Collaborative Gaussian Processes for Preference Learning – Supplementary Material

Jose Miguel Hernández-Lobato
University of Cambridge
jmh233@cam.ac.uk

Neil Houlsby
University of Cambridge
nmth2@cam.ac.uk

Ferenc Huszár
University of Cambridge
fh277@cam.ac.uk

Zoubin Ghahramani
University of Cambridge
zoubin@eng.cam.ac.uk

1 Taylor expansion on $\log \mathrm{h}[\Phi(x)]$

The function $\log \mathrm{h}[\Phi(x)]$ can be approximated using

$$
\begin{aligned}
f(x) &= f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} + \dots \,, \\
f(x) &= \log \mathrm{h}[\Phi(x)] \,, \\
f'(x) &= -\frac{1}{\log 2} \frac{\Phi'(x)}{\mathrm{h}[\Phi(x)]} \left[\log \Phi(x) - \log(1 - \Phi(x))\right] \,, \\
f''(x) &= \frac{1}{\log 2} \frac{\Phi'(x)^2}{\mathrm{h}[\Phi(x)]^2} \left[\log \Phi(x) - \log(1 - \Phi(x))\right] - \frac{1}{\log 2} \frac{\Phi''(x)}{\mathrm{h}[\Phi(x)]} \left[\log \Phi(x) - \log(1 - \Phi(x))\right] - \\
&\quad \frac{1}{\log 2} \frac{\Phi'(x)^2}{\mathrm{h}[\Phi(x)]} \left[\frac{1}{\Phi(x)} + \frac{1}{(1 - \Phi(x)}\right] \,. \\
\therefore \log \mathrm{h}[\Phi(x)] &= 1 - \frac{1}{\pi \log 2} x^2 + \mathcal{O}(x^4) \,.
\end{aligned}
$$

Note that the x^3 term will be zero because the function is even. By exponentiating, we obtain

$$
\mathrm{h}[\Phi(x)] \approx \exp\left(-\frac{x^2}{\pi \log 2}\right) \,. \tag{1}
$$

2 Update Operations for Expectation Propagation and Variational Bayes

The first factor to be refined is $\hat{f}_4$. The update operations that minimize $\mathrm{KL}(\mathcal{Q}^{\backslash 4} f_4 \| \mathcal{Q}^{\backslash 4} \hat{f}_4)$ are given by

$$
[\hat{v}_{d,i}^{h,4}]_{\mathrm{new}} = \left\{ [v_{d,i}^{h}]_{\mathrm{new}}^{-1} - [\hat{v}_{d,i}^{h,2}]_{\mathrm{old}}^{-1} \right\}^{-1} \,, \tag{2}
$$

$$
[\hat{m}_{d,i}^{h,4}]_{\mathrm{new}} = [\hat{v}_{d,i}^{h,4}]_{\mathrm{new}} \left\{ [m_{d,i}^{h}]_{\mathrm{new}} [v_{d,i}^{h}]_{\mathrm{new}}^{-1} - [\hat{m}_{d,i}^{h,4}]_{\mathrm{old}} [\hat{v}_{d,i}^{h,4}]_{\mathrm{old}}^{-1} \right\}^{-1} \,, \tag{3}
$$

for $d = 1, \dots, D$ and $i = 1, \dots, P$, where the subscripts *new* and *old* denote the parameter value after and before the update, respectively, and the parameters $[v_{d,i}^{h}]_{\mathrm{new}}$ and $[m_{d,i}^{h}]_{\mathrm{new}}$ are the i-th entries in the vectors $[\mathbf{v}_d^h]_{\mathrm{new}}$ and $[\mathbf{m}_d^h]_{\mathrm{new}}$ given by

$$[\mathbf{v}_d^h]_{\text{new}} = \text{diag}\left[\mathbf{\Sigma}_d^h\right] , \tag{4}$$

$$[\mathbf{m}_d^h]_{\text{new}} = \mathbf{\Sigma}_d^h \text{diag}[\hat{\mathbf{v}}_d^{h,2}]^{-1} \hat{\mathbf{m}}_d^{h,2} , \tag{5}$$

where $[\mathbf{\Sigma}_d^h]^{-1} = \mathbf{K}_{\text{items}}^{-1} + \text{diag}[\hat{\mathbf{v}}_d^{h,2}]^{-1}$ and the vectors $\hat{\mathbf{m}}_d^{h,2}$ and $\hat{\mathbf{v}}_d^{h,2}$ are P-dimensional vectors given by $\hat{\mathbf{m}}_d^{h,2} = (\hat{m}_{1,d}^{h,2}, \ldots, \hat{m}_{P,d}^{h,2})^{\text{T}}$ and $\hat{\mathbf{v}}_d^{h,2} = (\hat{v}_{1,d}^{h,2}, \ldots, \hat{v}_{P,d}^{h,2})^{\text{T}}$.

