Supplementary Material to Probabilistic Matrix Factorization with Non-random Missing Data

José Miguel Hernández-Lobato, Neil Houlsby and Zoubin Ghahramani

Contents

1 Hyper-parameter values in the complete data model

Recall that the priors for i) the base boundary variables $\mathbf{b}_0 = (b_{0,1}, \ldots, b_{0,L-1})$ and ii) the factors γ_i^{row} and γ_j^{col} for the noise variance are

$$p(\mathbf{b}_0) = \prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|m_k^{\mathbf{b}_0}, v_0)\,, \qquad p(\gamma_i^{\text{row}}) = \mathcal{IG}(\gamma_i^{\text{row}}|a_0, b_0)\,, \qquad p(\gamma_j^{\text{col}}) = \mathcal{IG}(\gamma_j^{\text{col}}|a_0, b_0)\,, \tag{1}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, d$, $\mathcal{N}(x|m, v)$ denotes a Gaussian density with mean m and variance v and

$$\mathcal{IG}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left\{-\frac{b}{x}\right\} \tag{2}$$

denotes an inverse-gamma density with parameters a and b. We initialize the prior means $m_1^{\mathbf{b}_0}, \ldots, m_{L-1}^{\mathbf{b}_0}$ to form an evenly spaced grid in the interval $[-6, 6]$ as suggested in Paquet et al. (2012). For example, when $L = 5$, we have that $m_1^{\mathbf{b}_0} = -6$, $m_2^{\mathbf{b}_0} = -2$, $m_3^{\mathbf{b}_0} = -2$ and $m_4^{\mathbf{b}_0} = -6$. The prior variance v_0 for each component of $\mathbf{b}_0$ is initialized to $v_0 = 0.1$. The hyper-parameters a_0 and b_0 for the priors on γ_i^{row} and γ_j^{col} are initialized to $a_0 = 10/2$ and $b_0 = 10\sqrt{10}/2$. The strength of the resulting priors is then equivalent to having seen for each of these random variables a random sample of size 10 with empirical variance $\sqrt{10}$. The prior expectations for γ_i^{row} and γ_j^{col} are close to $\sqrt{10}$. This means that the product of γ_i^{row} and γ_j^{col} is close on average to 10, which is the recommended noise level in the ordinal matrix factorization model described in Paquet et al. (2012).

We use factorized standard Gaussian hyper-priors for the prior means $\mathbf{m}^{\mathbf{U}} = (m_1^{\mathbf{U}}, \ldots, m_h^{\mathbf{U}})$ and $\mathbf{m}^{\mathbf{V}} = (m_1^{\mathbf{V}}, \ldots, m_h^{\mathbf{V}})$, that is,

$$p(\mathbf{m}^{\mathbf{U}}) = \prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{U}}|0, 1)\,, \qquad p(\mathbf{m}^{\mathbf{V}}) = \prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{V}}|0, 1)\,. \tag{3}$$

Similarly, we use factorized inverse-gamma hyper-priors for the prior variances $\mathbf{v}^{\mathbf{U}} = (v_1^{\mathbf{U}}, \ldots, v_h^{\mathbf{U}})$ and $\mathbf{v}^{\mathbf{V}} = (v_1^{\mathbf{V}}, \ldots, v_h^{\mathbf{V}})$, that is,

$$p(\mathbf{v}^{\mathbf{U}}) = \prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{U}}|a_0', b_0')\,, \qquad p(\mathbf{v}^{\mathbf{V}}) = \prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{V}}|a_0', b_0')\,. \tag{4}$$

The hyper-parameters a_0' and b_0' are initialized to $a_0' = 10/2$ and $b_0' = 10/2$. The strength of the resulting priors is then equivalent to having seen for each of these random variables a random sample of size 10 with unit empirical variance.

2 Priors in the missing data model

We choose to use fully factorized Gaussian priors for all the parameters in the missing data model, that is,

$$p(\mathbf{E}) = \prod_{i=1}^{n} \prod_{k=1}^{h} \mathcal{N}(e_{i,k}|\bar{e}_{i,k}^0, \tilde{e}_{i,k}^0)\,, \tag{5}$$

$$p(\mathbf{F}) = \prod_{j=1}^{d} \prod_{k=1}^{h} \mathcal{N}(f_{j,k}|\bar{f}_{j,k}^0, \tilde{f}_{j,k}^0)\,, \tag{6}$$

$$p(\mathbf{\Lambda}^{\text{row}}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \mathcal{N}(\lambda_{i,l}^{\text{row}}|\bar{\lambda}_{i,l}^{\text{row}\,0}, \tilde{\lambda}_{i,l}^{\text{row}\,0})\,, \tag{7}$$

$$p(\mathbf{\Psi}^{\text{col}}) = \prod_{j=1}^{d}\prod_{l=1}^{L}\mathcal{N}(\psi_{j,l}^{\text{col}}|\bar{\psi}_{j,l}^{\text{col}\,0}, \tilde{\psi}_{j,l}^{\text{col}\,0}) \tag{8}$$

and $p(z) = \mathcal{N}(z|\bar{z}^0, \tilde{z}^0)$. We fix these priors to have zero-mean and unit variance. We also incorporate a local bias to each row and column. For example, the column h in $\mathbf{F}$ contains the biases for the columns of $\mathbf{X}$. In this case, $\bar{e}_{i,h}^0 = 1$ and $\tilde{e}_{i,h}^0 = \varepsilon$, for $i = 1, \ldots, n$, where ε is a small positive constant. Similarly, the column $h-1$ in $\mathbf{E}$ contains the biases for the rows and $\bar{f}_{j,h-1}^0 = 1$ and $\tilde{f}_{j,h-1}^0 = \varepsilon$, for $j = 1, \ldots, d$.

3 Description of each factor in the factor graph

We describe the form of each factor in the factor graph shown in Figure 1. This factor graph includes a total of 19 different factors. The first 13 factors belong to the complete data model. The last 6 factors belong to the missing data model. We first describe the factors for the complete data model. After that, we describe the factors for the missing data model.

3.1 Factors for the complete data model

The complete data model is formed by factors 1 to 13. The first four factors are given by the hyper-priors for the mean and variances of the Gaussian priors on the entries in the rows of $\mathbf{U}$ and $\mathbf{V}$, that is,

$$f_{1,k}(v_k^{\mathbf{V}}) = \mathcal{IG}(v_k^{\mathbf{V}}|a_0', b_0')\,, \qquad f_{2,k}(v_k^{\mathbf{U}}) = \mathcal{IG}(v_k^{\mathbf{U}}|a_0', b_0')\,, \tag{9}$$

$$f_{3,k}(m_k^{\mathbf{V}}) = \mathcal{N}(m_k^{\mathbf{V}}|0, 1)\,, \qquad f_{4,k}(m_k^{\mathbf{U}}) = \mathcal{N}(m_k^{\mathbf{U}}|0, 1)\,, \tag{10}$$

for $k = 1, \ldots, h$. Factors 5 and 6 are the priors for the factors γ_i^{row} and γ_j^{col} that form the variance of the additive noise on $c_{i,j}$, that is,

$$f_{5,j}(\gamma_j^{\text{col}}) = \mathcal{IG}(\gamma_i^{\text{row}}|a_0, b_0)\,, \qquad f_{6,i}(\gamma_i^{\text{row}}) = \mathcal{IG}(\gamma_j^{\text{col}}|a_0, b_0)\,, \tag{11}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Factor 7 is formed by the Gaussian prior for the base boundary variables $\mathbf{b}_0 = (b_{0,1}, \ldots, b_{0,L-1})$, that is,

$$f_{7,k}(b_{0,k}) = \mathcal{N}(b_{0,k}|m_k^{\mathbf{b}_0}, v_0)\,, \tag{12}$$

for $k = 1, \ldots, L-1$. Factor 8 is given by the conditional Gaussian prior for the vector of boundary variables $\mathbf{b}_j = (b_{j,1}, \ldots b_{j,L-1})$, that is,

$$f_{8,j,k}(b_{j,k}, b_{0,k}) = \mathcal{N}(b_{j,k}|b_{0,k}, v_0)\,, \tag{13}$$

for $j = 1, \ldots, d$, $k = 1, \ldots, L-1$. Factors 9 and 10 are the Gaussian priors for the entries of the low-rank latent matrices $\mathbf{U}$ and $\mathbf{V}$, namely

$$f_{9,j,k}(v_{j,k}, m_k^{\mathbf{V}}, v_k^{\mathbf{V}}) = \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, v_k^{\mathbf{V}})\,, \qquad f_{10,i,k}(u_{i,k}, m_k^{\mathbf{U}}, v_k^{\mathbf{U}}) = \mathcal{N}(u_{i,k}|m_k^{\mathbf{U}}, v_k^{\mathbf{U}})\,, \tag{14}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d$ and $k = 1, \ldots, h$. Factor 11 is formed by the delta functions that constrain each $c_{i,j}$ to be equal to $\mathbf{u}_i\mathbf{v}_j^{\text{T}}$, where $\mathbf{u}_i$ is the i-th row of $\mathbf{U}$ and $\mathbf{v}_j$ is the j-th row of $\mathbf{V}$, that is,

$$f_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j) = \delta(c_{i,j} - \mathbf{u}_i\mathbf{v}_j^{\text{T}})\,, \tag{15}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Factor 12 is the conditional prior for the variables $a_{i,j}$. These variables are obtained after adding Gaussian noise to $c_{i,j}$ with variance $\gamma_i^{\text{row}}\gamma_j^{\text{col}}$, that is,

$$f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})\,, \tag{16}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d$. Finally, factor 13 is given by

$$f_{13,i,j,k}(r_{i,j}, a_{i,j}, b_{j,k}) = \Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]\,, \tag{17}$$

where Θ is the Heaviside step function and sign$[x]$ is the sign function, which returns -1 if $x < 0$ and 1 otherwise.

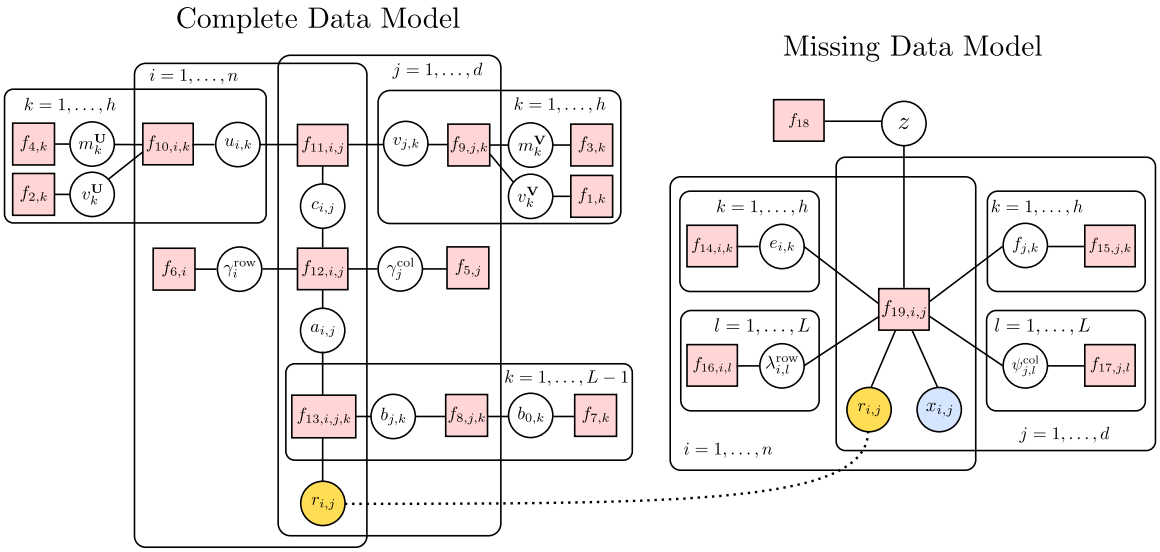

Figure 1: Factor graph for our Matrix Factorization model with data Missing Not At Random (MF-MNAR).

3.2 Factors for the missing data model

The missing data model is formed by factors 14 to 19. The first five factors in this model are formed by the Gaussian priors on the model parameters $\mathbf{E}$, $\mathbf{F}$, $\mathbf{\Lambda}^{\mathrm{row}}$ $\mathbf{\Psi}^{\mathrm{col}}$ and z, that is,

$$f_{14,i,k}(e_{i,k}) = \mathcal{N}(e_{i,k}|\bar{e}_{i,k}^0, \tilde{u}_{i,k}^0)\,, \qquad f_{15,j,k}(f_{j,k}) = \mathcal{N}(f_{j,k}|\bar{f}_{j,k}^0, \tilde{v}_{j,k}^0)\,, \tag{18}$$

$$f_{16,i,l}(\lambda_{i,l}^{\mathrm{row}}) = \mathcal{N}(\lambda_{i,l}^{\mathrm{row}}|\bar{\lambda}_{i,l}^{\mathrm{row}\,0}, \tilde{\lambda}_{i,l}^{\mathrm{row}\,0})\,, \qquad f_{17,j,l}(\psi_{j,l}^{\mathrm{col}}) = \mathcal{N}(\psi_{j,l}^{\mathrm{col}}|\bar{\psi}_{j,l}^{\mathrm{col}\,0}, \tilde{\psi}_{j,l}^{\mathrm{col}\,0})\,, \tag{19}$$

$$f_{18}(z) = \mathcal{N}(z|\bar{z}^0, \tilde{z}^0)\,, \tag{20}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d$, $k = 1, \ldots, h$, $l = 1, \ldots, L$. Finally, factor 19 is given by the likelihood of the missing data model, that is,

$$f_{19,i,j}(r_{i,j}, x_{i,j}, \mathbf{e}_i, \mathbf{f}_j, z, \boldsymbol{\lambda}_i^{\mathrm{row}}, \boldsymbol{\psi}_j^{\mathrm{col}}) = x_{i,j}\sigma\{\mathbf{e}_i\mathbf{f}_j^{\mathrm{T}} + z + \sum_{l=1}^{L}(\lambda_{i,l}^{\mathrm{row}} + \psi_{j,l}^{\mathrm{col}})\mathbf{I}[r_{i,j} = l]\}+$$

$$(1 - x_{i,j})\sigma\{-\mathbf{e}_i\mathbf{f}_j^{\mathrm{T}} - z - \sum_{l=1}^{L}(\lambda_{i,l}^{\mathrm{row}} + \psi_{j,l}^{\mathrm{col}})\mathbf{I}[r_{i,j} = l]\}\,, \tag{21}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, d$, $\mathbf{e}_i$ and $\mathbf{f}_j$ are the i-th and j-th rows of the latent low-rank matrices $\mathbf{E}$ and $\mathbf{F}$, respectively, $\boldsymbol{\lambda}_i^{\mathrm{row}}$ and $\boldsymbol{\psi}_j^{\mathrm{col}}$ are the are the i-th and j-th rows of matrices $\mathbf{\Lambda}^{\mathrm{row}}$ and $\mathbf{\Psi}^{\mathrm{col}}$, respectively, $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function and $\mathbf{I}[\cdot]$ is the indicator function that takes value 1 when its argument is true and 0 otherwise.

4 Approximate Inference

We now describe how to perform approximate Bayesian inference in the proposed Matrix Factorization model with data Missing Not At Random (MF-MNAR). Our approach is based on a combination of the methods expectation propagation (EP) Minka (2001) and variational Bayes (VB) Ghahramani and Beal (2001).

We approximate the exact posterior in MF-MNAR, that is, $p(\mathbf{\Omega}, \mathbf{\Theta}, \mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}}, \mathbf{X})$ given by Equation (8) in the main document, with the parametric distribution $\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}}) = \mathcal{Q}_1(\mathbf{\Theta})\mathcal{Q}_2(\mathbf{\Omega})\mathcal{Q}_3(\mathbf{R}^{\neg\mathcal{O}})$, where we

assume that $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$ are fully factorized distributions with inverse-gamma, Gaussian and multinomial factors, namely

$$\begin{aligned}
\mathcal{Q}_1(\boldsymbol{\Theta}) = &\left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}}}, a_k^{v^{\mathbf{U}}})\right]\left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}}}, b_k^{v^{\mathbf{V}}})\right]\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1} \mathcal{N}(b_{i,k}|m_{i,k}^{b}, v_{i,k}^{b})\right] \\
&\left[\prod_{i=1}^{n}\prod_{j=1}^{d} \mathcal{N}(a_{i,j}|m_{i,j}^{a}, v_{i,j}^{a})\right]\left[\prod_{i=1}^{n}\prod_{j=1}^{d} \mathcal{N}(c_{i,j}|m_{i,j}^{c}, v_{i,j}^{c})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u}, v_{i,k}^{u})\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_{j,k}^{v}, v_{j,k}^{v})\right]\left[\prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|m_k^{b_0}, v_k^{b_0})\right]\left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}}}, v_k^{m^{\mathbf{U}}})\right] \\
&\left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}}}, v_k^{m^{\mathbf{V}}})\right]\left[\prod_{i=1}^{n} \mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})\right]\left[\prod_{j=1}^{d} \mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})\right], \qquad (22)
\end{aligned}$$

$$\begin{aligned}
\mathcal{Q}_2(\boldsymbol{\Omega}) = &\mathcal{N}(z|m^z, v^z)\left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(e_{i,k}|m_{i,k}^{e}, v_{i,k}^{e})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(f_{j,k}|m_{j,k}^{f}, v_{j,k}^{f})\right] \\
&\left[\prod_{i=1}^{n}\prod_{l=1}^{L} \mathcal{N}(\lambda_{i,l}^{\text{row}}|m_{i,l}^{\lambda^{\text{row}}}, v_{i,l}^{\lambda^{\text{row}}})\right]\left[\prod_{j=1}^{d}\prod_{l=1}^{L} \mathcal{N}(\psi_{j,l}^{\text{col}}|m_{j,l}^{\psi^{\text{col}}}, v_{j,l}^{\psi^{\text{col}}})\right], \qquad (23)
\end{aligned}$$

$$\mathcal{Q}_3(\mathbf{R}^{\neg\mathcal{O}}) = \prod_{(i,j)\notin\mathcal{O}}\prod_{l=1}^{L} p_{i,j,l}^{\mathbf{I}[r_{i,j}=l]}, \qquad (24)$$

where $\sum_{l=1}^{L} p_{i,j,l} = 1$. Ideally, we want to adjust the parameters of $\mathcal{Q}$ so that this approximation is as close as possible to the exact posterior $p(\boldsymbol{\Omega}, \boldsymbol{\Theta}, \mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}}, \mathbf{X})$. Recall that this posterior can be written, up to its normalization constant, as the product of all the factors shown in the factor graph from Figure 1. Section 3 contains a description of each of these factors. We will approximate each exact factor in the factor graph with an approximate factor that has the same functional form as $\mathcal{Q}$. For example, the exact factor $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})$ will be approximated with the approximate factor $\tilde{f}_{12,i,j}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$ given by

$$\begin{aligned}
\tilde{f}_{12,i,j}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}}) = &\left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{U}}|\tilde{a}_k^{v^{\mathbf{U}},12,i,j}, \tilde{a}_k^{v^{\mathbf{U}},12,i,j})\right]\left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{V}}|\tilde{a}_k^{v^{\mathbf{V}},12,i,j}, \tilde{b}_k^{v^{\mathbf{V}},12,i,j})\right] \\
&\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1} \mathcal{N}(b_{i,k}|\tilde{m}_{i,k}^{b,12,i,j}, \tilde{v}_{i,k}^{b,12,i,j})\right]\left[\prod_{i=1}^{n}\prod_{j=1}^{d} \mathcal{N}(a_{i,j}|\tilde{m}_{i,j}^{a,12,i,j}, \tilde{v}_{i,j}^{a,12,i,j})\right] \\
&\left[\prod_{i=1}^{n}\prod_{j=1}^{d} \mathcal{N}(c_{i,j}|\tilde{m}_{i,j}^{c,12,i,j}, \tilde{v}_{i,j}^{c,12,i,j})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|\tilde{m}_{i,k}^{u,12,i,j}, \tilde{v}_{i,k}^{u,12,i,j})\right] \\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|\tilde{m}_{j,k}^{v,12,i,j}, \tilde{v}_{j,k}^{v,12,i,j})\right]\left[\prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|\tilde{m}_k^{b_0,12,i,j}, \tilde{v}_k^{b_0,12,i,j})\right] \\
&\left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{U}}|\tilde{m}_k^{m^{\mathbf{U}},12,i,j}, \tilde{v}_k^{m^{\mathbf{U}},12,i,j})\right]\left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{V}}|\tilde{m}_k^{m^{\mathbf{V}},12,i,j}, \tilde{v}_k^{m^{\mathbf{V}},12,i,j})\right] \\
&\left[\prod_{i=1}^{n} \mathcal{IG}(\gamma_i^{\text{row}}|\tilde{a}_i^{\gamma^{\text{row}},12,i,j}, \tilde{b}_i^{\gamma^{\text{row}},12,i,j})\right]\left[\prod_{j=1}^{d} \mathcal{IG}(\gamma_j^{\text{row}}|\tilde{a}_j^{\gamma^{\text{col}},12,i,j}, \tilde{b}_j^{\gamma^{\text{col}},12,i,j})\right]
\end{aligned}$$

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(e_{i,k}|\tilde{m}_{i,k}^{e,12,i,j},\tilde{v}_{i,k}^{e,12,i,j})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(f_{j,k}|\tilde{m}_{j,k}^{f},\tilde{v}_{j,k}^{f,12,i,j})\right]$$
$$\left[\prod_{i=1}^{n}\prod_{l=1}^{L}\mathcal{N}(\lambda_{i,l}^{\text{row}}|\tilde{m}_{i,l}^{\lambda^{\text{row}},12,i,j},\tilde{v}_{i,l}^{\lambda^{\text{row}},12,i,j})\right]\left[\prod_{j=1}^{d}\prod_{l=1}^{L}\mathcal{N}(\psi_{j,l}^{\text{col}}|\tilde{m}_{j,l}^{\psi^{\text{col}},12,i,j},\tilde{v}_{j,l}^{\psi^{\text{col}},12,i,j})\right]$$
$$\left[\prod_{(i,j)\notin\mathcal{O}}\prod_{l=1}^{L}[\tilde{p}_{i,j,l}^{12,i,j}]^{\mathbf{I}[r_{i,j}=l]}\right]\mathcal{N}(z|\tilde{m}^{z,12,i,j},\tilde{v}^{z,12,i,j})\tilde{s}_{12,i,j}\,, \tag{25}$$

where $\sum_{l=1}^{L}\tilde{p}_{i,j,l}^{12,i,j}$ and we have introduced the multiplicative constant $\tilde{s}_{12,i,j}$ because the approximate factors may not be normalized. The notation that we have used for the parameters of $\tilde{f}_{12,i,j}$ is the same that we used for the parameters of $\mathcal{Q}$, but adding to each parameter the tilde symbol $\tilde{}$ and the superscript $12,i,j$ with the indexes of the approximate factor. Note hat $\mathcal{Q}$ and all the approximate factors belong to the family of exponential distributions. This family is closed under the product operation. Therefore, the product of all the approximate factors still has the same functional form as $\mathcal{Q}$ and can be readily normalized. The exact posterior $p(\mathbf{\Omega},\mathbf{\Theta},\mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}},\mathbf{X})$ is the normalized product of all the exact factors in the factor graph from Figure 1. Similarly, we define the posterior approximation $\mathcal{Q}$ to be the normalized product of all the approximate factors. This means that, we can make $\mathcal{Q}$ be close to the exact posterior $p(\mathbf{\Omega},\mathbf{\Theta},\mathbf{R}^{\neg\mathcal{O}}|\mathbf{R}^{\mathcal{O}},\mathbf{X})$ by adjusting each approximate factor so that it is as close as possible to its corresponding exact factor in the factor graph. This is the approach followed by the method expectation propagation (EP) Minka (2001) and it will be the basis of our algorithm for approximate inference in MF-MNAR.

