Supplementary Material

In this appendix, we provide the proof of Lemma 3.1 in the main paper (Section 1), the details of CEP message updates in Bayesian tensor decomposition, logisitic regression and probit regression (Section 2), and more experimental results in Section 3.

1 Proof of Lemma 3.1

Lemma 3.1 When the conditional moment h is part of the sufficient statistics of $\boldsymbol{\theta}_{\backslash m}$, i.e., each element of h belongs to Φ_m, the fixed points of EP are also that of CEP without Taylor approximations.

Proof. Upon convergence, EP reaches a fixed point such that for $\forall i, m$,

$$\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\backslash m})}[\Phi_m] = \mathbb{E}_{q(\boldsymbol{\theta}_{\backslash m})}[\Phi_m], \quad (1)$$

$$\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}\big(\phi(\boldsymbol{\theta}_m)\big) = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\backslash m})}[h] = \mathbb{E}_{q(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)], \quad (2)$$

where the conditional moment $h = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\backslash m})}\big(\phi(\boldsymbol{\theta}_m)\big)$. When $h \subset \Phi_m$, we have $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\backslash m})}[h] = \mathbb{E}_{q(\boldsymbol{\theta}_{\backslash m})}[h]$ from (1). Then from (2), we further obtain $\mathbb{E}_{q(\boldsymbol{\theta}_{\backslash m})}[h] = \mathbb{E}_{q(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)]$, which is the fixed point when CEP converges without Taylor approximations. □

2 CEP for Bayesian Tensor Decomposition, Logistic Regression and Probit Regression

2.1 Bayesian Tensor Decomposition

2.1.1 Continuous Tensor

Let us first consider continuous entry values $\{y_{\mathbf{i}}\}_{\mathbf{i}\in S}$. The joint probability of the Bayesian tensor decomposition model, according to (12) in Section 4.1 of the main paper, is given by

$$p(\{y_{\mathbf{i}}\}_{\mathbf{i}\in S}, \mathcal{U}, \tau) = \mathrm{Gam}(\tau|a_0, b_0) \prod_{k=1}^{K}\prod_{s=1}^{d_k} \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^k, v\mathbf{I}) \cdot \prod_{\mathbf{i}\in S} \mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^\top(\mathbf{u}_{i_1}^1 \circ \ldots \circ \mathbf{u}_{i_K}^K), \tau^{-1}).$$

We first introduce an exponential family term to approximate each factor in the joint probability. Since the prior distributions of the embeddings $\mathcal{U}$ and τ are already inside the exponential family, we do not need any approximation. We then use factorized messages to approximate the likelihood of each observed entry value $y_{\mathbf{i}}$,

$$\mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^\top(\mathbf{u}_{i_1}^1 \circ \ldots \circ \mathbf{u}_{i_K}^K), \tau^{-1}) \approx \tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k),$$

where $\tilde{f}_{\mathbf{i}}(\tau) = \mathrm{Gam}(\tau|a_{\mathbf{i}}, b_{\mathbf{i}})$ and $\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{\mathbf{i}}^k, \mathbf{S}_{\mathbf{i}}^k)$. Therefore, the approximate posterior distribution is

$$q(\mathcal{U}, \tau) \propto \mathrm{Gam}(\tau|a_0, b_0) \prod_{k=1}^{K}\prod_{s=1}^{d_k} \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^k, v\mathbf{I}) \cdot \prod_{\mathbf{i}\in S} \tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k).$$

Obviously, the approximate posterior is factorized over all $\{\mathbf{u}_s^k\}_{1\le k\le K, 1\le s\le d_k}$ and τ,

$$q(\mathcal{U}, \tau) = q(\tau) \prod_{k=1}^{K}\prod_{s=1}^{d_k} q(\mathbf{u}_{i_k}^k)$$

where $q(\tau)$ is a Gamma distribution and each $q(\mathbf{u}_{i_k}^k)$ Gaussian. To update the messages for each entry $y_{\mathbf{i}}$, we need to first divide $q(\mathcal{U}, \tau)$ by them to obtain the calibrating distribution

$$q^{\backslash \mathbf{i}}(\mathcal{U}, \tau) \propto \frac{q(\mathcal{U}, \tau)}{\tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k)}$$

and then construct the tilted distribution

$$\hat{p}_{\mathbf{i}}(\mathcal{U}, \tau) \propto q^{\backslash \mathbf{i}}(\mathcal{U}, \tau)\mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^\top(\mathbf{u}_{i_1}^1 \circ \ldots \circ \mathbf{u}_{i_K}^K), \tau^{-1}).$$

Since we only require the moments for the inverse variance τ and the embedding vectors that associate with entry $\mathbf{i}$, $\mathbf{u}_{\mathbf{i}} = \{\mathbf{u}_{\mathbf{i}_1}^1, \ldots, \mathbf{u}_{\mathbf{i}_K}^K\}$, the other embeddings vectors will be marginalized out and we only need to consider the marginal titled distribution for $\{\mathbf{u}_{\mathbf{i}}, \tau\}$,

$$\hat{p}_{\mathbf{i}}(\mathbf{u}_{\mathbf{i}}, \tau) \propto q^{\backslash \mathbf{i}}(\tau) \prod_{k=1}^{K} q^{\backslash \mathbf{i}}(\mathbf{u}_{i_k}^k) \cdot \mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^\top(\mathbf{u}_{i_1}^1 \circ \ldots \circ \mathbf{u}_{i_K}^K), \tau^{-1}), \quad (3)$$

where

$$q^{\backslash \mathbf{i}}(\tau) = \mathrm{Gam}(\tau|a^{\backslash \mathbf{i}}, b^{\backslash \mathbf{i}}), \quad q^{\backslash \mathbf{i}}(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{i_k}^k, \mathbf{S}_{i_k}^k),$$

and

$$a^{\backslash \mathbf{i}} = a_0 + \sum_{\mathbf{j}\in S, \mathbf{j}\neq\mathbf{i}} a_{\mathbf{j}} - |S| + 1,$$

$$b^{\backslash \mathbf{i}} = b_0 + \sum_{\mathbf{j}\in S, \mathbf{j}\neq\mathbf{i}} b_{\mathbf{j}},$$

$$\mathbf{S}_{i_k}^k = \left(\sum_{\mathbf{j}\in S,\, \mathbf{j}\neq\mathbf{i},\, j_k=i_k} \mathbf{S}_{\mathbf{j}}^{k\,-1} + v\mathbf{I}\right)^{-1},$$

$$\mathbf{m}_{i_k}^k = \mathbf{S}_{i_k}^k \left(\sum_{\mathbf{j}\in S,\, \mathbf{j}\neq\mathbf{i},\, j_k=i_k} \mathbf{S}_{\mathbf{j}}^{k\,-1}\mathbf{m}_{\mathbf{j}}^k + v\boldsymbol{\mu}_{i_k}^k\right).$$

Here $|S|$ is the size of S, i.e., the number of observed entries.

