$$\mathbf{m}_{g_{ij}\to W_{ijl}}(W_{ijl}) = \mathbf{m}_{W_{ijl}\to f_{ijl}}(W_{ijl})$$

We will discuss in detail of the message involve factor g in next section.

factor to variable message

For factor h, it only connect to the single variable X_{il} or Y_{lj}, which works as prior knowledge for the sparsity of X and Y, where their information is passed through

$$h_{il}(X_{il}=1) - h_{il}(X_{il}=0) = log(\frac{p(X_{il}=1)}{p(X_{il}=0)})$$

$$h_{lj}(Y_{lj}=1) - h_{lj}(Y_{lj}=0) = log(\frac{p(Y_{lj}=1)}{p(Y_{lj}=0)})$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Factor f links X, Y with the auxiliary variable W, that ensures $W_{ijl} = X_{il} \wedge Y_{lj}$, i.e.,

$$f(X_{il}, Y_{lj}, W_{ijl}) = log(\mathcal{I}(W_{ijl} = X_{il} \wedge Y_{lj}))$$

Notably, $f(X_{il}, Y_{lj}, W_{ijl}) \rightarrow -\infty$ if $W_{ijl} \neq X_{il} \wedge Y_{lj}$. Such that it restrict the message scenarios when passing the information from f to X, Y. Here, we use $\mathbf{m}_{f_{ijl} \rightarrow X_{il}(X_{il})}$ as example, where $\mathbf{m}_{f_{ijl} \rightarrow Y_{lj}(Y_{lj})}$ can be similarly derived. For X_{il} to equal to 1, if $Y_{lj} = 1$, restricted by f, $W_{ijl} = 1$, and if $Y_{lj} = 0$, $W_{ijl} = 0$, thus,

$$\begin{aligned}&\mathbf{m}_{f_{ijl} \rightarrow X_{il}}(1)^{t+1} = max(\mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(1)^t + \\ &\mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(1)^t, \quad \mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(0)^t + \mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(0)^t)\end{aligned}$$

While if $X_{il} = 0$, $W_{ijl} = 0$ regardless the value of Y_{lj}, i.e.,

$$\begin{aligned}&\mathbf{m}_{f_{ijl} \rightarrow X_{il}}(0)^{t+1} = max(\mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(1)^t + \\ &\mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(0)^t, \quad \mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(1)^t + \mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(0)^t)\end{aligned}$$

Since $\hat{\Psi}_{ijl} = \mathbf{m}_{Y_{lj} \rightarrow f_{ijl}}(1) - \mathbf{m}_{Y_{lj} \rightarrow f_{ijl}}(0)$, and $\Gamma_{ijl} = \mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(1) - \mathbf{m}_{W_{ijl} \rightarrow f_{ijl}}(0)$ the message from f to X can be derived as

$$\begin{aligned}\Phi_{ijl} &= \mathbf{m}_{f_{ijl} \rightarrow X_{il}}(1) - \mathbf{m}_{f_{ijl} \rightarrow X_{il}}(0) \\ &= max(\Gamma_{ijl} + \hat{\Psi}_{ijl}, 0) - max(\hat{\Psi}_{ijl}, 0)\end{aligned}$$

Similarly, we have

$$\begin{aligned}\Psi_{ijl} &= \mathbf{m}_{f_{ijl} \rightarrow Y_{il}}(1) - \mathbf{m}_{f_{ijl} \rightarrow Y_{il}}(0) \\ &= max(\Gamma_{ijl} + \hat{\Phi}_{ijl}, 0) - max(\hat{\Phi}_{ijl}, 0)\end{aligned}$$

Following the same strategy, while considering the message from factor f to variable W, if $W_{ijl} = 1$, $X_{il} = Y_{lj} = 1$, whereas if $W_{ijl} = 0$, either X_{il} or Y_{lj} should equal to zero, i.e.,

$$\mathbf{m}_{f_{ijl} \rightarrow W_{ijl}}(1)^{t+1} = \mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(1)^t + \mathbf{m}_{X_{il} \rightarrow f_{ijl}}(1)^t$$

$$\begin{aligned}&\mathbf{m}_{f_{ijl} \rightarrow W_{ijl}}(0)^{t+1} = max(\mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(1)^t + \\ &\mathbf{m}_{X_{il} \rightarrow f_{ijl}}(0)^t, \mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(0)^t + \mathbf{m}_{X_{il} \rightarrow f_{ijl}}(1)^t, \\ &\mathbf{m}_{Y_{lj} \rightarrow f_{ill}}(0)^t + \mathbf{m}_{X_{il} \rightarrow f_{ijl}}(0)^t)\end{aligned}$$

Such that

$$\begin{aligned}\hat{\Gamma}_{ijk} &= \mathbf{m}_{f_{ijl} \rightarrow W_{ijl}}(1) - \mathbf{m}_{f_{ijl} \rightarrow W_{ijl}}(0) \\ &= min(\hat{\Phi}_{ijl} + \hat{\Psi}_{ijl}, \hat{\Phi}_{ijl}, \hat{\Psi}_{ijl})\end{aligned}$$