EP works by first, initializing all the approximate factors and $\mathcal{Q}$ to be non-informative or flat. This is done by setting i) the mean and variance parameters of the Gaussians to be zero and infinite, respectively, ii) the a and b parameters of the inverse gammas to be one and zero, respectively, and iii) the parameters of the multinomials to be $1/L$. After that, EP iteratively refines the parameters of the different approximate factors. We now describe how EP performs each of these operations. For example, let us assume that EP will refine the parameters of the approximate factor $\tilde{f}_{12,i,j}$. For this, EP computes the ratio of $\mathcal{Q}$ and $\tilde{f}_{12,i,j}$ and then normalizes the resulting distribution, which we denote by $\mathcal{Q}^{\backslash 12,i,j}$. Therefore, $\mathcal{Q}^{\backslash 12,i,j}$ is equal to the normalized product of all the approximate factors except $\tilde{f}_{12,i,j}$. The functional form of $\mathcal{Q}^{\backslash 12,i,j}$ is again the same as $\mathcal{Q}$ and all the other approximate factors. In particular, we have that

$$\begin{aligned}\mathcal{Q}^{\backslash 12,i,j}(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}}) = &\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}},\backslash 12,i,j},a_k^{v^{\mathbf{U}},\backslash 12,i,j})\right]\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}},\backslash 12,i,j},b_k^{v^{\mathbf{V}},\backslash 12,i,j})\right]\\
&\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1}\mathcal{N}(b_{i,k}|m_{i,k}^{b,\backslash 12,i,j},v_{i,k}^{b,\backslash 12,i,j})\right]\left[\prod_{i=1}^{n}\prod_{j=1}^{d}\mathcal{N}(a_{i,j}|m_{i,j}^{a,\backslash 12,i,j},v_{i,j}^{a,\backslash 12,i,j})\right]\\
&\left[\prod_{i=1}^{n}\prod_{j=1}^{d}\mathcal{N}(c_{i,j}|m_{i,j}^{c,\backslash 12,i,j},v_{i,j}^{c,\backslash 12,i,j})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u,\backslash 12,i,j},v_{i,k}^{u,\backslash 12,i,j}\right]\\
&\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 12,i,j},v_{j,k}^{v,\backslash 12,i,j}\right]\left[\prod_{k=1}^{L-1}\mathcal{N}(b_{0,k}|m_k^{b_0,\backslash 12,i,j},v_k^{b_0,\backslash 12,i,j})\right]\\
&\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}},\backslash 12,i,j},v_k^{m^{\mathbf{U}},\backslash 12,i,j})\right]\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\backslash 12,i,j},v_k^{m^{\mathbf{V}},\backslash 12,i,j})\right]\\
&\left[\prod_{i=1}^{n}\mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\backslash 12,i,j},b_i^{\gamma^{\text{row}},\backslash 12,i,j})\right]\left[\prod_{j=1}^{d}\mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}},\backslash 12,i,j},b_j^{\gamma^{\text{col}},\backslash 12,i,j})\right]\end{aligned}$$

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(e_{i,k}|m_{i,k}^{e,\backslash 12,i,j},v_{i,k}^{e,\backslash 12,i,j})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(f_{j,k}|m_{j,k}^{f},v_{j,k}^{f,\backslash 12,i,j})\right]$$
$$\left[\prod_{i=1}^{n}\prod_{l=1}^{L}\mathcal{N}(\lambda_{i,l}^{\text{row}}|m_{i,l}^{\lambda^{\text{row}},\backslash 12,i,j},v_{i,l}^{\lambda^{\text{row}},\backslash 12,i,j})\right]\left[\prod_{j=1}^{d}\prod_{l=1}^{L}\mathcal{N}(\psi_{j,l}^{\text{col}}|m_{j,l}^{\psi^{\text{col}},\backslash 12,i,j},v_{j,l}^{\psi^{\text{col}},\backslash 12,i,j})\right]$$
$$\left[\prod_{(i,j)\notin\mathcal{O}}\prod_{l=1}^{L}[p_{i,j,l}^{\backslash 12,i,j}]^{\mathbf{I}[r_{i,j}=l]}\right]\mathcal{N}(z|m^{z,\backslash 12,i,j},v^{z,\backslash 12,i,j}), \tag{26}$$

where notation for the parameters of $\mathcal{Q}^{\backslash 12,i,j}$ is the same that we used for the parameters of $\mathcal{Q}$, but adding to each parameter the superscript $\backslash 12,i,j$ with the indexes of the approximate factor that is removed from $\mathcal{Q}$ to obtain $\mathcal{Q}^{\backslash 12,i,j}$. EP refines the parameters of $\tilde{f}_{12,i,j}$ by minimizing the Kullback-Leibler (KL) divergence between $\mathcal{Q}^{\backslash 12,i,j}(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{12,i,j}(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}})$ and $\mathcal{Q}^{\backslash 12,i,j}(\mathbf{\Theta},\mathbf{\Omega},\mathbf{R}^{\neg\mathcal{O}})f_{12,i,j}(a_{i,j},c_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}})$ where $f_{12,i,j}$ is the exact factor in the factor graph that is been approximated by $\tilde{f}_{12,i,j}$. In particular, EP refines the parameters of $\tilde{f}_{12,i,j}$ by minimizing

$$\mathrm{D}_{\mathrm{KL}}(Q^{\backslash 12,i,j}f_{12,i,j}\|Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}) =$$
$$\sum_{\mathbf{R}^{\neg\mathcal{O}}}\int\left[Q^{\backslash 12,i,j}f_{12,i,j}\log\frac{Q^{\backslash 12,i,j}f_l}{Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}}+Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}-Q^{\backslash 12,i,j}f_{12,i,j}\right]d\mathbf{\Phi}\,d\mathbf{\Omega}, \tag{27}$$

where the arguments to $Q^{\backslash 12,i,j}f_{12,i,j}$ and $Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}$ have been omitted in the right-hand side of this equation to improve readability. The divergence above is minimized when the expectation of the sufficient statistics of $Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}$ with respect to $Q^{\backslash 12,i,j}\tilde{f}_{12,i,j}$ is the same as the expectation of those sufficient statistics with respect to $Q^{\backslash 12,i,j}f_{12,i,j}$. Note also that, when we refine the approximate factor $\tilde{f}_{12,i,j}$, we will only be modifying the parameters of $\tilde{f}_{12,i,j}$ that have an effect on the variables connected to the corresponding exact factor $f_{12,i,j}$ in the factor graph from Figure 1, that is, $a_{i,j}$, $c_{i,j}$, γ_i^{row}, and γ_j^{col}. This means that most of the parameters of $\tilde{f}_{12,i,j}$ will never be modified by EP and can be ignored.

The main loop of EP iterates over all the approximate factors, refining one after the other by minimizing the corresponding KL divergence. To simplify the exposition, we describe first how EP approximates the factors of the complete data model (factors 1 to 13 in Figure 1) when the data is assumed to be Missing At Random (MAR).

4.1 Approximate Inference in the complete data model with MAR data

In this section we describe the operations performed by EP to refine the approximate factors for the complete data model, that is, the approximate factors approximating the exact factors 1 to 13 in Figure 1. We will assume here that the data is Missing At Random (MAR). In this case the dotted line connecting the complete data model and the missing data model in Figure 1 does not exist and we can ignore the contribution of the missing data model. Furthermore, we can also ignore all the exact factors $f_{11,i,j}$, $f_{12,i,j}$ and $f_{13,i,j,k}$ with $(i,j)\notin\mathcal{O}$ since they do not have any effect in the posterior distribution in the MAR setting.

As described in Section 4, EP works by iteratively minimizing the KL divergence (27) with respect to each approximate factor. In the following sections we show the form of the resulting EP update operations.

4.1.1 EP updates for $\tilde{f}_{1,k}$

Recall that $f_{1,k}(v_k^{\mathbf{V}})=\mathcal{IG}(v_k^{\mathbf{V}}|a_0',b_0')$, where $k=1,\ldots,h$. In this case, $f_{1,k}$ has the same functional form as the inverse-gamma factor that specifies the distribution of $v_k^{\mathbf{V}}$ in $\tilde{f}_{1,k}$. Therefore, the EP update for $\tilde{f}_{1,k}$ sets the parameters of that inverse-gamma factor to be the same as the parameters of the the inverse-gamma

distribution in $f_{1,k}$, namely

$$[\tilde{a}_k^{v^{\mathbf{V}},1,k}]^{\text{new}} = a'_0\,, \qquad [\tilde{b}_k^{v^{\mathbf{V}},1,k}]^{\text{new}} = b'_0\,, \tag{28}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{1,k}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{1,k}$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[a_k^{v^{\mathbf{V}}}]^{\text{new}} = a'_0\,, \qquad [b_k^{v^{\mathbf{V}}}]^{\text{new}} = b'_0\,. \tag{29}$$

4.1.2 EP updates for $\tilde{f}_{2,k}$

Recall that $f_{2,k}(v_k^{\mathbf{U}}) = \mathcal{IG}(v_k^{\mathbf{U}}|a'_0, b'_0)$, where $k = 1, \ldots, h$. The EP update operations for $\tilde{f}_{2,k}$ are in this case the same as for the approximate factor $\tilde{f}_{1,k}$, namely,

$$[\tilde{a}_k^{v^{\mathbf{U}},2,k}]^{\text{new}} = a'_0\,, \qquad [\tilde{b}_k^{v^{\mathbf{U}},2,k}]^{\text{new}} = b'_0\,, \tag{30}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{2,k}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{2,k}$, we update $\mathcal{Q}$ by setting

$$[a_k^{v^{\mathbf{U}}}]^{\text{new}} = a'_0\,, \qquad [b_k^{v^{\mathbf{U}}}]^{\text{new}} = b'_0\,. \tag{31}$$

4.1.3 EP updates for $\tilde{f}_{3,k}$

Recall that $f_{3,k}(m_k^{\mathbf{V}}) = \mathcal{N}(m_k^{\mathbf{V}}|0, 1)$, where $k = 1, \ldots, h$. In this case, $f_{3,k}$ has the same functional form as the Gaussian factor that specifies the distribution of $m_k^{\mathbf{V}}$ in $\tilde{f}_3$. Therefore, the EP update for $\tilde{f}_{3,k}$ sets the parameters of that Gaussian factor to be the same as the parameters of the the Gaussian in $f_{3,k}$, namely,

$$[\tilde{m}_k^{m^{\mathbf{V}},3,k}]^{\text{new}} = 0\,, \qquad [\tilde{v}_k^{m^{\mathbf{V}},3,k}]^{\text{new}} = 1\,, \tag{32}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{3,k}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{3,k}$, we update $\mathcal{Q}$ by setting

$$[m_k^{m^{\mathbf{V}}}]^{\text{new}} = 0\,, \qquad [v_k^{m^{\mathbf{V}}}]^{\text{new}} = 1\,. \tag{33}$$

4.1.4 EP updates for $\tilde{f}_{4,k}$

Recall that $f_{4,k}(m_k^{\mathbf{U}}) = \mathcal{N}(m_k^{\mathbf{U}}|0, 1)$, where $k = 1, \ldots, h$. The EP update operations for $\tilde{f}_{4,k}$ are in this case the same as for the approximate factor $\tilde{f}_{3,k}$, namely,

$$[\tilde{m}_k^{m^{\mathbf{U}},4,k}]^{\text{new}} = 0\,, \qquad [\tilde{v}_k^{m^{\mathbf{U}},4,k}]^{\text{new}} = 1\,, \tag{34}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{4,k}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{4,k}$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[m_k^{m^{\mathbf{U}}}]^{\text{new}} = 0\,, \qquad [v_k^{m^{\mathbf{U}}}]^{\text{new}} = 1\,. \tag{35}$$

4.1.5 EP updates for $\tilde{f}_{5,j}$

Recall that $f_{5,j}(\gamma_j^{\text{col}}) = \mathcal{IG}(\gamma_i^{\text{row}}|a_0, b_0)$, where $j = 1, \ldots, d$. In this case, $f_{5,j}$ has the same functional form as the inverse-gamma factor that specifies the distribution of γ_j^{col} in $\tilde{f}_{5,j}$. Therefore, the EP update for $\tilde{f}_{5,j}$ sets the parameters of that inverse-gamma factor to be the same as the parameters of the the inverse gamma in $p(\boldsymbol{\gamma}^{\text{col}})$, namely

$$[\tilde{a}_j^{\gamma^{\text{col}},5,j}]^{\text{new}} = a_0\,, \qquad [\tilde{b}_j^{\gamma^{\text{col}},5,j}]^{\text{new}} = b_0\,, \tag{36}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{5,j}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{5,j}$, we update $\mathcal{Q}$ by setting

$$[a_j^{\gamma^{\text{col}}}]^{\text{new}} = a_0\,, \qquad [b_j^{\gamma^{\text{col}}}]^{\text{new}} = b_0\,. \tag{37}$$

4.1.6 EP updates for $\tilde{f}_{6,i}$

Recall that $f_{6,i}(\gamma_i^{\text{row}}) = \mathcal{IG}(\gamma_j^{\text{col}}|a_0, b_0)$, where $i = 1, \ldots, n$. The EP update operation for $\tilde{f}_{6,i}$ are in this case the same as for the approximate factor $\tilde{f}_{5,j}$, namely,

$$[\tilde{a}_i^{\gamma^{\text{row}},6,i}]^{\text{new}} = a_0\,, \qquad [\tilde{b}_i^{\gamma^{\text{row}},6,i}]^{\text{new}} = b_0\,, \tag{38}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{6,i}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{6,i}$, we update $\mathcal{Q}$ by setting

$$[a_i^{\gamma^{\text{row}}}]^{\text{new}} = a_0\,, \qquad [b_i^{\gamma^{\text{row}}}]^{\text{new}} = b_0\,. \tag{39}$$

4.1.7 EP updates for $\tilde{f}_{7,k}$

Recall that $f_{7,k}(b_{0,k}) = \mathcal{N}(b_{0,k}|m_k^{\mathbf{b}_0}, v_0)$, where $k = 1, \ldots, L-1$. In this case, $f_{7,k}$ has the same functional form as the Gaussian factor that specifies the distribution of $b_{0,k}$ in $\tilde{f}_{7,k}$. Therefore, the EP update for $\tilde{f}_{7,k}$ sets the parameters of that Gaussian factor to be the same as the parameters of the Gaussian in $f_{7,k}$, namely

$$[\tilde{m}_k^{b_0,7,k}]^{\text{new}} = m_k^{\mathbf{b}_0}\,, \qquad [\tilde{v}_k^{b_0,7,k}]^{\text{new}} = v_0\,, \tag{40}$$

Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_{7,k}$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_{7,k}$, we update $\mathcal{Q}$ by setting

$$[m_k^{b_0}]^{\text{new}} = m_k^{\mathbf{b}_0}\,, \qquad [v_k^{b_0}]^{\text{new}} = v_0\,. \tag{41}$$

4.1.8 EP updates for $\tilde{f}_{8,j,k}$

Recall that $f_{8,j,k}(b_{j,k}, b_{0,k}) = \mathcal{N}(b_{j,k}|b_{0,k}, v_0)$, where $j = 1, \ldots, d$ and $k = 1, \ldots, L-1$. In this case, we firstly compute the parameters of $\mathcal{Q}^{\setminus 8,j,k}$, which is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{8,j,k}$. This leads to

$$v_k^{b_0,\setminus 8,j,k} = \left[[v_k^{b_0}]^{-1} - [\tilde{v}_k^{b_0,8,j,k}]^{-1}\right]^{-1}\,, \quad m_k^{b_0,\setminus 8,j,k} = v_k^{b_0,\setminus 8,j,k}\left[m_k^{b_0}[v_k^{b_0}]^{-1} - \tilde{m}_k^{b_0,8,j,k}[\tilde{v}_k^{b_0,8,j,k}]^{-1}\right]\,, \tag{42}$$

$$v_{j,k}^{b,\setminus 8,j,k} = \left[[v_{j,k}^{b}]^{-1} - [\tilde{v}_{j,k}^{b,8,j}]^{-1}\right]^{-1}\,, \quad m_{j,k}^{b,\setminus 8,j,k} = v_{j,k}^{b,\setminus 8,j,k}\left[m_{j,k}^{b}[v_{j,k}^{b}]^{-1} - \tilde{m}_{j,k}^{b,8,j,k}[\tilde{v}_{j,k}^{b,8,j,k}]^{-1}\right]\,, \tag{43}$$

After that, we refine $\tilde{f}_{8,j,k}$ by setting

$$[\tilde{m}_k^{b_0,8,j,k}]^{\text{new}} = m_{j,k}^{b,\setminus 8,j,k}\,, \qquad [\tilde{v}_k^{b_0,8,j,k}]^{\text{new}} = v_{j,k}^{b,\setminus 8,j,k} + v_0\,, \tag{44}$$

$$[\tilde{m}_k^{b,8,j,k}]^{\text{new}} = m_k^{b_0,\setminus 8,j,k}\,, \qquad [\tilde{v}_k^{b,8,j,k}]^{\text{new}} = v_k^{b_0,\setminus 8,j,k} + v_0\,, \tag{45}$$

These update equations guarantee that the normalized versions of $\mathcal{Q}^{\setminus 8,j,k}(\boldsymbol{\Theta},\boldsymbol{\Omega},\mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{8,j,k}(\boldsymbol{\Theta},\boldsymbol{\Omega},\mathbf{R}^{\neg\mathcal{O}})$ and $\mathcal{Q}^{\setminus 8,j,k}(\boldsymbol{\Theta},\boldsymbol{\Omega},\mathbf{R}^{\neg\mathcal{O}})\mathcal{N}(b_{j,k}|b_{0,k},v_0)$ have the same expectations of sufficient statistics. Finally, we recompute Q as the normalized product of the updated $\tilde{f}_{8,j,k}$ and $\mathcal{Q}^{\setminus 8,j,k}$, that is,

$$[v_k^{b_0}]^{\text{new}} = \left[[v_k^{b_0,\setminus 8,j,k}]^{-1} + [\tilde{v}_k^{b_0,8,j,k}]^{-1}\right]^{-1}, \tag{46}$$

$$[m_k^{b_0}]^{\text{new}} = [v_k^{b_0}]^{\text{new}}\left[m_k^{b_0,\setminus 8,j,k}[v_k^{b_0,\setminus 8,j,k}]^{-1} + \tilde{m}_k^{b_0,8,j,k}[\tilde{v}_k^{b_0,8,j,k}]^{-1}\right], \tag{47}$$

$$[v_{j,k}^{b}]^{\text{new}} = \left[[v_{j,k}^{b,\setminus 8,j,k}]^{-1} + [\tilde{v}_{j,k}^{b,8,j,k}]^{-1}\right]^{-1}, \tag{48}$$

$$[m_{j,k}^{b}]^{\text{new}} = [v_{j,k}^{b}]^{\text{new}}\left[m_{j,k}^{b,\setminus 8,j,k}[v_{j,k}^{b,\setminus 8,j,k}]^{-1} + \tilde{m}_{j,k}^{b,8,j}[\tilde{v}_{j,k}^{b,8,j,k}]^{-1}\right]. \tag{49}$$