Due to the production term in the Gaussian likelihood, The moments w.r.t $\hat{p}_{\mathbf{i}}(\mathbf{u}_{\mathbf{i}}, \tau)$ (see (3)) are intractable. To overcome this barrier, we use CEP. Specifically, to update each message $\tilde{f}^k_{\mathbf{i}}(\mathbf{u}^k_{i_k})$, we first compute the conditional moments w.r.t the conditional tilted distribution given τ and $\mathbf{u}_{\mathbf{i}}^{\setminus k} = \{\mathbf{u}^1_{i_1}, \ldots, \mathbf{u}^{k-1}_{i_{k-1}}, \mathbf{u}^{k+1}_{i_{k+1}}, \ldots, \mathbf{u}^K_{i_K}\}$ fixed,

$$\hat{p}_{\mathbf{i}}(\mathbf{u}^k_{i_k} | \mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau) \propto \mathcal{N}(\mathbf{u}^k_{i_k} | \mathbf{m}^k_{i_k}, \mathbf{S}^k_{i_k}) \mathcal{N}(y_{\mathbf{i}} | \mathbf{z}_{\mathbf{i}}^{\setminus k\top} \mathbf{u}^k_{i_k}, \tau^{-1}),$$

where $\mathbf{z}_{\mathbf{i}}^{\setminus k}$ is the Hadmard product of the vectors in $\mathbf{u}_{\mathbf{i}}^{\setminus k}$. It is easy to see that this is a Gaussian distribution, and the conditional moments can be calculated from

$$\mathrm{cov}(\mathbf{u}^k_{i_k} | \mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau) = \big[\mathbf{S}^{k\ -1}_{i_k} + \tau(\mathbf{z}_{\mathbf{i}}^{\setminus k} \mathbf{z}_{\mathbf{i}}^{\setminus k\top})\big]^{-1}, \tag{4}$$

$$\mathbb{E}(\mathbf{u}^k_{i_k} | \mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau) = \mathrm{cov}(\mathbf{u}^k_{i_k} | \mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau) \big[\mathbf{S}^{k\ -1}_{i_k} \mathbf{m}^k_{i_k} + \tau y_{\mathbf{i}} \mathbf{z}_{\mathbf{i}}^{\setminus k}\big]. \tag{5}$$

Note that for Gaussian random variables, we only need the first and second moments. The mean is the first moment, and the covariance second central moment. To obtain the second raw moment, we can simply add the outer-product of the mean to the covariance. To be concise, we will stick our presentation to the mean and variance.

Next, we compute the expectation of the conditional moments (4)(5) w.r.t the current approximate posterior for $\mathbf{u}_{\mathbf{i}}^{\setminus k}$ and τ, i.e., $q(\mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau)$. To this end, we will use the first-order Taylor expansions at the expectation of the statistics (i.e., moments) that appear in (4)(5), including τ, $\mathbf{z}^k_{\mathbf{i}}$ and $\mathbf{z}^k_{\mathbf{i}} \mathbf{z}^{k\top}_{\mathbf{i}}$ (see (8) in the main paper). Their expectations are given by

$$\mathbb{E}_q(\tau) = \frac{a_0 + \sum_{\mathbf{i} \in S} a_{\mathbf{i}} - |S|}{b_0 + \sum_{\mathbf{i} \in S} b_{\mathbf{i}}},$$

$$\mathbb{E}_q(\mathbf{z}_{\mathbf{i}}^{\setminus k}) = \mathbb{E}_q(\mathbf{u}^1_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^{k-1}_{i_{k-1}}) \circ \mathbb{E}_q(\mathbf{u}^{k+1}_{i_{k+1}}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K}),$$

$$\mathbb{E}_q(\mathbf{z}_{\mathbf{i}}^{\setminus k} \mathbf{z}_{\mathbf{i}}^{\setminus k\top}) = \mathbb{E}_q(\mathbf{u}^1_{i_1} \mathbf{u}^{1\top}_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^{k-1}_{i_{k-1}} \mathbf{u}^{k-1\top}_{i_{k-1}}) \circ \mathbb{E}_q(\mathbf{u}^{k+1}_{i_{k+1}} \mathbf{u}^{k+1\top}_{i_{k+1}}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K} \mathbf{u}^{K\top}_{i_K}),$$

where for $\forall t \neq k$,

$$\mathbb{E}_q(\mathbf{u}^t_{i_t}) = \mathrm{cov}_q(\mathbf{u}^t_{i_t}) \left(\sum_{\mathbf{j} \in S,\ j_t = i_t} \mathbf{S}^{t\ -1}_{\mathbf{j}} \mathbf{m}^t_{\mathbf{j}} + v \boldsymbol{\mu}^k_{i_t} \right),$$

$$\mathbb{E}_q(\mathbf{u}_{i_t} \mathbf{u}^\top_{i_t}) = \mathrm{cov}_q(\mathbf{u}^t_{i_t}) + \mathbb{E}_q(\mathbf{u}^t_{i_t}) \mathbb{E}_q(\mathbf{u}^t_{i_t})^\top,$$

$$\mathrm{cov}_q(\mathbf{u}^t_{i_t}) = \left(\sum_{\mathbf{j} \in S,\ j_t = i_t} \mathbf{S}^{t\ -1}_{\mathbf{j}} + v \mathbf{I} \right)^{-1}. \tag{6}$$