The second factor to be refined by EP is $\hat{f}_3$. The update operations that minimize $\text{KL}(\mathcal{Q}^{\backslash 3} f_3 \| \mathcal{Q}^{\backslash 3} \hat{f}_3)$ are

$$[\hat{v}_{u,d}^{w,3}]_{\text{new}} = \left\{ [v_{u,d}^w]_{\text{new}}^{-1} - [\hat{v}_{u,d}^{w,2}]_{\text{old}}^{-1} \right\}^{-1} , \tag{6}$$

$$[\hat{m}_{u,d}^{w,3}]_{\text{new}} = [\hat{v}_{u,d}^{w,3}]_{\text{new}} \left\{ [m_{u,d}^w]_{\text{new}} [v_{u,d}^w]_{\text{new}}^{-1} - [\hat{m}_{u,d}^{w,3}]_{\text{old}} [\hat{v}_{u,d}^{w,3}]_{\text{old}}^{-1} \right\}^{-1} , \tag{7}$$

for $u = 1, \ldots, U$ and $d = 1, \ldots, D$, where the parameters $[v_{u,d}^w]_{\text{new}}$ and $[m_{u,d}^w]_{\text{new}}$ are the u-th entries in the vectors $[\mathbf{v}_d^w]_{\text{new}}$ and $[\mathbf{m}_d^w]_{\text{new}}$ given by

$$[\mathbf{v}_d^w]_{\text{new}} = \text{diag}\left[\mathbf{\Sigma}_d^w\right] , \tag{8}$$

$$[\mathbf{m}_d^w]_{\text{new}} = \mathbf{\Sigma}_d^w \text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1} \hat{\mathbf{m}}_d^{w,2} , \tag{9}$$

where $[\mathbf{\Sigma}_d^w]^{-1} = \mathbf{K}_{\text{items}}^{-1} + \text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}$ and the vectors $\hat{\mathbf{m}}_d^{w,2}$ and $\hat{\mathbf{v}}_d^{w,2}$ are given by $\hat{\mathbf{m}}_d^{w,2} = (\hat{m}_{1,d}^{w,2}, \ldots, \hat{m}_{U,d}^{w,2})^{\text{T}}$ and $\hat{\mathbf{v}}_d^{w,2} = (\hat{v}_{1,d}^{w,2}, \ldots, \hat{v}_{U,d}^{w,2})^{\text{T}}$.

The third factor to be refined by EP is $\hat{f}_2$. For this, we follow the approach used by Stern et al. (2009) and first marginalize $f_2 \mathcal{Q}^{\backslash 2}$ with respect to $\mathbf{G}^{(\mathcal{D})}$. The result of this operation is the auxiliary un-normalized distribution $\mathcal{S}(\mathbf{W}, \mathbf{H})$ given by

$$\begin{aligned}
\mathcal{S}(\mathbf{W}, \mathbf{H}) &= \int \prod_{u=1}^{U} \prod_{i=1}^{M_u} \delta[g_{u,z_{u,i}} - \mathbf{w}_u \mathbf{h}_{\cdot,z_{u,i}}] \mathcal{Q}^{\backslash 2}(\mathbf{G}^{(\mathcal{D})}, \mathbf{W}, \mathbf{H}) \, d\mathbf{G}^{(\mathcal{D})} \\
&= \left[\prod_{u=1}^{U} \prod_{i=1}^{M_u} \mathcal{N}(\mathbf{w}_u \mathbf{h}_{\cdot,z_{u,i}} | \hat{m}_{u,i}^{g,1}, \hat{v}_{u,i}^{g,1}) \right] \left[\prod_{u=1}^{U} \prod_{d=1}^{D} \mathcal{N}(w_{u,d} | \hat{m}_{u,d}^{w,3}, \hat{v}_{u,d}^{w,3}) \right] \\
&\quad \left[\prod_{d=1}^{D} \prod_{i=1}^{P} \mathcal{N}(h_{d,i} | \hat{m}_{d,i}^{h,4}, \hat{v}_{d,i}^{h,4}) \right] .
\end{aligned} \tag{10}$$

Let $\mathcal{Q}_{\mathbf{W},\mathbf{H}}$ be the posterior approximation ($\mathcal{Q}(\mathbf{G}^{(\mathcal{D})}, \mathbf{W}, \mathbf{H})$) after marginalizing out $\mathbf{G}^{(\mathcal{D})}$. The parameters of $\mathcal{Q}_{\mathbf{W},\mathbf{H}}$, that is, $m_{d,i}^h$, $v_{d,i}^h$, $m_{u,d}^w$ and $v_{u,d}^w$, for $d = 1, \ldots, D$, $u = 1, \ldots, U$ and $i = 1, \ldots, P$, are then optimized to minimize $\text{KL}(\mathcal{Q}_{\mathbf{W},\mathbf{H}} \| \mathcal{S})$. This can be done very efficiently using the gradient descent method described by Raiko et al. (2007). The resulting EP updates for $\hat{f}_2$ are given by

$$[\hat{v}^{h,2}_{d,i}]_{\text{new}} = \left\{ [v^{h}_{d,i}]^{-1}_{\text{new}} - [\hat{v}^{h,2}_{d,i}]^{-1}_{\text{old}} \right\}^{-1} , \tag{11}$$

$$[\hat{m}^{h,2}_{d,i}]_{\text{new}} = [\hat{v}^{h,2}_{d,i}]_{\text{new}} \left\{ [m^{h}_{d,i}]_{\text{new}} [v^{h}_{d,i}]^{-1}_{\text{new}} - [\hat{m}^{h,2}_{d,i}]_{\text{old}} [\hat{v}^{h,2}_{d,i}]^{-1}_{\text{old}} \right\}^{-1} , \tag{12}$$