4.1.9 EP updates for $\tilde{f}_{9,j,k}$

Recall that $f_{9,j,k}(v_{j,k},m_k^{\mathbf{V}},v_k^{\mathbf{V}}) = \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}},v_k^{\mathbf{V}})$, where $j=1,\ldots,d$ and $k=1,\ldots,h$. We firstly compute the parameters of $\mathcal{Q}^{\setminus 9,j,k}$ which is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{9,j,k}$. This leads to

$$[v_k^{m^{\mathbf{V}},\setminus 9,j,k}]^{\text{new}} = \left[[v_k^{m^{\mathbf{V}}}]^{-1} - [\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right]^{-1}, \tag{50}$$

$$[m_k^{m^{\mathbf{V}},\setminus 9,j,k}]^{\text{new}} = [v_k^{m^{\mathbf{V}},\setminus 9,j,k}]^{\text{new}}\left[m_k^{m^{\mathbf{V}}}[v_k^{m^{\mathbf{V}}}]^{-1} - \tilde{m}_k^{m^{\mathbf{V}},9,j,k}[\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right], \tag{51}$$

$$[v_{j,k}^{v,\setminus 9,j,k}]^{\text{new}} = \left[[v_{j,k}^{v}]^{-1} - [\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right]^{-1}, \tag{52}$$

$$[m_{j,k}^{v,\setminus 9,j,k}]^{\text{new}} = [v_{j,k}^{v,\setminus 9,j,k}]^{\text{new}}\left[m_{j,k}^{v}[v_{j,k}^{v}]^{-1} - \tilde{m}_{j,k}^{v,9,j,k}[\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right], \tag{53}$$

$$[a_k^{v^{\mathbf{V}},\setminus 9,j,k}]^{\text{new}} = a_k^{v^{\mathbf{V}}} - \tilde{a}_k^{v^{\mathbf{V}},9,j,k} + 1\,, \tag{54}$$

$$[b_k^{v^{\mathbf{V}},\setminus 9,j,k}]^{\text{new}} = b_k^{v^{\mathbf{V}}} - \tilde{b}_k^{v^{\mathbf{V}},9,j,k}\,. \tag{55}$$

After this, we refine the approximate factor $\tilde{f}_{9,j,k}$. For this, we have to find the expectation of sufficient statistics with respect to $h(\boldsymbol{\Theta},\boldsymbol{\Omega},\mathbf{R}^{\neg\mathcal{O}}) = \mathcal{Q}^{\setminus 9,j,k}(\boldsymbol{\Theta},\boldsymbol{\Omega},\mathbf{R}^{\neg\mathcal{O}})\mathcal{N}(v_{j,k}|m_k^{\mathbf{V}},v_k^{\mathbf{V}})$. After summing out $\mathbf{R}^{\neg\mathcal{O}}$ and integrating out $\boldsymbol{\Omega}$ and $\boldsymbol{\Theta}\setminus\{v_{j,k},m_k^{\mathbf{V}},v_k^{\mathbf{V}}\}$ in h, we obtain

$$\begin{aligned} h(v_{j,k},m_k^{\mathbf{V}},v_k^{\mathbf{V}}) = {} & \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}},v_k^{\mathbf{V}})\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\setminus 9,j,k},v_k^{m^{\mathbf{V}},\setminus 9,j,k}) \\ & \mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus 9,j,k},v_{j,k}^{v,\setminus 9,j,k})\text{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}},\setminus 9,j,k},b_k^{v^{\mathbf{V}},\setminus 9,j,k})\,. \end{aligned} \tag{56}$$

The normalization constant of $h(v_{j,k},m_k^{\mathbf{V}},v_k^{\mathbf{V}})$ is then

$$Z = \int h(v_{j,k},m_k^{\mathbf{V}},v_k^{\mathbf{V}})\,dv_{j,k}\,dm_k^{\mathbf{V}}\,dv_k^{\mathbf{V}} \tag{57}$$

$$\begin{aligned} = {} & \int \mathcal{T}(v_{j,k}|m_k^{\mathbf{V}},\frac{b_k^{v^{\mathbf{V}},\setminus 9,j,k}}{a_k^{v^{\mathbf{V}},\setminus 9,j,k}},2a_k^{v^{\mathbf{V}},\setminus 9,j,k}) \\ & \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\setminus 9,j,k},v_k^{m^{\mathbf{V}},\setminus 9,j,k})\mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus 9,j,k},v_{j,k}^{v,\setminus 9,j,k})\,dv_{j,k}\,dm_k^{\mathbf{V}} \end{aligned} \tag{58}$$

$$\begin{aligned} \approx {} & \int \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}},\frac{2b_k^{v^{\mathbf{V}},\setminus 9,j,k}}{2a_k^{v^{\mathbf{V}},\setminus 9,j,k}-2}) \\ & \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\setminus 9,j,k},v_k^{m^{\mathbf{V}},\setminus 9,j,k})\mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus 9,j,k},v_{j,k}^{v,\setminus 9,j,k})\,dv_{j,k}\,dm_k^{\mathbf{V}} \end{aligned} \tag{59}$$

$$\approx \mathcal{N}(m_k^{m^{\mathbf{V}},\setminus 9,j,k} | m_{j,k}^{v,\setminus 9,j,k}, v_{j,k}^{v,\setminus 9,j,k} + v_k^{m^{\mathbf{V}},\setminus 9,j,k} + \frac{2b_k^{v^{\mathbf{V}},\setminus 9,j,k}}{2a_k^{v^{\mathbf{V}},\setminus 9,j,k} - 2}) \,, \tag{60}$$

where

$$\mathcal{T}(x|\mu,\lambda,\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu\lambda}\Gamma(\nu/2)} \left[1 + \frac{(x-\mu)^2}{\lambda\nu}\right]^{-(\nu+1)/2} \tag{61}$$

denotes a Student's t distribution with mean μ, variance parameter λ and degrees of freedom ν and in equation (59) we have approximated a Student's t distribution with a Gaussian distribution that has the same mean and variance as the original Student's t distribution. The expectation of the sufficient statistics $v_{j,k}$, $[v_{j,k}]^2$, $m_k^{\mathbf{V}}$, $[m_k^{\mathbf{V}}]^2$, $v_k^{\mathbf{V}}$ and $[v_k^{\mathbf{V}}]^2$ with respect to $h(v_{j,k}, m_k^{\mathbf{V}}, v_k^{\mathbf{V}})$ can be approximated in a similar way as the previous normalization constant. We describe below how to do this. For the random variable $v_k^{\mathbf{V}}$, the KL-divergence is actually minimized by matching the first moment and the expectation of $\log v_k^{\mathbf{V}}$. However, matching the expectation of $\log v_k^{\mathbf{V}}$ would require computing the inverse of the Digamma function, which has no analytical solution. To avoid this, we match the first and second moments of $v_k^{\mathbf{V}}$ which is expected to produce reasonably good results.

We approximate the moments of $v_k^{\mathbf{V}}$ using the following property of inverse gammas, see (2). Let $H(a,b)$ be the normalization constant of $f(x)\mathcal{IG}(x|a,b)$ for a particular f, that is, $H(a,b) = \int f(x)\mathcal{IG}(x|a,b)\,dx$. Then we have that $\int x f(x)\mathcal{IG}(x|a,b)\,dx = H(a+1,b)a/b$ and $\int x^2\mathcal{IG}(x|a,b)\,dx = H(a+2,b)a(a+1)/b^2$. Thus, each moment can be easily approximated given a procedure to approximate the normalization constant $H(a,b)$. For this, we only have to replace $H(a+1,b)$ and $H(a+2,b)$ in the previous equations with their corresponding approximations. In a similar way, we can compute approximations for the moments of $v_{j,k}$ and $m_k^{\mathbf{V}}$. In particular, we use the following property of the Gaussian distribution. Let $H(m,v)$ be the normalization constant of $f(x)\mathcal{N}(x|m,v)$ for a particular function f, that is, $H(m,v) = \int f(x)\mathcal{N}(x|m,v)\,dx$. Then we have that $[\mathrm{H}(m,v)]^{-1}\int x f(x)\mathcal{N}(x|m,v)\,dx = m + v\frac{d\log H(m,v)}{dm}$ and $[H(m,v)]^{-1}\int x^2\mathcal{N}(x|m,v)\,dx - [[H(m,v)]^{-1}\int x\mathcal{N}(x|m,v)\,dx]^2 = v - v^2([\frac{d\log H(m,v)}{dm}]^2 - 2\frac{d\log H(m,v)}{dv})$.

The resulting updates for $\tilde{f}_{9,j,k}$ are

$$[\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{\text{new}} = 2b_k^{v^{\mathbf{V}},\setminus 9,j,k}/(2a_k^{v^{\mathbf{V}},\setminus 9,j,k} - 2) + v_{j,k}^{v,\setminus 9,j,k} \,, \tag{62}$$

$$[\tilde{m}_k^{m^{\mathbf{V}},9,j,k}]^{\text{new}} = m_{j,k}^{v,\setminus 9,j,k} \,, \tag{63}$$

$$[\tilde{v}_{j,k}^{v,9,j,k}]^{\text{new}} = 2b_k^{v^{\mathbf{V}},\setminus 9,j,k}/(2a_k^{v^{\mathbf{V}},\setminus 9,j,k} - 2) + v_k^{m^{\mathbf{V}},\setminus 9,j,k} \,, \tag{64}$$

$$[\tilde{m}_{j,k}^{v,9,j,k}]^{\text{new}} = m_k^{m^{\mathbf{V}},\setminus 9,j,k} \,, \tag{65}$$

$$[\tilde{a}_k^{v^{\mathbf{V}},9,j,k}]^{\text{new}} = a' - a_k^{v^{\mathbf{V}},\setminus 9,j,k} + 1 \,, \tag{66}$$

$$[\tilde{b}_k^{v^{\mathbf{V}},9,j,k}]^{\text{new}} = b' - b_k^{v^{\mathbf{V}},\setminus 9,j,k} \,, \tag{67}$$

and we define a' and b' as

$$a' = \frac{a_k^{v^{\mathbf{V}},\setminus 9,j,k} Z_1^2}{(a_k^{v^{\mathbf{V}},\setminus 9,j,k} + 1)ZZ_2 - a_k^{v^{\mathbf{V}},\setminus 9,j,k} Z_1^2} \,, \qquad b' = \frac{b_k^{v^{\mathbf{V}},\setminus 9,j,k} ZZ_1}{(a_k^{v^{\mathbf{V}},\setminus 9,j,k} + 1)ZZ_2 - a_k^{v^{\mathbf{V}},\setminus 9,j,k} Z_1^2} \,, \tag{68}$$

where Z_1 and Z_2 are obtained in the same way as Z, but increasing $a_k^{v^{\mathbf{V}},\setminus 9,j,k}$ in one and two units during the computations, respectively. Once we have updated $\tilde{f}_{9,j,k}$, we recompute $\mathcal{Q}$ using

$$[v_k^{m^{\mathbf{V}}}]^{\text{new}} = \left[[v_k^{m^{\mathbf{V}},\setminus 9,j,k}]^{-1} + [\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right]^{-1} \,, \tag{69}$$

$$[m_k^{m^{\mathbf{V}}}]^{\text{new}} = [v_k^{m^{\mathbf{V}}}]^{\text{new}} \left[m_k^{m^{\mathbf{V}},\setminus 9,j,k}[v_k^{m^{\mathbf{V}},\setminus 9,j,k}]^{-1} + \tilde{m}_k^{m^{\mathbf{V}},9,j,k}[\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right] \,, \tag{70}$$

$$[v_{j,k}^{v}]^{\text{new}} = \left[[v_{j,k}^{v,\backslash 9,j,k}]^{-1} + [\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right]^{-1}, \tag{71}$$

$$[m_{j,k}^{v}]^{\text{new}} = [v_{j,k}^{v}]^{\text{new}} \left[m_{j,k}^{v,\backslash 9,j,k}[v_{j,k}^{v,\backslash 9,j,k}]^{-1} + \tilde{m}_{j,k}^{v,9,j,k}[\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right], \tag{72}$$

$$[a_{k}^{v^{\mathbf{V}}}]^{\text{new}} = a_{k}^{v^{\mathbf{V}},\backslash 9,j,k} + \tilde{a}_{k}^{v^{\mathbf{V}},9,j,k} - 1, \tag{73}$$

$$[b_{k}^{v^{\mathbf{V}}}]^{\text{new}} = b_{k}^{v^{\mathbf{V}},\backslash 9,j,k} + \tilde{b}_{k}^{v^{\mathbf{V}},9,j,k}, \tag{74}$$

Finally, note that we only update $\tilde{f}_{9,j,k}$ when $b_{k}^{v^{\mathbf{V}},\backslash 9,j,k} > 0$, $2a_{k}^{v^{\mathbf{V}},\backslash 9,j,k} - 2 > 0$, $v_{k}^{\mathbf{V},\backslash 9,j,k} > 0$ and $v_{j,k}^{v,\backslash 9,j,k} > 0$.

4.1.10 EP updates for $\tilde{f}_{10,i,k}$

Recall that $f_{10,i,k}(u_{i,k}, m_{k}^{\mathbf{U}}, v_{k}^{\mathbf{U}}) = \mathcal{N}(u_{i,k}|m_{k}^{\mathbf{U}}, v_{k}^{\mathbf{U}})$, where $i = 1, \ldots, n$, and $k = 1, \ldots, h$. The EP update equations for each $\tilde{f}_{10,i,k}$ are similar to those for each $\tilde{f}_{9,j,k}$ and therefore we do not include them here.

4.1.11 EP updates for $\tilde{f}_{11,i,j}$

Recall that $f_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j) = \delta(c_{i,j} - \mathbf{u}_i \mathbf{v}_j^{\mathrm{T}})$. Since we are assuming MAR data, we only have to consider the factors $f_{11,i,j}$ corresponding to those entries of the rating matrix $\mathbf{R}$ that are actually observed, that is, the factors $f_{11,i,j}$ such that $(i,j) \in \mathcal{O}$. We approximate all these exact factors in a single step. For this, we work with the extended exact factor $f_{11}(\mathbf{C}^{\mathcal{O}}, \mathbf{U}, \mathbf{V}) = \prod_{(i,j)\in\mathcal{O}} \delta(c_{i,j} - \mathbf{u}_i \mathbf{v}_j^{\mathrm{T}})$, where $\mathbf{C}^{\mathcal{O}}$ is the set of variables $c_{i,j}$ such that $(i,j) \in \mathcal{O}$. We approximate this exact factor with an approximate factor $\tilde{f}_{11}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$ that has the same functional form as the posterior approximation $\mathcal{Q}$. We now show how to refine the parameters of $\tilde{f}_{11}$ so that it is as similar as possible to f_{11}.

To refine $\tilde{f}_{11}$ we do not follow the standard EP algorithm. The reason for this is that the extended exact factor f_{11} is invariant to rotations or changes of sign in the matrices $\mathbf{U}$ and $\mathbf{V}$. This creates multiple modes in the posterior distribution and the KL divergence minimized by EP will attempt to cover the support of those modes. Covering all the modes is undesirable. Ideally, we would like to focus on a single mode locally to break the symmetry. To achieve this, we follow the approach used by Stern et al. (2009) and minimize the KL divergence with the arguments swapped. For this, we first marginalize $f_{11}(\mathbf{C}^{\mathcal{O}}, \mathbf{U}, \mathbf{V})\mathcal{Q}^{\backslash 11}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$ with respect to $\mathbf{\Omega}$, $\mathbf{R}^{\neg\mathcal{O}}$ and $\mathbf{\Theta} \setminus \{\mathbf{U}, \mathbf{V}\}$, where $\mathcal{Q}^{\backslash 11}$ is given by the ratio of $\mathcal{Q}$ and $\tilde{f}_{11}$. The parameters of $\mathcal{Q}^{\backslash 11}$ are obtained using

$$[v_{j,k}^{v,\backslash 11}]^{\text{new}} = \left[[v_{j,k}^{v}]^{-1} - [\tilde{v}_{j,k}^{v,11}]^{-1}\right]^{-1}, \tag{75}$$

$$[m_{j,k}^{v,\backslash 11}]^{\text{new}} = [v_{j,k}^{v,\backslash 11}]^{\text{new}} \left[m_{j,k}^{v}[v_{j,k}^{v}]^{-1} - \tilde{m}_{j,k}^{v,11}[\tilde{v}_{j,k}^{v,11}]^{-1}\right], \tag{76}$$

$$[v_{i,k}^{u,\backslash 11}]^{\text{new}} = \left[[v_{i,k}^{u}]^{-1} - [\tilde{v}_{i,k}^{u,11}]^{-1}\right]^{-1}, \tag{77}$$

$$[m_{i,k}^{u,\backslash 11}]^{\text{new}} = [v_{i,k}^{u,\backslash 11}]^{\text{new}} \left[m_{i,k}^{u}[v_{i,k}^{u}]^{-1} - \tilde{m}_{i,k}^{u,11}[\tilde{v}_{i,k}^{u,11}]^{-1}\right], \tag{78}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d$ and $k = 1, \ldots, k$ and

$$[v_{i,j}^{c,\backslash 11}]^{\text{new}} = \left[[v_{i,j}^{c}]^{-1} - [\tilde{v}_{i,j}^{c,11}]^{-1}\right]^{-1}, \tag{79}$$

$$[m_{i,j}^{c,\backslash 11}]^{\text{new}} = [v_{i,j}^{c,\backslash 11}]^{\text{new}} \left[m_{i,j}^{c}[v_{i,j}^{c}]^{-1} - \tilde{m}_{i,j}^{c,11}[\tilde{v}_{i,j}^{c,11}]^{-1}\right], \tag{80}$$

for $(i,j) \in \mathcal{O}$. Let $\mathcal{S}(\mathbf{U}, \mathbf{V})$ be the result of summing out $\mathbf{R}^{\neg\mathcal{O}}$ and integrating out $\mathbf{\Omega}$ and $\mathbf{\Theta} \setminus \{\mathbf{U}, \mathbf{V}\}$ in $f_{11}(\mathbf{C}^{\mathcal{O}}, \mathbf{U}, \mathbf{V})\mathcal{Q}^{\backslash 11}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$. Then

$$\mathcal{S}(\mathbf{U}, \mathbf{V}) = \int \prod_{(i,j)\in\mathcal{O}} \delta(c_{i,j} - \mathbf{u}_i^{\mathrm{T}}\, \mathbf{v}_j) \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(c_{i,j}|m_{i,j}^{c,\backslash 11}, v_{i,j}^{c,\backslash 11})\right]$$

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u,\backslash 11},v_{i,k}^{u,\backslash 11})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 11},v_{j,k}^{v,\backslash 11})\right]\,d\mathbf{C}^{\mathcal{O}} \tag{81}$$

$$=\left[\prod_{(i,j)\in\mathcal{O}}\mathcal{N}(\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j|m_{i,j}^{c,\backslash 11},v_{i,j}^{c,\backslash 11})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u,\backslash 11},v_{i,k}^{u,\backslash 11})\right]$$
$$\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 11},v_{j,k}^{v,\backslash 11})\right]. \tag{82}$$

Let $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ be the posterior approximation $\mathcal{Q}$ after summing out $\mathbf{R}^{\neg\mathcal{O}}$ and integrating out $\mathbf{\Omega}$ and $\mathbf{\Theta}\setminus\{\mathbf{U},\mathbf{V}\}$, that is,

$$\mathcal{Q}_{\mathbf{U},\mathbf{V}}=\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u},v_{i,k}^{u})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v},v_{j,k}^{v})\right]. \tag{83}$$

The parameters of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, that is, $m_{i,k}^{u}$, $v_{i,k}^{u}$, $m_{j,k}^{v}$ and $v_{j,k}^{v}$, for $i=1,\ldots,n$, $j=1,\ldots,d$ and $k=1,\ldots,h$, are then optimized to minimize $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$. We describe how to do this in Section 4.1.14. Once $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ has been updated, we update the parameters of $\mathcal{Q}$ for $\mathbf{U}$ and $\mathbf{V}$ to be the same as those of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$. We also update the parameters of $\mathcal{Q}$ for $\mathbf{C}^{\mathcal{O}}$. To do this, we note that in the exact posterior $c_{i,j}$ is always equal to $\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j$ because of the delta function $\delta(c_{i,j}-\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j)$. Therefore, we set the mean and variance of each $c_{i,j}$ in $\mathcal{Q}$ with $(i,j)\in\mathcal{O}$ to be the same as the mean and variance of the corresponding $\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j$ according to the newly updated $\mathcal{Q}$. This leads to the update

$$[m_{i,j}^{c}]^{\mathrm{new}}=\sum_{k=1}^{h}m_{i,k}^{u}m_{j,k}^{v}\,,\qquad\qquad[v_{i,j}^{c}]^{\mathrm{new}}=\sum_{k=1}^{h}[m_{i,k}^{u}]^{2}v_{j,k}^{v}+v_{i,k}^{u}[m_{j,k}^{v}]^{2}+v_{i,k}^{u}v_{j,k}^{v}\,. \tag{84}$$

for $(i,j)\in\mathcal{O}$. After updating $\mathcal{Q}$, we refine $\tilde{f}_{11}$ so that it is the ratio of $\mathcal{Q}$ and $\mathcal{Q}^{\backslash 11}$, that is,

$$[\tilde{v}_{j,k}^{v,11}]^{\mathrm{new}}=\left[[v_{j,k}^{v}]^{-1}-[v_{j,k}^{v,\backslash 11}]^{-1}\right]^{-1}, \tag{85}$$

$$[\tilde{m}_{j,k}^{v,11}]^{\mathrm{new}}=[\tilde{v}_{j,k}^{v,11}]^{\mathrm{new}}\left[m_{j,k}^{v}[v_{j,k}^{v}]^{-1}-m_{j,k}^{v,\backslash 11}[v_{j,k}^{v,\backslash 11}]^{-1}\right], \tag{86}$$

$$[\tilde{v}_{i,k}^{u,11}]^{\mathrm{new}}=\left[[v_{i,k}^{u}]^{-1}-[v_{i,k}^{u,\backslash 11}]^{-1}\right]^{-1}, \tag{87}$$

$$[\tilde{m}_{i,k}^{u,11}]^{\mathrm{new}}=[\tilde{v}_{i,k}^{u,11}]^{\mathrm{new}}\left[m_{i,k}^{u}[v_{i,k}^{u}]^{-1}-m_{i,k}^{u,\backslash 11}[v_{i,k}^{u,\backslash 11}]^{-1}\right], \tag{88}$$

for $i=1,\ldots,n$, $j=1,\ldots,d$ and $k=1,\ldots,k$ and

$$[\tilde{v}_{i,j}^{c,11}]^{\mathrm{new}}=\left[[v_{i,j}^{c}]^{-1}-[v_{i,j}^{c,\backslash 11}]^{-1}\right]^{-1}, \tag{89}$$

$$[\tilde{m}_{i,j}^{c,11}]^{\mathrm{new}}=[\tilde{v}_{i,j}^{c,11}]^{\mathrm{new}}\left[m_{i,j}^{c}[v_{i,j}^{c}]^{-1}-m_{i,j}^{c,\backslash 11}[v_{i,j}^{c,\backslash 11}]^{-1}\right], \tag{90}$$

for $(i,j)\in\mathcal{O}$. Note that, when performing these EP updates, some of the variances $\tilde{v}_{j,k}^{v,11}$, $\tilde{v}_{i,k}^{u,11}$ and $\tilde{v}_{i,j}^{c,11}$ in $\tilde{f}_{11}$ can become negative. In our experiments, this sometimes created problems when updating other approximate factors. To avoid this, whenever one of the variances of a Gaussian factor in $\tilde{f}_{11}$ is going to become negative, we do not perform the EP update of that Gaussian factor. When this happens, we have to eliminate the EP update in the corresponding factor of $\mathcal{Q}$ since we are first updating $\mathcal{Q}$ and then $\tilde{f}_{11}$ as a function of $\mathcal{Q}$.