By taking expectation over the first-order Taylor expansion w.r.t $q(\mathbf{u}_{\mathbf{i}}^{\setminus k}, \tau)$, we obtain the approximated expectation of the conditional moments (see (10) in the main paper). This can be done by simply replacing τ, $\mathbf{z}_{\mathbf{i}}^{\setminus k}$ and $\mathbf{z}_{\mathbf{i}}^{\setminus k} \mathbf{z}_{\mathbf{i}}^{\setminus k\top}$ in (4)(5) with $\mathbb{E}(\tau)$, $\mathbb{E}(\mathbf{z}_{\mathbf{i}}^{\setminus k})$ and $\mathbb{E}(\mathbf{z}_{\mathbf{i}}^{\setminus k} \mathbf{z}_{\mathbf{i}}^{\setminus k\top})$, respectively. We then use the expected conditional moments to build a new posterior for $\mathbf{u}^k_{i_k}$ and then obtain the updated message $\tilde{f}^k_{\mathbf{i}}(\mathbf{u}^k_{i_k}) = \mathcal{N}(\mathbf{u}^k_{i_k} | \mathbf{m}^{k*}_{\mathbf{i}}, \mathbf{S}^{k*}_{\mathbf{i}})$ accordingly:

$$\mathbf{S}^{k*}_{\mathbf{i}} = \left(\mathbb{E}_q(\tau) \mathbb{E}_q(\mathbf{z}_{\mathbf{i}}^{\setminus k} \mathbf{z}_{\mathbf{i}}^{\setminus k\top}) \right)^{-1},$$

$$\mathbf{m}^{k*}_{\mathbf{i}} = \mathbf{S}^{k*}_{\mathbf{i}} \left(y_{\mathbf{i}} \mathbb{E}_q(\tau) \mathbb{E}_q(\mathbf{z}_{\mathbf{i}}^{\setminus k}) \right).$$

We follow the same procedure to update $\tilde{f}_{\mathbf{i}}(\tau)$. The conditional tilted distribution of τ is

$$\hat{p}_{\mathbf{i}}(\tau | \mathbf{u}_{\mathbf{i}}) = \mathrm{Gam}(\tau | \hat{a}, \hat{b}) \propto \mathrm{Gam}(\tau | a^{\setminus \mathbf{i}}, b^{\setminus \mathbf{i}}) \mathcal{N}(y_{\mathbf{i}} | \mathbf{1}^\top (\mathbf{u}^1_{i_1} \circ \ldots \circ \mathbf{u}^K_{i_K}), \tau^{-1})$$

where

$$\hat{a} = a^{\setminus \mathbf{i}} + \frac{1}{2},$$

$$\hat{b} = b^{\setminus \mathbf{i}} + \frac{1}{2} (y_{\mathbf{i}} - \mathbf{1}^\top (\mathbf{u}^1_{i_1} \circ \ldots \circ \mathbf{u}^K_{i_K}))^2.$$

The conditional moments are given by

$$\mathbb{E}_{\hat{p}_{\mathbf{i}}(\tau | \mathbf{u}_{\mathbf{i}})}(\tau) = \frac{\hat{a}}{\hat{b}}, \tag{7}$$

$$\mathbb{E}_{\hat{p}_{\mathbf{i}}(\tau | \mathbf{u}_{\mathbf{i}})}\big(\log(\tau)\big) = \psi(\hat{a}) - \log(\hat{b}), \tag{8}$$

where $\psi(\cdot)$ is the digamma function. To compute the expectation of the conditional moments, we can use the first-order Taylor expansion. This can be done by substituting for $\hat{a}$ and $\hat{b}$ in the conditional moments (7)(8) their expectation w.r.t $q(\mathbf{u}_{\mathbf{i}})$, i.e., $\mathbb{E}_q(\hat{a})$ and $\mathbb{E}_q(\hat{b})$. The computation of $\mathbb{E}_q(\hat{a})$ and $\mathbb{E}_q(\hat{b})$ is straightforward and analytical (see (18) in the main paper),

$$\mathbb{E}_q(\hat{a}) = \hat{a},$$

$$\mathbb{E}_q(\hat{b}) = b^{\setminus \mathbf{i}} + \frac{1}{2} y^2_{\mathbf{i}} - y_{\mathbf{i}} \mathbf{1}^\top \big[\mathbb{E}_q(\mathbf{u}^1_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K})\big] + \frac{1}{2} \mathrm{tr}\big[\mathbb{E}_q(\mathbf{u}^1_{i_1} \mathbf{u}^{1\top}_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K} \mathbf{u}^{K\top}_{i_K})\big]. \tag{9}$$

We then use the expected conditional moments to build a new posterior for τ and obtain the updated message $\tilde{f}_{\mathbf{i}}(\tau) = \mathrm{Gam}(\tau | a^*_{\mathbf{i}}, \mathbf{b}^*_{\mathbf{i}})$, where

$$a^*_{\mathbf{i}} = \frac{1}{2},$$

$$b^*_{\mathbf{i}} = \frac{1}{2} y^2_{\mathbf{i}} - y_{\mathbf{i}} \mathbf{1}^\top \big[\mathbb{E}_q(\mathbf{u}^1_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K})\big] + \frac{1}{2} \mathrm{tr}\big[\mathbb{E}_q(\mathbf{u}^1_{i_1} \mathbf{u}^{1\top}_{i_1}) \circ \ldots \circ \mathbb{E}_q(\mathbf{u}^K_{i_K} \mathbf{u}^{K\top}_{i_K})\big].$$

2.1.2 Binary Tensor

When entry values are binary, according to (13) in Section 4.1 of the main paper, the joint probability is

$$p(\{y_{\mathbf{i}}\}_{\mathbf{i}\in S}, \mathcal{U}) = \prod_{k=1}^{K}\prod_{s=1}^{d_k}\mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^k, v\mathbf{I}) \cdot \prod_{\mathbf{i}\in S}\psi\big((2y_{\mathbf{i}}-1)\mathbf{1}^\top(\mathbf{u}_{i_1}^1\circ\ldots\circ\mathbf{u}_{i_K}^K)\big)$$

where $\psi(\cdot)$ is the cumulative density function (CDF) of the standard Gaussian distribution. We use factorized Gaussian messages to approximate each likelihood,

$$\psi\big((2y_{\mathbf{i}}-1)\mathbf{1}^\top(\mathbf{u}_{i_1}^1\circ\ldots\circ\mathbf{u}_{i_K}^K)\big) \approx \prod_{k=1}^{K}\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k),$$

where $\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{\mathbf{i}}^k, \mathbf{S}_{\mathbf{i}}^k)$. The approximate posterior is given by

$$q(\mathcal{U}) = \prod_{k=1}^{K}\prod_{s=1}^{d_k}\mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^k, v\mathbf{I})\prod_{\mathbf{i}\in S}\prod_{k=1}^{K}\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k).$$