$$[\hat{v}^{w,2}_{u,d}]_{\text{new}} = \left\{ [v^{w}_{u,d}]^{-1}_{\text{new}} - [\hat{v}^{w,2}_{u,l}]^{-1}_{\text{old}} \right\}^{-1} , \tag{13}$$

$$[\hat{m}^{w,2}_{u,d}]_{\text{new}} = [\hat{v}^{w,2}_{u,d}]_{\text{new}} \left\{ [m^{w}_{u,d}]_{\text{new}} [v^{w}_{u,d}]^{-1}_{\text{new}} - [\hat{m}^{w,2}_{u,d}]_{\text{old}} [\hat{v}^{w,2}_{u,d}]^{-1}_{\text{old}} \right\}^{-1} , \tag{14}$$

$$[\hat{v}^{g,2}_{u,j}]_{\text{new}} = \left\{ [v^{g}_{u,j}]^{-1}_{\text{new}} - [\hat{v}^{g,2}_{u,j}]^{-1}_{\text{old}} \right\}^{-1} , \tag{15}$$

$$[\hat{m}^{g,2}_{u,j}]_{\text{new}} = [\hat{v}^{g,2}_{u,j}]_{\text{new}} \left\{ [m^{g}_{u,j}]_{\text{new}} [v^{g}_{u,j}]^{-1}_{\text{new}} - [\hat{m}^{g,2}_{u,j}]_{\text{old}} [\hat{v}^{g,2}_{u,j}]^{-1}_{\text{old}} \right\}^{-1} , \tag{16}$$

for $d = 1, \ldots, D$, $u = 1, \ldots, U$, $j = 1, \ldots, M_u$ and $i = 1, \ldots, P$ where $[m^{h}_{d,i}]_{\text{new}}$, $[v^{h}_{d,i}]_{\text{new}}$, $[m^{w}_{u,d}]_{\text{new}}$ and $[v^{w}_{u,d}]_{\text{new}}$, are the parameters of $\mathcal{Q}$ that minimize $\text{KL}(\mathcal{Q}_{\mathbf{W},\mathbf{H}} \| \mathcal{S})$ and

$$[m^{g}_{u,j}]_{\text{new}} = \sum_{d=1}^{D} [m^{w}_{u,d}]_{\text{new}} [m^{h}_{d,z_{u,j}}]_{\text{new}} , \tag{17}$$

$$[v^{g}_{u,j}]_{\text{new}} = \sum_{d=1}^{D} [m^{w}_{u,d}]^{2}_{\text{new}} [v^{h}_{d,z_{u,j}}]_{\text{new}} + \sum_{d=1}^{D} [v^{w}_{u,d}]_{\text{new}} [m^{h}_{d,z_{u,j}}]^{2}_{\text{new}} + \sum_{d=1}^{D} [v^{w}_{u,d}]_{\text{new}} [v^{h}_{d,z_{u,j}}]_{\text{new}} . \tag{18}$$

The last factor to be refined on each cycle of EP is $\hat{f}_1$. The EP update operations for this factor are

$$[\hat{m}^{g,1}_{u,i}]_{\text{new}} = \hat{m}^{g,2}_{u,i} + \hat{v}^{g,2}_{u,i} [m_{u,i}]^{-1}_{\text{new}} , \tag{19}$$

$$[\hat{v}^{g,1}_{u,i}]_{\text{new}} = \hat{v}^{g,2}_{u,i} \left[\alpha^{-1}_{u,i} [m_{u,i}]^{-1}_{\text{new}} - 1 \right] , \tag{20}$$

for $u = 1, \ldots, U$ and $i = 1, \ldots, M_u$, where

$$[m_{u,i}]_{\text{new}} = \hat{m}^{g,2}_{u,i} + \hat{v}^{g,2}_{u,i} \alpha_{u,i} , \tag{21}$$

$$\alpha_{u,i} = \Phi[\beta_{u,i}]^{-1} \phi[\beta_{u,i}] t_{u,i} [\hat{v}^{g,2}_{u,i} + 1]^{-\frac{1}{2}} , \tag{22}$$

$$\beta_{u,i} = t_{u,i} \hat{m}^{g,2}_{u,i} [\hat{v}^{g,2}_{u,i} + 1]^{-\frac{1}{2}} \tag{23}$$

and ϕ and Φ are the density and the cumulative probability functions of a standard Gaussian distribution, respectively.

2.1 The EP Approximation of the Model Evidence

The model evidence is given by $\mathcal{P}(\mathbf{T}^{(\mathcal{D})} | \mathbf{X}, \ell)$. Once EP has converged, we can approximate it using

$$\mathcal{P}(\mathbf{T}^{(\mathcal{D})} | \mathbf{X}, \ell) \approx \int \prod_{a=1}^{4} \hat{f}_a(\mathbf{G}^{(\mathcal{D})}, \mathbf{W}, \mathbf{H}) \, d\mathbf{G}^{(\mathcal{D})} \, d\mathbf{H} \, d\mathbf{W} . \tag{24}$$