4.1.12 EP updates for $\tilde{f}_{12,i,j}$

Recall that $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})$. Since we are assuming MAR data, we only have to consider the factors $f_{12,i,j}$ corresponding to those entries of the rating matrix $\mathbf{R}$ that are actually observed, that is, the factors $f_{12,i,j}$ such that $(i,j) \in \mathcal{O}$. To refine each $\tilde{f}_{12,i,j}$ such that $(i,j) \in \mathcal{O}$ we firstly compute the parameters of $\mathcal{Q}^{\setminus 12,i,j}$. This distribution is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{12,i,j}$. This leads to

$$[v_{i,j}^{a,\setminus 12,i,j}]^{\text{new}} = \left[[v_{i,j}^{a}]^{-1} - [\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]^{-1}, \tag{91}$$

$$[m_{i,j}^{a,\setminus 12,i,j}]^{\text{new}} = [v_{i,j}^{a,\setminus 12,i,j}]^{\text{new}}\left[m_{i,j}^{a}[v_{i,j}^{a}]^{-1} - \tilde{m}_{i,j}^{a,12,i,j}[\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right], \tag{92}$$

$$[v_{i,j}^{c,\setminus 12,i,j}]^{\text{new}} = \left[[v_{i,j}^{c}]^{-1} - [\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right]^{-1}, \tag{93}$$

$$[m_{i,j}^{c,\setminus 12,i,j}]^{\text{new}} = [v_{i,j}^{c,\setminus 12,i,j}]^{\text{new}}\left[m_{i,j}^{c}[v_{i,j}^{c}]^{-1} - \tilde{m}_{i,j}^{c,12,i,j}[\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right], \tag{94}$$

$$[a_i^{\gamma^{\text{row}},\setminus 12,i,j}]^{\text{new}} = a_i^{\gamma^{\text{row}}} - \tilde{a}_i^{\gamma^{\text{row}},12,i,j} + 1, \tag{95}$$

$$[b_i^{\gamma^{\text{row}},\setminus 12,i,j}]^{\text{new}} = b_i^{\gamma^{\text{row}}} - \tilde{b}_i^{\gamma^{\text{row}},12,i,j}, \tag{96}$$

$$[a_j^{\gamma^{\text{col}},\setminus 12,i,j}]^{\text{new}} = a_j^{\gamma^{\text{col}}} - \tilde{a}_j^{\gamma^{\text{col}},12,i,j} + 1, \tag{97}$$

$$[b_j^{\gamma^{\text{col}},\setminus 12,i,j}]^{\text{new}} = b_j^{\gamma^{\text{col}}} - \tilde{b}_j^{\gamma^{\text{col}},12,i,j}. \tag{98}$$

After this, to refine the approximate factor $\tilde{f}_{12,i,j}$, we have to find the expectation of the sufficient statistics with respect to $h(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}}) = \mathcal{Q}^{\setminus 12,j,k}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})$. After summing out $\mathbf{R}^{\neg\mathcal{O}}$ and integrating out $\boldsymbol{\Omega}$ and $\boldsymbol{\Theta} \setminus \{a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}\}$ in h, we obtain

$$\begin{aligned} h(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = {} & \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})\mathcal{N}(a_{i,j}|m_{i,j}^{a,\setminus 12,i,j}, v_{i,j}^{a,\setminus 12,i,j})\mathcal{N}(c_{i,j}|m_{i,j}^{c,\setminus 12,i,j}, v_{i,j}^{c,\setminus 12,i,j}) \\ & \text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\setminus 12,i,j}, b_i^{\gamma^{\text{row}},\setminus 12,i,j})\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\setminus 12,i,j}, b_j^{\gamma^{\text{col}},\setminus 12,i,j}). \end{aligned} \tag{99}$$

The normalization constant of $h(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})$ is then

$$Z = \int h(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})\, da_{i,j}\, dc_{i,j}\, d\gamma_i^{\text{row}}\, d\gamma_j^{\text{col}} \tag{100}$$

$$= \int \mathcal{N}(m_{i,j}^{a,\setminus 12,i,j}|m_{i,j}^{c,\setminus 12,i,j}, v_{i,j}^{a,\setminus 12,i,j} + v_{i,j}^{c,\setminus 12,i,j} + \gamma_i^{\text{row}}\gamma_j^{\text{col}}) \tag{101}$$

$$\text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\setminus 12,i,j}, b_i^{\gamma^{\text{row}},\setminus 12,i,j})\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\setminus 12,i,j}, b_j^{\gamma^{\text{col}},\setminus 12,i,j}) d\gamma_i^{\text{row}}\, d\gamma_j^{\text{col}} \tag{102}$$

$$\begin{aligned} \approx {} & \mathcal{N}(m_{i,j}^{a,\setminus 12,i,j}|m_{i,j}^{c,\setminus 12,i,j}, v_{i,j}^{a,\setminus 12,i,j} + v_{i,j}^{c,\setminus 12,i,j} + \\ & b_i^{\gamma^{\text{row}},\setminus 12,i,j} b_j^{\gamma^{\text{col}},\setminus 12,i,j}/[(a_i^{\gamma^{\text{row}},\setminus 12,i,j} + 1)(a_j^{\gamma^{\text{col}},\setminus 12,i,j} + 1)]), \end{aligned} \tag{103}$$

where in (103) we have approximated $\text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\setminus 12,i,j}, b_i^{\gamma^{\text{row}},\setminus 12,i,j})$ and $\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\setminus 12,i,j}, b_j^{\gamma^{\text{col}},\setminus 12,i,j})$ with point probability masses located at the modes of these factors. The expectation of the sufficient statistics $a_{i,j}$, $[a_{i,j}]^2$, $c_{i,j}$, $[c_{i,j}]^2$, γ_i^{row}, $[\gamma_i^{\text{row}}]^2$, γ_j^{col} and $[\gamma_j^{\text{col}}]^2$ with respect to $h(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})$ can be approximated in a similar way as the previous normalization constant, as we describe below. For the random variables γ_i^{row} and γ_j^{col}, the KL-divergence is actually minimized by matching the first moments and the expectations of $\log \gamma_i^{\text{row}}$ and $\log \gamma_j^{\text{col}}$. However, matching the expectation of $\log \gamma_i^{\text{row}}$ and $\log \gamma_j^{\text{col}}$ would require computing the inverse of the Digamma function, which has no analytical solution. To avoid this, we match the first and second moments of γ_i^{row} and γ_j^{col}, which is expected to produce reasonably good results.

We approximate the moments of the random variables γ_i^{row} and γ_j^{col}, using the following property of inverse gammas, see (2). Let $H(a,b)$ be the normalization constant of $f(x)\mathcal{IG}(x|a,b)$ for a function f, that is,

$H(a,b) = \int f(x)\mathcal{IG}(x|a,b)\,dx$. Then we have $\int x f(x)\mathcal{IG}(x|a,b)\,dx = H(a+1,b)a/b$ and $\int x^2\mathcal{IG}(x|a,b)\,dx = H(a+2,b)a(a+1)/b^2$. Therefore, each moment can be easily approximated given a procedure to approximate the normalization constant $H(a,b)$. For this, we only have to replace $H(a+1,b)$ and $H(a+2,b)$ in the previous equations with their corresponding approximations. Following a similar approach, we can compute approximations for the moments of $a_{i,j}$ and $c_{i,j}$. In particular, we use the following property of the Gaussian distribution. Let $H(m,v)$ be the normalization constant of $f(x)\mathcal{N}(x|m,v)$ for a particular function f, that is, $H(m,v) = \int f(x)\mathcal{N}(x|m,v)\,dx$. Then $[\mathrm{H}(m,v)]^{-1}\int x f(x)\mathcal{N}(x|m,v)\,dx = m + v\frac{d\log H(m,v)}{dm}$ and $[H(m,v)]^{-1}\int x^2\mathcal{N}(x|m,v)\,dx - [[H(m,v)]^{-1}\int x\mathcal{N}(x|m,v)\,dx]^2 = v - v^2([\frac{d\log H(m,v)}{dm}]^2 - 2\frac{d\log H(m,v)}{dv})$.

The updates for $\tilde{f}_{12,i,j}$ are then

$$[\tilde{m}_{i,j}^{a,12,i,j}]^{\text{new}} = m_{i,j}^{c,\backslash 12,i,j}\,, \tag{104}$$

$$[\tilde{v}_{i,j}^{a,12,i,j}]^{\text{new}} = v_{i,j}^{c,\backslash 12,i,j} + b_i^{\gamma^{\text{row}},\backslash 12,i,j} b_j^{\gamma^{\text{col}},\backslash 12,i,j}/[(a_i^{\gamma^{\text{row}},\backslash 12,i,j}+1)(a_j^{\gamma^{\text{col}},\backslash 12,i,j}+1)]\,, \tag{105}$$

$$[\tilde{m}_{i,j}^{c,12,i,j}]^{\text{new}} = m_{i,j}^{a,\backslash 12,i,j}\,, \tag{106}$$

$$[\tilde{v}_{i,j}^{c,12,i,j}]^{\text{new}} = v_{i,j}^{a,\backslash 12,i,j} + b_i^{\gamma^{\text{row}},\backslash 12,i,j} b_j^{\gamma^{\text{col}},\backslash 12,i,j}/[(a_i^{\gamma^{\text{row}},\backslash 12,i,j}+1)(a_j^{\gamma^{\text{col}},\backslash 12,i,j}+1)]\,, \tag{107}$$

$$[\tilde{a}_i^{\gamma^{\text{row}},12,i,j}]^{\text{new}} = a'_{\text{row}} - a_i^{\gamma^{\text{row}},\backslash 12,i,j} + 1\,, \tag{108}$$

$$[\tilde{b}_i^{\gamma^{\text{row}},12,i,j}]^{\text{new}} = b'_{\text{row}} - b_i^{\gamma^{\text{row}},\backslash 12,i,j}\,, \tag{109}$$

$$[\tilde{a}_j^{\gamma^{\text{col}},12,i,j}]^{\text{new}} = a'_{\text{col}} - a_j^{\gamma^{\text{col}},\backslash 12,i,j} + 1\,, \tag{110}$$

$$[\tilde{b}_j^{\gamma^{\text{col}},12,i,j}]^{\text{new}} = b'_{\text{col}} - b_j^{\gamma^{\text{col}},\backslash 12,i,j}\,, \tag{111}$$

where we define a'_{row}, b'_{row}, a'_{col}, b'_{col} as

$$a'_{\text{row}} = \frac{a_i^{\gamma^{\text{row}},\backslash 12,i,j}[Z_1^{\text{row}}]^2}{(a_i^{\gamma^{\text{row}},\backslash 12,i,j}+1)ZZ_2^{\text{row}} - a_i^{\gamma^{\text{row}},\backslash 12,i,j}[Z_1^{\text{row}}]^2}\,, \tag{112}$$

$$b'_{\text{row}} = \frac{b_i^{\gamma^{\text{row}},\backslash 12,i,j} ZZ_1^{\text{row}}}{(a_i^{\gamma^{\text{row}},\backslash 12,i,j}+1)ZZ_2^{\text{row}} - a_i^{\gamma^{\text{row}},\backslash 12,i,j}[Z_1^{\text{row}}]^2}\,, \tag{113}$$

$$a'_{\text{col}} = \frac{a_j^{\gamma^{\text{col}},\backslash 12,i,j}[Z_1^{\text{col}}]^2}{(a_j^{\gamma^{\text{col}},\backslash 12,i,j}+1)ZZ_2^{\text{col}} - a_j^{\gamma^{\text{col}},\backslash 12,i,j}[Z_1^{\text{col}}]^2}\,, \tag{114}$$

$$b'_{\text{col}} = \frac{b_j^{\gamma^{\text{col}},\backslash 12,i,j} ZZ_1^{\text{col}}}{(a_j^{\gamma^{\text{col}},\backslash 12,i,j}+1)ZZ_2^{\text{col}} - a_j^{\gamma^{\text{col}},\backslash 12,i,j}[Z_1^{\text{col}}]^2}\,, \tag{115}$$

Z_1^{row} and Z_2^{row} are obtained in the same way as the normalization constant Z, but increasing $a_i^{\gamma^{\text{row}},\backslash 12,i,j}$ in one and two units, respectively, and similarly, Z_1^{col} and Z_2^{col} are obtained by increasing $a_j^{\gamma^{\text{col}},\backslash 12,i,j}$ in one and two units, respectively.

Note that, in these EP update equations, some of the variances $\tilde{v}_{i,j}^{a,12,i,j}$ and, $\tilde{v}_{i,j}^{c,12,i,j}$ and can become negative. To avoid this, whenever one of the variances of a Gaussian factor in $\tilde{f}_{12,i,j}$ is going to become negative, we do not perform the EP update of that Gaussian factor. Furthermore, we only refine the approximate factor $\tilde{f}_{12,i,j}$ if all the conditions $a_j^{\gamma^{\text{col}},\backslash 12,i,j} > 2$, $b_j^{\gamma^{\text{col}},\backslash 12,i,j} > 0$, $a_i^{\gamma^{\text{row}},\backslash 12,i,j} > 2$, $b_i^{\gamma^{\text{ro}},\backslash 12,i,j} > 0$, $v_{i,j}^{a,\backslash 12,i,j} > 0$ and $v_{i,j}^{c,\backslash 12,i,j} > 0$ are satisfied.

Once we have updated $\tilde{f}_{12,i,j}$, we recompute $\mathcal{Q}$ using

$$[v_{i,j}^{a}]^{\text{new}} = \left[[v_{i,j}^{a,\backslash 12,i,j}]^{-1} + [\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]^{-1}\,, \tag{116}$$

$$[m_{i,j}^{a}]^{\text{new}} = [v_{i,j}^{a}]^{\text{new}}\left[m_{i,j}^{a,\backslash 12,i,j}[v_{i,j}^{a,\backslash 12,i,j}]^{-1} + \tilde{m}_{i,j}^{a,12,i,j}[\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]\,, \tag{117}$$

$$[v_{i,j}^{c}]^{\text{new}} = \left[[v_{i,j}^{c,\backslash 12,i,j}]^{-1} + [\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right]^{-1} , \tag{118}$$

$$[m_{i,j}^{c}]^{\text{new}} = [v_{i,j}^{c}]^{\text{new}} \left[m_{i,j}^{c,\backslash 12,i,j}[v_{i,j}^{c,\backslash 12,i,j}]^{-1} + \tilde{m}_{i,j}^{c,12,i,j}[\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right] , \tag{119}$$

$$[a_{i}^{\gamma^{\text{row}}}]^{\text{new}} = a_{i}^{\gamma^{\text{row}},\backslash 12,i,j} + \tilde{a}_{i}^{\gamma^{\text{row}},12,i,j} - 1 , \tag{120}$$

$$[b_{i}^{\gamma^{\text{row}}}]^{\text{new}} = b_{i}^{\gamma^{\text{row}},\backslash 12,i,j} + \tilde{b}_{i}^{\gamma^{\text{row}},12,i,j} , \tag{121}$$

$$[a_{j}^{\gamma^{\text{col}}}]^{\text{new}} = a_{j}^{\gamma^{\text{col}},\backslash 12,i,j} + \tilde{a}_{j}^{\gamma^{\text{col}},12,i,j} - 1 , \tag{122}$$

$$[b_{j}^{\gamma^{\text{col}}}]^{\text{new}} = b_{j}^{\gamma^{\text{col}},\backslash 12,i,j} + \tilde{b}_{j}^{\gamma^{\text{col}},12,i,j} . \tag{123}$$

In our experiments we observed that, if we refine the approximate factors $\tilde{f}_{12,i,j}$ during the first iterations of EP, the proposed model gets stuck in solutions in which the the components of the noise variables $\boldsymbol{\gamma}^{\text{row}}$ and $\boldsymbol{\gamma}^{\text{col}}$ take very large values. The reason for this is that during the first iterations of EP, the posterior approximation for the latent variables $\mathbf{U}$ and $\mathbf{V}$ is not yet very good and consequently the EP update equations explain this by assuming that there is large additive noise. The result is that the EP approximation $\mathcal{Q}$ gets stuck in solutions in which the components of $\boldsymbol{\gamma}^{\text{row}}$ and $\boldsymbol{\gamma}^{\text{col}}$ are too large. To avoid this, we do not refine the approximate factors $\tilde{f}_{12,i,j}$ during the second iteration of EP. Note that in the first iteration, when we refine the approximate factors $\tilde{f}_{12,i,j}$, we do not modify the factors of $\mathcal{Q}$ for $\boldsymbol{\gamma}^{\text{row}}$ and $\boldsymbol{\gamma}^{\text{col}}$. This means that we can always safely refine the approximate factors $\tilde{f}_{12,i,j}$ during the first EP iteration, even though the current posterior approximation for $\mathbf{U}$ and $\mathbf{V}$ is not yet good.

4.1.13 EP updates for $\tilde{f}_{13,i,j,k}$

Recall that $f_{13,i,j,k}(r_{i,j}, a_{i,j}, b_{j,k}) = \Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]$. Since we are assuming MAR data, we only have to consider the factors $f_{12,i,j}$ corresponding to those entries of the rating matrix $\mathbf{R}$ that are actually observed, that is, the factors $f_{13,i,j,k}$ such that $(i,j) \in \mathcal{O}$. To refine each $\tilde{f}_{13,i,j,k}$ such that $(i,j) \in \mathcal{O}$, we firstly compute the parameters of $\mathcal{Q}^{\backslash 13,i,j,k}$. This distribution is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{13,i,j,k}$. This leads to

$$[v_{j,k}^{b,\backslash 13,i,j,k}]^{\text{new}} = \left[[v_{j,k}^{b}]^{-1} - [\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right]^{-1} , \tag{124}$$

$$[m_{j,k}^{b,\backslash 13,i,j,k}]^{\text{new}} = [v_{j,k}^{b,\backslash 13,i,j,k}]^{\text{new}} \left[m_{j,k}^{b}[v_{j,k}^{b}]^{-1} - \tilde{m}_{j,k}^{b,13,i,j,k}[\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right] , \tag{125}$$

$$[v_{i,j}^{a,\backslash 13,i,j,k}]^{\text{new}} = \left[[v_{i,j}^{a}]^{-1} - [\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right]^{-1} , \tag{126}$$

$$[m_{i,j}^{a,\backslash 13,i,j,k}]^{\text{new}} = [v_{i,j}^{a,\backslash 13,i,j,k}]^{\text{new}} \left[m_{i,j}^{a}[v_{i,j}^{a}]^{-1} - \tilde{m}_{i,j}^{a,13,i,j,k}[\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right] . \tag{127}$$

After this, we update the approximate factor $\tilde{f}_{13,i,j,k}$ by matching expected sufficient statistics between $\mathcal{Q}^{\backslash 13,i,j,k}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]$ and $\mathcal{Q}^{\backslash 13,i,j,k}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{13,i,j,k}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$. This leads to the updates

$$\tilde{m}_{j,k}^{b,13,i,j,k} = m_{j,k}^{b,\backslash 13,i,j,k} + \kappa \qquad \tilde{v}_{j,k}^{b,13,i,j,k} = -v_{j,k}^{b,\backslash 13,i,j,k} - 1/\beta \tag{128}$$

$$\tilde{m}_{i,j}^{a,13,i,j,k} = m_{i,j}^{a,\backslash 13,i,j,k} - \kappa \qquad \tilde{v}_{i,j}^{a,13,i,j,k} = -v_{i,j}^{a,\backslash 13,i,j,k} - 1/\beta \tag{129}$$

where β and κ are given by

$$\beta = -\frac{\phi(\alpha)}{\Phi(\alpha)}\left(\alpha + \frac{\phi(\alpha)}{\Phi(\alpha)}\right)\left[v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}\right]^{-1} , \tag{130}$$

$$\kappa = -\frac{\text{sign}[r_{i,j} - k - 0.5]}{\sqrt{v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}}}\left[\alpha + \frac{\phi(\alpha)}{\Phi(\alpha)}\right]^{-1} , \tag{131}$$

with

$$\alpha = \text{sign}[r_{i,j} - k - 0.5] \frac{m_{j,k}^{a,\backslash 13,i,j,k} - m_{j,k}^{b,\backslash 13,i,j,k}}{\sqrt{v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}}} \tag{132}$$

and ϕ and Φ denote the standard Gaussian density and cdf functions, respectively.

