Supplementary Material to Cold-Start Active Learning with Robust Ordinal Matrix Factorization

Neil Houlsby[1], José Miguel Hernández-Lobato[1] and Zoubin Ghahramani

1 Approximate Inference in HOMF

We describe the implementation of approximate inference in the proposed heteroskedastic ordinal matrix factorization (HOMF) model. Approximate Bayesian inference is performed using expectation propagation (EP) Minka (2001) and variational Bayes Ghahramani and Beal (2001). We use variational Bayes to approximate some operations within the execution of EP. First, we describe the hyper-parameter values used in the prior distributions of the HOMF model. After that, we describe in detail the implementation of expectation propagation in HOMF, how to make predictions using the EP approximation to the posterior distribution and the specific form the EP update operations.

1.1 Description of the hyper-parameter values used in the priors of HOMF

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 for the noise variance γ_i^{row} and γ_j^{col} 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 $i = 1, \ldots, n$, $j = 1, \ldots, d$ 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 distribution 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 on 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 a random sample of each of these variables of size 10 with empirical variance $\sqrt{10}$. The prior expectations for γ_i^{row} and γ_j^{col} are $\sqrt{10}$. This means that the product of γ_i^{row} and γ_j^{col} is on average 10. This 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,

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

[1] Equal contributors.

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,

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

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 a random sample of each of these variables of size 10 with unit empirical variance.

1.2 Expectation propagation in HOMF

Recall that the latent variables in HOMF are given by $\boldsymbol{\Xi} = \{\mathbf{U}, \mathbf{V}, \mathbf{B}, \mathbf{A}, \mathbf{C}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}}, \mathbf{b}_0, \mathbf{m}^{\mathbf{U}}, \mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{U}}, \mathbf{v}^{\mathbf{V}}\}$. As described in the main document, the posterior distribution for $\boldsymbol{\Xi}$ given the set of entries in the rating matrix $\mathbf{R}$ that are observed, that is, $\mathbf{R}^{\mathcal{O}}$, is

$$\begin{aligned}
p(\boldsymbol{\Xi}|\mathbf{R}^{\mathcal{O}}) = {} & p(\mathbf{R}^{\mathcal{O}}|\mathbf{A}, \mathbf{B}) p(\mathbf{A}|\mathbf{C}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}}) p(\mathbf{C}|\mathbf{U}, \mathbf{V}) p(\mathbf{U}|\mathbf{m}^{\mathbf{U}}, \mathbf{v}^{\mathbf{U}}) \\
& p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}}) p(\mathbf{B}|\mathbf{b}_0) p(\mathbf{b}_0) p(\boldsymbol{\gamma}^{\text{row}}) p(\boldsymbol{\gamma}^{\text{col}}) \\
& p(\mathbf{m}^{\mathbf{U}}) p(\mathbf{m}^{\mathbf{V}}) p(\mathbf{v}^{\mathbf{U}}) p(\mathbf{v}^{\mathbf{V}}) [p(\mathbf{R}^{\mathcal{O}})]^{-1}\,,
\end{aligned} \tag{5}$$

where $p(\mathbf{R}^{\mathcal{O}})$ is the normalization constant. Expectation propagation (EP) approximates this posterior distribution with the following parametric approximation within the exponential family of distributions:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\Xi}) = {} & \left[\prod_{j=1}^{d}\prod_{k=1}^{L-1} \mathcal{N}(b_{j,k}|m_{j,k}^{b}, v_{j,k}^{b})\right] \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(a_{i,j}|m_{i,j}^{a}, v_{i,j}^{a})\right] \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(c_{i,j}|m_{i,j}^{c}, v_{i,j}^{c})\right] \\
& \left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u}, v_{i,k}^{u}\right] \left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_{j,k}^{v}, v_{j,k}^{v}\right] \left[\prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|m_{k}^{b_0}, v_{k}^{b_0})\right] \\
& \left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}}}, v_k^{m^{\mathbf{U}}})\right] \left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}}}, v_k^{m^{\mathbf{V}}})\right] \left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}}}, b_k^{v^{\mathbf{U}}})\right] \\
& \left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}}}, b_k^{v^{\mathbf{V}}})\right] \left[\prod_{i=1}^{n} \mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}}}, b_i^{\gamma^{\text{row}}})\right] \left[\prod_{j=1}^{d} \mathcal{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}}}, b_j^{\gamma^{\text{col}}})\right]\,.
\end{aligned} \tag{6}$$

where the parameters of each of the factors that form $\mathcal{Q}$ will be fixed during the execution of the EP algorithm. If we ignore the normalization constant, the exact posterior distribution (5) includes 13 different factors, namely, $p(\mathbf{R}^{\mathcal{O}}|\mathbf{A}, \mathbf{B})$, $p(\mathbf{A}|\mathbf{C}, \boldsymbol{\gamma}^{\text{row}}, \boldsymbol{\gamma}^{\text{col}})$, $p(\mathbf{C}|\mathbf{U}, \mathbf{V})$, $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}}, \mathbf{v}^{\mathbf{U}})$, $p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}})$, $p(\mathbf{B}|\mathbf{b}_0)$, $p(\mathbf{b}_0)$, $p(\boldsymbol{\gamma}^{\text{row}})$, $p(\boldsymbol{\gamma}^{\text{col}})$, $p(\mathbf{m}^{\mathbf{U}})$, $p(\mathbf{m}^{\mathbf{V}})$, $p(\mathbf{v}^{\mathbf{U}})$ and $p(\mathbf{v}^{\mathbf{V}})$. EP works by approximating each of these exact factors with a corresponding approximate factor $\tilde{f}_l$, where $l = 1, \ldots, 13$ and each $\tilde{f}_l$ has the same functional form as the posterior approximation $\mathcal{Q}$, namely

$$\begin{aligned}
\tilde{f}_l(\boldsymbol{\Xi}) = \tilde{s}_l & \left[\prod_{j=1}^{d}\prod_{k=1}^{L-1} \mathcal{N}(b_{j,k}|\tilde{m}_{j,k}^{b,l}, \tilde{v}_{j,k}^{b,l})\right] \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(a_{i,j}|\tilde{m}_{i,j}^{a,l}, \tilde{v}_{i,j}^{a,l})\right] \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(c_{i,j}|\tilde{m}_{i,j}^{c,l}, \tilde{v}_{i,j}^{c,l})\right] \\
& \left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|\tilde{m}_{i,k}^{u,l}, \tilde{v}_{i,k}^{u,l}\right] \left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|\tilde{m}_{j,k}^{v,l}, \tilde{v}_{j,k}^{v,l}\right] \left[\prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|\tilde{m}_{k}^{b_0,l}, \tilde{v}_{k}^{b_0,l})\right] \\
& \left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{U}}|\tilde{m}_k^{m^{\mathbf{U}},l}, \tilde{v}_k^{m^{\mathbf{U}},l})\right] \left[\prod_{k=1}^{h} \mathcal{N}(m_k^{\mathbf{V}}|\tilde{m}_k^{m^{\mathbf{V}},l}, \tilde{v}_k^{m^{\mathbf{V}},l})\right] \left[\prod_{k=1}^{h} \mathcal{IG}(v_k^{\mathbf{U}}|\tilde{a}_k^{v^{\mathbf{U}},l}, \tilde{b}_k^{v^{\mathbf{U}},l})\right]
\end{aligned}$$

$$\left[\prod_{k=1}^{h}\mathcal{IG}(\tilde{v}_k^{\mathbf{V},l}|\tilde{a}_k^{v^{\mathbf{V}},l},\tilde{b}_k^{v^{\mathbf{V}},l})\right]\left[\prod_{i=1}^{n}\mathcal{IG}(\gamma_i^{\text{row}}|\tilde{a}_i^{\gamma^{\text{row}},l},\tilde{b}_i^{\gamma^{\text{row}},l})\right]\left[\prod_{j=1}^{d}\mathcal{IG}(\gamma_j^{\text{col}}|\tilde{a}_j^{\gamma^{\text{col}},l},\tilde{b}_j^{\gamma^{\text{col}},l})\right]\,, \qquad (7)$$

where we have introduced the multiplicative constant $\tilde{s}_l$ in $\tilde{f}_l$ because the approximate factors $\tilde{f}_1,\ldots,\tilde{f}_{13}$ may not be normalized. EP will adjust the parameters of $\tilde{f}_1,\ldots,\tilde{f}_{13}$ so that the l-th approximate factor $\tilde{f}_l$ is similar to the corresponding l-th exact factor in (5), for $l=1,\ldots,13$. Note that the exact posterior is given by the product of all the exact factors in (5) and then normalizing the resulting function. Similarly, the posterior approximation $\mathcal{Q}$ is obtained by computing the product of all the approximate factors $\tilde{f}_1,\ldots,\tilde{f}_{13}$ and then normalizing the result of this operation. The family of exponential distributions is closed under the product operation. Therefore, the product of $\tilde{f}_1,\ldots,\tilde{f}_{13}$ has the same functional form as $\mathcal{Q}$ and can be readily normalized. In particular, at any step of the EP algorithm we have that $\mathcal{Q}(\mathbf{\Xi})\propto\prod_{l=1}^{13}\tilde{f}_l(\mathbf{\Xi})$.

EP works by first initializing all the approximate factors $\tilde{f}_1,\ldots,\tilde{f}_{13}$ and $\mathcal{Q}$ to be non-informative or flat. This is done by setting the mean and variance parameters of the Gaussian factors in $\tilde{f}_1,\ldots,\tilde{f}_{13}$ and $\mathcal{Q}$ to take value zero and infinity, respectively, and setting the a and b parameters of the inverse-gamma factors to take values one and zero, respectively. After that, EP iteratively refines the parameters of the approximate factors. For this, let $\mathcal{Q}^{\setminus l}$ denote the distribution obtained by computing the ratio of $\mathcal{Q}$ and $\tilde{f}_l$ and then normalizing the resulting function. That is, $\mathcal{Q}^{\setminus l}$ is equal to the normalized product of all the approximate factors except the l-th one: $\mathcal{Q}^{\setminus l}(\mathbf{\Xi})\propto\mathcal{Q}(\mathbf{\Xi})/\tilde{f}_l(\mathbf{\Xi})$. The functional form of $\mathcal{Q}^{\setminus l}$ is again the same as that of $\mathcal{Q}$ and all the $\tilde{f}_1,\ldots,\tilde{f}_{13}$, that is,

$$\begin{aligned}\mathcal{Q}^{\setminus l}(\mathbf{\Xi}) = &\left[\prod_{i=1}^{d}\prod_{k=1}^{L-1}\mathcal{N}(b_{j,k}|m_{j,k}^{b,\setminus l},v_{j,k}^{b,\setminus l})\right]\left[\prod_{(i,j)\in\mathcal{O}}\mathcal{N}(a_{i,j}|m_{i,j}^{a,\setminus l},v_{i,j}^{a,\setminus l})\right]\left[\prod_{(i,j)\in\mathcal{O}}\mathcal{N}(c_{i,j}|m_{i,j}^{c,\setminus l},v_{i,j}^{c,\setminus l})\right]\\ &\left[\prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{i,k}|m_{i,k}^{u,\setminus l},v_{i,k}^{u,\setminus l}\right]\left[\prod_{j=1}^{d}\prod_{k=1}^{h}\mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus l},v_{j,k}^{v,\setminus l}\right]\left[\prod_{k=1}^{L-1}\mathcal{N}(b_{0,k}|m_k^{b_0,\setminus l},v_k^{b_0,\setminus l})\right]\\ &\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{U}}|m_k^{m^{\mathbf{U}},\setminus l},v_k^{m^{\mathbf{U}},\setminus l})\right]\left[\prod_{k=1}^{h}\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\setminus l},v_k^{m^{\mathbf{V}},\setminus l})\right]\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{U}}|a_k^{v^{\mathbf{U}},\setminus l},b_k^{v^{\mathbf{U}},\setminus l})\right]\\ &\left[\prod_{k=1}^{h}\mathcal{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}},\setminus l},b_k^{v^{\mathbf{V}},\setminus l})\right]\left[\prod_{i=1}^{n}\mathcal{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\setminus l},b_i^{\gamma^{\text{row}},\setminus l})\right]\left[\prod_{j=1}^{d}\mathcal{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\setminus l},b_j^{\gamma^{\text{col}},\setminus l})\right]\,. \qquad (8)\end{aligned}$$

EP refines the parameters of $\tilde{f}_l$ by minimizing the Kullback-Leibler (KL) divergence between $\mathcal{Q}^{\setminus l}(\mathbf{\Xi})\tilde{f}_l(\mathbf{\Xi})$ and $\mathcal{Q}^{\setminus l}(\mathbf{\Xi})f_l(\mathbf{\Xi})$, where $f_l(\mathbf{\Xi})$ denotes the l-th exact factor in the exact posterior (5), that is, EP refines the parameters of $\tilde{f}_l$ by minimizing

$$\mathrm{D}_{\mathrm{KL}}(Q^{\setminus l}f_l\|Q^{\setminus l}\tilde{f}_l)=\int\left[Q^{\setminus l}f_l\log\frac{Q^{\setminus l}f_l}{Q^{\setminus l}\tilde{f}_l}+Q^{\setminus l}\tilde{f}_l-Q^{\setminus l}f_l\right]d\mathbf{\Xi}\,, \qquad (9)$$

where the arguments to $Q^{\setminus l}f_l$ and $Q^{\setminus l}\tilde{f}_l$ have been omitted in the right-hand side of this equation to improve readability. It can be shown that (9) is minimized when the expectation of the sufficient statistics of $Q^{\setminus l}\tilde{f}_l$ with respect to $Q^{\setminus l}\tilde{f}_l$ is the same as the expectation of those sufficient statistics with respect to $Q^{\setminus l}f_l$. The main loop of EP iterates over all the approximate factors $\tilde{f}_l$, $l=1,\ldots,13$, refining one after the other by minimizing (9). In our experiments, we run the 30 iterations of the main loop of the EP algorithm.