A.1 UPDATE μ, ν, W

In the previous section, we have derived the messages passing between the X, Y and W. In this section, we derive the message passing between μ, $\mathbf{v}$ and W, where they all related to the likelihood factor g. Also, different with binary variable W, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are Bernoulli variable, that the simplified singleton message does not applied for their message update. We first reinstate the log likelihood function of each element (A_{ij}) that represent the factor g.

$$\begin{aligned}&p(A_{ij} = 1 | Z_{ij} = 0) = 1 - (1 - p_f)(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j) \\ &p(A_{ij} = 0 | Z_{ij} = 0) = (1 - p_f)(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j) \\ &p(A_{ij} = 1 | Z_{ij} = 1) = 1 - p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j) \\ &p(A_{ij} = 0 | Z_{ij} = 1) = p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)\end{aligned}$$

In the case of Bernoulli variables $\boldsymbol{\mu}_i$, the incoming message from factor g to $\boldsymbol{\mu}_i$ is certainly the likelihood information,

$$\Omega_i = log(\prod_{j=1}^{n} p(A_{ij}))$$

while the message from $\boldsymbol{\mu}_i$ to g would be the MAP of the posterior distribution, i.e.,

$$\begin{aligned}\hat{\Omega}_i &= \arg\max_{\boldsymbol{\mu}_i} log(\prod_{j=1}^{n} p(A_{ij}) p(\boldsymbol{\mu}_i)) \\ &= \arg\max_{\boldsymbol{\mu}_i} (\sum_{j=1}^{n} log(p(A_{ij})) + b_i(\boldsymbol{\mu}_i))\end{aligned}$$

Given no knowledge on the bias before hand, here we impose a uniform prior on the Bernoulli variable, such that $b_i(\boldsymbol{\mu}_i) = 0$. In addition, the log posterior is related to 4 situations,

$$\begin{aligned}\Omega_i = &\sum_{j=1, A_{ij}=1, Z_{ij}=0}^{n} log(1 - (1 - p_f)(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)) \\ &+ \sum_{j=1, A_{ij}=1, Z_{ij}=1}^{n} log(1 - p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)) \\ &+ \sum_{j=1, A_{ij}=0, Z_{ij}=0}^{n} log((1 - p_f)(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)) \\ &+ \sum_{j=1, A_{ij}=0, Z_{ij}=1}^{n} log(p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)\end{aligned}$$

Here we assume $P_f \rightarrow 0$, such that $p_f(1 - \boldsymbol{\mu}_i \boldsymbol{v}_i) \rightarrow 0$, and both $\sum_{j=1, A_{ij}=1, Z_{ij}=1}^{n} log(1 - p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j))$ and $\sum_{j=1, A_{ij}=0, Z_{ij}=1}^{n} log(p_f(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)$ can be approximate by a constant that does not contribute to the inference of $\hat{\Omega}_i$. Also $(1 - p_f)(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)$ can be approximated by $(1 - \boldsymbol{\mu}_i \boldsymbol{\nu}_j)$. It also has practical meanings, that for the inference of background bias, we only consider the values that are not covered by the latent pattern X, Y. While our objective is to infer $\boldsymbol{\mu}_i$ that better reflect the background information of $A_{i:}$. However, it is still non-trivial to derive $\hat{\Omega}_i$ as every observation is related to a different $\boldsymbol{v}_j$. Instead of deriving exact MAP of

likelihood, we treat this as an optimization problem, where we could utilize conventional loss function to achieve the same objective that optimize the difference between $\boldsymbol{\mu}_i$ with $A_{i:}$. Here, we apply a modified mean square loss, i.e.,

$$\Omega = \sum_{j=1,Z_{ij}=0}^{n} \boldsymbol{v}_j (A_{ij} - \boldsymbol{\mu}_i)^2$$

The most important benefit of this modified loss is that it ensures the probability of each $\boldsymbol{\mu}_i$ would be from $[0, 1]$ and still consider the impact from $\boldsymbol{v}_j$ for each observations. Conveniently, $\hat{\Omega}_i$ is inferred from the derivative of Ω, i.e.,

$$\hat{\Omega} = \arg\max_{\boldsymbol{\mu}_i} \Omega = \frac{\sum_{j=1,Z_{ij}=0}^{n} A_{ij} v_j}{\sum_{j=1,Z_{ij}=0}^{n} v_j}$$

Similarly, we have

$$\Theta_j = \sum_{i=1,Z_{ij}=0}^{m} \boldsymbol{\mu}_i (A_{ij} - \boldsymbol{v}_j)^2$$

$$\hat{\Theta}_j = \frac{\sum_{i=1,Z_{ij}=0}^{m} A_{ij} \boldsymbol{\mu}_i}{\sum_{i=1,Z_{ij}=0}^{n} \boldsymbol{\mu}_i}$$