To update messages for entry $\mathbf{i}$, we first obtain the calibrating distribution

$$q^{\backslash\mathbf{i}}(\mathcal{U}) \propto \frac{q(\mathcal{U},\tau)}{\prod_{k=1}^{K}\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k)},$$

and then the tilted distribution

$$\hat{p}_{\mathbf{i}}(\mathcal{U}) \propto q^{\backslash\mathbf{i}}(\mathcal{U})\psi\big((2y_{\mathbf{i}}-1)\mathbf{1}^\top(\mathbf{u}_{i_1}^1\circ\ldots\circ\mathbf{u}_{i_K}^K)\big).$$

Again, due to the production term in the likelihood, the moments w.r.t the tilted distribution are intractable to calculate. Therefore, we use CEP. Following the same procedure as in Section 2.1.1, to update each message $\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k)$ for entry $y_{\mathbf{i}}$, we first obtain the conditional tilted distribution by

$$\hat{p}_{\mathbf{i}}(\mathbf{u}_{i_k}^k|\mathbf{u}_{\mathbf{i}}^{\backslash k}) \propto \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{i_k}^k, \mathbf{S}_{i_k}^k)\psi\big((2y_{\mathbf{i}}-1)\mathbf{1}^\top(\mathbf{u}_{i_1}^1\circ\ldots\circ\mathbf{u}_{i_K}^K)\big),$$

where

$$\mathbf{S}_{i_k}^k = \left(\sum_{\mathbf{j}\in S,\,\mathbf{j}\neq\mathbf{i},\,j_k=i_k}{\mathbf{S}_{\mathbf{j}}^k}^{-1} + v\mathbf{I}\right)^{-1},$$

$$\mathbf{m}_{i_k}^k = \mathbf{S}_{i_k}^k\left(\sum_{\mathbf{j}\in S,\,\mathbf{j}\neq\mathbf{i},\,j_k=i_k}{\mathbf{S}_{\mathbf{j}}^k}^{-1}\mathbf{m}_{\mathbf{j}}^k + v\boldsymbol{\mu}_{i_k}^k\right).$$

To compute the conditional moments of $\mathbf{u}_{i_k}^k$, we can use the same trick as applying EP for Bayesian probit regression (Dusek, 2013). We first derive the normalizer of the conditional tilted distribution,

$$Z = \int\mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{i_k}^k, \mathbf{S}_{i_k}^k)\psi\big((2y_{\mathbf{i}}-1)\mathbf{1}^\top(\mathbf{u}_{i_1}^1\circ\ldots\circ\mathbf{u}_{i_K}^K)\big)d\mathbf{u}_{i_k}^k = \psi\left(\frac{(2y_{\mathbf{i}}-1){\mathbf{z}_{\mathbf{i}}^{\backslash k}}^\top\mathbf{m}_{i_k}^k}{\sqrt{1+{\mathbf{z}_{\mathbf{i}}^{\backslash k}}^\top\mathbf{S}_{i_k}^k\mathbf{z}_{\mathbf{i}}^{\backslash k}}}\right).$$

Then we can compute the conditional moments through the derivatives of the logarithm of the normalizer,

$$\mathrm{cov}(\mathbf{u}_{i_k}^k|\mathbf{u}_{\mathbf{i}}^{\backslash k}) = \mathbf{S}_{i_k}^k - \mathbf{S}_{i_k}^k\mathbf{A}\mathbf{S}_{i_k}^k, \tag{10}$$

$$\mathbb{E}(\mathbf{u}_{i_k}^k|\mathbf{u}_{\mathbf{i}}^{\backslash k}) = \mathbf{m}_{i_k}^k + \mathbf{S}_{i_k}^k\frac{\partial\log Z}{\partial\mathbf{m}_{i_k}^k}, \tag{11}$$

where

$$\mathbf{A} = \frac{\partial\log Z}{\partial\mathbf{m}_{i_k}^k}(\frac{\partial\log Z}{\partial\mathbf{m}_{i_k}^k})^\top - 2\frac{\partial\log Z}{\partial\mathbf{S}_{i_k}^k}.$$

Next, we compute the expectation of the conditional moments (10)(11) w.r.t $q(\mathbf{u}_{\mathbf{i}}^{\backslash k})$, the current posterior of $\mathbf{u}_{\mathbf{i}}^{\backslash k}$. To enable tractable computation, we use the first-order Taylor expansion of (10)(11) at the moments of $\mathbf{u}_{\mathbf{i}}^{\backslash k}$, and then take expectation. Again, it is equivalent to substituting $\mathbb{E}(\mathbf{z}_{\mathbf{i}}^{\backslash k})$ and $\mathrm{tr}\big(\mathbf{S}_{i_k}^k\mathbb{E}(\mathbf{z}_{\mathbf{i}}^{\backslash k}{\mathbf{z}_{\mathbf{i}}^{\backslash k}}^\top)\big)$ for $\mathbf{z}_{\mathbf{i}}^{\backslash k}$ and ${\mathbf{z}_{\mathbf{i}}^{\backslash k}}^\top\mathbf{S}_{i_k}^k\mathbf{z}_{\mathbf{i}}^{\backslash k}$, respectively. Denote the expected conditional mean and covariance by $\boldsymbol{\eta}$ and $\boldsymbol{\Omega}$. We use them to construct a new posterior of $\mathbf{u}_{i_k}^k$ and obtain the updated message $\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|{\mathbf{m}_{\mathbf{i}}^k}^*, {\mathbf{S}_{\mathbf{i}}^k}^*)$ accordingly, where

$${\mathbf{S}_{\mathbf{i}}^k}^* = \left(\boldsymbol{\Omega}^{-1} - {\mathbf{S}_{i_k}^k}^{-1}\right)^{-1},$$

$${\mathbf{m}_{\mathbf{i}}^k}^* = {\mathbf{S}_{\mathbf{i}}^k}^*\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} - {\mathbf{S}_{i_k}^k}^{-1}\mathbf{m}_{i_k}^k\right).$$

2.2 Logistic Regression

In this section, we provide the details of updating messages in EP and CEP for Bayesian logistic regression. Both methods are based on quadrature rules.