1.3 The EP predictive distribution

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 $r^\star_{i,j}$.

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 $\mathbf{R}^{\mathcal{O}}$. The exact posterior distribution for $r^\star_{i,j}$ given $\mathbf{R}^{\mathcal{O}}$ is then

$$p(r^\star_{i,j}|\mathbf{R}^{\mathcal{O}}) = \int p(r^\star_{i,,j}|a^\star_{i,j},\mathbf{b}_j)p(a^\star_{i,j}|c^\star_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}})p(c^\star_{i,j}|\mathbf{u}_i,\mathbf{v}_j)p(\Xi|\mathbf{R}^{\mathcal{O}})\,d\Xi\,da^\star_{i,j}\,dc^\star_{i,j}\,, \tag{10}$$

with $p(r^\star_{i,,j}|a^\star_{i,j},\mathbf{b}_j) = \prod_{k=1}^{L-1}\Theta\left[\text{sign}[r^\star_{i,j}-k-0.5](a^\star_{i,j}-b_{j,k})\right]$, $p(a^\star_{i,j}|c^\star_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}}) = \mathcal{N}(a^\star_{i,j}|c^\star_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})$, and $p(c^\star_{i,j}|\mathbf{u}_i,\mathbf{v}_j) = \delta(c^\star_{i,j}-\mathbf{u}_i^{\text{T}}\,\mathbf{v}_j)$. Recall that δ is a point mass at zero. To obtain an approximation to (10) we first replace the exact posterior $p(\Xi|\mathbf{R}^{\mathcal{O}})$ in (10) with the EP approximation $\mathcal{Q}$. However, even after making this approximation, the resulting integral is not analytically tractable. We therefore perform an additional approximation. We replace $\int\delta(c^\star_{i,j}-\mathbf{u}_i^{\text{T}}\,\mathbf{v}_j)\mathcal{Q}(\Xi)\,d\Xi$ 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 $\mathbf{u}_i^{\text{T}}\,\mathbf{v}_j$ is 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 $\int\mathcal{N}(a^\star_{i,j}|c^\star_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})\mathcal{N}(c^\star_{i,j}|m^{c,\star}_{i,j},v^{c,\star}_{i,j})\mathcal{Q}(\Xi)d\Xi$ with an additional Gaussian with mean $m^{c,\star}_{i,j}$ and variance $v^{c,\star}_{i,j}+v^\gamma_{i,j}$ where $v^\gamma_{i,j} = [b^{\gamma^{\text{row}}}b^{\gamma^{\text{col}}}][(a^{\gamma^{\text{row}}}+1)(a^{\gamma^{\text{col}}}+1)]^{-1}$. 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 $\mathbf{R}^{\mathcal{O}}$ is then approximated by

$$\begin{aligned}\tilde{p}(r^\star_{i,j}|\mathbf{R}^{\mathcal{O}}) &= \int\prod_{k=1}^{L-1}\Theta\left[\text{sign}[r^\star_{i,j}-k-0.5](a^\star_{i,j}-b_{j,k})\right]\mathcal{N}(a^\star_{i,j}|m^{c,\star}_{i,j},v^{c,\star}_{i,j}+v^\gamma_{i,j})\mathcal{Q}(\Xi)\,d\Xi\,da^\star_{i,j}\\ &= \Phi\left\{\zeta(r^\star_{i,j})\right\} - \Phi\left\{\zeta(r^\star_{i,j}-1)\right\}\,,\end{aligned} \tag{11}$$

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.

1.4 The EP update operations

As described in Section 1.2, EP works by iteratively minimizing (9) with respect to each approximate factor $\tilde{f}_l$. Note that we have one approximate factor $\tilde{f}_l$ for each of the 13 exact factors in the posterior distribution (5), namely, $p(\mathbf{R}^{\mathcal{O}}|\mathbf{A},\mathbf{B})$, $p(\mathbf{A}|\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}})$, $p(\mathbf{C}|\mathbf{U},\mathbf{V})$, $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}})$, $p(\mathbf{V}|\mathbf{m}^{\mathbf{V}},\mathbf{v}^{\mathbf{V}})$, $p(\mathbf{B}|\mathbf{b}_0)$, $p(\mathbf{b}_0)$, $p(\boldsymbol{\gamma}^{\text{row}})$, $p(\boldsymbol{\gamma}^{\text{col}})$, $p(\mathbf{m}^{\mathbf{U}})$, $p(\mathbf{m}^{\mathbf{V}})$, $p(\mathbf{v}^{\mathbf{U}})$ and $p(\mathbf{v}^{\mathbf{V}})$. In the following sections we show the form of the EP updates for refining the parameters of each approximate factor $\tilde{f}_l$.

1.4.1 EP updates for $\tilde{f}_1$

In our mapping between approximate factors and exact factors, we specify that $\tilde{f}_1$ approximates the exact factor $p(\mathbf{v}^{\mathbf{V}}) = \prod_{k=1}^h \mathcal{IG}(v_k^{\mathbf{V}}|a_0',b_0')$. In this case, $p(\mathbf{v}^{\mathbf{V}})$ has the same functional form as the inverse-gamma factors which specify the distribution of $\mathbf{v}^{\mathbf{V}}$ in $\tilde{f}_1$. Therefore, the EP update for $\tilde{f}_1$ sets the parameters of those inverse-gamma factors to be the same as the parameters of the the inverse-gamma factors in $p(\mathbf{v}^{\mathbf{V}})$, namely

$$[\tilde{a}_k^{v^{\mathbf{V}},1}]^{\text{new}} = a_0'\,, \qquad [\tilde{b}_k^{v^{\mathbf{V}},1}]^{\text{new}} = b_0'\,, \tag{12}$$

for $k=1,\ldots,h$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_1$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_1$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[a_k^{v^{\mathbf{V}}}]^{\text{new}} = a_0'\,, \qquad [b_k^{v^{\mathbf{V}}}]^{\text{new}} = b_0'\,, \tag{13}$$

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

1.4.2 EP updates for $\tilde{f}_2$

In our mapping between approximate factors and exact factors, $\tilde{f}_2$ approximates the exact factor $p(\mathbf{v}^\mathbf{U}) = \prod_{k=1}^h \mathcal{IG}(v_k^\mathbf{U}|a_0', b_0')$. The EP update operations in this case are the same as for the approximate factor $\tilde{f}_1$, namely,

$$[\tilde{a}_k^{v^\mathbf{U},2}]^\text{new} = a_0'\,, \qquad [\tilde{b}_k^{v^\mathbf{U},2}]^\text{new} = b_0'\,, \tag{14}$$

for $k = 1, \ldots, h$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_2$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_2$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[a_k^{v^\mathbf{U}}]^\text{new} = a_0'\,, \qquad [b_k^{v^\mathbf{U}}]^\text{new} = b_0'\,, \tag{15}$$

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

1.4.3 EP updates for $\tilde{f}_3$

In our mapping between approximate factors and exact factors, $\tilde{f}_3$ approximates the exact factor $p(\mathbf{m}^\mathbf{V}) = \prod_{k=1}^h \mathcal{N}(m_k^\mathbf{V}|0,1)$. In this case, $p(\mathbf{m}^\mathbf{V})$ has the same functional form as the Gaussian factors which specify the distribution of $\mathbf{m}^\mathbf{V}$ in $\tilde{f}_3$. Therefore, the EP update for $\tilde{f}_3$ sets the parameters of those Gaussian factors to be the same as the parameters of the the Gaussians in $p(\mathbf{m}^\mathbf{V})$, namely

$$[\tilde{m}_k^{m^\mathbf{V},3}]^\text{new} = 0\,, \qquad [\tilde{v}_k^{m^\mathbf{V},3}]^\text{new} = 1\,, \tag{16}$$

for $k = 1, \ldots, h$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_3$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_3$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[m_k^{m^\mathbf{V}}]^\text{new} = 0\,, \qquad [v_k^{m^\mathbf{V}}]^\text{new} = 1\,, \tag{17}$$

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

1.4.4 EP updates for $\tilde{f}_4$

In our mapping between approximate factors and exact factors, $\tilde{f}_4$ approximates the exact factor $p(\mathbf{m}^\mathbf{U}) = \prod_{k=1}^h \mathcal{N}(m_k^\mathbf{U}|0,1)$. The EP update operation are in this case the same as for the approximate factor $\tilde{f}_3$, namely,

$$[\tilde{m}_k^{m^\mathbf{U},4}]^\text{new} = 0\,, \qquad [\tilde{v}_k^{m^\mathbf{U},4}]^\text{new} = 1\,, \tag{18}$$

for $k = 1, \ldots, h$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_4$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_4$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[m_k^{m^\mathbf{U}}]^\text{new} = 0\,, \qquad [v_k^{m^\mathbf{U}}]^\text{new} = 1\,, \tag{19}$$

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

1.4.5 EP updates for $\tilde{f}_5$

In our mapping between approximate factors and exact factors, $\tilde{f}_5$ approximates the exact factor $p(\boldsymbol{\gamma}^\text{col}) = \prod_{j=1}^d \mathcal{IG}(\gamma_j^\text{col}|a_0, b_0)$, In this case, $p(\boldsymbol{\gamma}^\text{col})$ has the same functional form as the inverse-gamma factors which specify the distribution of $\boldsymbol{\gamma}^{\text{col}}$ in $\tilde{f}_5$. Therefore, the EP update for $\tilde{f}_5$ sets the parameters of those inverse-gamma factors to be the same as the parameters of the the factors in $p(\boldsymbol{\gamma}^{\text{col}})$, namely

$$[\tilde{a}_j^{\gamma^{\text{col}},5}]^{\text{new}} = a_0\,, \qquad [\tilde{b}_j^{\gamma^{\text{col}},5}]^{\text{new}} = b_0\,, \tag{20}$$

for $j = 1, \ldots, d$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_5$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_5$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[a_j^{\gamma^{\text{col}}}]^{\text{new}} = a_0\,, \qquad [b_j^{\gamma^{\text{col}}}]^{\text{new}} = b_0\,, \tag{21}$$

for $j = 1, \ldots, d$.