Now we have derived all messages in the likelihood despite $\Gamma_{ijl} : \mathbf{m}_{g_{ij} \to W_{ijl}}$ that passed the information from the likelihood factor to each of auxiliary variable W_{ijl}. Overall, the message take the form of

$$\begin{aligned}\mathbf{m}_{g_{ij} \to W_{ijl}}(W_{ijl})^{t+1} = &\max_{W_{ijl'}, l' \neq l} (g_{ij}(Z_{ij}, \boldsymbol{\mu}_i, \boldsymbol{v}_j) \\ &+ \sum_{l' \neq l} \mathbf{m}_{W_{ijl'} \to g_{ij}} (W_{ijl'})^t)\end{aligned}$$

When updating W_{ijl}, we consider two scenarios: 1. $Z_{ij} = \vee_{l=1}^{k} W_{ijl} = 1$ with likelihood factor $p(A_{ij}|Z_{ij} = 1)$ and 2. $\vee_{l=1}^{k} W_{ijl} = 0$, $p(A_{ij}|Z_{ij} = 0)$.

$W_{ijl} = 1$ falls into the situation of scenarios 1, that no matter the value of $W_{ijl'}$, $Z_{ij} = \vee W_{ijl} = 1$. The message for $W_{ijl} = 1$ can be derived as

$$\begin{aligned}\mathbf{m}_{g_{ij} \to W_{ijl}}(1) = &\max_{W_{ijl'}, l' \neq l} (g_{ij}(Z_{ij}, \boldsymbol{\mu}_i, \boldsymbol{v}_j) \\ &+ \sum_{l' \neq l} \mathbf{m}_{W_{ijl'} \to g_{ij}} (W_{ijl'})^t) \\ &= log(p_(A_{ij}|1)) \\ &+ \sum_{l' \neq l} max(\mathbf{m}_{W_{ijl'} \to g_{ij}}(1), \mathbf{m}_{W_{ijl'} \to g_{ij}}(0))\end{aligned}$$

$W_{ijl} = 0$ could involve both cases. If $Z_{ij} = 0$, all $W_{ijl'} = 0$, i.e.,

$$\mathbf{m}_{g_{ij} \to W_{ijl}}(0) = log(p_(A_{ij}|0)) + \sum_{l' \neq l} \mathbf{m}_{W_{ijl'} \to g_{ij}}(0)$$

If $Z_{ij} = 1$, at least one of $W_{ijl'}$ equal to zero. To achieve the maximum likelihood, the $W_{ijl'}$ with the maximum likelihood difference on 0 or 1 should be set as 1, we denote it as W_{ijl^*}, where $l^* = \arg\max_{l' \neq l} (\mathbf{m}_{W_{ijl'} \to g_{ij}}(1) - \mathbf{m}_{W_{ijl'} \to g_{ij}}(0))$, such that we have

$$\begin{aligned}&\mathbf{m}_{g_{ij} \to W_{ijl}}(0) = log(p_(A_{ij}|1)) \\ &+ \sum_{l' \neq l} max(\mathbf{m}_{W_{ijl'} \to g_{ij}}(1), \mathbf{m}_{W_{ijl'} \to g_{ij}}(0)) \\ &- \mathbf{m}_{W_{ijl^*} \to g_{ij}}(0)\end{aligned}$$

Taken together,

$$\begin{aligned}&\mathbf{m}_{g_{ij} \to W_{ijl}}(0) = max(log(p_(A_{ij}|0)) \\ &+ \sum_{l' \neq l} \mathbf{m}_{W_{ijl'} \to g_{ij}}(0), log(p_(A_{ij}|1)) \\ &+ \sum_{l' \neq l} max(\mathbf{m}_{W_{ijl'} \to g_{ij}}(1), \mathbf{m}_{W_{ijl'} \to g_{ij}}(0)) \\ &- \mathbf{m}_{W_{ijl^*} \to g_{ij}}(0))\end{aligned}$$

Therefore

$$\begin{aligned}&\Gamma_{ijl} = \mathbf{m}_{g_{ij} \to W_{ijl}}(1) - \mathbf{m}_{g_{ij} \to W_{ijl}}(0) \\ &= log(p_(A_{ij}|1)) \\ &+ \sum_{l' \neq l} max(\mathbf{m}_{W_{ijl'} \to g_{ij}}(1), \mathbf{m}_{W_{ijl'} \to g_{ij}}(0)) \\ &- max(log(p_(A_{ij}|0)) \\ &+ \sum_{l' \neq l} \mathbf{m}_{W_{ijl'} \to g_{ij}}(0), log(p_(A_{ij}|1)) \\ &+ \sum_{l' \neq l} max(\mathbf{m}_{W_{ijl'} \to g_{ij}}(1), \mathbf{m}_{W_{ijl'} \to g_{ij}}(0)) \\ &- \mathbf{m}_{W_{ijl^*} \to g_{ij}}(0)) \\ &= min(log(\frac{p(A_{ij}|1)}{p(A_{ij}|0)}) + \sum_{l' \neq l} max(0, \hat{\Gamma}^t_{ijl'}), \\ &max(0, -max_{l' \neq l} \hat{\Gamma}^t_{ijl'}))\end{aligned}$$

References

Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *International Conference on Machine Learning*, pages 945–954. PMLR, 2016.