First, according to Section 4.2 in the main paper, the joint probability is

$$p(\mathbf{y},\mathbf{w}|\mathbf{X}) = p(\mathbf{w})\prod_{i=1}^{n}1/\big(1+\exp(-(2y_i-1)\mathbf{w}^\top\mathbf{x}_i)\big),$$

where $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda\mathbf{I})$. We use factorized messages to approximate each logistic likelihood,

$$1/\big(1+\exp(-(2y_i-1)\mathbf{w}^\top \mathbf{x}_i)\big) \approx \prod_m \tilde{f}_{im}(w_m)$$

where $\tilde{f}_{im}(w_m) = \mathcal{N}(w_m|\mu_{im}, v_{im})$.

To update the messages for i-th sample, we first obtain the calibrating distribution,

$$q^{\backslash i}(\mathbf{w}) \propto p(\mathbf{w}) \prod_{j\neq i}\prod_m \tilde{f}_{jm}(w_m) = \prod_m \mathcal{N}(w_m|\mu_m^{\backslash i}, v_m^{\backslash i}),$$

where

$$v_m^{\backslash i} = \frac{1}{1/\lambda + \sum_{j\neq i} 1/v_{jm}}, \qquad \mu_m^{\backslash i} = v_m^{\backslash i} \sum_{j\neq i} \frac{\mu_{jm}}{v_{jm}},$$

and the tilted distribution,

$$\hat{p}_i(\mathbf{w}) \propto \frac{1}{1+\exp(-(2y_i-1)\mathbf{w}^\top \mathbf{x}_i)} q^{\backslash i}(\mathbf{w}).$$

2.2.1 Moment Matching in EP

To update each message $\tilde{f}_{im}(w_m)$, we first consider the moment matching in EP. Due to the logistic likelihood, the moments of w_m w.r.t the tilted distribution $\hat{p}_i(\mathbf{w})$ is intractable. To overcome this problem, we use an idea similar to (Gelman et al., 2013). We first split $\mathbf{w}$ into w_m and $\mathbf{w}_{\backslash m}$, and project them onto x_{im} and $\mathbf{x}_{i\backslash m}$, respectively. Here $\mathbf{x}_{i\backslash m}$ are $\mathbf{x}_i$ excluding the m-th element. Consequently, we obtain two random variables, $\eta_1 = w_{im}x_{im}$ and $\eta_2 = \mathbf{w}_{\backslash m}^\top \mathbf{x}_{i\backslash m}$, and their calibrating distributions,

$$q^{\backslash i}(\eta_1) = \mathcal{N}(\eta_1|x_{im}\mu_m^{\backslash i}, x_{im}^2 v_m^{\backslash i}),$$
$$q^{\backslash i}(\eta_2) = \mathcal{N}(\eta_2|\sum_{l\neq m} x_{il}\mu_l^{\backslash i}, \sum_{l\neq m} x_{il}^2 v_l^{\backslash i}).$$

Instead of matching the moments of $\mathbf{w}$ that may demand a high-dimensional integration, we can match the moments for η_1 and η_2, because the integration only involves two variables, and can be accurately approximated by a two-dimensional quadrature. Specifically, the tilted distribution of $\{\eta_1, \eta_2\}$ is

$$\hat{p}_i(\eta_1,\eta_2) \propto q^{\backslash i}(\eta_1)q^{\backslash i}(\eta_2) \cdot \frac{1}{1+\exp(-(2y_i-1)(\eta_1+\eta_2))}.$$

Since w_m is related to η_1, we need to compute the moments of η_1 w.r.t $\hat{p}_i(\eta_1, \eta_2)$. To this end, we use 9 Gauss-Hermite quadrature nodes and weights for both η_1 and η_2, denoted by $\{\gamma_{1j}, \alpha_j\}$ and $\{\gamma_{2j}, \alpha_j\}$, respectively. Note that the nodes are adjusted according to the mean and variance in $q^{\backslash \mathbf{i}}(\eta_1)$ and $q^{\backslash \mathbf{i}}(\eta_2)$, while the weights are the same. We then compute the zeroth, first and second moment of η_1 w.r.t the unnormalized titled distribution,

$$E_0 \approx \sum_k \sum_j \frac{\alpha_k\alpha_j}{1+\exp(-(2y_i-1)(\gamma_{1k}+\gamma_{2j}))},$$
$$E_1 \approx \sum_k \sum_j \frac{\alpha_k\alpha_j\gamma_{1k}}{1+\exp(-(2y_i-1)(\gamma_{1k}+\gamma_{2j}))},$$
$$E_2 \approx \sum_k \sum_j \frac{\alpha_k\alpha_j\gamma_{1k}^2}{1+\exp(-(2y_i-1)(\gamma_{1k}+\gamma_{2j}))}.$$

The mean and variance of η_1 can then be calculated by

$$\beta = \frac{E_1}{E_0}, \quad \sigma^2 = \frac{E_2}{E_0} - \big(\frac{E_1}{E_0}\big)^2.$$

It is easy to verify that σ^2 is always non-negative and therefore valid. Next, we construct a new posterior for η_1, $q^*(\eta_1) = \mathcal{N}(\eta_1|\beta, \sigma^2)$. Dividing $q^*(\eta_1)$ by the calibrating distribution $q^{\backslash i}(\eta_1)$, we then obtain the updated message for $\eta_1, \mathcal{N}(\eta_1|\beta_i, \sigma_i^2)$.