1.4.6 EP updates for $\tilde{f}_6$

In our mapping between approximate factors and exact factors, $\tilde{f}_6$ approximates the exact factor $p(\boldsymbol{\gamma}^{\text{row}}) = \prod_{i=1}^n \mathcal{IG}(\gamma_i^{\text{row}}|a_0, b_0)$. The EP update operation are in this case the same as for the approximate factor $\tilde{f}_6$, namely,

$$[\tilde{a}_i^{\gamma^{\text{row}},6}]^{\text{new}} = a_0\,, \qquad [\tilde{b}_i^{\gamma^{\text{row}},6}]^{\text{new}} = b_0\,, \tag{22}$$

for $i = 1, \ldots, n$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_6$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_6$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[a_i^{\gamma^{\text{row}}}]^{\text{new}} = a_0\,, \qquad [b_i^{\gamma^{\text{row}}}]^{\text{new}} = b_0\,, \tag{23}$$

for $i = 1, \ldots, n$.

1.4.7 EP updates for $\tilde{f}_7$

In our mapping between approximate factors and exact factors, $\tilde{f}_7$ approximates the exact factor $p(\mathbf{b}_0) = \prod_{k=1}^{L-1} \mathcal{N}(b_{0,k}|m_k^{\mathbf{b}_0}, v_0)$. In this case, $p(\mathbf{b}_0)$ has the same functional form as the Gaussian factors which specify the distribution of $\mathbf{b}_0$ in $\tilde{f}_7$. Therefore, the EP update for $\tilde{f}_7$ sets the parameters of those Gaussian factors to be the same as the parameters of the factors in $p(\mathbf{b}_0)$, namely

$$[\tilde{m}_k^{b_0,7}]^{\text{new}} = m_k^{\mathbf{b}_0}\,, \qquad [\tilde{v}_k^{b_0,7}]^{\text{new}} = v_0\,, \tag{24}$$

for $k = 1, \ldots, L-1$. Since these update equations do not depend on the parameters of any other approximate factor, we have that $\tilde{f}_7$ has to be refined only once, during the first iteration of the main loop of EP. After refining $\tilde{f}_7$, we update $\mathcal{Q}$ (which is initially uniform) by setting

$$[m_k^{b_0}]^{\text{new}} = m_k^{\mathbf{b}_0}\,, \qquad [v_k^{b_0}]^{\text{new}} = v_0\,, \tag{25}$$

for $k = 1, \ldots, L-1$.

1.4.8 EP updates for $\tilde{f}_8$

In our mapping between approximate factors and exact factors, $\tilde{f}_8$ approximates the exact factor $p(\mathbf{B}|\mathbf{b}_0) = \prod_{j=1}^d \prod_{k=1}^{L-1} \mathcal{N}(b_{j,k}|b_{0,k}, v_0)$. Because $p(\mathbf{B}|\mathbf{b}_0)$ has a complicated form, we approximate individually each of its d internal factors of the form $\prod_{k=1}^{L-1} \mathcal{N}(b_{j,k}|b_{0,k}, v_0)$, where $j = 1, \ldots, d$, with extra approximate factors $\tilde{f}_{8,1}, \ldots, \tilde{f}_{8,d}$. In this case, $\tilde{f}_8$ is given by the product of $\tilde{f}_{8,1}, \ldots, \tilde{f}_{8,d}$, where all these additional approximate factors also have the same functional form as $\mathcal{Q}$. Initially, all the $\tilde{f}_{8,1}, \ldots, \tilde{f}_{8,d}$ and $\tilde{f}_8$ are non-informative or flat. EP will iteratively refine each of the extra approximate factors as follows. To refine $\tilde{f}_{8,j}$, where

$j = 1, \ldots, d$, we firstly compute the parameters of $\mathcal{Q}^{\backslash 8,j}$ which is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{8,j}$. This leads to

$$v_k^{b_0,\backslash 8,j} = \left[[v_k^{b_0}]^{-1} - [\tilde{v}_k^{b_0,8,j}]^{-1} \right]^{-1} , \qquad m_k^{b_0,\backslash 8,j} = v_k^{b_0,\backslash 8,j} \left[m_k^{b_0} [v_k^{b_0}]^{-1} - \tilde{m}_k^{b_0,8,j} [\tilde{v}_k^{b_0,8,j}]^{-1} \right] , \tag{26}$$

$$v_{j,k}^{b,\backslash 8,j} = \left[[v_{j,k}^{b}]^{-1} - [\tilde{v}_{j,k}^{b,8,j}]^{-1} \right]^{-1} , \qquad m_{j,k}^{b,\backslash 8,j} = v_{j,k}^{b,\backslash 8,j} \left[m_{j,k}^{b} [v_{j,k}^{b}]^{-1} - \tilde{m}_{j,k}^{b,8,j} [\tilde{v}_{j,k}^{b,8,j}]^{-1} \right] , \tag{27}$$

for $k = 1, \ldots, L-1$. After that, we refine the approximate factor $\tilde{f}_{8,j}$ by setting

$$[\tilde{m}_k^{b_0,8,j}]^{\text{new}} = m_{j,k}^{b,\backslash 8,j} , \qquad [\tilde{v}_k^{b_0,8,j}]^{\text{new}} = v_{j,k}^{b,\backslash 8,j} + v_0 , \tag{28}$$

$$[\tilde{m}_k^{b,8,j}]^{\text{new}} = m_k^{b_0,\backslash 8,j} , \qquad [\tilde{v}_k^{b,8,j}]^{\text{new}} = v_k^{b_0,\backslash 8,j} + v_0 , \tag{29}$$

for $k = 1, \ldots, L-1$. These update equations guarantee that the normalized versions of $\mathcal{Q}^{\backslash 8,j}(\boldsymbol{\Xi})\tilde{f}_{8,j}(\boldsymbol{\Xi})$ and $\mathcal{Q}^{\backslash 8,j}(\boldsymbol{\Xi}) \prod_{k=1}^{L-1} \mathcal{N}(b_{j,k}|b_{0,k}, v_0)$ have the same expected sufficient statistics. Finally, we recompute Q as the normalized product of the updated $\tilde{f}_{8,j}$ and $\mathcal{Q}^{\backslash 8,j}$, that is,

$$[v_k^{b_0}]^{\text{new}} = \left[[v_k^{b_0,\backslash 8,j}]^{-1} + [\tilde{v}_k^{b_0,8,j}]^{-1} \right]^{-1} , \tag{30}$$

$$[m_k^{b_0}]^{\text{new}} = [v_k^{b_0}]^{\text{new}} \left[m_k^{b_0,\backslash 8,j} [v_k^{b_0,\backslash 8,j}]^{-1} + \tilde{m}_k^{b_0,8,j} [\tilde{v}_k^{b_0,8,j}]^{-1} \right] , \tag{31}$$

$$[v_{j,k}^{b}]^{\text{new}} = \left[[v_{j,k}^{b,\backslash 8,j}]^{-1} + [\tilde{v}_{j,k}^{b,8,j}]^{-1} \right]^{-1} , \tag{32}$$

$$[m_{j,k}^{b}]^{\text{new}} = [v_{j,k}^{b}]^{\text{new}} \left[m_{j,k}^{b,\backslash 8,j} [v_{j,k}^{b,\backslash 8,j}]^{-1} + \tilde{m}_{j,k}^{b,8,j} [\tilde{v}_{j,k}^{b,8,j}]^{-1} \right] , \tag{33}$$

for $k = 1, \ldots, L-1$.

1.4.9 EP updates for $\tilde{f}_9$

In our mapping between approximate factors and exact factors, $\tilde{f}_9$ approximates the factor $p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}}) = \prod_{j=1}^{d} \prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, v_k^{\mathbf{V}})$. Because $p(\mathbf{V}|\mathbf{m}^{\mathbf{V}}, \mathbf{v}^{\mathbf{V}})$ has a complicated form, we approximate individually each of its $d \times h$ internal factors of the form $\mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, v_k^{\mathbf{V}})$, where $j = 1, \ldots, d$ and $k = 1, \ldots, h$, with extra approximate factors $\tilde{f}_{9,1,1}, \ldots, \tilde{f}_{9,d,h}$. In this case, $\tilde{f}_9$ is given by the product of $\tilde{f}_{9,1,1}, \ldots, \tilde{f}_{9,d,h}$, where all these additional approximate factors also have the same functional form as $\mathcal{Q}$. Initially, all the $\tilde{f}_{9,1,1}, \ldots, \tilde{f}_{9,d,h}$ and $\tilde{f}_9$ are non-informative or flat. EP will iteratively refine each of the extra approximate factors as follows. To refine $\tilde{f}_{9,j,k}$, where $j = 1, \ldots, d$ and $k = 1, \ldots, h$, we firstly compute the parameters of $\mathcal{Q}^{\backslash 9,j,k}$ which is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{9,j,k}$. This leads to

$$[v_k^{m^{\mathbf{V}},\backslash 9,j,k}]^{\text{new}} = \left[[v_k^{m^{\mathbf{V}}}]^{-1} - [\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1} \right]^{-1} , \tag{34}$$

$$[m_k^{m^{\mathbf{V}},\backslash 9,j,k}]^{\text{new}} = [v_k^{m^{\mathbf{V}},\backslash 9,j,k}]^{\text{new}} \left[m_k^{m^{\mathbf{V}}} [v_k^{m^{\mathbf{V}}}]^{-1} - \tilde{m}_k^{m^{\mathbf{V}},9,j,k} [\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1} \right] , \tag{35}$$

$$[v_{j,k}^{v,\backslash 9,j,k}]^{\text{new}} = \left[[v_{j,k}^{v}]^{-1} - [\tilde{v}_{j,k}^{v,9,j,k}]^{-1} \right]^{-1} , \tag{36}$$

$$[m_{j,k}^{v,\backslash 9,j,k}]^{\text{new}} = [v_{j,k}^{v,\backslash 9,j,k}]^{\text{new}} \left[m_{j,k}^{v} [v_{j,k}^{v}]^{-1} - \tilde{m}_{j,k}^{v,9,j,k} [\tilde{v}_{j,k}^{v,9,j,k}]^{-1} \right] , \tag{37}$$

$$[a_k^{v^{\mathbf{V}},\backslash 9,j,k}]^{\text{new}} = a_k^{v^{\mathbf{V}}} - \tilde{a}_k^{v^{\mathbf{V}},9,j,k} + 1 , \tag{38}$$

$$[b_k^{v^{\mathbf{V}},\backslash 9,j,k}]^{\text{new}} = b_k^{v^{\mathbf{V}}} - \tilde{b}_k^{v^{\mathbf{V}},9,j,k} . \tag{39}$$

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 $h(\boldsymbol{\Xi}) = \mathcal{Q}^{\backslash 9,j,k}(\boldsymbol{\Xi})\mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, v_k^{\mathbf{V}})$. After integrating out $\boldsymbol{\Xi} \setminus \{v_{j,k}, m_k^{\mathbf{V}}, v_k^{\mathbf{V}}\}$ in h, we obtain

$$h(v_{j,k}, m_k^{\mathbf{V}}, v_k^{\mathbf{V}}) = \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, v_k^{\mathbf{V}})\mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\backslash 9,j,k}, v_k^{m^{\mathbf{V}},\backslash 9,j,k})$$

$$\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 9,j,k}, v_{j,k}^{v,\backslash 9,j,k})\mathrm{IG}(v_k^{\mathbf{V}}|a_k^{v^{\mathbf{V}},\backslash 9,j,k}, b_k^{v^{\mathbf{V}},\backslash 9,j,k})\,. \tag{40}$$

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

$$\begin{align}
Z &= \int h(v_{j,k}, m_k^{\mathbf{V}}, v_k^{\mathbf{V}})\, dv_{j,k}\, dm_k^{\mathbf{V}}\, dv_k^{\mathbf{V}} \tag{41}\\
&= \int \mathcal{T}(v_{j,k}|m_k^{\mathbf{V}}, \frac{b_k^{v^{\mathbf{V}},\backslash 9,j,k}}{a_k^{v^{\mathbf{V}},\backslash 9,j,k}}, 2a_k^{v^{\mathbf{V}},\backslash 9,j,k}) \nonumber\\
&\qquad\qquad \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\backslash 9,j,k}, v_k^{m^{\mathbf{V}},\backslash 9,j,k})\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 9,j,k}, v_{j,k}^{v,\backslash 9,j,k})\, dv_{j,k}\, dm_k^{\mathbf{V}} \tag{42}\\
&\approx \int \mathcal{N}(v_{j,k}|m_k^{\mathbf{V}}, \frac{2b_k^{v^{\mathbf{V}},\backslash 9,j,k}}{2a_k^{v^{\mathbf{V}},\backslash 9,j,k}-2}) \nonumber\\
&\qquad\qquad \mathcal{N}(m_k^{\mathbf{V}}|m_k^{m^{\mathbf{V}},\backslash 9,j,k}, v_k^{m^{\mathbf{V}},\backslash 9,j,k})\mathcal{N}(v_{j,k}|m_{j,k}^{v,\backslash 9,j,k}, v_{j,k}^{v,\backslash 9,j,k})\, dv_{j,k}\, dm_k^{\mathbf{V}} \tag{43}\\
&\approx \mathcal{N}(m_k^{m^{\mathbf{V}},\backslash 9,j,k}|m_{j,k}^{v,\backslash 9,j,k}, v_{j,k}^{v,\backslash 9,j,k} + v_k^{m^{\mathbf{V}},\backslash 9,j,k} + \frac{2b_k^{v^{\mathbf{V}},\backslash 9,j,k}}{2a_k^{v^{\mathbf{V}},\backslash 9,j,k}-2})\,, \tag{44}
\end{align}$$

where

$$\mathcal{T}(x|\mu,\lambda,\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu\lambda}\Gamma(\nu/2)}\left[1+\frac{(x-\mu)^2}{\lambda\nu}\right]^{-(\nu+1)/2} \tag{45}$$

denotes a Student's t distribution with mean μ, variance parameter λ and degrees of freedom ν and in equation (43) 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 Cowell et al. (1996).