For this, we have to compute the value of the parameters $\hat{s}_1, \ldots, \hat{s}_4$. The value of $\hat{s}_1$ is

$$\log \hat{s}_1 = \sum_{u=1}^{U} \sum_{i=1}^{M_u} \left[\log \Phi[\beta_{u,i}] + \frac{1}{2} \log(2\pi) + \frac{1}{2} \log \frac{\hat{v}^{g,1}_{u,i} \hat{v}^{g,2}_{u,i}}{v^{g}_{u,i}} - \frac{[m^{g}_{u,i}]^2}{2 v^{g}_{u,i}} + \frac{[\hat{m}^{g,1}_{u,i}]^2}{2 \hat{v}^{g,1}_{u,i}} + \frac{[\hat{m}^{g,2}_{u,i}]^2}{2 \hat{v}^{g,2}_{u,i}} \right] . \tag{25}$$

The value of $\hat{s}_2$ is given by

$$\log \hat{s}_2 = \log Z_2 + \sum_{u=1}^{U}\sum_{i=1}^{M_u}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\frac{\hat{v}_{u,i}^{g,1}\hat{v}_{u,i}^{g,2}}{v_{u,i}^{g}} - \frac{[m_{u,i}^{g}]^2}{2v_{u,i}^{g}} + \frac{[\hat{m}_{u,i}^{g,1}]^2}{2\hat{v}_{u,i}^{g,1}} + \frac{[\hat{m}_{u,i}^{g,2}]^2}{2\hat{v}_{u,i}^{g,2}}\right] + \sum_{d=1}^{D}\sum_{i=1}^{P}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\frac{\hat{v}_{d,i}^{h,2}\hat{v}_{d,i}^{h,4}}{v_{d,i}^{h}} - \frac{[m_{d,i}^{h}]^2}{2v_{d,i}^{h}} + \frac{[\hat{m}_{d,i}^{h,2}]^2}{2\hat{v}_{d,i}^{h,2}} + \frac{[\hat{m}_{d,i}^{h,4}]^2}{2\hat{v}_{d,i}^{h,4}}\right] + \sum_{u=1}^{U}\sum_{d=1}^{D}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\frac{\hat{v}_{u,d}^{w,2}\hat{v}_{u,d}^{w,3}}{v_{u,d}^{w}} - \frac{[m_{u,d}^{w}]^2}{2v_{u,d}^{w}} + \frac{[\hat{m}_{u,d}^{w,2}]^2}{2\hat{v}_{u,d}^{w,2}} + \frac{[\hat{m}_{u,d}^{w,3}]^2}{2\hat{v}_{u,d}^{w,3}}\right], \tag{26}$$

where Z_2 is the variational lower bound obtained in the update of $\hat{f}_2$, that is,

$$Z_2 = \int \mathcal{Q}_{\mathbf{W},\mathbf{H}} \log \frac{\mathcal{S}(\mathbf{W},\mathbf{H})}{\mathcal{Q}_{\mathbf{W},\mathbf{H}}(\mathbf{W},\mathbf{H})}\, d\mathbf{W}, d\mathbf{H}\,. \tag{27}$$

The value of $\tilde{s}_3$ is given by

$$\log \hat{s}_3 = \log Z_3 + \sum_{d=1}^{D}\sum_{u=1}^{U}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\frac{\hat{v}_{u,d}^{w,3}\hat{v}_{u,d}^{w,2}}{v_{u,d}^{w}} - \frac{[m_{u,d}^{w}]^2}{2v_{u,d}^{w}} + \frac{[\hat{m}_{u,d}^{w,3}]^2}{2\hat{v}_{u,d}^{w,3}} + \frac{[\hat{m}_{u,d}^{w,2}]^2}{2\hat{v}_{u,d}^{w,2}}\right], \tag{28}$$

where Z_3 is computed using

$$\begin{aligned}
\log Z_3 &= \log \int \mathcal{P}(\mathbf{W}|\mathbf{U}) \left[\prod_{u=1}^{U}\prod_{d=1}^{D}\mathcal{N}(w_{u,d}|\hat{m}_{u,d}^{w,2}, \hat{m}_{u,d}^{w,2})\right] d\mathbf{W} \\
&= -\frac{DP}{2}\log(2\pi) + \frac{1}{2}\sum_{d=1}^{D}\log|\mathbf{\Sigma}_d^w| - \frac{D}{2}\log|\mathbf{K}_{\text{users}}| - \frac{1}{2}\sum_{u=1}^{U}\sum_{d=1}^{D}\log\hat{v}_{u,d}^{w,2} - \\
&\quad \frac{1}{2}\sum_{u=1}^{U}\sum_{d=1}^{D}\frac{[\hat{m}_{u,d}^{w,2}]^2}{\hat{v}_{u,d}^{w,2}} + \frac{1}{2}\sum_{d=1}^{D}[\mathbf{m}_d^w]^{\mathrm{T}}[\Sigma_d^w]^{-1}\mathbf{m}_d^w\,,
\end{aligned} \tag{29}$$

and $[\mathbf{\Sigma}_d^w]^{-1} = \mathbf{K}_{\text{users}}^{-1} + \text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}$, $\mathbf{m}_d^w = \mathbf{\Sigma}_d^w \text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}\hat{\mathbf{m}}_d^{w,2}$ and the vectors $\hat{\mathbf{m}}_d^{w,2}$ and $\hat{\mathbf{v}}_d^{w,2}$ are given by $\hat{\mathbf{m}}_d^{w,2} = (\hat{m}_{1,d}^{w,2}, \ldots, \hat{m}_{U,d}^{w,2})^{\mathrm{T}}$ and $\hat{\mathbf{v}}_d^{w,2} = (\hat{v}_{1,d}^{w,2}, \ldots, \hat{v}_{U,d}^{w,2})^{\mathrm{T}}$. Finally, the value of $\tilde{s}_4$ is given by