Finally, we map η_1 back to w_m to obtain the updated message for w_m. Specifically, we have

$$\begin{aligned}
&\mathcal{N}(\eta_1|\beta_i, \sigma_i^2) = \mathcal{N}(w_m x_{im}|\beta_i, \sigma_i^2) \\
&\propto \exp(-\frac{w_m^2 x_{im}^2}{2\sigma_i^2} + \frac{w_m x_{im}\beta_i}{\sigma_i^2}) \\
&\propto \mathcal{N}(w_m|\mu_{im}^*, v_{im}^*) = \tilde{f}_{im}(w_m)
\end{aligned} \tag{12}$$

where

$$v_{im}^* = \frac{\sigma_i^2}{x_{im}^2}, \qquad \mu_{im}^* = v_{im}^* \frac{x_{im}\beta_i}{\sigma_i^2}.$$

2.2.2 Moment Matching with CEP

We now consider to use CEP to match moments to update each message $\tilde{f}_{im}(w_m)$. To this end, we first obtain the conditional tilted distribution,

$$\hat{p}_i(w_m|\mathbf{w}_{\backslash m}) \propto q^{\backslash i}(w_m) g_{im}(w_m|\mathbf{w}_{\backslash m}),$$

where $q^{\backslash i}(w_m) = \mathcal{N}(w_m|\mu_m^{\backslash i}, v_m^{\backslash i})$ and

$$g_{im}(w_m|\mathbf{w}_{\backslash m}) = \big(1+\exp((2y_i-1)(w_m x_{im} + \mathbf{w}_{\backslash m}^\top \mathbf{x}_{i\backslash m}))\big)^{-1}.$$

To obtain the conditional moments, we use Gauss-Hermite quadrature with 9 nodes and weights $\{\gamma_j, \alpha_j\}$. As in Section 2.2.1, we first compute the zeroth, first and second moment of w_m w.r.t the unnormalized conditional

tilted distribution,

$$
\begin{aligned}
E_0 &= \int q^{\setminus i}(w_m) g_{im}(w_m|\mathbf{w}_{\setminus m}) dw_m \\
&\approx \sum_j \alpha_j g_{im}(\gamma_j|\mathbf{w}_{\setminus m}), \\
E_1 &= \int w_m q^{\setminus i}(w_m) g_{im}(w_m|\mathbf{w}_{\setminus m}) dw_m \\
&\approx \sum_j \alpha_j \gamma_j g_{im}(\gamma_j|\mathbf{w}_{\setminus m}), \\
E_2 &= \int w_m^2 q^{\setminus i}(w_m) g_{im}(w_m|\mathbf{w}_{\setminus m}) dw_m \\
&\approx \sum_j \alpha_j \gamma_j^2 g_{im}(\gamma_j|\mathbf{w}_{\setminus m}).
\end{aligned}
$$

We then obtain the conditional mean and variance of w_m by

$$
\begin{aligned}
\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) &= \frac{E_1}{E_0}, \\
\mathrm{var}(w_m|\mathbf{w}_{\setminus m}) &= \frac{E_2}{E_0} - \left(\frac{E_1}{E_0}\right)^2.
\end{aligned}
$$

We can also derive the Hessian matrix w.r.t $\mathbf{w}_{\setminus m}$,

$$
\begin{aligned}
&\nabla\nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) \\
&= \frac{E_1 \sum_j t_j c_j - E_0 \sum_j t_j \gamma_j c_j}{E_0^3} \mathbf{x}_{i\setminus m} \mathbf{x}_{i\setminus m}^\top, \\
&\nabla\nabla\mathrm{var}(w_m|\mathbf{w}_{\setminus m}) \\
&= \nabla\nabla\mathbb{E}(w_m^2|\mathbf{w}_{\setminus m}) - 2\nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) \nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m})^\top \\
&- 2\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) \nabla\nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m}),
\end{aligned}
$$

where

$$
\begin{aligned}
&\nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) = \frac{E_1 \sum_j t_j - E_0 \sum_j t_j \gamma_j}{E_0^2} (2y_i - 1) \mathbf{x}_{i\setminus m}, \\
&\nabla\nabla\mathbb{E}(w_m^2|\mathbf{w}_{\setminus m}) = \frac{E_2 \sum_j t_j c_j - E_0 \sum_j t_j \gamma_j^2 c_j}{E_0^3} \mathbf{x}_{i\setminus m} \mathbf{x}_{i\setminus m}^\top, \\
&t_j = \alpha_j g_{im}^2(\gamma_j|\mathbf{w}_{\setminus m}), \\
&c_j = E_0(1 - 2g_{im}(\gamma_j|\mathbf{w}_{\setminus m})) + 2\sum_j \alpha_j g_{im}^2(\gamma_j|\mathbf{w}_{\setminus m}).
\end{aligned}
$$

Next, to compute the expected conditional moments, we derive their first or second order Taylor expansions at the expectation of $\mathbf{w}_{\setminus m}$ and then take expectation w.r.t $q(\mathbf{w}_{\setminus m})$ (see (10)(11) in the main paper). We then use the expected conditional moments to construct a new posterior of w_m, $\mathcal{N}(w_m|\mu_m, v_m)$. When the first order Taylor expansion is used, we obtain

$$
\mu_m = h_1(\mathbb{E}_q(\mathbf{w}_{\setminus m})), \quad v_m = h_2(\mathbb{E}_q(\mathbf{w}_{\setminus m}))
$$

where

$$
h_1(\mathbf{w}_{\setminus m}) = \mathbb{E}(w_m|\mathbf{w}_{\setminus m}), \quad h_2(\mathbf{w}_{\setminus m}) = \mathrm{var}(w_m|\mathbf{w}_{\setminus m}).
$$

When the second order Taylor expansion is used, we obtain

$$
\begin{aligned}
\mu_m &= h_1(\mathbb{E}_q(\mathbf{w}_{\setminus m})) + \frac{1}{2}\mathrm{tr}\left(\mathrm{var}_q(\mathbf{w}_{\setminus m}) h_3(\mathbb{E}_q(\mathbf{w}_{\setminus m}))\right), \\
v_m &= h_2(\mathbb{E}_q(\mathbf{w}_{\setminus m})) + \frac{1}{2}\mathrm{tr}\left(\mathrm{var}_q(\mathbf{w}_{\setminus m}) h_4(\mathbb{E}_q(\mathbf{w}_{\setminus m}))\right).
\end{aligned}
$$

where

$$
\begin{aligned}
h_3(\mathbf{w}_{\setminus m}) &= \nabla\nabla\mathbb{E}(w_m|\mathbf{w}_{\setminus m}), \\
h_4(\mathbf{w}_{\setminus m}) &= \nabla\nabla\mathrm{var}(w_m|\mathbf{w}_{\setminus m}).
\end{aligned}
$$

Finally, dividing $\mathcal{N}(w_m|\mu_m, v_m)$ by the calibrating distribution $\mathcal{N}(w_m|\mu_m^{\setminus i}, v_m^{\setminus i})$, we obtain the updated message $\tilde{f}_{im}(w_m) = \mathcal{N}(w_m|\mu_{im}^*, v_{im}^*)$, where

$$
\begin{aligned}
v_{im}^* &= \left(\frac{1}{v_m} - \frac{1}{v_m^{\setminus i}}\right)^{-1}, \\
\mu_{im}^* &= v_{im}^* \left(\frac{\mu_m}{v_m} - \frac{\mu_m^{\setminus i}}{v_m^{\setminus i}}\right).
\end{aligned}
$$