We approximately compute the moments of $v_k^{\mathbf{V}}$ using the following property of inverse-gamma distributions, see (2). Let $H(a,b)$ be the normalization constant of $f(x)\mathcal{IG}(x|a,b)$ for a particular function f, that is, $H(a,b) = \int f(x)\mathcal{IG}(x|a,b)\,dx$. Then we have that $\int xf(x)\mathcal{IG}(x|a,b)\,dx = H(a+1,b)a/b$ 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 $[\mathrm{H}(m,v)]^{-1}\int xf(x)\mathcal{N}(x|m,v)\,dx = m + v\frac{d\log H(m,v)}{dm}$ 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

$$[\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{\mathrm{new}} = 2b_k^{v^{\mathbf{V}},\backslash 9,j,k}/(2a_k^{v^{\mathbf{V}},\backslash 9,j,k}-2) + v_{j,k}^{v,\backslash 9,j,k}\,, \tag{46}$$

$$[\tilde{m}_k^{m^{\mathbf{V}},9,j,k}]^{\mathrm{new}} = m_{j,k}^{v,\backslash 9,j,k}\,, \tag{47}$$

$$[\tilde{v}_{j,k}^{v,9,j,k}]^{\mathrm{new}} = 2b_k^{v^{\mathbf{V}},\backslash 9,j,k}/(2a_k^{v^{\mathbf{V}},\backslash 9,j,k}-2) + v_k^{m^{\mathbf{V}},\backslash 9,j,k}\,, \tag{48}$$

$$[\tilde{m}_{j,k}^{v,9,j,k}]^{\mathrm{new}} = m_k^{m^{\mathbf{V}},\backslash 9,j,k}\,, \tag{49}$$

$$[\tilde{a}_k^{v^{\mathbf{V}},9,j,k}]^{\mathrm{new}} = a' - a_k^{v^{\mathbf{V}},\backslash 9,j,k} + 1\,, \tag{50}$$

$$[\tilde{b}_k^{v^{\mathbf{V}},9,j,k}]^{\text{new}} = b' - b_k^{v^{\mathbf{V}},\backslash 9,j,k}\,, \tag{51}$$

and we define a' and b' as

$$a' = \frac{a_k^{v^{\mathbf{V}},\backslash 9,j,k} Z_1^2}{(a_k^{v^{\mathbf{V}},\backslash 9,j,k}+1)ZZ_2 - a_k^{v^{\mathbf{V}},\backslash 9,j,k} Z_1^2}\,, \qquad b' = \frac{b_k^{v^{\mathbf{V}},\backslash 9,j,k} ZZ_1}{(a_k^{v^{\mathbf{V}},\backslash 9,j,k}+1)ZZ_2 - a_k^{v^{\mathbf{V}},\backslash 9,j,k} Z_1^2}\,, \tag{52}$$

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

$$[v_k^{m^{\mathbf{V}}}]^{\text{new}} = \left[[v_k^{m^{\mathbf{V}},\backslash 9,j,k}]^{-1} + [\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right]^{-1}\,, \tag{53}$$

$$[m_k^{m^{\mathbf{V}}}]^{\text{new}} = [v_k^{m^{\mathbf{V}}}]^{\text{new}}\left[m_k^{m^{\mathbf{V}},\backslash 9,j,k}[v_k^{m^{\mathbf{V}},\backslash 9,j,k}]^{-1} + \tilde{m}_k^{m^{\mathbf{V}},9,j,k}[\tilde{v}_k^{m^{\mathbf{V}},9,j,k}]^{-1}\right]\,, \tag{54}$$

$$[v_{j,k}^{v}]^{\text{new}} = \left[[v_{j,k}^{v,\backslash 9,j,k}]^{-1} + [\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right]^{-1}\,, \tag{55}$$

$$[m_{j,k}^{v}]^{\text{new}} = [v_{j,k}^{v}]^{\text{new}}\left[m_{j,k}^{v,\backslash 9,j,k}[v_{j,k}^{v,\backslash 9,j,k}]^{-1} + \tilde{m}_{j,k}^{v,9,j,k}[\tilde{v}_{j,k}^{v,9,j,k}]^{-1}\right]\,, \tag{56}$$

$$[a_k^{v^{\mathbf{V}}}]^{\text{new}} = a_k^{v^{\mathbf{V}},\backslash 9,j,k} + \tilde{a}_k^{v^{\mathbf{V}},9,j,k} - 1\,, \tag{57}$$

$$[b_k^{v^{\mathbf{V}}}]^{\text{new}} = b_k^{v^{\mathbf{V}},\backslash 9,j,k} + \tilde{b}_k^{v^{\mathbf{V}},9,j,k}\,, \tag{58}$$

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

1.4.10 EP updates for $\tilde{f}_{10}$

In our mapping between approximate factors and exact factors, $\tilde{f}_{10}$ approximates the factor $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}}) = \prod_{i=1}^{n}\prod_{k=1}^{h}\mathcal{N}(u_{j,k}|m_k^{\mathbf{U}},v_k^{\mathbf{U}})$. Because $p(\mathbf{U}|\mathbf{m}^{\mathbf{U}},\mathbf{v}^{\mathbf{U}})$ has a complicated form, we approximate individually each of its $n \times h$ internal factors of the form $\mathcal{N}(u_{i,k}|m_k^{\mathbf{U}},v_k^{\mathbf{U}})$, where $i = 1,\ldots,d$ and $k = 1,\ldots,h$, with extra approximate factors $\tilde{f}_{10,1,1},\ldots,\tilde{f}_{9,n,h}$. In this case, $\tilde{f}_{10}$ is given by the product of $\tilde{f}_{10,1,1},\ldots,\tilde{f}_{10,n,h}$, where all these additional approximate factors also have the same functional form as $\mathcal{Q}$. 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.

1.4.11 EP updates for $\tilde{f}_{11}$

In our mapping between approximate factors and exact factors, $\tilde{f}_{11}$ approximates the factor $p(\mathbf{C}|\mathbf{U},\mathbf{V}) = \prod_{(i,j)\in\mathcal{O}}\delta(c_{i,j} - \mathbf{u}_i^{\text{T}}\,\mathbf{v}_j)$. To refine $\tilde{f}_{11}$ we do not follow the standard EP algorithm. Instead, we use the approach used by Stern et al. (2009) and first marginalize $p(\mathbf{C}|\mathbf{U},\mathbf{V})\mathcal{Q}^{\backslash 11}(\mathbf{\Xi})$ with respect to $\mathbf{\Xi}\setminus\{\mathbf{U},\mathbf{V}\}$. To do this, we first compute the parameters of $\mathcal{Q}^{\backslash 11}(\mathbf{\Xi})$, that is,

$$[v_{j,k}^{v,\backslash 11}]^{\text{new}} = \left[[v_{j,k}^{v}]^{-1} - [\tilde{v}_{j,k}^{v,11}]^{-1}\right]^{-1}\,, \tag{59}$$

$$[m_{j,k}^{v,\backslash 11}]^{\text{new}} = [v_{j,k}^{v,\backslash 11}]^{\text{new}}\left[m_{j,k}^{v}[v_{j,k}^{v}]^{-1} - \tilde{m}_{j,k}^{v,11}[\tilde{v}_{j,k}^{v,11}]^{-1}\right]\,, \tag{60}$$

$$[v_{i,k}^{u,\backslash 11}]^{\text{new}} = \left[[v_{i,k}^{u}]^{-1} - [\tilde{v}_{i,k}^{u,11}]^{-1}\right]^{-1}\,, \tag{61}$$

$$[m_{i,k}^{u,\backslash 11}]^{\text{new}} = [v_{i,k}^{u,\backslash 11}]^{\text{new}}\left[m_{i,k}^{u}[v_{i,k}^{u}]^{-1} - \tilde{m}_{i,k}^{u,11}[\tilde{v}_{i,k}^{u,11}]^{-1}\right]\,, \tag{62}$$

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

$$[v_{i,j}^{c,\backslash 11}]^{\text{new}} = \left[[v_{i,j}^{c}]^{-1} - [\tilde{v}_{i,j}^{c,11}]^{-1}\right]^{-1}\,, \tag{63}$$

$$[m_{i,j}^{c,\setminus 11}]^{\text{new}} = [v_{i,j}^{c,\setminus 11}]^{\text{new}} \left[m_{i,j}^{c} [v_{i,j}^{c}]^{-1} - \tilde{m}_{i,j}^{c,11} [\tilde{v}_{i,j}^{c,11}]^{-1} \right] , \tag{64}$$

for $(i,j) \in \mathcal{O}$. The result of marginalizing $p(\mathbf{C}|\mathbf{U},\mathbf{V})\mathcal{Q}^{\setminus 11}(\boldsymbol{\Xi})$ with respect to $\boldsymbol{\Xi} \setminus \{\mathbf{U},\mathbf{V}\}$ is then

$$\begin{aligned} \mathcal{S}(\mathbf{U},\mathbf{V}) = \int \prod_{(i,j)\in\mathcal{O}} \delta(c_{i,j} - \mathbf{u}_i^{\text{T}}\, \mathbf{v}_j) \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(c_{i,j}|m_{i,j}^{c,\setminus 11}, v_{i,j}^{c,\setminus 11}) \right] & \\ \left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u,\setminus 11}, v_{i,k}^{u,\setminus 11} \right] \left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus 11}, v_{j,k}^{v,\setminus 11} \right] d\mathbf{C} & \qquad (65) \\ = \left[\prod_{(i,j)\in\mathcal{O}} \mathcal{N}(\mathbf{u}_i^{\text{T}}\, \mathbf{v}_j|m_{i,j}^{c,\setminus 11}, v_{i,j}^{c,\setminus 11}) \right] \left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u,\setminus 11}, v_{i,k}^{u,\setminus 11} \right] & \\ \left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_{j,k}^{v,\setminus 11}, v_{j,k}^{v,\setminus 11} \right] . & \qquad (66) \end{aligned}$$

Let $\mathcal{Q}_{\mathbf{U},\mathbf{V}}$ be the posterior approximation $\mathcal{Q}$ after marginalizing $\boldsymbol{\Xi} \setminus \{\mathbf{U},\mathbf{V}\}$ out, that is,

$$\mathcal{Q}_{\mathbf{U},\mathbf{V}} = \left[\prod_{i=1}^{n}\prod_{k=1}^{h} \mathcal{N}(u_{i,k}|m_{i,k}^{u}, v_{i,k}^{u} \right] \left[\prod_{j=1}^{d}\prod_{k=1}^{h} \mathcal{N}(v_{j,k}|m_{j,k}^{v}, v_{j,k}^{v} \right] . \tag{67}$$