$$\log \hat{s}_4 = \log Z_4 + \sum_{d=1}^{D}\sum_{i=1}^{P}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\frac{\hat{v}_{d,i}^{h,4}\hat{v}_{d,i}^{h,2}}{v_{d,i}^{h}} - \frac{[m_{d,i}^{h}]^2}{2v_{d,i}^{h}} + \frac{[\hat{m}_{d,i}^{h,4}]^2}{2\hat{v}_{d,i}^{h,4}} + \frac{[\hat{m}_{d,i}^{h,2}]^2}{2\hat{v}_{d,i}^{h,2}}\right], \tag{30}$$

where Z_4 is computed using

$$\begin{aligned}
\log Z_4 &= \log \int \mathcal{P}(\mathbf{H}|\mathbf{X},\ell) \left[\prod_{d=1}^{D}\prod_{i=1}^{P}\mathcal{N}(h_{d,i}|\hat{m}_{d,i}^{h,2}, \hat{m}_{d,i}^{h,2})\right] d\mathbf{H} \\
&= -\frac{DP}{2}\log(2\pi) + \frac{1}{2}\sum_{d=1}^{D}\log|\mathbf{\Sigma}_d^h| - \frac{D}{2}\log|\mathbf{K}_{\text{items}}| - \frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{P}\log\hat{v}_{d,i}^{h,2} - \\
&\quad \frac{1}{2}\sum_{d=1}^{D}\sum_{i=1}^{p}\frac{[\hat{m}_{d,i}^{h,2}]^2}{\hat{v}_{d,i}^{h,2}} + \frac{1}{2}\sum_{d=1}^{D}[\mathbf{m}_d^h]^{\mathrm{T}}[\mathbf{\Sigma}_d^h]^{-1}\mathbf{m}_d^h\,,
\end{aligned} \tag{31}$$

and $[\boldsymbol{\Sigma}_d^h]^{-1} = \mathbf{K}_{\text{items}}^{-1} + \text{diag}[\hat{\mathbf{v}}_d^{h,2}]^{-1}$, $\mathbf{m}_d^h = \boldsymbol{\Sigma}_d \text{diag}[\hat{\mathbf{v}}_d^{h,2}]^{-1} \hat{\mathbf{m}}_d^{h,2}$ and the vectors $\hat{\mathbf{m}}_d^{h,2}$ and $\hat{\mathbf{v}}_d^{h,2}$ are given by $\hat{\mathbf{m}}_d^{h,2} = (\hat{m}_{1,d}^{h,2}, \ldots, \hat{m}_{P,d}^{h,2})^{\text{T}}$ and $\hat{\mathbf{v}}_d^{h,2} = (\hat{v}_{1,d}^{h,2}, \ldots, \hat{v}_{P,d}^{h,2})^{\text{T}}$. Given $\hat{s}_1, \ldots, \hat{s}_4$, we approximate $\mathcal{P}(\mathbf{T}^{(\mathcal{D})}|\mathbf{X}, \ell)$ using

$$
\begin{aligned}
\log \mathcal{P}(\mathbf{T}^{(\mathcal{D})}|\mathbf{X}, \ell) \quad \approx \quad & \sum_{i=a}^{4} \log \hat{s}_a - \sum_{u=1}^{U} \sum_{i=1}^{M_u} \left[\frac{1}{2} \log(2\pi) + \frac{1}{2} \log \frac{\hat{v}_{u,i}^{g,1} \hat{v}_{u,i}^{g,2}}{v_{u,i}^{g}} - \frac{[m_{u,i}^{g}]^2}{2v_{u,i}^{g}} + \frac{[\hat{m}_{u,i}^{g,1}]^2}{2\hat{v}_{u,i}^{g,1}} + \frac{[\hat{m}_{u,i}^{g,2}]^2}{2\hat{v}_{u,i}^{g,2}} \right] - \\
& \sum_{d=1}^{D} \sum_{i=1}^{P} \left[\frac{1}{2} \log(2\pi) + \frac{1}{2} \log \frac{\hat{v}_{d,i}^{h,4} \hat{v}_{d,i}^{h,2}}{v_{d,i}^{h}} - \frac{[m_{d,i}^{h}]^2}{2v_{d,i}^{h}} + \frac{[\hat{m}_{d,i}^{h,4}]^2}{2\hat{v}_{d,i}^{h,4}} + \frac{[\hat{m}_{d,i}^{h,2}]^2}{2\hat{v}_{d,i}^{h,2}} \right] - \\
& \sum_{u=1}^{U} \sum_{d=1}^{D} \left[\frac{1}{2} \log(2\pi) + \frac{1}{2} \log \frac{\hat{v}_{u,d}^{w,2} \hat{v}_{u,d}^{w,3}}{v_{u,d}^{w}} - \frac{[m_{u,d}^{w}]^2}{2v_{u,d}^{w}} + \frac{[\hat{m}_{u,d}^{w,2}]^2}{2\hat{v}_{u,d}^{w,2}} + \frac{[\hat{m}_{u,d}^{w,3}]^2}{2\hat{v}_{u,d}^{2,3}} \right] . \qquad (32)
\end{aligned}
$$