2.3 Probit Regression

In this section, we provide the details of message updating in CEP for Bayesian probit regression. Given the data, the joint probability of the model is

$$
p(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p(\mathbf{w}) \prod_{i=1}^n \psi\left((2y_i - 1)\mathbf{w}^\top \mathbf{x}_i\right),
$$

where $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda\mathbf{I})$ and $\psi(\cdot)$ is the cumulative density function (CDF) of the standard Gaussian distribution. We use factorized messages to approximate each likelihood,

$$
\psi\left((2y_i - 1)\mathbf{w}^\top \mathbf{x}_i\right) \approx \prod_m \tilde{f}_{im}(w_m).
$$

The approximate posterior is therefore

$$
q(\mathbf{w}) \propto p(\mathbf{w}) \prod_{i=1}^n \prod_m \tilde{f}_{im}(w_m).
$$

To update the messages for the i-th sample, we derive the calibrating distribution

$$
q^{\setminus i}(\mathbf{w}) \propto p(\mathbf{w}) \prod_{j\neq i} \prod_m \tilde{f}_{jm}(w_m) = \prod_m \mathcal{N}(w_m|\mu_m^{\setminus i}, v_m^{\setminus i}),
$$

where

$$
v_m^{\setminus i} = \frac{1}{1/\lambda + \sum_{j\neq i} 1/v_{jm}}, \quad \mu_m^{\setminus i} = v_m^{\setminus i} \sum_{j\neq i} \frac{\mu_{jm}}{v_{jm}},
$$

and the tilted distribution

$$\hat{p}_i(\mathbf{w}) \propto \psi\left((2y_i-1)\mathbf{w}^\top \mathbf{x}_i\right) q^{\setminus i}(\mathbf{w}).$$

To update each message $\tilde{f}_{im}(w_m)$, we first derive the conditional tilted distribution,

$$\hat{p}_i(w_m|\mathbf{w}_{\setminus m}) \propto \psi\left((2y_i-1)\mathbf{w}^\top \mathbf{x}_i\right) \mathcal{N}(w_m|\mu_m^{\setminus i}, v_m^{\setminus i}).$$

We can use the same method in (Dusek, 2013) to obtain the conditional moments,

$$\mathbb{E}(w_m|\mathbf{w}_{\setminus m}) = \mu_m^{\setminus i} + v_m^{\setminus i}\frac{\partial \log Z_{im}}{\partial \mu_m^{\setminus i}},$$

$$\mathrm{cov}(w_m|\mathbf{w}_{\setminus m}) = v_m^{\setminus i} - v_m^{\setminus i 2}\left((\frac{\partial \log Z_{im}}{\partial \mu_m^{\setminus i}})^2 - 2\frac{\partial \log Z_{im}}{\partial v_m^{\setminus i}}\right)$$

where

$$\begin{aligned} Z_{im} &= \int \psi\left((2y_i-1)\mathbf{w}^\top \mathbf{x}_i\right)\mathcal{N}(w_m|\mu_m^{\setminus i}, v_m^{\setminus i}) \\ &= \psi\left(\frac{(2y_i-1)(x_{im}\mu_m^{\setminus i} + \mathbf{w}_{\setminus m}^\top \mathbf{x}_{i\setminus m})}{\sqrt{1+x_{im}^2 v_m^{\setminus i}}}\right). \end{aligned}$$

For simplicity, we define

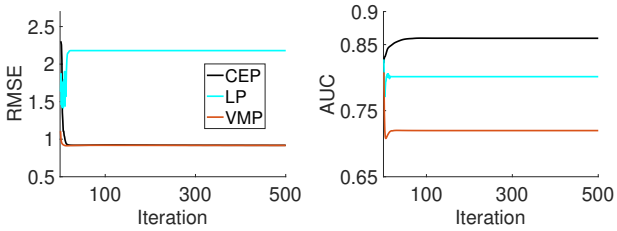
$$\begin{aligned} \mathcal{N} &= \mathcal{N}\left(\frac{(2y_i-1)(x_{im}\mu_m^{\setminus i} + \mathbf{w}_{\setminus m}^\top \mathbf{x}_{i\setminus m})}{\sqrt{1+x_{im}^2 v_m^{\setminus i}}}|0,1\right), \\ \psi &= \psi\left(\frac{(2y_i-1)(x_{im}\mu_m^{\setminus i} + \mathbf{w}_{\setminus m}^\top \mathbf{x}_{i\setminus m})}{\sqrt{1+x_{im}^2 v_m^{\setminus i}}}\right), \\ c_1 &= \frac{(2y_i-1)}{\sqrt{1+x_{im}^2 v_m^{\setminus i}}}, \\ c_2 &= x_{im}\mu_m^{\setminus i} + \mathbf{w}_{\setminus m}^\top \mathbf{x}_{i\setminus m}. \end{aligned}$$

We can then derive the Hessian matrix of the conditional moments w.r.t $\mathbf{w}_{\setminus m}$,

$$\begin{aligned} \nabla\nabla\mathbb{E}(w_m|\mathbf{w}_{i\setminus i}) &= T_1 c_1^3 v_m^{\setminus i} x_{im}\mathbf{x}_{i\setminus m}\mathbf{x}_{i\setminus m}^\top, \\ \nabla\nabla\mathrm{cov}(w_m|\mathbf{w}_{i\setminus i}) &= T_2 c_1^4 v_m^{\setminus i 2} x_{im}^2 \mathbf{x}_{i\setminus m}\mathbf{x}_{i\setminus m}^\top, \end{aligned}$$

where

$$\begin{aligned} T_1 &= (c_1^2c_2^2-1)\frac{\mathcal{N}}{\psi} + 3c_1c_2\frac{\mathcal{N}^2}{\psi^2} + 2\frac{\mathcal{N}^3}{\psi^3}, \\ T_2 &= c_1c_2(3-c_1^2c_2^2)\frac{\mathcal{N}}{\psi} + (4-7c_1^2c_2^2)\frac{\mathcal{N}^2}{\psi^2} \\ &\quad - 12c_1c_2\frac{\mathcal{N}^3}{\psi^3} - 6\frac{\mathcal{N}^4}{\psi^4}. \end{aligned}$$



(a) *Alog* (b) *Enron*

Figure 1: Average prediction accuracy *v.s.* running iteration. The rank of the embeddings is 3.