Note that, when performing these EP updates, the variances $\tilde{v}_{i,j}^{a,13,i,j,k}$ or $\tilde{v}_{j,k}^{b,13,i,j,k}$ can become negative. In our experiments, this sometimes created problems when updating other approximate factors. To avoid this, whenever one of the variances of a Gaussian factor in $\tilde{f}_{13,i,j,k}$ is going to become negative, we do not perform the EP update of that Gaussian factor. Similarly, we do not update $\tilde{f}_{13,i,j,k}$ when $v_{i,j}^{a,\backslash 13,i,j,k}$ or $v_{j,k}^{b,\backslash 13,i,j,k}$ are negative.

Finally, once we have updated $\tilde{f}_{13,i,j,k}$, we recompute $\mathcal{Q}$ by setting

$$[v_{j,k}^{b}]^{\text{new}} = \left[[v_{j,k}^{b,\backslash 13,i,j,k}]^{-1} + [\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1} \right]^{-1} , \tag{133}$$

$$[m_{j,k}^{b}]^{\text{new}} = [v_{j,k}^{b}]^{\text{new}} \left[m_{j,k}^{b,\backslash 13,i,j,k} [v_{j,k}^{b,\backslash 13,i,j,k}]^{-1} + \tilde{m}_{j,k}^{b,13,i,j,k} [\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1} \right] , \tag{134}$$

$$[v_{i,j}^{a}]^{\text{new}} = \left[[v_{i,j}^{a,\backslash 13,i,j,k}]^{-1} + [\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1} \right]^{-1} , \tag{135}$$

$$[m_{i,j}^{a}]^{\text{new}} = [v_{i,j}^{a}]^{\text{new}} \left[m_{i,j}^{a,\backslash 13,i,j,k} [v_{i,j}^{a,\backslash 13,i,j,k}]^{-1} + \tilde{m}_{i,j}^{a,13,i,j,k} [\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1} \right] . \tag{136}$$

4.1.14 Minimizing the reversed KL divergence when refining $\tilde{f}_{11}$

In Section 4.1.11 we had to minimize $\text{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}} \| \mathcal{S})$, where

$$\mathcal{S}(\mathbf{U},\mathbf{V}) = \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(\mathbf{u}_i^{\text{T}} \mathbf{v}_j | m_{i,j}^{c,\backslash 11}, v_{i,j}^{c,\backslash 11}) \right] \left[\prod_{i=1}^{n} \prod_{k=1}^{h} \mathcal{N}(u_{i,k} | m_{i,k}^{u,\backslash 11}, v_{i,k}^{u,\backslash 11}) \right] \left[\prod_{j=1}^{d} \prod_{k=1}^{h} \mathcal{N}(v_{j,k} | m_{j,k}^{v,\backslash 11}, v_{j,k}^{v,\backslash 11}) \right] . \tag{137}$$

and

$$\mathcal{Q}_{\mathbf{U},\mathbf{V}} = \left[\prod_{i=1}^{n} \prod_{k=1}^{h} \mathcal{N}(u_{i,k} | m_{i,k}^{u}, v_{i,k}^{u}) \right] \left[\prod_{j=1}^{d} \prod_{k=1}^{h} \mathcal{N}(v_{j,k} | m_{j,k}^{v}, v_{j,k}^{v}) \right] , \tag{138}$$

with respect to the parameters of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, that is, $m_{i,k}^{u}$, $v_{i,k}^{u}$, $m_{j,k}^{v}$ and $v_{j,k}^{v}$, for $i = 1,\ldots,n$, $j = 1,\ldots,d$ and $k = 1,\ldots,h$, For this, we follow Paquet and Koenigstein (2013) and make use of the intermediate approximation $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ that does not factorize across columns, where

$$\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}} = \left[\prod_{i=1}^{n} \mathcal{N}(\mathbf{u}_i | \hat{\mathbf{m}}_i^{u}, \hat{\mathbf{V}}_i^{u}) \right] \left[\prod_{j=1}^{d} \mathcal{N}(\mathbf{v}_j | \hat{\mathbf{m}}_j^{v}, \hat{\mathbf{V}}_j^{v}) \right] \tag{139}$$

and $\hat{\mathbf{m}}_i^{u}$ and $\hat{\mathbf{m}}_j^{v}$ are the mean vectors for the i-th row of $\mathbf{U}$ and the j-th row of $\mathbf{V}$, respectively, and $\mathbf{V}_i^{u}$ and $\mathbf{V}_j^{v}$ are the corresponding covariance matrices. To update the variational parameters in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the i-th row of $\mathbf{U}$ we first equate the gradient of $\text{KL}(\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}} \| \mathcal{S})$ with respect to the parameters $\hat{\mathbf{m}}_i^{u}$ and $\hat{\mathbf{V}}_i^{u}$ of the full (not factorized) Gaussian approximation $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ to zero. Then we adjust $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ so that $\text{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}} \| \hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}})$ is minimized with respect to the parameters $m_{i,k}^{u}$ and $v_{i,k}^{u}$ of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, for $k = 1,\ldots,h$. This is achieved when,

for the i-th row of $\mathbf{U}$, the means of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ match the means of $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ and the precisions of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ match the diagonal precisions of $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$. Following Lim and Teh (2007), we update $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ with respect to $\hat{\mathbf{m}}_i^u$ and $\hat{\mathbf{V}}_i^u$ by setting

$$[\hat{\mathbf{V}}_i^u]^{-1} = \text{diag}(v_{i,1}^{u,\backslash 11}, \ldots, v_{i,h}^{u,\backslash 11})^{-1} + \sum_{j:(i,j)\in\mathcal{O}} \frac{\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j]}{v_{i,j}^{c,\backslash 11}}\,, \tag{140}$$

$$\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1} = (m_{i,1}^{u,\backslash 11}, \ldots, m_{i,h}^{u,\backslash 11})\text{diag}(v_{i,1}^{u,\backslash 11}, \ldots, v_{i,h}^{u,\backslash 11})^{-1} + \sum_{j:(i,j)\in\mathcal{O}} \frac{m_{i,j}^{c,\backslash 11}\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}]}{v_{i,j}^{c,\backslash 11}}\,. \tag{141}$$

After this, we update $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ by setting

$$m_{i,k}^u = [\hat{\mathbf{m}}_i^u]_k\,, \qquad\qquad v_{i,k}^u = 1/\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{k,k}\,, \tag{142}$$

for $k = 1, \ldots, h$. The corresponding parameters for the j-th row of $\mathbf{V}$, that is, $m_{j,k}^v$ and $v_{j,k}^v$, where $k = 1, \ldots, h$, are updated in a similar way. In practice, we first iterate over $i = 1, \ldots, n$, updating the $m_{i,k}^v$ and $v_{j,k}^v$ in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the i-th row of $\mathbf{U}$, and then we iterate over $j = 1, \ldots, d$, updating the $m_{i,k}^v$ and $v_{j,k}^v$ in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the j-th column of $\mathbf{V}$. We repeat this process a total of 3 times each time we want to refine the approximate factor $\tilde{f}_{11}$. Furthermore, we use as initial solution for $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ the value obtained during the previous iteration of EP. On the first EP iteration, we initialize $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ by randomly sampling all the mean parameters $m_{i,k}^u$ and $m_{j,k}^v$ from a standard Gaussian distribution and then setting all the variance parameters $v_{i,k}^u$ and $v_{j,k}^v$ to one.

4.1.15 The predictive distribution of the complete data model

Once the parameters of $\mathcal{Q}$ have been fixed by running the EP method, we can use $\mathcal{Q}$ to estimate the posterior probability that the entry in the i-th row and j-th column of the rating matrix $\mathbf{R}$ may have taken value $r_{i,j}^\star$. Here, we assume that the entry in the i-th row and j-th column of $\mathbf{R}$ is not contained in the set of observed ratings $\mathbf{R}^{\mathcal{O}}$. When the data is Missing At Random (MAR), the exact posterior distribution for $r_{i,j}^\star$ given $\mathbf{R}^{\mathcal{O}}$ is then

$$p(r_{i,j}^\star|\mathbf{R}^{\mathcal{O}}) = \int p(r_{i,,j}^\star|a_{i,j}^\star, \mathbf{b}_j)p(a_{i,j}^\star|c_{i,j}^\star, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})p(c_{i,j}^\star|\mathbf{u}_i, \mathbf{v}_j)p(\boldsymbol{\Theta}|\mathbf{R}^{\mathcal{O}})\, d\boldsymbol{\Theta}\, da_{i,j}^\star\, dc_{i,j}^\star\,, \tag{143}$$

with $p(r_{i,,j}^\star|a_{i,j}^\star, \mathbf{b}_j) = \prod_{k=1}^{L-1} \Theta\left[\text{sign}[r_{i,j}^\star - k - 0.5](a_{i,j}^\star - b_{j,k})\right]$, $p(a_{i,j}^\star|c_{i,j}^\star, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}^\star|c_{i,j}^\star, \gamma_i^{\text{row}}\gamma_j^{\text{col}})$, $p(c_{i,j}^\star|\mathbf{u}_i, \mathbf{v}_j) = \delta(c_{i,j}^\star - \mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j)$ and $p(\boldsymbol{\Theta}|\mathbf{R}^{\mathcal{O}})$ is the posterior distribution for $\boldsymbol{\Theta}$ given $\mathbf{R}^{\mathcal{O}}$ in the complete data model under the MAR assumption, that is,

$$\begin{aligned} p(\boldsymbol{\Theta}|\mathbf{R}^{\mathcal{O}}) = {} & p(\mathbf{R}^{\mathcal{O}}|\mathbf{A}^{\mathcal{O}}, \mathbf{B})p(\mathbf{A}^{\mathcal{O}}|\mathbf{C}^{\mathcal{O}}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}})p(\mathbf{C}^{\mathcal{O}}|\mathbf{U}, \mathbf{V}) \\ & p(\mathbf{U}|\mathbf{m}^{\mathbf{U}}, \mathbf{v}^{\mathbf{U}})p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}})p(\mathbf{B}|\mathbf{b}_0)p(\mathbf{b}_0) \\ & p(\boldsymbol{\gamma}^{\text{row}})p(\boldsymbol{\gamma}^{\text{col}})p(\mathbf{m}^{\mathbf{U}})p(\mathbf{m}^{\mathbf{V}})p(\mathbf{v}^{\mathbf{U}})p(\mathbf{v}^{\mathbf{V}})/p(\mathbf{R}^{\mathcal{O}})\,. \end{aligned} \tag{144}$$

where $p(\mathbf{R}^{\mathcal{O}})$ is a normalization constant and

$$p(\mathbf{R}^{\mathcal{O}}|\mathbf{A}^{\mathcal{O}}, \mathbf{B}) = \prod_{(i,j)\in\mathcal{O}} \prod_{k=1}^{L-1} \Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]\,, \tag{145}$$

$$p(\mathbf{A}^{\mathcal{O}}|\mathbf{C}^{\mathcal{O}}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}}) = \prod_{(i,j)\in\mathcal{O}} \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})\,, \tag{146}$$

$$p(\mathbf{C}^{\mathcal{O}}|\mathbf{U}, \mathbf{V}) = \prod_{(i,j)\in\mathcal{O}} \delta(c_{i,j} - \mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j)\,. \tag{147}$$

To obtain an approximation to (143) we first replace the exact posterior $p(\mathbf{\Theta}|\mathbf{R}^{\mathcal{O}})$ in (143) with the EP approximation $\mathcal{Q}$. However, even after doing this approximation, the resulting integral is not analytical. We therefore, perform an additional approximation. We replace $\int \delta(c^{\star}_{i,j} - \mathbf{u}_i^{\mathrm{T}}\ \mathbf{v}_j)\mathcal{Q}(\mathbf{\Theta})\, d\mathbf{\Theta}$ with a Gaussian with mean $m^{c,\star}_{i,j} = \sum_{k=1}^{h} m^{u}_{i,k} m^{v}_{j,k}$ and variance $v^{c,\star}_{i,j} = \sum_{k=1}^{h} [m^{u}_{i,k}]^2 v^{v}_{j,k} + v^{u}_{i,k}[m^{v}_{j,k}]^2 + v^{u}_{i,k} v^{v}_{j,k}$. Note that $\mathbf{u}_i^{\mathrm{T}}\ \mathbf{v}_j$ is a a random variable with mean $m^{c,\star}_{i,j}$ and variance $v^{c,\star}_{i,j}$ under $\mathcal{Q}$. Again, we still need to perform an additional approximation. We replace $\int \mathcal{N}(a^{\star}_{i,j}|c^{\star}_{i,j}, \gamma_i^{\mathrm{row}}\gamma_j^{\mathrm{col}})\mathcal{N}(c^{\star}_{i,j}|m^{c,\star}_{i,j}, v^{c,\star}_{i,j})\mathcal{Q}(\mathbf{\Theta})d\mathbf{\Theta}$ with an additional Gaussian with mean $m^{c,\star}_{i,j}$ and variance $v^{c,\star}_{i,j} + v^{\gamma}_{i,j}$ where $v^{\gamma}_{i,j} = [b^{\gamma^{\mathrm{row}}} b^{\gamma^{\mathrm{col}}}][(a^{\gamma^{\mathrm{row}}}+1)(a^{\gamma^{\mathrm{col}}}+1)]^{-1}$. In this case, we are approximating the inverse-gamma factors for γ_i^{row} and γ_j^{col} in $\mathcal{Q}$ with point masses located at the modes of those factors. The posterior distribution for $r^{\star}_{i,j}$ given by the complete data model (CDM) once we have observed $\mathbf{R}^{\mathcal{O}}$ is then approximated by

$$\begin{aligned}\tilde{p}_{CDM}(r^{\star}_{i,j}|\mathbf{R}^{\mathcal{O}}) &= \int \prod_{k=1}^{L-1} \Theta\left[\mathrm{sign}[r^{\star}_{i,j} - k - 0.5](a^{\star}_{i,j} - b_{j,k})\right] \mathcal{N}(a^{\star}_{i,j}|m^{c,\star}_{i,j}, v^{c,\star}_{i,j} + v^{\gamma}_{i,j})\mathcal{Q}(\mathbf{\Theta})\, d\mathbf{\Theta}\, da^{\star}_{i,j} \\ &= \Phi\left\{\zeta(r^{\star}_{i,j})\right\} - \Phi\left\{\zeta(r^{\star}_{i,j} - 1)\right\}\,, \end{aligned} \tag{148}$$

where $\zeta(r^{\star}_{i,j}) = (m^{b}_{i,r^{\star}_{i,j}} - m^{c,\star}_{i,j})(v^{c,\star}_{i,j} + v^{b}_{j,r^{\star}_{i,j}} + v^{\gamma}_{i,j})^{-0.5}$ and $\Phi(\cdot)$ is the standard Gaussian cdf.

4.2 Approximate Inference in the missing data model

In this section we describe the operations performed to approximate the exact factors for the missing data model, that is, the exact factors 14 to 19 in Figure 1. We approximate all these factors in a single step. For this, we define the extended factor for the missing data model (MDM) as

$$\begin{aligned} f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}}) = &\left[\prod_{i=1}^{n}\prod_{k=1}^{h} f_{14,i,k}(e_{i,k})\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h} f_{15,j,k}(f_{j,k})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{L} f_{16,i,l}(\lambda^{\mathrm{row}}_{i,l})\right] \\ &\left[\prod_{j=1}^{d}\prod_{k=1}^{L} f_{17,j,l}(\psi^{\mathrm{col}}_{j,l})\right] f_{18}(z)\left[\prod_{i=1}^{n}\prod_{j=1}^{d} f_{19,i,j}(r_{i,j}, x_{i,j}, \mathbf{e}_i, \mathbf{f}_j, z, \boldsymbol{\lambda}^{\mathrm{row}}_i, \boldsymbol{\psi}^{\mathrm{col}}_j)\right]\,, \end{aligned} \tag{149}$$

where $\mathbf{\Omega}$ is the set of variables $\mathbf{\Omega} = \{\mathbf{E}, \mathbf{F}, z, \mathbf{\Lambda}^{\mathrm{row}}, \mathbf{\Psi}^{\mathrm{col}}\}$ and $\mathbf{R}^{\neg\mathcal{O}}$ denotes the set with the entries of the raining matrix $\mathbf{R}$ that are not observed. We approximate the above extended factor with an approximate factor $\tilde{f}_{MDM}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$ that has the same functional form as the posterior approximation $\mathcal{Q}$. We now show how to refine the parameters of $\tilde{f}_{MDM}$ so that it is as similar as possible to f_{MDM}. The first step is to compute $\mathcal{Q}^{\backslash MDM}$ as the ratio between $\mathcal{Q}$ and $\tilde{f}_{MDM}$. Note that the exact factor f_{MDM} does not depend on $\mathbf{\Theta}$. This means that we can ignore in $\mathcal{Q}^{\backslash MDM}$ any factor for any variable in $\mathbf{\Theta}$. Furthermore, the factors in the complete data model, that is, factors 1 to 13 in Figure 1, do not depend on $\mathbf{\Omega}$. This means that $\mathcal{Q}^{\backslash MDM}$ is uniform and non-informative on $\mathbf{\Omega}$ and consequently we can ignore in $\mathcal{Q}^{\backslash MDM}$ any factor for any variable in $\mathbf{\Omega}$. Therefore, we are only interested in knowing the parameters of the factors in $\mathcal{Q}^{\backslash MDM}$ for $\mathbf{R}^{\neg\mathcal{O}}$. In particular, we have that the parameters of $\mathcal{Q}^{\backslash MDM}$ are

$$p^{\backslash MDM}_{i,j,l} = \tilde{p}_{CDM}(r^{\star}_{i,j} = l|\mathbf{R}^{\mathcal{O}})\,, \tag{150}$$

for $(i,j) \notin \mathcal{O}$ and $l = 1, \ldots, L$, where $\tilde{p}_{CDM}(r^{\star}_{i,j} = l|\mathbf{R}^{\mathcal{O}})$ is given by (148), that is, the prediction of the complete data model for the probability that the entry in the i-th row and j-th column of the rating matrix $\mathbf{R}$ may have taken value l, where $(i,j) \notin \mathcal{O}$. In practice we do not store all the $p^{\backslash MDM}_{i,j,l}$, for $(i,j) \notin \mathcal{O}$ and $l = 1, \ldots, L$, in memory, since their number scales as $nd - |\mathcal{O}|$. When n and d are very large, storing all of them is infeasible. As a solution, we will compute them only when needed and we will discard them afterwards.

The standard EP algorithm would minimize the KL-divergence between $f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\backslash MDM}(\mathbf{R}^{\neg\mathcal{O}})$ and $\tilde{f}_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\backslash MDM}(\mathbf{R}^{\neg\mathcal{O}})$, where $\mathcal{Q}^{\backslash MDM}(\mathbf{R}^{\neg\mathcal{O}})$ is the marginal of $\mathcal{Q}^{\backslash MDM}$ on $\mathbf{R}^{\neg\mathcal{O}}$. However,

this is infeasible in practice. Instead, we follow the approach used by Stern et al. (2009) and minimize the KL divergence with the arguments swapped. That is, we minimize the KL divergence between $\mathcal{Q}$ and $f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\setminus MDM}(\mathbf{R}^{\neg\mathcal{O}})$. The following section shows how to minimize this divergence with respect to the parameters of $\mathcal{Q}$.

Once we have adjusted $\mathcal{Q}$, we would have to update $\tilde{f}_{MDM}$ so that it is the ratio of $\mathcal{Q}$ and $\mathcal{Q}^{\setminus MDM}$. However, in practice, this is not necessary since we always work with $\mathcal{Q}$ and never access $\tilde{f}_{MDM}$ directly.