Finally, some of the EP updates may generate a negative value for $\hat{v}_{u,i}^{g,a}$, $\hat{v}_{u,d}^{w,a}$ or $\hat{v}_{d,j}^{h,a}$, where $u = 1, \ldots, U$, $i = 1, \ldots, M_u$, $j = 1, \ldots, P$ and $i = 1, \ldots, 4$. Negative variances in Gaussian approximate factors are common in many EP implementations (Minka, 2001; Minka and Lafferty, 2002). When this happens, the marginals of the approximate factor with negative variances are not density functions. Instead, they are correction factors that compensate the errors in the corresponding marginals of other approximate factors. However, these negative variances can lead to failure of the proposed EP algorithm. This may happen when we have to compute $\log |\boldsymbol{\Sigma}_d^h|$ in (31) and some of the $\hat{v}_{d,i}^{h,2}$ are negative. In this case, Σ_d^h may not be positive definite and $|\boldsymbol{\Sigma}_d^h|$ may be negative. The result is that EP may no longer be able to approximate the model evidence since $\log |\boldsymbol{\Sigma}_d^h|$ may not be defined in (31). The same may occur for $\log |\boldsymbol{\Sigma}_d^w|$ in (29). To address this problem, whenever an EP update yields a negative number for any of the $\hat{v}_{u,i}^{g,a}$, $\hat{v}_{u,d}^{w,a}$ or $\hat{v}_{d,j}^{h,a}$, we do not update this parameter, nor the corresponding $\hat{m}_{u,i}^{g,a}$, $\hat{m}_{u,d}^{w,a}$ or $\hat{m}_{d,j}^{h,a}$.

2.2 Details of the Sparse Approximations

The computational cost of EP is determined by the operations needed to refine the approximate factors $\hat{f}_3$ and $\hat{f}_4$. In particular, computing the vectors $[\mathbf{v}_d^h]_{\text{new}}$ and $[\mathbf{m}_d^h]_{\text{new}}$ in (4) and (5), for $d = 1, \ldots, D$, has cost $\mathcal{O}(DP^3)$. Similarly, the computation of the vectors $[\mathbf{v}_d^w]_{\text{new}}$ and $[\mathbf{m}_d^w]_{\text{new}}$ in (8) and (9), for $d = 1, \ldots, D$, has cost $\mathcal{O}(DU^3)$. These costs can be prohibitive when P or U are very large. Nevertheless, they can be reduced by using sparse approximations to the covariance matrices $\mathbf{K}_{\text{users}}$ and $\mathbf{K}_{\text{items}}$. We use the fully independent training conditional or FITC approximation, also known as the sparse pseudo-input GP (SPGP) Snelson and Ghahramani (2005). With FITC, the $U \times U$ covariance matrix $\mathbf{K}_{\text{users}}$ is approximated by $\mathbf{K}'_{\text{users}} = \mathbf{Q}_{\text{users}} + \text{diag}(\mathbf{K}_{\text{users}} - \mathbf{Q}_{\text{users}})$, where $\mathbf{Q}_{\text{users}} = \mathbf{K}_{\text{users},U,U_0} \mathbf{K}_{\text{users},U_0,U_0}^{-1} \mathbf{K}_{\text{users},U,U_0}^{\text{T}}$. In this expression, $\mathbf{K}_{\text{users},U_0,U_0}$ is an $U_0 \times U_0$ covariance matrix given by the evaluation of the covariance function for the users at all possible pairs of $U_0 < U$ locations or *user pseudo-inputs* $\{\mathbf{u}'_1, \ldots, \mathbf{u}'_{U_0}\}$, where $\mathbf{u}'_i \in \mathcal{U}$ for $i = 1, \ldots, U_0$, and $\mathbf{K}_{\text{users},U,U_0}$ is an $U \times U_0$ matrix with the evaluation of the covariance function for the users at all possible pairs of original user feature vectors and user pseudo-inputs, that is, $(\mathbf{u}_i, \mathbf{u}'_j)$, for $i = 1, \ldots, U$ and $j = 1, \ldots, U_0$. Similarly, the $P \times P$ covariance matrix $\mathbf{K}_{\text{items}}$ is also approximated by $\mathbf{K}'_{\text{items}} = \mathbf{Q}_{\text{items}} + \text{diag}(\mathbf{K}_{\text{items}} - \mathbf{Q}_{\text{items}})$, where $\mathbf{Q}_{\text{items}} = \mathbf{K}_{\text{items},P,P_0} \mathbf{K}_{\text{items},P_0,P_0}^{-1} \mathbf{K}_{\text{items},P,P_0}^{\text{T}}$, $\mathbf{K}_{\text{items},P_0,P_0}$ is a $P_0 \times P_0$ covariance matrix given by the evaluation of the preference kernel at all possible pairs of $P_0 < P$ locations or *item-pair pseudo-inputs* $\{(\mathbf{x}'_1, \mathbf{x}''_1), \ldots, (\mathbf{x}'_{P_0}, \mathbf{x}''_{P_0})\}$, where $\mathbf{x}'_i, \mathbf{x}''_i \in \mathcal{X}$ for $i = 1, \ldots, P_0$, and $\mathbf{K}_{\text{items},P,P_0}$ is a $P \times P_0$ matrix with the evaluation of the preference kernel at all possible combinations of feature vectors for the original item pairs and item-pair pseudo-inputs, that is, $((\mathbf{x}_{\alpha(i)}, \mathbf{x}_{\beta(i)}), (\mathbf{x}'_j, \mathbf{x}''_j))$, for $i = 1, \ldots, P$ and $j = 1, \ldots, P_0$.