Next, to compute the expected conditional moments, we use their first or second order Taylor expansions at the expectation of $\mathbf{w}_{\setminus m}$ and then take expectation w.r.t $q(w_{\setminus m})$ (see (10)(11) in the main paper). We then use the expected conditional moments to construct a new posterior of w_m, from which we can update the message $\tilde{f}_{im}(w_m)$. The process is the same as in Section 2.2.2.

3 Experiment

In this section, we supplement more experimental results.

3.1 Bayesian Probit and Logistic Regression

We first in Table 1 list the results of test AUC on six real-world datasets from UCI machine learning repository[1], *australian*, *breast*, *crab*,*ionos*, *pima* and *sonar*. As we can see, for Bayesian probit regression, CEP-1 is better than or close to EP; CEP-2 always obtains the best test AUC among all the methods. For Bayesian logistic regression, both CEP-1 and CEP-2 are close to EP. CEP-2 obtains the highest AUC on *brerast*, *ionos*, *pima* and *sonar* while CEP-1 on the remaining two, *australian* and *crab*. Note that KJIT performs the best on *crab* and the worst on *pima* and *sonar*. These results are consist with the test log-likelihoods in Table 1 of the main paper.

3.2 Bayesian Tensor Decomposition

Next, we show in Fig. 1, 2 and 3 how the predictive performance of CEP, LP and VMP vary along with the running iterations when rank of the embeddings is 3, 5 and 8 respectively. Note that in the main paper, we only show the performance when the rank is 10 (due to the space limit). As we can see, the prediction accuracy of all the three methods converge rapidly and remain stable with more iterations. The predictive performance of CEP is close to or slightly better than VMP on *Alog* dataset. On *Enron* dataset, CEP significantly outperforms VMP. On

[1] https://archive.ics.uci.edu/ml/index.php

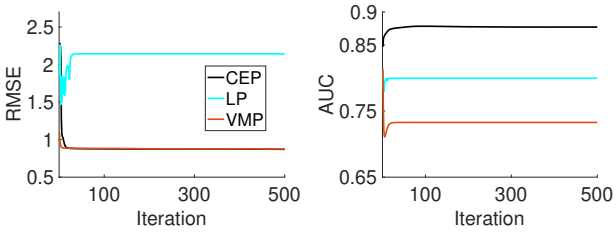
Dataset	CEP-1	CEP-2	LP	EP	VB	KJIT
australian	$\mathbf{0.873 \pm 0.008}$	0.873 ± 0.010	0.873 ± 0.009	0.873 ± 0.010	0.873 ± 0.009	0.873 ± 0.009
breast	0.675 ± 0.021	$\mathbf{0.683 \pm 0.020}$	0.675 ± 0.021	0.677 ± 0.020	0.676 ± 0.019	0.676 ± 0.019
crab	0.993 ± 0.001	0.992 ± 0.001	0.993 ± 0.001	0.993 ± 0.001	0.993 ± 0.001	$\mathbf{0.994 \pm 0.001}$
ionos	0.912 ± 0.011	$\mathbf{0.925 \pm 0.010}$	0.912 ± 0.012	0.914 ± 0.010	0.908 ± 0.010	0.902 ± 0.010
pima	0.830 ± 0.005	$\mathbf{0.831 \pm 0.005}$	$\mathbf{0.831 \pm 0.005}$	0.830 ± 0.005	0.830 ± 0.005	0.819 ± 0.011
sonar	0.820 ± 0.015	$\mathbf{0.831 \pm 0.016}$	0.820 ± 0.016	0.827 ± 0.016	0.825 ± 0.014	0.651 ± 0.041

(a) Bayesian logistic regression

Dataset	CEP-1	CEP-2	LP	EP	VB
australian	$\mathbf{0.883 \pm 0.009}$	$\mathbf{0.883 \pm 0.009}$	$\mathbf{0.883 \pm 0.009}$	0.880 ± 0.009	0.878 ± 0.010
breast	0.686 ± 0.016	$\mathbf{0.690 \pm 0.016}$	0.681 ± 0.020	0.677 ± 0.020	0.672 ± 0.022
crab	$\mathbf{0.995 \pm 0.002}$	$\mathbf{0.995 \pm 0.002}$	$\mathbf{0.995 \pm 0.002}$	$\mathbf{0.995 \pm 0.002}$	$\mathbf{0.995 \pm 0.002}$
ionos	0.926 ± 0.006	$\mathbf{0.929 \pm 0.006}$	0.896 ± 0.006	0.903 ± 0.005	0.883 ± 0.010
pima	$\mathbf{0.825 \pm 0.004}$	$\mathbf{0.825 \pm 0.004}$	$\mathbf{0.825 \pm 0.004}$	$\mathbf{0.825 \pm 0.004}$	$\mathbf{0.825 \pm 0.004}$
sonar	0.813 ± 0.023	$\mathbf{0.831 \pm 0.022}$	0.804 ± 0.020	0.813 ± 0.022	0.797 ± 0.021

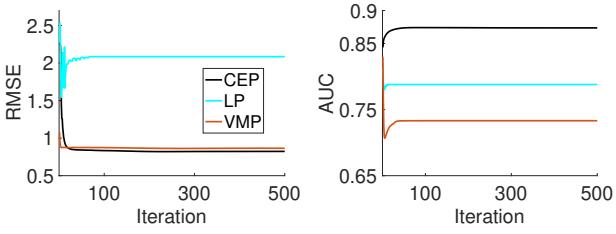
(b) Bayesian probit regression

Table 1: Average test AUC on six real datasets.



(a) *Alog* (b) *Enron*

Figure 2: Average prediction accuracy *v.s.* running iteration. The rank of the embeddings is 5.



(a) *Alog* (b) *Enron*

Figure 3: Average prediction accuracy *v.s.* running iteration. The rank of the embeddings is 8.

both datasets, CEP improves upon LP by a large margin. These results are consistent with that when the rank of the embeddings is 10 (see Fig. 4 in the main paper).