4.2.1 The variational objective function

The KL divergence between $\mathcal{Q}$ and $f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\setminus MDM}(\mathbf{R}^{\neg\mathcal{O}})$ is minimized when we maximize the following objective with respect to the parameters of $\mathcal{Q}$

$$\mathcal{L} = \mathcal{H}[\mathcal{Q}] + \sum_{\mathbf{R}^{\neg\mathcal{O}}} \int \mathcal{Q}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}}) \log \left[f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\setminus MDM}(\mathbf{R}^{\neg\mathcal{O}}) \right] d\mathbf{\Omega}\,, \tag{151}$$

where $\mathcal{H}[\cdot]$ denotes the entropy of a distribution. However, maximizing this objective is problematic because we cannot analytically integrate the logarithm of the logistic functions in $f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$. These logistic functions have their origin in the exact factors $f_{19,i,j}$ from Figure 1. To solve this problem, we approximate the logistic function with a Gaussian lower bound Jaakkola and Jordan (1997). In particular, we lower bound $x_{i,j}\sigma(a) + (1 - x_{i,j})\sigma(-a)$ in (21) with

$$\tau(a, \xi) = \sigma(\xi) \exp\left\{ -\frac{a(1 - 2x_{i,j}) + \xi}{2} - \lambda(\xi)(a^2 - \xi^2) \right\}, \tag{152}$$

where $\lambda(\xi) = (\sigma(\xi) - 0.5)/(2\xi)$, $\sigma(x) = 1/(1 + \exp(-1))$ is the logistic function and ξ is adjusted to make the lower bound tight at $a = \pm\xi$. When we replace each $f_{19,i,j}$ in (21) with an instantiation of (152) that includes its own variational parameter $\xi_{i,j}$, we obtain the new objective function

$$\mathcal{L}' = \sum_{i=1}^{n}\sum_{j=1}^{d} \alpha_{i,j} + \sum_{i=1}^{n}\sum_{k=1}^{h} \beta_{i,k} + \sum_{j=1}^{d}\sum_{k=1}^{h} \gamma_{j,k} + \kappa + \sum_{(i,j)\notin\mathcal{O}} \sum_{l=1}^{L} \varphi_{i,j,l}\,, \tag{153}$$

where

$$\alpha_{i,j} = \log\sigma(\xi_{i,j}) - \frac{\mu_{i,j}(1 - 2x_{i,j}) + \xi_{i,j}}{2} - \lambda(\xi_{i,j})(\mu_{i,j}^2 + s_{i,j}^2 - \xi_{i,j}^2)\,, \tag{154}$$

$$\varphi_{i,j,l} = -p_{i,j,l} \log p_{i,j,l} + p_{i,j,l} \log p_{i,j,l}^{\setminus MDM}\,, \tag{155}$$

$\beta_{i,d} = \rho(\tilde{e}_{i,d}, \tilde{e}^0_{i,d}, \bar{e}_{i,d}, \bar{e}^0_{i,d})$, $\gamma_{j,d} = \rho(\tilde{f}_{j,d}, \tilde{f}^0_{j,d}, \bar{f}_{j,d}, \bar{f}^0_{j,d})$, $\kappa = \rho(\tilde{z}, \tilde{z}, \bar{z}^0, \bar{z}^0)$, we define ρ as $\rho(a, b, c, d) = -0.5 - 0.5 \log a/b + [(c - d)^2 + a][2b]^{-1}$ and

$$\mu_{i,j} = \begin{cases} \left(\sum_{k=1}^{h} m^e_{i,k} m^f_{j,k}\right) + m^z + \sum_{l=1}^{L} (m^{\lambda^{\text{row}}}_{i,l} + m^{\psi^{\text{col}}}_{j,l}) p_{i,j,l} & : (i,j) \notin \mathcal{O} \\ \left(\sum_{k=1}^{h} m^e_{i,k} m^f_{j,k}\right) + m^z + (m^{\lambda^{\text{row}}}_{i,r_{i,j}} + m^{\psi^{\text{col}}}_{j,r_{i,j}}) & : (i,j) \in \mathcal{O} \end{cases} \tag{156}$$

$$s^2_{i,j} = \begin{cases} \left(\sum_{k=1}^{h} [m^e_{i,k}]^2 v^f_{j,k} + v^e_{i,k}[m^f_{j,k}]^2 + v^e_{i,k} v^f_{j,k}\right) + v^z + \sum_{l=1}^{L} (v^{\lambda^{\text{row}}}_{i,l} + v^{\psi^{\text{col}}}_{j,l}) p_{i,j,l} & : (i,j) \notin \mathcal{O} \\ \left(\sum_{k=1}^{h} [m^e_{i,k}]^2 v^f_{j,k} + v^e_{i,k}[m^f_{j,k}]^2 + v^e_{i,k} v^f_{j,k}\right) + v^z + (v^{\lambda^{\text{row}}}_{i,r_{i,j}} + v^{\psi^{\text{col}}}_{j,r_{i,j}}) & : (i,j) \in \mathcal{O} \end{cases} \tag{157}$$

In the following section we show how to optimize the cost function (153) with respect to the parameters of $\mathcal{Q}$ and the variational parameters $\xi_{i,j}$.

4.2.2 Optimality conditions and batch inference

The cost function (153) is optimized with respect to $\xi_{i,j}$ by setting

$$\xi_{i,j} = \sqrt{\mu_{i,j}^2 + s_{i,j}^2}\,, \tag{158}$$

where $\mu_{i,j}$ and $s_{i,j}^2$ are given by (156) and (157). The optimal value for m^z and v^z are

$$[v^z]^{-1} = [\tilde{z}^0]^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{d} 2\lambda(\xi_{i,j})\,, \tag{159}$$

$$m^z[v^z]^{-1} = \bar{z}^0[\tilde{z}^0]^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{d} [x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - m^z)]\,, \tag{160}$$

The optimal value for $m_{i,l}^{\lambda^{\text{row}}}$ and $v_{i,l}^{\lambda^{\text{row}}}$ are

$$[v_{i,l}^{\lambda^{\text{row}}}]^{-1} = [\tilde{\lambda}^{\text{row}\,0}]^{-1} + \sum_{j=1}^{d} 2\lambda(\xi_{i,j})\hat{p}_{i,j,l}\,, \tag{161}$$

$$m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1} = \bar{\lambda}^{\text{row}\,0}[\tilde{\lambda}^{\text{row}\,0}]^{-1} + \sum_{j=1}^{d}\left[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - m_{i,l}^{\lambda^{\text{row}}}\hat{p}_{i,j,l})\right]\hat{p}_{i,j,l}\,, \tag{162}$$

where $\hat{p}_{i,j,l} = p_{i,j,l}$ if $(i,j) \notin \mathcal{O}$ and $\hat{p}_{i,j,l} = \mathbf{I}[r_{i,j} = l]$, otherwise, and $\mathbf{I}[r_{i,j} = l]$ is the indicator function that takes value 1 when $r_{i,j} = l$ and 0 otherwise. The optimal values for $m_{j,l}^{\psi^{\text{col}}}$ and $v_{j,l}^{\psi^{\text{col}}}$ have similar expressions.

To obtain the optimal values for the mean and variance parameters $m_{i,k}^e$, $v_{i,k}^e$, $m_{i,k}^f$ and $v_{i,k}^f$, for $i = 1,\ldots,n$, $j = 1,\ldots,d$ and $k = 1,\ldots,h$, we proceed as in Section 4.1.14. In particular, we follow Paquet and Koenigstein (2013) and make use of the intermediate approximation $\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}}$ that does not factorize across columns, where

$$\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}} = \left[\prod_{i=1}^{n}\mathcal{N}(\mathbf{e}_i|\hat{\mathbf{m}}_i^e, \hat{\mathbf{V}}_i^e)\right]\left[\prod_{j=1}^{d}\mathcal{N}(\mathbf{f}_j|\hat{\mathbf{m}}_j^f, \hat{\mathbf{V}}_j^f)\right] \tag{163}$$

and $\hat{\mathbf{m}}_i^e$ and $\hat{\mathbf{m}}_j^f$ are the mean vectors for the i-th row of $\mathbf{E}$ and the j-th row of $\mathbf{F}$, respectively, and $\mathbf{V}_i^e$ and $\mathbf{V}_j^f$ are the corresponding covariance matrices. To update the variational parameters in $\mathcal{Q}$ for the i-th row of $\mathbf{E}$ we first equate the gradient of the objective (153) with respect to the parameters $\hat{\mathbf{m}}_i^e$ and $\hat{\mathbf{V}}_i^e$ of the full (not factorized) Gaussian approximation $\hat{\mathcal{Q}}$ to zero. Then we adjust $\mathcal{Q}$ so that $\text{KL}(\mathcal{Q}\|\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}})$ is minimized with respect to the parameters $m_{i,k}^e$ and $v_{i,k}^f$ of $\mathcal{Q}$, for $k = 1,\ldots,h$. This is achieved when, for the i-th row of $\mathbf{E}$, the means of $\mathcal{Q}$ match the means of $\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}}$ and the precisions of $\mathcal{Q}$ match the diagonal precisions of $\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}}$. We update $\hat{\mathcal{Q}}_{\mathbf{E},\mathbf{F}}$ with respect to $\hat{\mathbf{m}}_i^e$ and $\hat{\mathbf{V}}_i^e$ using

$$[\hat{\mathbf{V}}_i^e]^{-1} = \text{diag}(\tilde{e}_{i,k}^0,\ldots,\tilde{e}_{i,k}^0)^{-1} + \sum_{j=1}^{d} 2\lambda(\xi_{i,j})\mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\mathrm{T}}\mathbf{f}_j]\,, \tag{164}$$

$$\begin{aligned}\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1} = {} & (\tilde{e}_{i,1}^0,\ldots,\tilde{e}_{i,h}^0)\text{diag}(\tilde{e}_{i,k}^0,\ldots,\tilde{e}_{i,k}^0)^{-1} + \\ & \sum_{j=1}^{d}\left[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - \sum_{k=1}^{h} m_{i,k}^e m_{j,k}^f)\right]\mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\mathrm{T}}]\,.\end{aligned} \tag{165}$$

After this, we update $\mathcal{Q}$ by setting

$$m_{i,k}^e = [\hat{\mathbf{m}}_i^e]_k\,, \qquad\qquad v_{i,k}^e = 1/\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{k,k}\,, \tag{166}$$

for $k = 1, \ldots, h$. The corresponding parameters for the j-th row of $\mathbf{F}$, that is, $m^f_{j,k}$ and $v^f_{j,k}$, where $k = 1, \ldots, h$, are updated in a similar way.

Unfortunately, there is no analytical solution for the optimal value of $p_{i,j,l}$ as a function of all the other variational parameters. Because of this, when we refine $\tilde{f}_{MDM}$, we do not fully optimize $p_{i,j,l}$ and instead, just fix $p_{i,j,l}$ to be equal to $p^{\setminus MDM}_{i,j,l}$, that is,

$$p_{i,j,l} = p^{\setminus MDM}_{i,j,l}\,, \tag{167}$$

for $l = 1, \ldots, L$. In practice, this approximation is expected to produce reasonably good results since $f_{19,i,j}(r_{i,j}, x_{i,j}, \mathbf{e}_i, \mathbf{f}_j, z, \boldsymbol{\lambda}^{\text{row}}_i, \boldsymbol{\psi}^{\text{col}}_j)$ will most of the times be rather flat as a function of $r_{i,j}$. However, note that at prediction time, we do adjust $p_{i,j,l}$ by using the prediction formula described in Section 4.5.

To optimize $\mathcal{Q}$, we could follow the batch procedure described in Algorithm 1. However, this method is infeasible in practice. For example, to update the variational parameters $\hat{\mathbf{V}}^e_1, \ldots, \hat{\mathbf{V}}^e_n$ just once, we need to examine the whole matrix $\mathbf{X}$, which has dimension $n \times d$. When n and d are massive, the computational cost of that operation is to high. In practice $\mathbf{X}$ is a very sparse matrix since only a very reduced number of the ratings in $\mathbf{R}$ are actually observed. Ideally, we would like to optimize $\mathcal{Q}$ using a method that scales with the number of ones in $\mathbf{X}$, that is, with the number of observed ratings. The following section describes an stochastic optimization method that has this property.

Algorithm 1 Batch method for updating $\mathcal{Q}$ when refining $\tilde{f}_{MDM}$.

Input: Current $\mathcal{Q}$, $\mathcal{Q}^{\setminus MDM}$ and binary matrix $\mathbf{X}$.
for $t = 1$ **to** T **do**
  {Update the variational parameters for the rows and global bias.}
  for $i = 1$ **to** n **do**
    Update $p_{i,1,1}, \ldots, p_{i,d,L}$ using (167)
    Update $\xi_{i,1}, \ldots, \xi_{i,d}$ using (158).
    Update $[\hat{\mathbf{V}}^e_i]^{-1}$ and $\hat{\mathbf{m}}^e_i[\hat{\mathbf{V}}^e_i]^{-1}$ using (164) and (165).
    Update $m^e_{i,1}, \ldots, m^e_{i,h}$ and $v^e_{i,1}, \ldots, v^e_{i,h}$ using (166).
    Update m^z and v^z using (159) and (160).
    Update $m^{\lambda^{\text{row}}}_{i,1}, \ldots, m^{\lambda^{\text{row}}}_{i,h}$ and $v^{\lambda^{\text{row}}}_{i,1}, \ldots, v^{\lambda^{\text{row}}}_{i,h}$ using (161) and (162).
  end for
  {Update the variational parameters for the columns and global bias.}
  for $j = 1$ **to** d **do**
    Perform updates similar to the ones in the previous loop.
  end for
end for
Output: Updated $\mathcal{Q}$.

4.2.3 Stochastic inference in the missing data model

We describe how to minimize the KL divergence between $\mathcal{Q}$ and $f_{MDM}(\boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\mathcal{Q}^{\setminus MDM}(\mathbf{R}^{\neg\mathcal{O}})$ in an efficient way. Our approach is based on the method stochastic variational inference (SVI) Hoffman et al. (2013). SVI works by sub-sampling the data and doing small partial updates of the variational parameters. This allows us to obtain an accurate approximation $\mathcal{Q}$ when we have only examined a reduced fraction of the entries in $\mathbf{X}$. Note that the SVI updates are performed on the natural parameters of distributions in the exponential family Hoffman et al. (2013).

The SVI updates for $[\hat{\mathbf{V}}^e_i]^{-1}$ and $\hat{\mathbf{m}}^e_i[\hat{\mathbf{V}}^e_i]^{-1}$ at time t are given by

$$\{[\hat{\mathbf{V}}^e_i]^{-1}\}_t = (1 - \rho^e_i)\{[\hat{\mathbf{V}}^e_i]^{-1}\}_{t-1} + \rho^e_i\{[\hat{\mathbf{V}}^e_i]^{-1}\}_{\text{noisy}}\,, \tag{168}$$

$$\{\hat{\mathbf{m}}^e_i[\hat{\mathbf{V}}^e_i]^{-1}\}_{t-1} = (1 - \rho^e_i)\{\hat{\mathbf{m}}^e_i[\hat{\mathbf{V}}^e_i]^{-1}\}_{t-1} + \rho^e_i\{\hat{\mathbf{m}}^e_i[\hat{\mathbf{V}}^e_i]^{-1}\}_{\text{noisy}}\,, \tag{169}$$

where the subscripts $t-1$, t and "*noisy*" denote, respectively, the previous value of the variational parameter, the new value of the variational parameter and a noisy estimate of the optimal value for the variational

parameter, that is, the optimal value given by (164) or (165). The parameter $\rho_i^e \in [0,1]$ is a learning rate that should converge to zero as t increases. The value of ρ_i^e can be specified each time that we update $[\hat{\mathbf{V}}_i^e]^{-1}$ and $\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}$ using a Robins-Monroe update schedule (Robbins and Monro, 1951), that is, $\rho_i^e = (1 + \text{updateCounter})^{-\kappa}$ where $\kappa \in (0.5, 1]$ and "updateCounter" is the number of times that $[\hat{\mathbf{V}}_i^e]^{-1}$ and $\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}$ have been updated so far. In our experiments, we fix $\kappa = 0.7$.

We compute the noisy estimates of (164) and (165) by considering all the entries $x_{i,j}$ with value one in the i-th row of $\mathbf{X}$ and then randomly subsampling the same number of entries in that row with value zero. In particular, we have that

$$\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}} = \text{diag}(\tilde{e}_{i,k}^0, \ldots, \tilde{e}_{i,k}^0)^{-1} + \sum_{j \in \mathcal{I}_i^{\text{row},1}} 2\lambda(\xi_{i,j}) \mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\text{T}} \mathbf{f}_j] + \eta_i^e \sum_{j \in \mathcal{I}_i^{\text{row},0}} 2\lambda(\xi_{i,j}) \mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\text{T}} \mathbf{f}_j] \,, \tag{170}$$

$$\begin{aligned}\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}} = (\bar{e}_{i,1}^0, \ldots, \bar{e}_{i,h}^0)\text{diag}(\tilde{e}_{i,k}^0, \ldots, \tilde{e}_{i,k}^0)^{-1} + \\ \sum_{j \in \mathcal{I}_i^{\text{row},1}} \left[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - \sum_{k=1}^{h} m_{i,k}^e m_{j,k}^f) \right] \mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\text{T}}] + \\ \eta_i^e \sum_{j \in \mathcal{I}_i^{\text{row},0}} \left[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - \sum_{k=1}^{h} m_{i,k}^e m_{j,k}^f) \right] \mathbf{E}_{\mathcal{Q}}[\mathbf{f}_j^{\text{T}}] \,,\end{aligned} \tag{171}$$

where $\mathcal{I}_i^{\text{row},1}$ is a deterministic set with the column indexes of the entries in the i-th row of $\mathbf{X}$ that take value one, that is, $\mathcal{I}_i^{\text{row},1} = \{j : j \in \{1, \ldots, d\} \text{ and } (i,j) \in \mathcal{O}\}$, and $\mathcal{I}_i^{\text{row},0}$ is a random set with the column indexes of some entries in the i-th row of $\mathbf{X}$ that take value zero. In particular, $\mathcal{I}_i^{\text{row},0}$ satisfies that a) for any $j \in \mathcal{I}_i^{\text{row},0}$ we have that $x_{i,j} = 0$, b) the size of $\mathcal{I}_i^{\text{row},0}$ is the number of variables $x_{i,j}$ with value one in the i-th row of $\mathbf{X}$ and c) all the elements in $\mathcal{I}_i^{\text{row},0}$ are chosen randomly from the set $\{j : j \in \{1, \ldots, d\} \text{ and } (i,j) \notin \mathcal{O}\}$ with equal probability and with replacement. Note that $|\mathcal{I}_i^{\text{row},0}| = |\mathcal{I}_i^{\text{row},1}|$. Finally, the constant η_i^e takes value $\eta_i^e = (d - |\mathcal{I}_i^{\text{row},0}|)/|\mathcal{I}_i^{\text{row},0}|$. This scaling constant guarantees that the expectations of $\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ are the same as the exact optimal values given by (164) and (165), respectively.

The stochastic update for $[\hat{\mathbf{V}}_j^f]^{-1}$ and $\hat{\mathbf{m}}_j^f[\hat{\mathbf{V}}_j^f]^{-1}$ at time t are computed in a similar way.

We will update $[v^z]^{-1}$ and $m^z[v^z]^{-1}$ each time that we access the value of an entry in the binary matrix $\mathbf{X}$. The SVI updates for $[v^z]^{-1}$ and $m^z[v^z]^{-1}$ are given by

$$\{[v^z]^{-1}\}_t = (1 - \rho^z)\{[v^z]^{-1}\}_{t-1} + \rho^z \{[v^z]^{-1}\}_{\text{noisy}} \,, \tag{172}$$

$$\{m^z[v^z]^{-1}\}_t = (1 - \rho^z)\{m^z[v^z]^{-1}\}_{t-1} + \rho^z \{m^z[v^z]^{-1 m^z[v^z]^{-1}}\}_{\text{noisy}} \,, \tag{173}$$

where ρ^z is specified in a similar way as ρ_i^e and $\{[v^z]^{-1}\}_{\text{noisy}}$ and $\{m^z[v^z]^{-1 m^z[v^z]^{-1}}\}_{\text{noisy}}$ are given by

$$\{[v^z]^{-1}\}_{\text{noisy}} = [\tilde{z}^0]^{-1} + \eta^z 2\lambda(\xi_{i,j}) \,, \tag{174}$$

$$\{m^z[v^z]^{-1}\}_{\text{noisy}} = \tilde{z}^0[\tilde{z}^0]^{-1} + \eta^z \left[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - m^z) \right] \,, \tag{175}$$

$x_{i,j}$ is the entry of $\mathbf{X}$ accessed during the update of $[v^z]^{-1}$ and $m^z[v^z]^{-1}$ and the constant η^z takes value

$$\eta^z = \begin{cases} 2|\mathcal{O}| & : x_{i,j} = 1 \\ 4|\mathcal{O}| / \left[|\mathcal{I}_i^{\text{row},0}|/(d - |\mathcal{I}_i^{\text{row},0}|) + |\mathcal{I}_j^{\text{col},0}|/(n - |\mathcal{I}_j^{\text{col},0}|) \right] & : (i,j) \notin \mathcal{O} \end{cases} . \tag{176}$$

This constant guarantees that the expectations of $\{[v^z]^{-1}\}_{\text{noisy}}$ and $\{m^z[v^z]^{-1 m^z[v^z]^{-1}}\}_{\text{noisy}}$ are the same as their optimal values given by (159) and (160).

The SVI updates for $m_{i,l}^{\lambda^{\text{row}}}$ and $v_{i,l}^{\lambda^{\text{row}}}$, for $l = 1, \ldots, L$, are given by

$$\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_t = (1 - \rho_i^e)\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{t-1} + \rho_i^e \{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}} \,, \tag{177}$$

Algorithm 2 Stochastic method for updating $\mathcal{Q}$ when refining $\tilde{f}_{MDM}$.