We now describe how to refine the third and fourth approximate factors when $\mathbf{K}_{\text{users}}$ and $\mathbf{K}_{\text{items}}$ are replaced by $\mathbf{K}'_{\text{users}}$ and $\mathbf{K}'_{\text{items}}$, respectively. The required operations are can be efficiently implemented using the formulas described in (Naish-Guzman and Holden, 2007) and (Gredilla, 2010). In particular, let $\mathbf{K}'_{\text{users}} = \mathbf{D} + \mathbf{P}\mathbf{R}^{\text{T}}\mathbf{R}\mathbf{P}^{\text{T}}$, where $\mathbf{D} = \text{diag}(\mathbf{K}_{\text{users}} - \mathbf{Q}_{\text{users}})$, $\mathbf{P} = \mathbf{K}_{\text{users},U,U_0}$ and $\mathbf{R}$ is the upper Cholesky factor of $\mathbf{K}_{\text{users},U_0,U_0}^{-1}$, that is, $\mathbf{K}_{\text{users},U_0,U_0}^{-1} = \mathbf{R}^{\text{T}}\mathbf{R}$. This Cholesky factor can be computed using

$$\mathbf{R} = \text{rot180}(\text{chol}(\text{rot180}(\mathbf{K}_{\text{users},U_0,U_0}))^{\text{T}} \setminus \mathbf{I}), \tag{33}$$

where $\mathbf{I}$ is the identity matrix, rot180$(\cdot)$ rotates an $m \times m$ square matrix 180° so that the element in position (i, j) is moved to position $(m - i + 1, m - j + 1)$, $\mathbf{A} \setminus \mathbf{a}$ denotes the solution to the linear system $\mathbf{Ax} = \mathbf{a}$ and chol$(\cdot)$ returns the upper Cholesky factor of its argument. The matrix $\mathbf{\Sigma}_d^w$, required to compute the vectors $[\mathbf{v}_d^w]_{\text{new}}$ and $[\mathbf{m}_d^w]_{\text{new}}$ in (8) and (9), can the be encoded efficiently using $\mathbf{\Sigma}_d^w = \mathbf{D}_d^{\text{new}} + \mathbf{P}_d^{\text{new}}[\mathbf{R}_d^{\text{new}}]^{\text{T}}\mathbf{R}_d^{\text{new}}[\mathbf{P}_d^{\text{new}}]^{\text{T}}$, where

$$\mathbf{D}_d^{\text{new}} = \left(\mathbf{I} + \mathbf{D}\text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}\right)^{-1}\mathbf{D}, \tag{34}$$

$$\mathbf{P}_d^{\text{new}} = \left(\mathbf{I} + \mathbf{D}\text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}\right)^{-1}\mathbf{P}, \tag{35}$$

$$\mathbf{R}_d^{\text{new}} = \text{rot180}(\text{chol}(\text{rot180}(\mathbf{I} + \mathbf{R}\mathbf{P}^{\text{T}}\text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1}(\mathbf{I} + \mathbf{D}\text{diag}[\hat{\mathbf{v}}_d^{w,2}]^{-1})^{-1}\mathbf{P}\mathbf{R}^{\text{T}})^{\text{T}}) \setminus \mathbf{R} \tag{36}$$

and $\hat{\mathbf{v}}_d^{w,2}$ is given by $\hat{\mathbf{v}}_d^{w,2} = (\hat{v}_{1,d}^{w,2}, \ldots, \hat{v}_{U,d}^{w,2})^{\text{T}}$. The matrix $\mathbf{\Sigma}_d^h$, required to compute the vectors $[\mathbf{v}_d^h]_{\text{new}}$ and $[\mathbf{m}_d^h]_{\text{new}}$ in (4) and (5), can the be efficiently encoded in a similar manner. For this, we only have to replace $\hat{\mathbf{v}}_d^{w,2}$ by $\hat{\mathbf{v}}_d^{h,2} = (\hat{v}_{d,1}^{h,2}, \ldots, \hat{v}_{d,P}^{h,2})^{\text{T}}$ and $\mathbf{K}_{\text{users},U_0,U_0}$ and $\mathbf{K}_{\text{users},U,U_0}$ by $\mathbf{K}_{\text{items},P_0,P_0}$ and $\mathbf{K}_{\text{items},P,P_0}$, respectively. These alternative representations of $\mathbf{\Sigma}_d^w$ and $\mathbf{\Sigma}_d^h$ allow us to update $\hat{f}_3$ and $\hat{f}_4$ in $\mathcal{O}(DU_0^2U)$ and $\mathcal{O}(DP_0^2P)$ operations, respectively.