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 $\text{KL}(\mathcal{Q}_{\mathbf{U},\mathbf{V}} \| \mathcal{S})$. This can be done very efficiently using the gradient descent method described by Raiko et al. (2007). 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 $\mathbf{C}$. To do this, we note that in the exact posterior $c_{i,j}$ is always equal to $\mathbf{u}_i^{\text{T}}\, \mathbf{v}_j$ because of the delta function $\delta(c_{i,j} - \mathbf{u}_i^{\text{T}}\, \mathbf{v}_j)$. Therefore, we set the mean and variance of each $c_{i,j}$ in $\mathcal{Q}$ to be the same as the mean and variance of the corresponding $\mathbf{u}_i^{\text{T}}\, \mathbf{v}_j$ according to the newly updated $\mathcal{Q}$. This leads to the update

$$[m_{i,j}^{c}]^{\text{new}} = \sum_{k=1}^{h} m_{i,k}^{u} m_{j,k}^{v} \,, \qquad [v_{i,j}^{c}]^{\text{new}} = \sum_{k=1}^{h} [m_{i,k}^{u}]^2 v_{j,k}^{v} + v_{i,k}^{u} [m_{j,k}^{v}]^2 + v_{i,k}^{u} v_{j,k}^{v} \,. \tag{68}$$

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 $\mathcal{Q}^{\setminus 11}$, that is,

$$[\tilde{v}_{j,k}^{v,11}]^{\text{new}} = \left[[v_{j,k}^{v}]^{-1} - [v_{j,k}^{v,\setminus 11}]^{-1} \right]^{-1} , \tag{69}$$

$$[\tilde{m}_{j,k}^{v,11}]^{\text{new}} = [\tilde{v}_{j,k}^{v,11}]^{\text{new}} \left[m_{j,k}^{v} [v_{j,k}^{v}]^{-1} - m_{j,k}^{v,\setminus 11} [v_{j,k}^{v,\setminus 11}]^{-1} \right] , \tag{70}$$

$$[\tilde{v}_{i,k}^{u,11}]^{\text{new}} = \left[[v_{i,k}^{u}]^{-1} - [v_{i,k}^{u,\setminus 11}]^{-1} \right]^{-1} , \tag{71}$$

$$[\tilde{m}_{i,k}^{u,11}]^{\text{new}} = [\tilde{v}_{i,k}^{u,11}]^{\text{new}} \left[m_{i,k}^{u} [v_{i,k}^{u}]^{-1} - m_{i,k}^{u,\setminus 11} [v_{i,k}^{u,\setminus 11}]^{-1} \right] , \tag{72}$$

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

$$[\tilde{v}_{i,j}^{c,11}]^{\text{new}} = \left[[v_{i,j}^{c}]^{-1} - [v_{i,j}^{c,\setminus 11}]^{-1} \right]^{-1} , \tag{73}$$

$$[\tilde{m}_{i,j}^{c,11}]^{\text{new}} = [\tilde{v}_{i,j}^{c,11}]^{\text{new}} \left[m_{i,j}^{c} [v_{i,j}^{c}]^{-1} - m_{i,j}^{c,\setminus 11} [v_{i,j}^{c,\setminus 11}]^{-1} \right] , \tag{74}$$

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 $\mathcal{Q}$.

1.4.12 EP updates for $\tilde{f}_{12}$

The approximate factor $\tilde{f}_{12}$ approximates the exact factor $p(\mathbf{A}|\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}}) = \prod_{(i,j)\in\mathcal{O}} \mathcal{N}(a_{i,j}|c_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})$. Because $p(\mathbf{A}|\mathbf{C},\boldsymbol{\gamma}^{\text{row}},\boldsymbol{\gamma}^{\text{col}})$ has a complicated form, we approximate individually each of its internal factors of the form $\mathcal{N}(a_{i,j}|c_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})$ with an extra approximate factor $\tilde{f}_{12,i,j}$, for $(i,j) \in \mathcal{O}$. In this case, $\tilde{f}_{12}$ is given by the product of all the additional approximate factors $\tilde{f}_{12,i,j}$, which also have the same functional form as $\mathcal{Q}$. Initially, all the $\tilde{f}_{12,i,j}$ and $\tilde{f}_{12}$ are non-informative or flat. EP will iteratively refine each of the extra approximate factors as follows. To refine $\tilde{f}_{12,i,j}$, we firstly compute the parameters of $\mathcal{Q}^{\backslash 12,i,j}$. This distribution is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{12,i,j}$. This leads to

$$
\begin{aligned}
[v_{i,j}^{a,\backslash 12,i,j}]^{\text{new}} &= \left[[v_{i,j}^{a}]^{-1} - [\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]^{-1}\,, &(75)\\
[m_{i,j}^{a,\backslash 12,i,j}]^{\text{new}} &= [v_{i,j}^{a,\backslash 12,i,j}]^{\text{new}}\left[m_{i,j}^{a}[v_{i,j}^{a}]^{-1} - \tilde{m}_{i,j}^{a,12,i,j}[\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]\,, &(76)\\
[v_{i,j}^{c,\backslash 12,i,j}]^{\text{new}} &= \left[[v_{i,j}^{c}]^{-1} - [\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right]^{-1}\,, &(77)\\
[m_{i,j}^{c,\backslash 12,i,j}]^{\text{new}} &= [v_{i,j}^{c,\backslash 12,i,j}]^{\text{new}}\left[m_{i,j}^{c}[v_{i,j}^{c}]^{-1} - \tilde{m}_{i,j}^{c,12,i,j}[\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right]\,, &(78)\\
[a_i^{\gamma^{\text{row}},\backslash 12,i,j}]^{\text{new}} &= a_i^{\gamma^{\text{row}}} - \tilde{a}_i^{\gamma^{\text{row}},12,i,j} + 1\,, &(79)\\
[b_i^{\gamma^{\text{row}},\backslash 12,i,j}]^{\text{new}} &= b_i^{\gamma^{\text{row}}} - \tilde{b}_i^{\gamma^{\text{row}},12,i,j}\,, &(80)\\
[a_j^{\gamma^{\text{col}},\backslash 12,i,j}]^{\text{new}} &= a_j^{\gamma^{\text{col}}} - \tilde{a}_j^{\gamma^{\text{col}},12,i,j} + 1\,, &(81)\\
[b_j^{\gamma^{\text{col}},\backslash 12,i,j}]^{\text{new}} &= b_j^{\gamma^{\text{col}}} - \tilde{b}_j^{\gamma^{\text{col}},12,i,j}\,. &(82)
\end{aligned}
$$

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

$$
\begin{aligned}
h(a_{i,j},c_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}}) = {}& \mathcal{N}(a_{i,j}|c_{i,j},\gamma_i^{\text{row}}\gamma_j^{\text{col}})\mathcal{N}(a_{i,j}|m_{i,j}^{a,\backslash 12,i,j},v_{i,j}^{a,\backslash 12,i,j})\mathcal{N}(c_{i,j}|m_{i,j}^{c,\backslash 12,i,j},v_{i,j}^{c,\backslash 12,i,j})\\
& \text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\backslash 12,i,j},b_i^{\gamma^{\text{row}},\backslash 12,i,j})\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\backslash 12,i,j},b_j^{\gamma^{\text{col}},\backslash 12,i,j})\,. \qquad (83)
\end{aligned}
$$

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

$$
\begin{aligned}
Z &= \int h(a_{i,j},c_{i,j},\gamma_i^{\text{row}},\gamma_j^{\text{col}})\,da_{i,j}\,dc_{i,j}\,d\gamma_i^{\text{row}}\,d\gamma_j^{\text{col}} &(84)\\
&= \int \mathcal{N}(m_{i,j}^{a,\backslash 12,i,j}|m_{i,j}^{c,\backslash 12,i,j},v_{i,j}^{a,\backslash 12,i,j} + v_{i,j}^{c,\backslash 12,i,j} + \gamma_i^{\text{row}}\gamma_j^{\text{col}}) &(85)\\
&\quad \text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\backslash 12,i,j},b_i^{\gamma^{\text{row}},\backslash 12,i,j})\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\backslash 12,i,j},b_j^{\gamma^{\text{col}},\backslash 12,i,j})d\gamma_i^{\text{row}}\,d\gamma_j^{\text{col}} &(86)\\
&\approx \mathcal{N}(m_{i,j}^{a,\backslash 12,i,j}|m_{i,j}^{c,\backslash 12,i,j},v_{i,j}^{a,\backslash 12,i,j} + v_{i,j}^{c,\backslash 12,i,j} + \\
&\qquad b_i^{\gamma^{\text{row}},\backslash 12,i,j}b_j^{\gamma^{\text{col}},\backslash 12,i,j}/[(a_i^{\gamma^{\text{row}},\backslash 12,i,j}+1)(a_j^{\gamma^{\text{col}},\backslash 12,i,j}+1)])\,, &(87)
\end{aligned}
$$

where in (87) we have approximated $\text{IG}(\gamma_i^{\text{row}}|a_i^{\gamma^{\text{row}},\backslash 12,i,j},b_i^{\gamma^{\text{row}},\backslash 12,i,j})$ and $\text{IG}(\gamma_j^{\text{col}}|a_j^{\gamma^{\text{col}},\backslash 12,i,j},b_j^{\gamma^{\text{col}},\backslash 12,i,j})$ 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 Cowell et al. (1996).

We approximately compute the moments of the random variables γ_i^{row} and γ_j^{col}, using the following property of inverse-gamma distributions, see (2). Let $H(a,b)$ be the normalization constant of $f(x)\mathcal{IG}(x|a,b)$ for a particular function 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 $\int x^2 \mathcal{IG}(x|a,b)\,dx = H(a+2,b)a(a+1)/b^2$. 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 it can be shown 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 updates for $\tilde{f}_{12,i,j}$ are then

$$[\tilde{m}_{i,j}^{a,12,i,j}]^{\text{new}} = m_{i,j}^{c,\setminus 12,i,j}\,, \tag{88}$$

$$[\tilde{v}_{i,j}^{a,12,i,j}]^{\text{new}} = v_{i,j}^{c,\setminus 12,i,j} + b_i^{\gamma^{\text{row}},\setminus 12,i,j} b_j^{\gamma^{\text{col}},\setminus 12,i,j} / [(a_i^{\gamma^{\text{row}},\setminus 12,i,j}+1)(a_j^{\gamma^{\text{col}},\setminus 12,i,j}+1)]\,, \tag{89}$$

$$[\tilde{m}_{i,j}^{c,12,i,j}]^{\text{new}} = m_{i,j}^{a,\setminus 12,i,j}\,, \tag{90}$$

$$[\tilde{v}_{i,j}^{c,12,i,j}]^{\text{new}} = v_{i,j}^{a,\setminus 12,i,j} + b_i^{\gamma^{\text{row}},\setminus 12,i,j} b_j^{\gamma^{\text{col}},\setminus 12,i,j} / [(a_i^{\gamma^{\text{row}},\setminus 12,i,j}+1)(a_j^{\gamma^{\text{col}},\setminus 12,i,j}+1)]\,, \tag{91}$$

$$[\tilde{a}_i^{\gamma^{\text{row}},12,i,j}]^{\text{new}} = a'_{\text{row}} - a_i^{\gamma^{\text{row}},\setminus 12,i,j} + 1\,, \tag{92}$$

$$[\tilde{b}_i^{\gamma^{\text{row}},12,i,j}]^{\text{new}} = b'_{\text{row}} - b_i^{\gamma^{\text{row}},\setminus 12,i,j}\,, \tag{93}$$

$$[\tilde{a}_j^{\gamma^{\text{col}},12,i,j}]^{\text{new}} = a'_{\text{col}} - a_j^{\gamma^{\text{col}},\setminus 12,i,j} + 1\,, \tag{94}$$

$$[\tilde{b}_j^{\gamma^{\text{col}},12,i,j}]^{\text{new}} = b'_{\text{col}} - b_j^{\gamma^{\text{col}},\setminus 12,i,j}\,, \tag{95}$$