Input: Current $\mathcal{Q}$, $\mathcal{Q}^{\setminus MDM}$ and binary matrix $\mathbf{X}$.
for $t = 1$ **to** T **do**
  {Update the variational parameters for the rows and global bias.}
  for $i = 1$ **to** n **do**
    Initialize $\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$, $\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ and $\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$ and $\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$, for $l = 1, \ldots, L$, with the contribution from the prior.
    Generate set of indexes $\mathcal{I}_i^{\text{row},1}$.
    for $j \in \mathcal{I}_i^{\text{row},1}$ **do**
      Update $p_{i,j,1}, \ldots, p_{i,j,L}$ using (167).
      Compute $\mu_{i,j}$ and $s_{i,j}^2$ using (156) and (157).
      Update $\xi_{i,j}$ using (158).
      Update $\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ using (170) and (171).
      Update $\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$ and $\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$, for $l = 1, \ldots, L$, using (179) and (180).
      Update m^z and v^z using first (174), (175) and then (172) and (173).
    end for
    Generate set of indexes $\mathcal{I}_j^{\text{row},0}$.
    for $j \in \mathcal{I}_i^{\text{row},0}$ **do**
      Update $p_{i,j,1}, \ldots, p_{i,j,L}$ using (167).
      Compute $\mu_{i,j}$ and $s_{i,j}^2$ using (156) and (157).
      Update $\xi_{i,j}$ using (158).
      Update $\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ using (170) and (171).
      Update $\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$ and $\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$, for $l = 1, \ldots, L$, using (179) and (180).
      Update m^z and v^z using first (174), (175) and then (172) and (173).
    end for
    Update $[\hat{\mathbf{V}}_i^e]^{-1}$ and $\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}$ using (168) and (169).
    Update $m_{i,1}^e, \ldots, m_{i,h}^e$ and $v_{i,1}^e, \ldots, v_{i,h}^e$ using (166).
    Update $v_{i,l}^{\lambda^{\text{row}}}$ and $m_{i,l}^{\lambda^{\text{row}}}$, for $l = 1, \ldots, L$, using (177) and (178).
    {Update the variational parameters for the columns and global bias.}
  end for
  for $j = 1$ **to** d **do**
    Perform updates similar to the ones in the previous loop.
  end for
end for
Output: Updated $\mathcal{Q}$.

$$\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_t = (1-\rho_i^e)\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{t-1} + \rho_i^e\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}\,, \tag{178}$$

where we are using the same learning rate ρ_i^e as for the stochastic updates of $[\hat{\mathbf{V}}_i^e]^{-1}$ and $\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}$ since we will be updating all these parameters at the same time. We compute $\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$ and $\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}}$ in a similar way as $\{[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^e]^{-1}\}_{\text{noisy}}$, that is,

$$\{[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}} = [\tilde{\lambda}^{\text{row}\,0}]^{-1} + \sum_{j\in\mathcal{I}_i^{\text{row},1}} \lambda(\xi_{i,j})\hat{p}_{i,j,l} + \eta_i^e \sum_{j\in\mathcal{I}_i^{\text{row},0}} 2\lambda(\xi_{i,j})\hat{p}_{i,j,l}\,, \tag{179}$$

$$\begin{aligned}\{m_{i,l}^{\lambda^{\text{row}}}[v_{i,l}^{\lambda^{\text{row}}}]^{-1}\}_{\text{noisy}} = \bar{\lambda}^{\text{row}\,0}[\tilde{\lambda}^{\text{row}\,0}]^{-1} + \sum_{j\in\mathcal{I}_i^{\text{row},1}} \Big[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - m_{i,l}^{\lambda^{\text{row}}}\hat{p}_{i,j,l})\Big]\hat{p}_{i,j,l} + \\ \eta_i^e \sum_{j\in\mathcal{I}_i^{\text{row},0}} \Big[x_{i,j} - 0.5 - 2\lambda(\xi_{i,j})(\mu_{i,j} - m_{i,l}^{\lambda^{\text{row}}}\hat{p}_{i,j,l})\Big]\hat{p}_{i,j,l}\,,\end{aligned} \tag{180}$$

The SVI updates for $m_{j,l}^{\psi^{\text{col}}}$ and $v_{j,l}^{\psi^{\text{col}}}$, for $l = 1, \ldots, L$, are computed in a similar way.

Finally, note that computing the optimal update for $p_{i,j,l}$ without using numerical methods is not feasible in practice. As mentioned in the previous section, we do not fully optimize $p_{i,j,l}$ and instead, just fix $p_{i,j,l}$ to be equal to $p_{i,j,l}^{\setminus MDM}$, which is expected to produce reasonably good results.

Algorithm 2 shows our stochastic method for adjusting $\mathcal{Q}$. The computational cost of this method scales linearly with respect to $|\mathcal{O}|$. This is a significant gain when n and d are very large but the number of observed entries from $\mathbf{R}$ is small. In practice, we do not keep in memory all the variables $\xi_{i,j}$, $\mu_{i,j}$, $s_{i,j}^2$ or $p_{i,j,l}$. We compute them when needed and then, we discard them afterwards.

4.2.4 The predictive distribution of the missing data model

Once $\mathcal{Q}$ as been updated to incorporate the contribution of the extended factor for the missing data model (MDM) $f_{MDM}(\mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$, that is, (149), we can use $\mathcal{Q}$ to compute the approximation $\tilde{p}_{i,j,l}^{\mathrm{MDM}}(x_{i,j})$ to the posterior probability $p_{i,j,l}^{\mathrm{MDM}}(x_{i,j})$ that the entry in the i-th row and j-th column of the rating matrix $\mathbf{R}$ has taken value l, as a function of $x_{i,j}$. This posterior probability is computed ignoring the contribution from the complete data model. In particular, we have that

$$\begin{aligned}\tilde{p}_{i,j,l}^{\mathrm{MDM}}(x_{i,j}) &\propto \int f_{19,i,j}(r_{i,j} = l, x_{i,j}, \mathbf{e}_i, \mathbf{f}_j, z, \boldsymbol{\lambda}_i^{\mathrm{row}}, \boldsymbol{\psi}_j^{\mathrm{col}}) \mathcal{Q}(\mathbf{\Omega})\, d\mathbf{\Omega} \\ &\propto \sigma\left\{(2x_{i,j} - 1)\left[\mathbf{e}_i \mathbf{f}_j^{\mathrm{T}} + z + \lambda_{i,l}^{\mathrm{row}} + \psi_{j,l}^{\mathrm{col}}\right]\right\} \mathcal{Q}(\mathbf{\Omega})\, d\mathbf{\Omega}\,. \end{aligned} \tag{181}$$

We can approximate the integral above by replacing the logistic function with a rescaled probit function that has the same slope at the origin as the logistic function $\sigma(\cdot)$ MacKay (1992). This leads to

$$\tilde{p}_{i,j,l}^{\mathrm{MDM}}(x_{i,j}) \propto \sigma\{(2x_{i,j} - 1)\varphi(s_{i,j,l}^2)\mu_{i,j,l}\}\,, \tag{182}$$

where $\varphi(x) = (1 + \pi x/8)^{-1/2}$ and

$$\mu_{i,j,l} = \left(\sum_{k=1}^{h} m_{i,k}^e m_{j,k}^f\right) + m^z + m_{i,l}^{\lambda^{\mathrm{row}}} + m_{j,l}^{\psi^{\mathrm{col}}}\,, \tag{183}$$

$$s_{i,j,l}^2 = \left(\sum_{k=1}^{h} [m_{i,k}^e]^2 v_{j,k}^f + v_{i,k}^e [m_{j,k}^f]^2 + v_{i,k}^e v_{j,k}^f\right) + v^z + v_{i,l}^{\lambda^{\mathrm{row}}} + v_{j,l}^{\psi^{\mathrm{col}}}\,. \tag{184}$$

4.3 Approximate Inference in the complete data model with MNAR data

With MNAR data, approximate inference in the complete data model is challenging. The reason for this is that we now have to deal with the exact factors $f_{11,i,j}$, $f_{12,i,j}$ and $f_{13,i,j,k}$ for which $(i,j) \notin \mathcal{O}$. Ideally, we would approximate each of these exact factors with an approximate factor that would be iteratively refined by EP. However, keeping all those approximate factors in memory is infeasible in practice since their number scales as $nd - |\mathcal{O}|$, where n and d can be very large. Instead, we will just generate an approximation to these exact factors that will be computed when needed and then discarded afterwards. Whenever we have to find an approximation to any of the $f_{11,i,j}$, $f_{12,i,j}$ and $f_{13,i,j,k}$ with $(i,j) \notin \mathcal{O}$, we proceed as follows. First, we approximate $f_{11,i,j}$ with a Gaussian factor on $c_{i,j}$. Given this approximation, we then approximate $f_{12,i,j}$ with another Gaussian factor on $a_{i,j}$. After this, we approximate $f_{13,i,j,k}$ with an additional Gaussian factor on $a_{i,j}$ and finally, we approximate again $f_{13,i,j,k}$ with a Gaussian factor on $c_{i,j}$. The following sections describe how to do this.

4.3.1 Approximating $f_{11,i,j}$ as a function of $c_{i,j}$

Recall that $f_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j) = \delta(c_{i,j} - \mathbf{u}_i \mathbf{v}_j^{\mathrm{T}})$. We can easily approximate this exact factor as a function $c_{i,j}$ with an approximate factor $\tilde{f}_{11,i,j}$. For this, we only have to compute the marginal mean and variance of $c_{i,j}$ with respect to $f_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j)\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$. Note that we are using $\mathcal{Q}$ and not $\mathcal{Q}^{\setminus 11,i,j}$ (the ratio

between $\mathcal{Q}$ and the approximation to $f_{11,i,j}$) to compute the aforementioned marginal. The reason for this is that we do not store any of the $\tilde{f}_{11,i,j}$ in memory and furthermore, we never include the contribution of any of these approximate factors, as a function of $c_{i,j}$, into $\mathcal{Q}$, where $(i,j) \notin \mathcal{O}$. The EP update for the parameters of the Gaussian factor for $c_{i,j}$ in $\tilde{f}_{11,i,j}$ is then given by the mean and variance of $c_{i,j}$ with respect to $f_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j)\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})$ (because $\mathcal{Q}$ is uniform on $c_{i,j}$ for $(i,j) \notin \mathcal{O}$ for the reasons mentioned above), that is,

$$[\tilde{m}_{i,j}^{c,11,i,j}]^{\text{new}} = \sum_{k=1}^{h} m_{i,k}^{u} m_{j,k}^{v}\,, \qquad [\tilde{v}_{i,j}^{c,11,i,j}]^{\text{new}} = \sum_{k=1}^{h} [m_{i,k}^{u}]^2 v_{j,k}^{v} + v_{i,k}^{u}[m_{j,k}^{v}]^2 + v_{i,k}^{u} v_{j,k}^{v}\,. \tag{185}$$

4.3.2 Approximating $f_{12,i,j}$ as a function of $a_{i,j}$

Recall that $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})$. Furthermore, recall that $\mathcal{Q}$ is uniform in $a_{i,j}$ for $(i,j) \notin \mathcal{O}$ because we do not store any of the $\tilde{f}_{12,i,j}$ or $\tilde{f}_{13,i,j}$ with $(i,j) \notin \mathcal{O}$ in memory and never include the contribution of these approximate factors into $\mathcal{Q}$. To update $\tilde{f}_{12,i,j}$ we only have to compute the mean and variance of the marginal of $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j)$ with respect to $a_{i,j}$. This marginal is given by

$$\begin{aligned}
&\sum_{\mathbf{R}^{\neg\mathcal{O}}} \int f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{11,i,j}(c_{i,j}, \mathbf{u}_i, \mathbf{v}_j)\, dc_{i,j}\, d\mathbf{\Theta}\, d\mathbf{\Omega} = \\
&= \int \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})\mathcal{N}(c_{i,j}|\tilde{m}_{i,j}^{c,11,i,j}, \tilde{v}_{i,j}^{c,11,i,j}) \\
&\qquad \mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})\mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})\, dc_{i,j}, d\gamma_i^{\text{row}}\, d\gamma_j^{\text{col}} \\
&\approx \mathcal{N}(a_{i,j}|\tilde{m}_{i,j}^{c,11,i,j}, \tilde{v}_{i,j}^{c,11,i,j} + b_i^{\gamma^{\text{row}}} b_j^{\gamma^{\text{col}}}/[(a_i^{\gamma^{\text{row}}}+1)(a_j^{\gamma^{\text{col}}}+1)])\,,
\end{aligned} \tag{186}$$

where in (186) we have approximated $\text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})$ and $\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})$ with point probability masses located at the modes of these factors.

The resulting update for $\tilde{f}_{12,i,j}$ is then

$$[\tilde{m}_{i,j}^{a,12,i,j}]^{\text{new}} = \tilde{m}_{i,j}^{c,11,i,j}\,, \qquad [\tilde{v}_{i,j}^{a,12,i,j}]^{\text{new}} = \tilde{v}_{i,j}^{c,11,i,j} + b_i^{\gamma^{\text{row}}} b_j^{\gamma^{\text{col}}}/[(a_i^{\gamma^{\text{row}}}+1)(a_j^{\gamma^{\text{col}}}+1)]\,. \tag{187}$$

4.3.3 Approximating $f_{13,i,j}$ as a function of $a_{i,j}$

We now approximate the extended factor $f_{13,i,j}(r_{i,j}, a_{i,j}, \mathbf{b}_j) = \prod_{k=1}^{L-1} f_{13,i,j,k}(r_{i,j}, a_{i,j}, b_{j,k})$ as a function of $a_{i,j}$, when $r_{i,j}$ and $\mathbf{b}_j$ are marginalized out. For this, we need to match the marginal mean and variance of $a_{i,j}$ between $f_{13,i,j}(r_{i,j}, a_{i,j}, \mathbf{b}_j)\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{12,i,j}(a_{i,j})$ and $\tilde{f}_{13,i,j}(r_{i,j}, a_{i,j}, \mathbf{b}_j)\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{12,i,j}(a_{i,j})$. The first step is to compute the normalization constant of $f_{13,i,j}(r_{i,j}, a_{i,j}, \mathbf{b}_j)\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{12,i,j}(a_{i,j})$. Let Z be such normalization constant. Then, it can be shown that

$$\begin{aligned}
Z &= \sum_{r_{i,j}} \int \prod_{k=1}^{L-1} \Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right] \left[\prod_{l=1}^{L} [\tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})]^{\mathbf{I}[r_{i,j}=l]}\right] \\
&\qquad \left[\prod_{k=1}^{L-1} \mathcal{N}(b_{i,k}|m_{i,k}^{b}, v_{i,k}^{b})\right] \mathcal{N}(a_{i,j}|\tilde{m}_{i,j}^{a,12,i,j}, \tilde{v}_{i,j}^{a,12,i,j})\, da_{i,j}\, dc_{i,j}\, d\mathbf{b}_j \\
&= \sum_{l=1}^{L} \tilde{p}_{i,j,l}^{\text{MDM}}(x_{i,j})\left[\Phi(\alpha_{l-1}) - \Phi(\alpha_l)\right]\,,
\end{aligned} \tag{188}$$

where Φ is the standard Gaussian cdf, $\alpha_l = (\tilde{m}_{i,j}^{a,12,i,j} - m_{i,l}^b)(\tilde{v}_{i,j}^{a,12,i,j} + v_{i,l}^b)^{-0.5}$, we define $\alpha_0 = \infty$ and $\alpha_L = -\infty$ and $\tilde{p}_{i,j,1}^{\text{MDM}}(x_{i,j}), \ldots, \tilde{p}_{i,j,L}^{\text{MDM}}(x_{i,j})$ are the probabilistic predictions for $r_{i,j}$ given by the missing data model, see Section 4.2.4.

The EP updates for $\tilde{f}_{13,i,j}(r_{i,j}, a_{i,j}, \mathbf{b}_j)$ are then given by

$$[\tilde{v}_{i,j}^{a,13,i,j}]^{\text{new}} = -\left[\frac{d^2 \log Z}{d[\tilde{m}_{i,j}^{a,12,i,j}]^2}\right]^{-1} - \tilde{v}_{i,j}^{a,12,i,j}\,, \tag{189}$$

$$[\tilde{m}_{i,j}^{a,13,i,j}]^{\text{new}} = \tilde{m}_{i,j}^{a,12,i,j} - \left[\frac{d \log Z}{d\tilde{m}_{i,j}^{a,12,i,j}}\right]\left[\frac{d^2 \log Z}{d[\tilde{m}_{i,j}^{a,12,i,j}]^2}\right]^{-1}\,. \tag{190}$$

where

$$\frac{d \log Z}{d\tilde{m}_{i,j}^{a,12,i,j}} = Z^{-1} \sum_{l=1}^{L} \tilde{p}_{i,j,l}^{\text{MDM}} \left[\frac{\phi(\alpha_{l-1})}{\sqrt{\beta_{l-1}}} - \frac{\phi(\alpha_l)}{\sqrt{\beta_l}}\right], \tag{191}$$

$$\frac{d^2 \log Z}{d[\tilde{m}_{i,j}^{a,12,i,j}]^2} = Z^{-1} \sum_{l=1}^{L} \tilde{p}_{i,j,l}^{\text{MDM}} \left[\frac{\phi(\alpha_{l-1})\alpha_{l-1}}{\beta_{l-1}} - \frac{\phi(\alpha_l)\alpha_l}{\beta_l}\right], \tag{192}$$

and $\beta_0 = 1$, $\beta_L = 1$ and $\beta_l = \tilde{v}_{i,j}^{a,12,i,j} + v_{i,l}^b$, for $l = 1, \ldots, L-1$.

4.3.4 Approximating $f_{12,i,j}$ as a function of $c_{i,j}$

Recall that $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}}) = \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})$. To update $\tilde{f}_{12,i,j}$ we only have to compute the mean and variance of the marginal of $f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{13,i,j}(a_{i,j})$ with respect to $a_{i,j}$. This marginal is given by

$$\begin{aligned}
&\sum_{\mathbf{R}^{\neg\mathcal{O}}} \int f_{12,i,j}(a_{i,j}, c_{i,j}, \gamma_i^{\text{row}}, \gamma_j^{\text{col}})\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{R}^{\neg\mathcal{O}})\tilde{f}_{13,i,j}(a_{i,j})\, da_{i,j}\, d\boldsymbol{\Theta}\, d\boldsymbol{\Omega} = \\
&= \int \mathcal{N}(a_{i,j}|c_{i,j}, \gamma_i^{\text{row}}\gamma_j^{\text{col}})\mathcal{N}(a_{i,j}|\tilde{m}_{i,j}^{a,13,i,j}, \tilde{v}_{i,j}^{a,13,i,j}) \\
&\qquad \mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})\mathcal{IG}(\gamma_j^{\text{row}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})\, da_{i,j}, d\gamma_i^{\text{row}}\, d\gamma_j^{\text{col}} \\
&\approx \mathcal{N}(c_{i,j}|\tilde{m}_{i,j}^{a,13,i,j}, \tilde{v}_{i,j}^{a,13,i,j} + b_i^{\gamma^{\text{row}}} b_j^{\gamma^{\text{col}}}/[(a_i^{\gamma^{\text{row}}} + 1)(a_j^{\gamma^{\text{col}}} + 1)])\,,
\end{aligned} \tag{193}$$

where in (193) we have approximated $\text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})$ and $\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})$ with point probability masses located at the modes of these factors.

The resulting update for $\tilde{f}_{12,i,j}$ is then

$$[\tilde{m}_{i,j}^{c,12,i,j}]^{\text{new}} = \tilde{m}_{i,j}^{a,13,i,j}\,, \qquad [\tilde{v}_{i,j}^{c,12,i,j}]^{\text{new}} = \tilde{v}_{i,j}^{a,13,i,j} + b_i^{\gamma^{\text{row}}} b_j^{\gamma^{\text{col}}}/[(a_i^{\gamma^{\text{row}}} + 1)(a_j^{\gamma^{\text{col}}} + 1)]\,. \tag{194}$$

4.3.5 Batch minimization of the reversed KL divergence when refining $\tilde{f}_{11}$

Recall that in Section 4.1.11 we had to minimize $\text{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$. The procedure for doing this with MAR data was described in Section 4.1.14. With MNAR data, we still have to perform the same operation. However $\mathcal{S}$ now includes a product of factors over all possible values of i and j and not only over those $(i,j) \in \mathcal{O}$. In particular, we have that $\mathcal{S}$ is now

$$\mathcal{S}(\mathbf{U}, \mathbf{V}) = \left[\prod_{i=1}^{n}\prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_i^{\text{T}}\ \mathbf{v}_j|m_{i,j}^{c,\setminus 11}, v_{i,j}^{c,\setminus 11})\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u,\setminus 11}, v_{i,k}^{u,\setminus 11})\right]$$

$$\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 11},v_{j,k}^{v,\backslash 11})\right], \tag{195}$$

where $m_{i,j}^{c,\backslash 11}$ and $v_{i,j}^{c,\backslash 11}$, for $(i,j)\notin\mathcal{O}$, are now given by

$$m_{i,j}^{c,\backslash 11}=\tilde{m}_{i,j}^{c,12,i,j}\,,\qquad\qquad v_{i,j}^{c,\backslash 11}=\tilde{v}_{i,j}^{c,12,i,j}\,. \tag{196}$$

and $\tilde{m}_{i,j}^{c,12,i,j}$ and $\tilde{v}_{i,j}^{c,12,i,j}$ can be computed when needed following the steps described in sections 4.3.1, 4.3.2 4.3.3 and 4.3.4.

To minimize $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$ with MNAR data, that is, when $\mathcal{S}$ is given by (195) and not (137), we again follow Paquet and Koenigstein (2013) and make use of the intermediate approximation $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ that does not factorize across columns, where

$$\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}=\left[\prod_{i=1}^{n}\mathcal{N}(\mathbf{u}_i|\hat{\mathbf{m}}_i^u,\hat{\mathbf{V}}_i^u)\right]\left[\prod_{j=1}^{d}\mathcal{N}(\mathbf{v}_j|\hat{\mathbf{m}}_j^v,\hat{\mathbf{V}}_j^v)\right] \tag{197}$$

and $\hat{\mathbf{m}}_i^u$ and $\hat{\mathbf{m}}_j^v$ are the mean vectors for the i-th row of $\mathbf{U}$ and the j-th row of $\mathbf{V}$, respectively, and $\mathbf{V}_i^u$ and $\mathbf{V}_j^v$ are the corresponding covariance matrices. To update the variational parameters in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the i-th row of $\mathbf{U}$ we first equate the gradient of $\mathrm{KL}(\hat{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}\|\mathcal{S})$ with respect to the parameters $\hat{\mathbf{m}}_i^u$ and $\hat{\mathbf{V}}_i^u$ of the full (not factorized) Gaussian approximation $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ to zero. Then we adjust $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ so that $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}})$ is minimized with respect to the parameters $m_{i,k}^u$ and $v_{i,k}^u$ of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, for $k=1,\ldots,h$. This is achieved when, for the i-th row of $\mathbf{U}$, the means of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ match the means of $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ and the precisions of $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ match the diagonal precisions of $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$. We update $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ with respect to $\hat{\mathbf{m}}_i^u$ and $\hat{\mathbf{V}}_i^u$ by setting

$$[\hat{\mathbf{V}}_i^u]^{-1}=\mathrm{diag}(v_{i,1}^{u,\backslash 11},\ldots,v_{i,h}^{u,\backslash 11})^{-1}+\sum_{j=1}^{d}\frac{\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j]}{v_{i,j}^{c,\backslash 11}}\,, \tag{198}$$

$$\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}=(m_{i,1}^{u,\backslash 11},\ldots,m_{i,h}^{u,\backslash 11})\mathrm{diag}(v_{i,1}^{u,\backslash 11},\ldots,v_{i,h}^{u,\backslash 11})^{-1}+\sum_{j=1}^{d}\frac{m_{i,j}^{c,\backslash 11}\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}]}{v_{i,j}^{c,\backslash 11}}\,. \tag{199}$$

After this, we update $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ by setting

$$m_{i,k}^u=[\hat{\mathbf{m}}_i^u]_k\,,\qquad\qquad v_{i,k}^u=1/\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{k,k}\,, \tag{200}$$

for $k=1,\ldots,h$. The corresponding parameters for the j-th row of $\mathbf{V}$, that is, $m_{j,k}^v$ and $v_{j,k}^v$, where $k=1,\ldots,h$, can be updated in a similar way.