We also describe the new update for $\log Z_3$. Instead of using (29), we now use the following expression

$$\begin{aligned}\log Z_3 = \sum_{d=1}^{D}\Bigg[&-\frac{U}{2}\log(2\pi) + \log|\mathbf{R}_d^{\text{new}}| - \log|\mathbf{R}| - \frac{1}{2}\sum_{u=1}^{U}\log\left(\hat{v}_{u,d}^{w,2} + d_u\right) + \\ &\frac{1}{2}\sum_{u=1}^{U}\hat{m}_{u,d}^{w,2}([m_d^w]_{\text{new}})_u - \frac{1}{2}\sum_{u=1}^{U}\frac{[\hat{m}_{u,d}^{w,2}]^2}{\hat{v}_{u,d}^{w,2}}\Bigg],\end{aligned} \tag{37}$$

where d_u is the u-th entry in the diagonal of $\mathbf{D}$ and $([m_d^w]_{\text{new}})_u$ is the u-th entry in the vector $[\mathbf{m}_d^w]_{\text{new}}$. The analogous update for $\log Z_4$ is given by

$$\begin{aligned}\log Z_4 = \sum_{d=1}^{D}\Bigg[&-\frac{P}{2}\log(2\pi) + \log|\mathbf{R}_d^{\text{new}}| - \log|\mathbf{R}| - \frac{1}{2}\sum_{i=1}^{P}\log\left(\hat{v}_{d,i}^{h,2} + d_i\right) + \\ &\frac{1}{2}\sum_{i=1}^{P}\hat{m}_{d,i}^{h,2}([m_d^h]_{\text{new}})_i - \frac{1}{2}\sum_{i=1}^{P}\frac{[\hat{m}_{d,i}^{h,2}]^2}{\hat{v}_{d,i}^{h,2}}\Bigg],\end{aligned} \tag{38}$$

where d_i is the i-th entry in the diagonal of $\mathbf{D}$ and $([m_d^h]_{\text{new}})_i$ is the i-th entry in the vector $[\mathbf{m}_d^h]_{\text{new}}$. Note that d_i, $\mathbf{R}$ and $\mathbf{R}_d^{\text{new}}$ in (38) refer to the matrices needed for working with the efficient encoding of $\mathbf{K}'_{\text{items}}$. By contrast, d_u, $\mathbf{R}$ and $\mathbf{R}_d^{\text{new}}$ in (37) refer to the same matrices, but for working with the efficient encoding of $\mathbf{K}'_{\text{users}}$.

Finally, to compute the predictive distribution, instead of using:

$$m_{d,P+1}^h = \mathbf{k}_\star^{\text{T}}\left[\mathbf{K}_{\text{items}} + \text{diag}[\hat{\mathbf{v}}_d^{h,2}]\right]^{-1}\hat{\mathbf{m}}_d^{h,2}, \tag{39}$$

$$v_{d,P+1}^h = k_\star - \mathbf{k}_\star^{\text{T}}\left[\mathbf{K}_{\text{items}} + \text{diag}[\hat{\mathbf{v}}_d^{h,2}]\right]^{-1}\mathbf{k}_\star, \tag{40}$$

as in the main text, we use

$$m^h_{d,P+1} = \mathbf{k}_\star^{\mathrm{T}} \boldsymbol{\gamma}_d^{\mathrm{new}}\,, \tag{41}$$

$$v^h_{d,P+1} = d_\star + \parallel \mathbf{R}_d^{\mathrm{new}} \mathbf{k}_\star \parallel^2\,, \tag{42}$$

where $\mathbf{k}_\star$ is a P_0-dimensional vector that contains the prior covariances between $h_d(\mathbf{x}_{\alpha(P+1)}, \mathbf{x}_{\beta(P+1)})$ and the value of latent function h_d at the item-pair pseudo-inputs, that is, $h_d(\mathbf{x}'_1, \mathbf{x}''_1), \ldots, h_d(\mathbf{x}'_{P_0}, \mathbf{x}''_{P_0})$, $\boldsymbol{\gamma}_d^{\mathrm{new}} = [\mathbf{R}_d^{\mathrm{new}}]^{\mathrm{T}} \mathbf{R}_d^{\mathrm{new}} [\mathbf{P}_d^{\mathrm{new}}]^{\mathrm{T}} \mathrm{diag}[\hat{\mathbf{v}}_d^{h,2}]^{-1} \hat{\mathbf{m}}_d^{h,2}$, $d_\star = k_\star - \mathbf{p}_\star^{\mathrm{T}} \mathbf{R}^{\mathrm{T}} \mathbf{R} \mathbf{p}_\star$ and finally, $k_\star$ is the prior variance of $h_d(\mathbf{x}_{\alpha(P+1)}, \mathbf{x}_{\beta(P+1)})$. Note that in all of these formulas, $\mathbf{R}_d^{\mathrm{new}}$, $\mathbf{R}_d^{\mathrm{new}}$ and $\mathbf{R}^{\mathrm{T}}$ refer to the matrices needed for working with the efficient encoding of $\mathbf{K}'_{\mathrm{items}}$.

References

Gredilla, M. L. (2010). *Sparse Gaussian Processes for Large-scale Machine Learning*. PhD thesis, Universidad Carlos III de Madrid.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359.

Naish-Guzman, A. and Holden, S. B. (2007). The generalized fitc approximation. In *Advances in Neural Information Processing Systems 20*.

Raiko, T., Ilin, A., and Juha, K. (2007). Principal component analysis for large scale problems with lots of missing values. In Kok, J., Koronacki, J., Mantaras, R., Matwin, S., Mladenic, D., and Skowron, A., editors, *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 691–698. Springer Berlin / Heidelberg.

Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*.

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.