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

$$a'_{\text{row}} = \frac{a_i^{\gamma^{\text{row}},\setminus 12,i,j}[Z_1^{\text{row}}]^2}{(a_i^{\gamma^{\text{row}},\setminus 12,i,j}+1)ZZ_2^{\text{row}} - a_i^{\gamma^{\text{row}},\setminus 12,i,j}[Z_1^{\text{row}}]^2}\,, \tag{96}$$

$$b'_{\text{row}} = \frac{b_i^{\gamma^{\text{row}},\setminus 12,i,j}ZZ_1^{\text{row}}}{(a_i^{\gamma^{\text{row}},\setminus 12,i,j}+1)ZZ_2^{\text{row}} - a_i^{\gamma^{\text{row}},\setminus 12,i,j}[Z_1^{\text{row}}]^2}\,, \tag{97}$$

$$a'_{\text{col}} = \frac{a_j^{\gamma^{\text{col}},\setminus 12,i,j}[Z_1^{\text{col}}]^2}{(a_j^{\gamma^{\text{col}},\setminus 12,i,j}+1)ZZ_2^{\text{col}} - a_j^{\gamma^{\text{col}},\setminus 12,i,j}[Z_1^{\text{col}}]^2}\,, \tag{98}$$

$$b'_{\text{col}} = \frac{b_j^{\gamma^{\text{col}},\setminus 12,i,j}ZZ_1^{\text{col}}}{(a_j^{\gamma^{\text{col}},\setminus 12,i,j}+1)ZZ_2^{\text{col}} - a_j^{\gamma^{\text{col}},\setminus 12,i,j}[Z_1^{\text{col}}]^2}\,, \tag{99}$$

Z_1^{row} and Z_2^{row} are obtained in the same way as the normalization constant Z, but increasing $a_i^{\gamma^{\text{row}},\setminus 12,i,j}$ in one and two units during the computations, respectively, and similarly, Z_1^{col} and Z_2^{col} are obtained by decreasing $a_j^{\gamma^{\text{col}},\setminus 12,i,j}$ 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 $a_j^{\gamma^{\text{col}},\setminus 12,i,j} > 2$, $b_j^{\gamma^{\text{col}},\setminus 12,i,j} > 0$, $a_i^{\gamma^{\text{row}},\setminus 12,i,j} > 2$, $b_i^{\gamma^{\text{ro}},\setminus 12,i,j} > 0$, $v_{i,j}^{a,\setminus 12,i,j} > 0$ and $v_{i,j}^{c,\setminus 12,i,j} > 0$ are satisfied.

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

$$[v_{i,j}^a]^{\text{new}} = \left[[v_{i,j}^{a,\setminus 12,i,j}]^{-1} + [\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right]^{-1} , \quad (100)$$

$$[m_{i,j}^a]^{\text{new}} = [v_{i,j}^a]^{\text{new}} \left[m_{i,j}^{a,\setminus 12,i,j}[v_{i,j}^{a,\setminus 12,i,j}]^{-1} + \tilde{m}_{i,j}^{a,12,i,j}[\tilde{v}_{i,j}^{a,12,i,j}]^{-1}\right] , \quad (101)$$

$$[v_{i,j}^c]^{\text{new}} = \left[[v_{i,j}^{c,\setminus 12,i,j}]^{-1} + [\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right]^{-1} , \quad (102)$$

$$[m_{i,j}^c]^{\text{new}} = [v_{i,j}^c]^{\text{new}} \left[m_{i,j}^{c,\setminus 12,i,j}[v_{i,j}^{c,\setminus 12,i,j}]^{-1} + \tilde{m}_{i,j}^{c,12,i,j}[\tilde{v}_{i,j}^{c,12,i,j}]^{-1}\right] , \quad (103)$$

$$[a_i^{\gamma^{\text{row}}}]^{\text{new}} = a_i^{\gamma^{\text{row}},\setminus 12,i,j} + \tilde{a}_i^{\gamma^{\text{row}},12,i,j} - 1 , \quad (104)$$

$$[b_i^{\gamma^{\text{row}}}]^{\text{new}} = b_i^{\gamma^{\text{row}},\setminus 12,i,j} + \tilde{b}_i^{\gamma^{\text{row}},12,i,j} , \quad (105)$$

$$[a_j^{\gamma^{\text{col}}}]^{\text{new}} = a_j^{\gamma^{\text{col}},\setminus 12,i,j} + \tilde{a}_j^{\gamma^{\text{col}},12,i,j} - 1 , \quad (106)$$

$$[b_j^{\gamma^{\text{col}}}]^{\text{new}} = b_j^{\gamma^{\text{col}},\setminus 12,i,j} + \tilde{b}_j^{\gamma^{\text{col}},12,i,j} . \quad (107)$$

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 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.

1.4.13 EP updates for $\tilde{f}_{13}$

In our mapping between approximate factors and exact factors, $\tilde{f}_{13}$ approximates the factor $p(\mathbf{R}^{\mathcal{O}}|\mathbf{A},\mathbf{B}) = \prod_{(i,j)\in\mathcal{O}}\prod_{k=1}^{L-1}\Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]$. Because $p(\mathbf{R}^{\mathcal{O}}|\mathbf{A},\mathbf{B})$ does not have a simple form, we approximate individually each of its $|\mathcal{O}|\times(L-1)$ internal factors of the form $\Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]$ with an extra approximate factor $\tilde{f}_{13,i,j,k}$, for $(i,j)\in\mathcal{O}$ and $k = 1,\ldots,L-1$. In this case, $\tilde{f}_{13}$ is given by the product of all the additional approximate factors $\tilde{f}_{13,i,j,k}$, which also have the same functional form as $\mathcal{Q}$. Initially, all the $\tilde{f}_{13,i,j,k}$ and $\tilde{f}_{13}$ are non-informative or flat. EP will iteratively refine each of the extra approximate factors as follows. To refine $\tilde{f}_{13,i,j,k}$, we firstly compute the parameters of $\mathcal{Q}^{\setminus 13,i,j,k}$. This distribution is defined as the normalized ratio of $\mathcal{Q}$ and $\tilde{f}_{13,i,j,k}$. This leads to

$$[v_{j,k}^{b,\setminus 13,i,j,k}]^{\text{new}} = \left[[v_{j,k}^b]^{-1} - [\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right]^{-1} , \quad (108)$$

$$[m_{j,k}^{b,\setminus 13,i,j,k}]^{\text{new}} = [v_{j,k}^{b,\setminus 13,i,j,k}]^{\text{new}} \left[m_{j,k}^b[v_{j,k}^b]^{-1} - \tilde{m}_{j,k}^{b,13,i,j,k}[\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right] , \quad (109)$$

$$[v_{i,j}^{a,\setminus 13,i,j,k}]^{\text{new}} = \left[[v_{i,j}^a]^{-1} - [\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right]^{-1} , \quad (110)$$

$$[m_{i,j}^{a,\setminus 13,i,j,k}]^{\text{new}} = [v_{i,j}^{a,\setminus 13,i,j,k}]^{\text{new}} \left[m_{i,j}^a[v_{i,j}^a]^{-1} - \tilde{m}_{i,j}^{a,13,i,j,k}[\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right] . \quad (111)$$

After this, we update the parameters of $\tilde{f}_{13,i,j,k}$ by minimizing the KL-divergence between the unnormalized distributions $\mathcal{Q}^{\setminus 13,i,j,k}(\boldsymbol{\Xi})\Theta\left[\text{sign}[r_{i,j} - k - 0.5](a_{i,j} - b_{j,k})\right]$ and $\mathcal{Q}^{\setminus 13,i,j,k}(\boldsymbol{\Xi})\tilde{f}_{13,i,j,k}(\boldsymbol{\Xi})$. This leads to the updates

$$\tilde{m}_{j,k}^{b,13,i,j,k} = m_{j,k}^{b,\backslash 13,i,j,k} + \kappa \qquad \tilde{v}_{j,k}^{b,13,i,j,k} = -v_{j,k}^{b,\backslash 13,i,j,k} - 1/\beta \tag{112}$$

$$\tilde{m}_{i,j}^{a,13,i,j,k} = m_{i,j}^{a,\backslash 13,i,j,k} - \kappa \qquad \tilde{v}_{i,j}^{a,13,i,j,k} = -v_{i,j}^{a,\backslash 13,i,j,k} - 1/\beta \tag{113}$$

where β and κ are given by

$$\beta = -\frac{\phi(\alpha)}{\Phi(\alpha)}\left(\alpha + \frac{\phi(\alpha)}{\Phi(\alpha)}\right)\left[v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}\right]^{-1}, \tag{114}$$

$$\kappa = -\frac{\mathrm{sign}[r_{i,j} - k - 0.5]}{\sqrt{v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}}}\left[\alpha + \frac{\phi(\alpha)}{\Phi(\alpha)}\right]^{-1}, \tag{115}$$

with

$$\alpha = \mathrm{sign}[r_{i,j} - k - 0.5]\frac{m_{j,k}^{a,\backslash 13,i,j,k} - m_{j,k}^{b,\backslash 13,i,j,k}}{\sqrt{v_{j,k}^{a,\backslash 13,i,j,k} + v_{j,k}^{b,\backslash 13,i,j,k}}} \tag{116}$$

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

$$[v_{j,k}^{b}]^{\mathrm{new}} = \left[[v_{j,k}^{b,\backslash 13,i,j,k}]^{-1} + [\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right]^{-1}, \tag{117}$$

$$[m_{j,k}^{b}]^{\mathrm{new}} = [v_{j,k}^{b}]^{\mathrm{new}}\left[m_{j,k}^{b,\backslash 13,i,j,k}[v_{j,k}^{b,\backslash 13,i,j,k}]^{-1} + \tilde{m}_{j,k}^{b,13,i,j,k}[\tilde{v}_{j,k}^{b,13,i,j,k}]^{-1}\right], \tag{118}$$

$$[v_{i,j}^{a}]^{\mathrm{new}} = \left[[v_{i,j}^{a,\backslash 13,i,j,k}]^{-1} + [\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right]^{-1}, \tag{119}$$

$$[m_{i,j}^{a}]^{\mathrm{new}} = [v_{i,j}^{a}]^{\mathrm{new}}\left[m_{i,j}^{a,\backslash 13,i,j,k}[v_{i,j}^{a,\backslash 13,i,j,k}]^{-1} + \tilde{m}_{i,j}^{a,13,i,j,k}[\tilde{v}_{i,j}^{a,13,i,j,k}]^{-1}\right]. \tag{120}$$

2 Data

2.1 Dataset descriptions

We perform experiments using seven datasets consisting of ratings. These come from a diverse set of domains. Unless otherwise stated, the ratings in each dataset were ordinal valued, in the range $1, \ldots, 5$.

- *MovieLens100K* and *MovieLens1M* are collections of ratings for movies, commonly used for benchmarking recommendation systems. These are available at `http://grouplens.org/datasets/movielens/`.

- *MovieTweets* has been released recently, and consists of ratings for movies collected from Tweets. Details of the dataset can be found in Dooms et al. (2013), and the data is available from `https://github.com/sidooms/MovieTweetings`. The original ratings took values in $\{0 \ldots, 10\}$. We map the original ratings to values in $\{1, \ldots, 5\}$ as follows: $\{0, 1, 2\} \to 1$, $\{3, 4\} \to 2$, $\{5, 6\} \to 3$, $\{7, 8\} \to 4$, $\{9, 10\} \to 5$.

- *Webscope* is a collection of ratings for songs. It is available for research upon request from Yahoo! Labs. We used the ‘R3’ dataset from `http://webscope.sandbox.yahoo.com/catalog.php?datatype`.

- *Jester* is a collection of ratings for jokes, available at `http://goldberg.berkeley.edu/jester-data/`. The ratings on this dataset are real valued $\in [-10, 10]$. We convert these to ordinal ratings by grouping the values into 5 bins with equal counts.
- *Book* is a set of ratings for books, publicly available from `http://www.informatik.uni-freiburg.de/~cziegler/BX/` The ratings take values $1, \dots, 10$. Most of the ratings take value higher than 6, so we merged the ratings $1, \dots, 6$ to yield 5 values in total.
- *IPIP* contains responses to a 336 item International Item Pool questionnaire Goldberg et al. (2006). These data were collected from Facebook Kosinski et al. (2013) and are available for research upon request at `http://mypersonality.org/wiki/doku.php?id=start`. This dataset is dense, that is, all of the ratings are observed. All of the other datasets contained many missing entries.