Note, however, that the update equations (198) and (199) include now a sum over $j=1,\ldots,d$, while the corresponding update equations in Section (4.1.14) included only a sum over only the j such that $(i,j)\in\mathcal{O}$. This last set of column indexes is usually much smaller than d since the number of observed entries in the rating matrix is often very small. The consequence is that minimizing $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$ with MNAR data has a cost that scales as nd instead of as $|\mathcal{O}|$. When n and d are very large, minimizing the divergence with MNAR data using the batch update equations (198) and (199) is infeasible in practice. Ideally, we would like to minimize $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$ using a method that scales with $|\mathcal{O}|$, that is, with the number of observed ratings. In the following section we describe an stochastic optimization method that has this property. In Section 4.2.3 we used a similar approach to approximate the exact factors in Figure 1 for the missing data model.

4.3.6 Stochastic minimization of the reversed KL divergence when refining $\tilde{f}_{11}$

We now describe how to minimize $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$ in an efficient way. Our approach is based on the method stochastic variational inference (SVI) Hoffman et al. (2013). SVI works by sub-sampling the data and doing

small partial updates of the variational parameters. This allows us to obtain an accurate approximation $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ when we have only examined a reduced fraction of the nd factors $\mathcal{N}(\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j|m_{i,j}^{c,\backslash 11}, v_{i,j}^{c,\backslash 11})$ that are included in $\mathcal{S}$ in (195). The SVI updates for $[\hat{\mathbf{V}}_i^u]^{-1}$ and $\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}$ at time t are given by

$$\{[\hat{\mathbf{V}}_i^u]^{-1}\}_t = (1-\rho_i^u)\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{t-1} + \rho_i^u\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}\,, \tag{201}$$

$$\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_t = (1-\rho_i^u)\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{t-1} + \rho_i^u\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}\,, \tag{202}$$

where the subscripts $t-1$, t and "*noisy*" denote, respectively, the previous value of the variational parameter, the new value of the variational parameter and a noisy estimate of the optimal value for the variational parameter, that is, the optimal value given by (198) or (199). The parameter $\rho_i^u \in [0,1]$ is a learning rate that should converge to zero as t increases. The value of ρ_i^u can be specified each time that we update $[\hat{\mathbf{V}}_i^u]^{-1}$ and $\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}$ using a Robins-Monroe update schedule (Robbins and Monro, 1951), that is, $\rho_i^u = (1+\text{updateCounter})^{-\kappa}$ where $\kappa \in (0.5, 1]$ and "updateCounter" is the number of times that $[\hat{\mathbf{V}}_i^u]^{-1}$ and $\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}$ have been updated so far. In our experiments, we fix $\kappa = 0.7$.

We compute the estimates $\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ of (198) and (199) by considering only a reduced fraction of the factors $\mathcal{N}(\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j|m_{i,j}^{c,\backslash 11}, v_{i,j}^{c,\backslash 11})$ $(j = 1,\ldots,d)$, instead of all of them, as it is actually done in (198) and (199). In particular, for each value of i, we only consider the factors $\mathcal{N}(\mathbf{u}_i^{\mathrm{T}}\,\mathbf{v}_j|m_{i,j}^{c,\backslash 11}, v_{i,j}^{c,\backslash 11})$ such that $(i,j) \in \mathcal{O}$ and a random subset of the factors with indexes (i,j) such that $(i,j) \notin \mathcal{O}$, where the two subsets of factors have the same size. In particular, we have that

$$\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}} = \text{diag}(v_{i,1}^{u,\backslash 11},\ldots,v_{i,h}^{u,\backslash 11})^{-1} + \sum_{j\in\mathcal{I}_i^{\text{row},1}} \frac{\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j]}{v_{i,j}^{c,\backslash 11}} + \eta_i^u \sum_{j\in\mathcal{I}_i^{\text{row},0}} \frac{\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j]}{v_{i,j}^{c,\backslash 11}}\,, \tag{203}$$

$$\begin{aligned}\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}} = {}& (m_{i,1}^{u,\backslash 11},\ldots,m_{i,h}^{u,\backslash 11})\text{diag}(v_{i,1}^{u,\backslash 11},\ldots,v_{i,h}^{u,\backslash 11})^{-1} + \\ & \sum_{j\in\mathcal{I}_i^{\text{row},1}} \frac{m_{i,j}^{c,\backslash 11}\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}]}{v_{i,j}^{c,\backslash 11}} + \eta_i^u \sum_{j\in\mathcal{I}_i^{\text{row},0}} \frac{m_{i,j}^{c,\backslash 11}\mathbf{E}_{\mathcal{Q}_{\mathbf{U},\mathbf{V}}}[\mathbf{v}_j^{\mathrm{T}}]}{v_{i,j}^{c,\backslash 11}}\,,\end{aligned} \tag{204}$$

where $\mathcal{I}_i^{\text{row},1}$ is a deterministic set with the column indexes of the observed entries in the i-th row of $\mathbf{R}$, that is, $\mathcal{I}_i^{\text{row},1} = \{j : j \in \{1,\ldots,d\} \text{ and } (i,j) \in \mathcal{O}\}$, and $\mathcal{I}_i^{\text{row},0}$ is a random set with the column indexes of some entries in the i-th row of $\mathbf{R}$ that are not observed. In particular, $\mathcal{I}_i^{\text{row},0}$ satisfies that a) for any $j \in \mathcal{I}_i^{\text{row},0}$ we have that $x_{i,j} = 0$, b) the size of $\mathcal{I}_i^{\text{row},0}$ is the number of variables $x_{i,j}$ with value one in the i-th row of $\mathbf{X}$ and c) all the elements in $\mathcal{I}_i^{\text{row},0}$ are chosen randomly from the set $\{j : j \in \{1,\ldots,d\} \text{ and } (i,j) \notin \mathcal{O}\}$ with equal probability and with replacement. Note that $|\mathcal{I}_i^{\text{row},0}| = |\mathcal{I}_i^{\text{row},1}|$. Finally, the constant η_i^u takes value $\eta_i^u = (d - |\mathcal{I}_i^{\text{row},0}|)/|\mathcal{I}_i^{\text{row},0}|$. This scaling constant guarantees that the expectations of $\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ are the same as the exact optimal values given by (198) and (199), respectively.

Once we have updated $[\hat{\mathbf{V}}_i^u]^{-1}$ and $\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}$ using (203) and (204), and (201) and (202), we update $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ by setting

$$m_{i,k}^u = [\hat{\mathbf{m}}_i^u]_k\,, \qquad\qquad v_{i,k}^u = 1/\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{k,k}\,, \tag{205}$$

for $k = 1,\ldots,h$.

The corresponding stochastic updates for $[\hat{\mathbf{V}}_j^v]^{-1}$ and $\hat{\mathbf{m}}_j^v[\hat{\mathbf{V}}_j^v]^{-1}$ in $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ and the $m_{i,k}^v$ and $v_{i,k}^v$ with $k = 1,\ldots,h$ in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, are computed in a similar way.

Algorithm 3 shows our stochastic method for minimizing $\text{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}}\|\mathcal{S})$. The computational cost of this method scales linearly with respect to $|\mathcal{O}|$. This is a significant gain when n and d are very large but the number $|\mathcal{O}|$ of observed rating entries is small. In practice, we only keep in memory the variables $m_{i,j}^{c,\backslash 11}$ and $v_{i,j}^{c,\backslash 11}$ with $(i,j) \in \mathcal{O}$. All the other variables $m_{i,j}^{c,\backslash 11}$ and $v_{i,j}^{c,\backslash 11}$ with $(i,j) \notin \mathcal{O}$ are computed when needed, as described in Sections 4.3.1, 4.3.2, 4.3.3 and 4.3.4, and then discarded afterwards.

Algorithm 3 Stochastic method for minimizing $\mathrm{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}} \| \mathcal{S})$.

Input: Current $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$, $\mathcal{Q}^{\backslash 11}$ and binary matrix $\mathbf{X}$.
for $t = 1$ **to** 3 **do**
  {Update the variational parameters for $\mathbf{U}$.}
  for $i = 1$ **to** n **do**
    Initialize $\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ with contribution from the prior.
    Generate set of indexes $\mathcal{I}_i^{\text{row},1}$.
    for $j \in \mathcal{I}_i^{\text{row},1}$ **do**
      Extract $m_{i,j}^{c,\backslash 11}$ and $v_{i,j}^{c,\backslash 11}$ from $\mathcal{Q}^{\backslash 11}$.
      Update $\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ using (203) and (204).
    end for
    Generate set of indexes $\mathcal{I}_j^{\text{row},0}$.
    for $j \in \mathcal{I}_i^{\text{row},0}$ **do**
      Compute $m_{i,j}^{c,\backslash 11}$ and $v_{i,j}^{c,\backslash 11}$ from $\mathcal{Q}^{\backslash 11}$ as described in sections 4.3.1, 4.3.2, 4.3.3 and 4.3.4.
      Update $\{[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ and $\{\hat{\mathbf{m}}_i^u[\hat{\mathbf{V}}_i^u]^{-1}\}_{\text{noisy}}$ using (203) and (204).
    end for
    Update $[\hat{\mathbf{V}}_i^u]^{-1}$ and $\hat{\mathbf{m}}_i^e[\hat{\mathbf{V}}_i^u]^{-1}$ using (168) and (169).
    Update $m_{i,1}^u, \ldots, m_{i,h}^u$ and $v_{i,1}^u, \ldots, v_{i,h}^u$ using (205).
  end for
  {Update the variational parameters for $\mathbf{V}$.}
  for $j = 1$ **to** d **do**
    Perform updates similar to the ones in the previous loop.
  end for
end for
Output: Updated $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$.

We first iterate over $i = 1, \ldots, n$, doing stochastic updates on the $m_{i,k}^v$ and $v_{j,k}^v$ in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the i-th row of $\mathbf{U}$, and then we iterate over $j = 1, \ldots, d$, doing stochastic updates for the $m_{i,k}^v$ and $v_{j,k}^v$ in $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ for the j-th column of $\mathbf{V}$. We repeat this process a total of 3 times each time we want to refine the approximate factor $\tilde{f}_{11}$. We always use as initial solution for $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ and $\hat{\mathcal{Q}}_{\mathbf{U},\mathbf{V}}$ the value obtained during the previous iteration of EP. Furthermore, the counter updateCounter that we use to compute the learning rates ρ_i^u counts the number of updates done during the whole execution of EP, not only during the individual executions of Algorithm 3.

4.4 Approximate Inference in the joint model

We describe how to perform approximate inference in the joint model formed by the combination of the complete data model (CDM) and the missing data model (MDM). Our algorithm first performs approximate inference in each of these models independently. First, we approximate the factors for the CDM assuming MAR data. For this, we run the expectation propagation (EP) method described in Section 4.1 for 40 iterations. Next, we approximate the factors for the MDM assuming that the predictions of the CDM are non-informative, that is, we assume that $\tilde{p}_{CDM}(r_{i,j}^\star = l|\mathbf{R}^{\mathcal{O}})$ in (148) is uniform across rating values. This disconnects the MDM from the CDM. The MDM is adjusted by running the stochastic variational inference (SVI) method described in Section 4.2.3 for 120 iterations. After this initial adjustment, we iterate refining the posterior of each model by taking into account the predictions of the other one. In this case, we adjust the MDM by running the SVI method from Section 4.2.3 again for 120 iterations, but this time without assuming that the predictions of the CDM are non-informative. Finally, the CDM is adjusted by running the EP-SVI method from Section (4.1) for 40 iterations. Algorithm 4 shows all the steps of our inference procedure.

Algorithm 4 Approximate Inference in the Joint Model

Input: Rating dataset $\mathcal{D}$.
Adjust $\mathcal{Q}$ by running EP on CDM with MAR data (Section 4.1).
Adjust $\mathcal{Q}$ by running SVI on MDM assuming CDM is uniform (Section 4.2.3).
for $i = 1$ **to** 2 **do**
  Adjust $\mathcal{Q}$ by running SVI on MDM (Section 4.2.3).
  Adjust $\mathcal{Q}$ by running EP-SVI on CDM with MNAR data (Section 4.3).
end for
Output: Posterior approximation $\mathcal{Q}$.

4.5 The predictive distribution of the joint model

Once we have adjusted $\mathcal{Q}$ by running the approximate inference method shown in Algorithm 4, we can use $\mathcal{Q}$ to make predictions. In particular, we can approximate the posterior probability $p^{\text{JM}}_{i,j,l}(x_{i,j})$ that the entry in the i-th row and j-th column of the rating matrix $\mathbf{R}$ may have taken value l, while conditioning to any specific value of $x_{i,j}$. When we fix $x_{i,j} = 0$, we assume that the entry was not selected by the missing data mechanism and it is actually missing. When we fix $x_{i,j} = 1$, we assume that the entry was selected by the missing data mechanism and should have been observed, but for some reason its value is unknown. The probability $p^{\text{JM}}_{i,j,l}(x_{i,j})$ is approximated by combining the predictions of the CDM and the MDM. In particular,

$$\tilde{p}^{\text{JM}}_{i,j,l}(x_{i,j}) \propto \tilde{p}_{CDM}(r^{\star}_{i,j} = l|\mathbf{R}^{\mathcal{O}})\tilde{p}^{\text{MDM}}_{i,j,l}(x_{i,j})\,, \tag{206}$$

where $\tilde{p}_{CDM}(r^{\star}_{i,j} = l|\mathbf{R}^{\mathcal{O}})$ is given by (148) and $\tilde{p}^{\text{MDM}}_{i,j,l}(x_{i,j})$ is given by (182).

5 Evaluation Using Other Metrics Besides Log-likelihood

In the main document, we have evaluated the performance of MF-MNAR by computing its average predictive log-likelihood (LL) on the standard and special test sets. In this section we report the results obtained by MF-MNAR using other evaluation metrics such as root mean squared error (RMSE), mean absolute error (MAS) and predictive accuracy (PA). The loss function used by PA takes value 1 if the predicted rating value is the same as the one actually found in the test set and 0 otherwise.

In the main document, we use LL as an evaluation metric because

1. LL is invariant to the arbitrary assignment of real values to ordinal ratings, while RMSE or MAE are not. For example, if we have three possible ordinal ratings with values "1", "2" and "3", these ratings could also have been labelled "low", "medium" and "high". In this latter case, it is not clear how to compute RMSE or MAE.

2. LL evaluates the quality of the whole predictive distribution. In this sense LL is more complete than RMSE, MAE or PA, which only evaluate the quality of point predictions. Having accurate predictive distributions, as measured by LL, can be very useful in practice, for example, to generate confidence intervals in the predictions.

3. LL is a better indicator of how accurately the method models the data. For example, RMSE is less sensitive than LL to how well the model captures the heteroskedasticity present in the data. In heteroskedastic and homoskedastic models the predictive means (used to minimize RMSE) are often very similar, while the predictive variances are not. In this sense, LL is a better performance metric than RMSE.

Tables 1 and 2 show the average test PA in the standard and special test sets, respectively. The best performing method is highlighted in bold and those results statistically indistinguishable according to a paired t-test are underlined. The results obtained for PA are similar to the ones shown in the main document for

LL. Regarding the standard test sets, MF-MNAR is on average the best method on the real-world datasets. Furthermore, in accordance with intuition, MF-MNAR is better than MF-MAR in the synthetic datasets with MNAR data, while the the opposite result occurs in the synthetic datasets with MAR data. Regarding the special test sets, Table 2 shows again that MF-MNAR is better than MF-MAR in the synthetic datasets with MNAR data. This table also shows that in the SMF-MAR and SRH-MAR datasets MF-MAR is better than MF-MNAR, as expected.

Tables 3 and 5 show for each method the average test MAE and RMSE in the standard test sets, respectively. MF-MNAR is on average the best method on the real-world datasets with respect to the MAE metric. However, the RMSE metric seems to favor Paquet's method which performs in this case better. Furthermore, in accordance with intuition, MF-MNAR is overall better than MF-MAR in the synthetic datasets with MNAR data, while the the opposite result occurs in the synthetic datasets with MAR data. Finally, tables 4 and 6 show the average test MAE and RMSE in the special test sets, respectively. In this case, we can observe that MF-MNAR is better than MF-MAR in the synthetic datasets with MNAR data, while the opposite results is obtained in the synthetic datasets with MAR data.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
ML100K	**0.479**	0.476	0.346	0.370	0.374	0.465	0.335
ML1M	**0.499**	0.495	0.416	0.397	0.400	0.486	0.348
MTweet	0.603	**0.603**	0.547	0.515	0.530	0.589	0.492
NIPS	**0.612**	0.609	0.455	0.495	0.494	0.594	0.398
Yahoo	**0.536**	0.522	0.493	0.437	0.481	0.503	0.314
SMF-MNAR	**0.638**	0.632	0.438	0.467	0.464	0.554	0.448
SMF-MAR	0.829	**0.835**	0.424	0.439	0.437	0.819	0.464
SRH-MNAR	0.477	0.476	**0.492**	0.474	0.480	0.475	0.318
SRH-MAR	0.419	0.430	0.422	0.419	0.416	**0.446**	0.293

Table 1: Average Test PA in Standard Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
Yahoo	0.400	0.403	0.363	0.454	0.341	0.425	**0.526**
SMF-MNAR	0.427	0.390	0.222	0.148	0.166	0.467	**0.479**
SMF-MAR	0.824	**0.825**	0.427	0.123	0.141	0.805	0.468
SRH-MNAR	**0.318**	0.300	0.306	0.180	0.183	0.264	0.245
SRH-MAR	0.438	0.440	0.426	0.354	0.341	**0.448**	0.252

Table 2: Average Test PA in Special Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
ML100K	**0.639**	0.642	0.826	0.806	0.801	0.645	0.900
ML1M	**0.595**	0.598	0.716	0.768	0.752	0.598	0.872
MTweet	0.451	**0.451**	0.534	0.553	0.540	0.457	0.593
NIPS	**0.532**	0.544	0.776	0.713	0.707	0.548	1.011
Yahoo	0.814	0.825	0.903	1.126	0.944	**0.800**	1.401
SMF-MNAR	**0.435**	0.459	0.832	0.824	0.819	0.466	0.935
SMF-MAR	0.170	**0.165**	0.675	0.623	0.620	0.183	0.602
SRH-MNAR	0.713	0.701	0.717	0.689	**0.684**	0.698	1.447
SRH-MAR	0.808	0.793	0.818	0.819	0.823	**0.782**	1.401

Table 3: Average Test MAE in Standard Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
Yahoo	1.180	1.177	1.249	**0.770**	0.881	1.094	0.819
SMF-MNAR	0.613	0.685	1.074	1.038	0.982	0.604	**0.562**
SMF-MAR	0.176	**0.175**	0.675	1.405	1.328	0.195	0.597
SRH-MNAR	1.253	1.318	1.231	1.206	**1.205**	1.310	1.427
SRH-MAR	0.801	**0.797**	0.816	0.847	0.869	0.799	1.450

Table 4: Average Test MAE in Special Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
ML100K	**0.883**	0.885	1.069	1.056	1.046	0.884	1.127
ML1M	0.828	0.829	0.954	1.025	0.997	**0.827**	1.117
MTweet	**0.690**	0.690	0.806	0.800	0.792	0.692	0.879
NIPS	**0.832**	0.834	1.014	0.968	0.963	0.837	1.163
Yahoo	1.180	1.186	1.265	1.427	1.286	**1.159**	1.583
SMF-MNAR	0.693	0.717	1.056	1.045	1.042	**0.659**	1.095
SMF-MAR	0.351	**0.349**	0.911	0.820	0.820	0.371	0.819
SRH-MNAR	0.949	0.957	0.903	0.896	**0.896**	0.937	1.604
SRH-MAR	1.136	1.129	1.134	1.134	1.134	**1.096**	1.552

Table 5: Average Test RMSE in Standard Test Sets.

Dataset	MF MNAR	MF MAR	MM MAR	CTPv MNAR	Logitvd MNAR	Paquet MAR	Oracle
Yahoo	1.483	1.480	1.500	**1.056**	1.141	1.404	1.057
SMF-MNAR-	0.793	0.857	1.163	1.181	1.152	0.812	**0.775**
SMF-MAR	0.362	**0.360**	0.912	1.402	1.350	0.385	0.816
SRH-MNAR	1.519	1.562	1.463	**1.430**	1.430	1.547	1.550
SRH-MAR	1.140	1.130	1.135	1.169	1.172	**1.101**	1.566

Table 6: Average Test RMSE in Special Test Sets.

References

Ghahramani, Z. and Beal, M. J. (2001). *Advanced Mean Field Method—Theory and Practice*, chapter Graphical models and variational methods, pages 161–177.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Jaakkola, T. and Jordan, M. (1997). A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*.

Lim, Y. J. and Teh, Y. W. (2007). Variational bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7. Citeseer.

MacKay, D. J. (1992). The evidence framework applied to classification networks. *Neural computation*, 4(5):720–736.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, pages 362–369.

Paquet, U. and Koenigstein, N. (2013). One-class collaborative filtering with random graphs. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 999–1008, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Paquet, U., Thomson, B., and Winther, O. (2012). A hierarchical model for ordinal matrix factorization. *Statistics and Computing*, 22(4):945–957.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Stern, D. H., Herbrich, R., and Graepel, T. (2009). Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, pages 111–120. ACM.