2.2 Data pre-processing

Some of the datasets are very sparse, so we selected only users and items that have 10 ratings or more, as proposed in Rendle et al. (2009). This formed the set of ratings that we used for the model-comparison experiments described in Section 5.1 of the main document. In these experiments, the ratings were split randomly into training and test sets containing 80% and 20% of the ratings, respectively.

For the active learning experiments in Section 5.2 with new users (the -U datasets) we selected the 2000 users and 1000 items with the most ratings, up to the maximum number available. This was to provide the largest possible pool for active sampling, since in a real-world setting the system can request any rating. As described in the main document, of these 2000 users, 75% were sampled randomly as the users 'already in the system' and all of their ratings were added to the training set. Then, one rating from each of the remaining 25% 'test users' was added to the training set. For each test user, 3 ratings were randomly held out in a test set for evaluation of predictive performance. The remaining ratings for the test users were added to the pool set. In each round of active sampling, a single item was selected from the pool for each user, these items were then added to the training set. After this, the model was incrementally retrained and evaluated on the test set. For the new-item experiments (-I), we followed the same procedure, except that the roles of the users and items were interchanged. In all of our experiments the dataset splits were re-sampled for each fold.

3 All Results

3.1 Learning curves

Figure 1 shows the log likelihood learning curves in the cold-start experiments when using the full heteroskedastic model (HOMF). This figure includes all the experiments in which we select items for new users (-U) and select users to rate new items (-I). Figure 2 shows the same for the homoskedastic model (OMF). All results are summarized in Table 3. Overall, with both models, BALD is the best performing algorithm. Entropy sampling and its model-free version, Emp-Ent, often perform poorly, and are even outperformed frequently by random sampling. This indicates that they often seek noisy, and hence uninformative, users or items. In many cases, such as Book-U, Movielens100k-U, Movielens100k-I, MovieTweets-I, IPIP-I, Webscope-U and Webscope-I, the performance gap between BALD and the alternative strategies is decreased substantially when using the homoskedastic model. This implies that to obtain robust performance with Bayesian active learning it is important to model all sources of uncertainty appropriately.

3.2 Root mean squared error

We also evaluated the performance of our model and active learning algorithm using the root mean squared error (RMSE) of the posterior mean prediction. RMSE is a popular metric for models that only produce point estimates. For probabilistic models it is a less informative metric than log likelihood because it measures only the quality of the mean rating, and does not consider the model's confidence in its predictions. Furthermore, unlike log likelihood, it is not invariant to the (normally arbitrary) assignment of numeric values $(1, \ldots, 5)$ to ordinal valued ratings. Tables 1 and 2 contain the log likelihood and RMSE for the model-comparison experiments, respectively. Tables 3 and 4 contain the results for the active learning experiments.

Table 2 shows that OMF performs best when evaluated using RMSE, although the improvement over HOMF is normally very small, the difference in RMSE is smaller than 0.005 in all but one case. Although learning heteroskedastic noise is crucial for assessing confidence correctly, as indicated by Table 1 where HOMF performs best in all cases, incorporating heteroskedasticity does not change the mean effect. We speculate that the small improvement of OMF over HOMF is due to the fast that OMF has fewer parameters to learn. Learning the bias parameters and using an ordinal rather than a Gaussian likelihood yields improved performance when evaluating with both RMSE and log likelihood.

A similar effect is observed in the active learning experiments. With log likelihood HOMF+BALD outperforms OMF+BALD in 15 out of 16 cases, and draws on the last. With RMSE, HOMF+BALD outperforms OMF+BALD 5 times, draws 9 times and loses twice. Again, this indicates that heteroskedasticity does not change the mean effect. However, with RMSE the heteroskedasticity is still important to achieve robust active learning. This is indicated by the fact that in Table 4, overall, HOMF+BALD outperforms OMF+BALD but HOMF+Rand loses to OMF+Rand. Furthermore, within the HOMF model, BALD outperforms Rand in more cases than it does with OMF. This indicates that although the heteroskedasticity does not improve the final evaluation, it is necessary for selecting good samples with BALD during active learning.

3.3 Approximation Losses

Figure 3 depicts, for all datasets, the information loss (Equation (8) in the main text) due to each of the approximations made to compute the second term in BALD, $\mathbb{E}_{\mathcal{Q}(\mathbf{u}_i)}\mathrm{H}[p(r^{\star}_{i,j}|\mathbf{u}_i)]$. In all 16 cases the information loss from using 100 Monte Carlo samples of $\mathbf{u}_i$ from $\mathcal{Q}$ to approximate the integral is smaller than 5%. In all but 3 cases using the unscented approximation results in less than 5% loss.

Table 1: Average test Log likelihood. Bold denotes best method, and those statistically indistinguishable.

Method	**HOMF**	**OMF**	**HOMF**	**OMF**	**Paquet**	**RBMF**	**BMF**	**BMM**
			-NoB	**-NoB**				
Books	**-1.415**	-1.436	-1.507	-1.439	-1.427	-1.545	-1.544	-1.622
Dating	**-0.867**	-0.906	-0.890	-1.028	-1.009	-1.045	-1.140	-0.948
IPIP	**-1.096**	-1.140	-1.131	-1.189	-1.188	-1.194	-1.225	-1.270
Jest	**-1.238**	-1.306	**-1.240**	-1.320	-1.320	-1.312	-1.368	-1.290
ML1M	**-1.136**	-1.165	-1.141	-1.177	-1.170	-1.173	-1.210	-1.324
ML100K	**-1.203**	-1.234	-1.208	-1.243	-1.232	-1.238	-1.277	-1.493
MTweet	**-0.956**	-0.991	-0.984	-1.025	-1.012	-1.014	-1.077	-1.115
WebSc.	**-1.207**	-1.253	-1.209	-1.257	-1.236	-1.529	-1.532	-1.298

Table 2: Average Test RMSE.

Method	**HOMF**	**OMF**	**HOMF-NoB**	**OMF-NoB**	**Paquet**	**RBMF**	**BMF**	**BMM**
Books	1.207	**1.204**	1.246	1.204	1.214	1.281	1.280	1.390
Dating	0.822	**0.821**	0.823	0.836	0.829	0.825	0.838	0.913
IPIP	0.886	**0.885**	0.887	0.887	0.887	0.893	0.895	1.046
Jester	1.019	**1.006**	1.015	1.008	1.009	1.016	1.015	1.078
MLens1M	0.838	0.836	0.839	0.837	**0.836**	0.842	0.847	0.965
MLens100K	**0.895**	**0.894**	**0.895**	0.895	**0.895**	0.898	0.903	1.077
MTweet	0.699	**0.698**	0.701	**0.699**	0.703	0.712	0.722	0.817
WebScope	1.200	1.195	1.201	1.195	**1.185**	1.215	1.218	1.283

Table 3: Log likelihood after collecting 10 active samples per user (-U) or item (-I). Underlining indicates the top performing active sampling algorithms for each model, and bold denotes the best overall method. The bottom row gives the number of datasets that each active learning strategy yields the best (or joint best) performance with each model.

	Heteroscedastic (HOMF)				**Homoscedastic (OMF)**				**BMM**			
Dataset	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**
Book-U	**-2122**	-2129	-2129	-2126	-2146	-2149	-2150	-2147	-2405	-2418	-2413	-2411
Dating-U	**-1214**	-1239	-1241	-1248	-1217	-1230	-1235	-1244	-1234	-1309	-1305	-1255
IPIP-U	**-1944**	-1977	-1960	-1967	**-1945**	-1978	-1964	-1973	-1964	-1988	-1983	-1987
Jester-U	-2051	-2095	-2070	-2064	-2080	-2119	-2100	-2099	**-2041**	-2075	-2054	**-2045**
MLens100k-U	**-918**	-928	-926	**-920**	-926	-927	-929	-926	-989	-1001	-997	-988
MLens1M-U	**-1831**	-1843	-1844	**-1835**	-1840	-1850	-1854	-1846	-1877	-1899	-1898	-1879
MTweets-U	**-1467**	-1475	-1475	**-1471**	-1503	-1508	-1508	-1503	-1608	-1624	-1622	-1613
Webscope-U	**-1837**	-1869	-1869	-1846	-1882	-1898	-1903	-1880	-1951	-1984	-1970	-1958
Book-I	**-2038**	-2039	**-2037**	**-2038**	-2095	-2094	-2094	-2095	-2186	-2198	-2202	-2195
Dating-I	-1630	-1720	-1655	**-1612**	-1672	-1722	-1684	-1643	**-1603**	-1691	-1631	**-1602**
IPIP-I	**-319**	-325	-339	-329	-325	-325	-339	-330	-335	-347	-346	-339
Jester-I	**-99**	**-99**	**-99**	**-100**	-102	-102	-101	-102	-104	-107	-106	-104
Mlens100k-I	**-1085**	-1103	-1095	-1099	-1110	-1112	-1111	-1113	-1160	-1186	-1171	-1170
Mlens1M-I	**-1831**	-1843	-1844	**-1835**	-1840	-1850	-1854	-1846	-1877	-1899	-1898	-1879
MTweets-I	**-1470**	-1479	-1475	-1476	-1519	-1520	-1520	-1520	-1605	-1617	-1613	-1608
Webscope-I	**-1837**	-1869	-1869	-1846	-1882	-1898	-1903	-1880	-1951	-1984	-1970	-1958
Wins / 16	15	1	2	7	15	7	5	12	16	1	2	12

Table 4: RMSE after collecting 10 active samples.

	Heteroscedastic (HOMF)				**Homoscedastic (OMF)**				**BMM**			
Dataset	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**	**BALD**	**Entro**	**Emp-Ent**	**Rand**
Book-U	**1.185**	1.188	1.189	**1.187**	**1.186**	1.188	1.189	**1.187**	1.345	1.352	1.353	1.350
Dating-U	**0.768**	0.788	0.790	0.795	**0.769**	0.783	0.788	0.794	0.789	0.841	0.838	0.807
IPIP-U	1.033	1.058	1.047	1.055	**1.030**	1.051	1.042	1.050	1.063	1.080	1.075	1.085
Jester-U	**1.089**	1.119	1.103	1.103	**1.086**	1.110	1.100	1.101	1.121	1.140	1.130	1.129
MLens100k-U	**0.968**	0.983	0.979	**0.974**	0.975	**0.973**	0.977	**0.973**	1.047	1.054	1.053	1.043
MLens1M-U	**0.888**	0.897	0.899	0.894	**0.890**	0.895	0.898	0.894	0.916	0.926	0.926	0.917
MTweets-U	**0.704**	0.708	0.708	0.706	**0.703**	**0.704**	**0.704**	**0.703**	0.777	0.775	0.783	0.774
Webscope-U	**1.192**	1.213	1.217	1.201	**1.199**	1.205	1.216	**1.198**	1.290	1.313	1.311	1.291
Book-I	1.175	1.175	**1.174**	**1.175**	**1.175**	**1.174**	**1.174**	**1.175**	1.250	1.256	1.258	1.254
Dating-I	**0.910**	0.962	0.941	0.914	0.924	0.951	0.937	**0.909**	0.966	1.019	0.989	0.963
IPIP-I	**1.039**	1.066	1.121	1.088	1.058	1.059	1.122	1.089	1.102	1.163	1.155	1.125
Jester-I	**1.086**	**1.100**	**1.095**	**1.108**	**1.105**	**1.101**	**1.096**	1.113	1.155	1.175	1.176	1.162
Mlens100k-I	**0.943**	0.960	0.955	0.957	0.953	0.954	0.954	0.957	1.004	1.030	1.016	1.015
Mlens1M-I	**0.888**	0.897	0.899	0.894	**0.890**	0.895	0.898	0.894	0.916	0.926	0.926	0.917
MTweets-I	**0.721**	0.725	0.724	0.724	0.725	0.725	0.725	0.725	0.768	0.774	0.774	0.769
Webscope-I	**1.192**	1.213	1.217	1.201	**1.199**	1.205	1.216	**1.198**	1.290	1.313	1.311	1.291
Wins /16	15	1	2	6	15	10	5	9	16	2	0	11

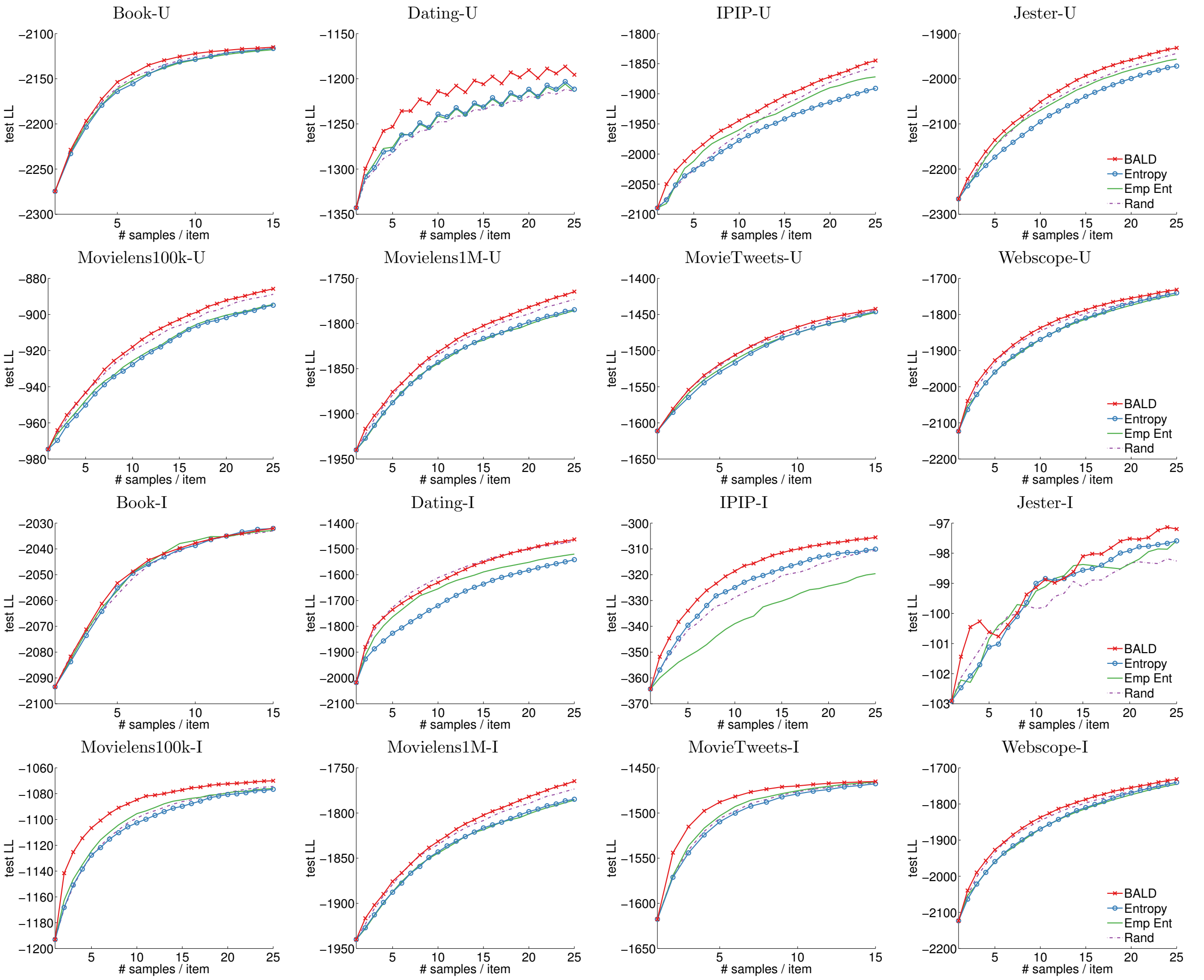

Figure 1: Log likelihood on the test users versus number of active samples selected by each algorithm with the HOMF model.

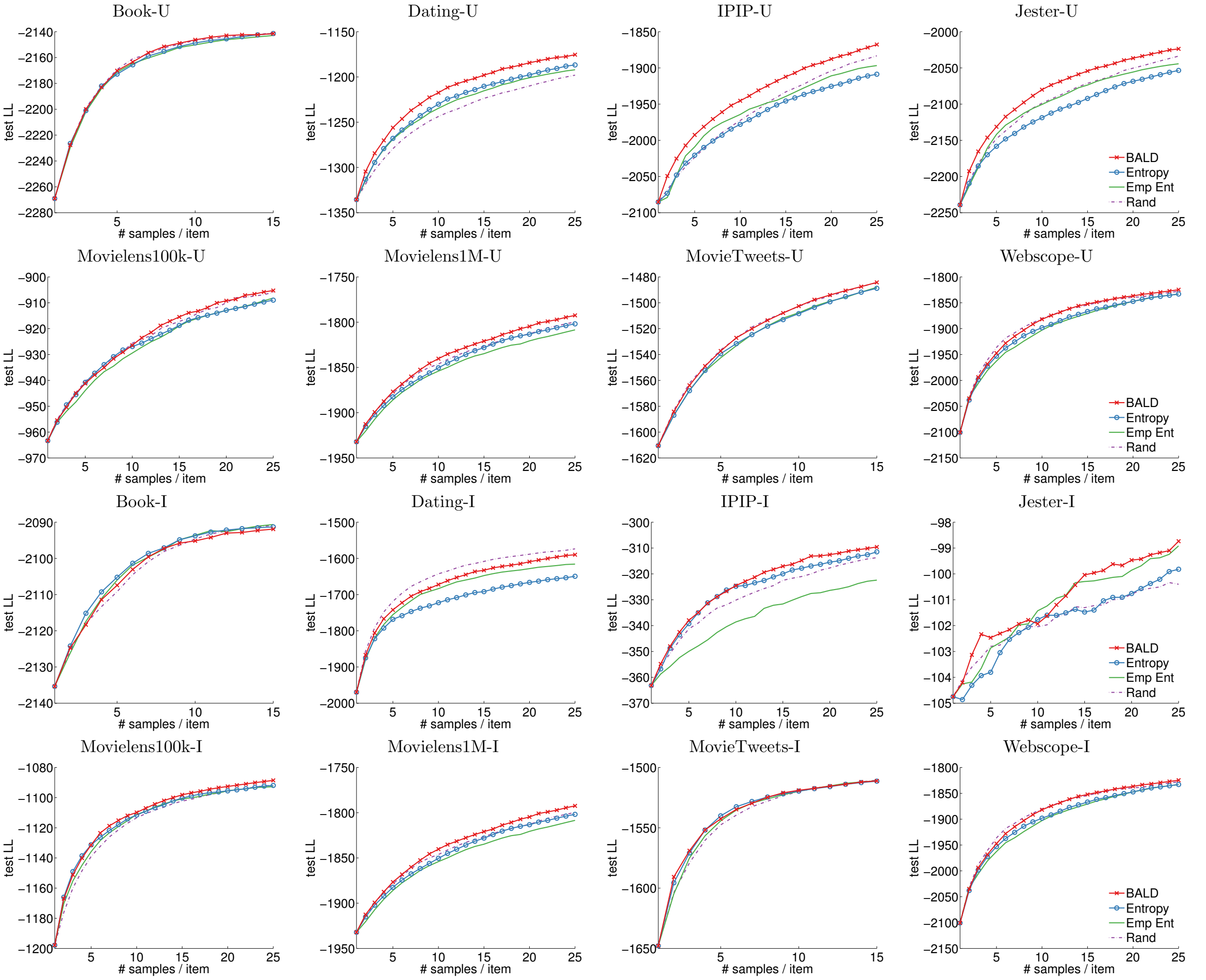

Figure 2: Log likelihood on the test users versus number of active samples selected by each algorithm with the OMF model.

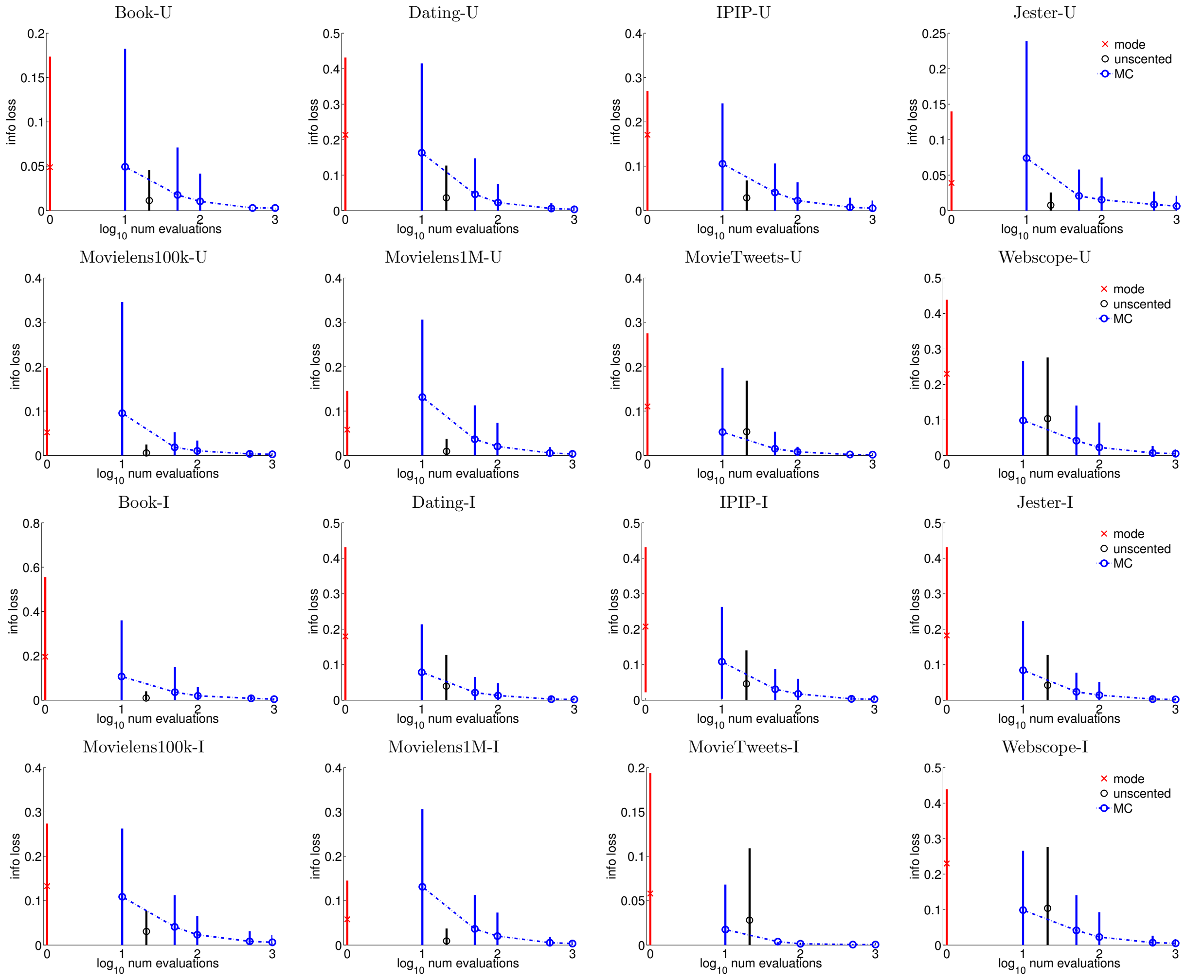

Figure 3: Information losses from sampling (blue), using the posterior mean (red), and the unscented approximation (black) to the integral over $\mathbf{u}_i$ in the second term of BALD, $\mathbb{E}_{\mathcal{Q}(\mathbf{u}_i)}\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)]$. x-axis is the number of evaluations of $\mathrm{H}[p(r^\star_{i,j}|\mathbf{u}_i)]$ required. Circles denote the mean loss, and vertical bars the 10th to 90th percentiles.

References

Cowell, R., Dawid, A., and Sebastiani, P. (1996). A comparison of sequential learning methods for incomplete data. *Bayesian statistics*, 5:533–542.

Dooms, S., De Pessemier, T., and Martens, L. (2013). Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013*.

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

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *UAI*, pages 362–369.

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

Raiko, T., Ilin, E., and Karhunen, J. (2007). Principal component analysis for large scale problems with lots of missing values. In *ECML*, pages 691–698.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461.

Stern, D. H., Herbrich, R., and Graepel, T. (2009). Matchbox: large scale online bayesian recommendations. In *WWW*, pages 111–120.