Perturbation Theory for Variational Inference

Manfred Opper
TU Berlin

Marco Fraccaro
Technical University of Denmark

Ulrich Paquet
Apple

Alex Susemihl
TU Berlin

Ole Winther
Technical University of Denmark

Abstract

The variational approximation is known to underestimate the true variance of a probability measure, like a posterior distribution. In latent variable models whose parameters are permutation-invariant with respect to the likelihood, the variational approximation might self-prune and ignore its parameters. We view the variational free energy and its accompanying evidence lower bound as a first-order term from a perturbation of the true log partition function and derive a power series of corrections. We sketch by means of a "variational matrix factorization" example how a further term could correct for predictions from a self-pruned approximation.

1 Introduction

We consider the probability measure

$$p(x) = \frac{\mu(x)e^{-H(x)}}{Z}$$

on a random variable x. We are often concerned with the computation of the negative log partition function

$$-\log Z = -\log \int d\mu(x)\, e^{-H(x)}\,,$$

or of expectations $E[f(x)] = \int d\mu(x)\, f(x)e^{-H(x)}$ under p. As these are typically not analytically tractable, the **variational approximation** introduces a tractable approximation measure $q(x) = \frac{1}{Z_q}\mu(x)e^{-H_q(x)}$, and adjusts the parameters in H_q in such a way that the free energy which is defined via the Kullback Leibler divergence

$$F[q] = D(q,p) - \log Z = E_q\left[\log \frac{q(x)}{p(x)}\right] - \log Z$$

is minimal. This yields the form

$$F[q] = -\log Z_q + E_q[V] \tag{1}$$

where $V = H - H_q$. The minimizer of F usually underestimates the variance of p, and symmetries are sometimes not broken [1, 3], so that the approximation self-prunes. In Sec. 3 we illustrate this effect with a small matrix factorization example, and then show how predictions (expectations $E[f(x)]$ of $f(x)$ under the posterior p) could be improved.

2 Perturbation theory

Perturbation theory aims to find approximate solutions to a problem given exact solutions of a simpler related sub-problem (the VB solution in our case). If we define $\widehat{H}_\lambda = H_q + \lambda V =$

$(1-\lambda)H_q + \lambda H$, we have $\widehat{H}_1 = H$ and $\widehat{H}_0 = H_q$. Let us now define the perturbation expansion using the cumulant expansion of $\log E_q[e^{-V}]$, where E_q denotes the expectation with respect to the variational distribution:

$$
\begin{aligned}
-\log \int d\mu(x)\; e^{-\widehat{H}_\lambda(x)} &= -\log Z_q - \log E_q[e^{-\lambda V(x)}] \\
&= \underbrace{-\log Z_q + \lambda E_q[V]}_{F[q] \text{ for } \lambda=1} - \frac{\lambda^2}{2} E_q\left[(V - E_q[V])^2\right] + \frac{\lambda^3}{3!} E_q\left[(V - E_q[V])^3\right] + \dots \qquad (2)
\end{aligned}
$$

At the end, of course, we set $\lambda = 1$. The *first order term* in (2) yields the variational free energy $F[q]$ in (1), i.e. the negative of the usual evidence lower bound (ELBO). It is therefore reasonable to correct it using higher orders. Note that this may not be a convergent series, but lead to an asymptotic expansion only. The approach is similar to corrections to Expectation Propagation [4, 6], but the computation of the correction terms here require less effort. A variational linear response correction can also be applied [5].

Expectations. In a similar way, we can compute corrections for expectations of functions. We first define E_λ as the expectation with respect to $p_\lambda(x) = \mu(x) e^{-\widehat{H}_\lambda(x)}$ (we then have $E_q = E_0$) and notice that

$$
\begin{aligned}
E_\lambda[f(x)] &= \frac{\int d\mu(x)\; f(x)\; e^{-\widehat{H}_\lambda(x)}}{\int d\mu(x)\; e^{-\widehat{H}_\lambda(x)}} = \frac{\int d\mu(x)\; f(x)\; e^{-\widehat{H}_0(x) - \lambda V(x)}}{\int d\mu(x)\; e^{-\widehat{H}_0(x) - \lambda V(x)}} \\
&= \frac{\int d\mu(x)\; f(x)\; e^{-\widehat{H}_0(x)} \left(1 - \lambda V(x) + \frac{\lambda^2}{2} V^2(x) - \frac{\lambda^3}{3!} V^3(x) \pm \dots\right)}{\int d\mu(x)\; e^{-\widehat{H}_0(x)} \left(1 - \lambda V(x) + \frac{\lambda^2}{2} V^2(x) - \frac{\lambda^3}{3!} V^3(x) \pm \dots\right)} \\
&= \frac{E_0\left[f(x)\left(1 - \lambda V + \frac{\lambda^2}{2} V^2 - \frac{\lambda^3}{3!} V^3 \pm \dots\right)\right]}{E_0\left[\left(1 - \lambda V + \frac{\lambda^2}{2} V^2 - \frac{\lambda^3}{3!} V^3 \pm \dots\right)\right]} \,. \qquad (3)
\end{aligned}
$$

Using $\frac{1}{1-z} = 1 + z + z^2 + \dots$ we can re-expand the part from the denominator[1] in (3) to get

$$
\frac{1}{1 - (\lambda E_0[V] - \frac{\lambda^2}{2} E_0[V^2] + \frac{\lambda^3}{3!} E_0[V^3] \pm \dots)} = 1 + \lambda E_0[V] - \frac{\lambda^2}{2} E_0[V^2] + \lambda^2 E_0[V]^2 \pm \dots \qquad (4)
$$

Putting this together with the numerator in (3) yields

$$
\begin{aligned}
E_\lambda[f(x)] &= \left(E_0[f(x)] - \lambda E_0[f(x)V] + \frac{\lambda^2}{2} E_0[f(x)V^2] \pm \dots\right) \\
&\quad \times \left(1 + \lambda E_0[V] - \frac{\lambda^2}{2} E_0[V^2] + \lambda^2 E_0[V]^2 \pm \dots\right) \\
&= E_0[f(x)] - \lambda E_0[f(x)V] + \lambda E_0[f(x)] E_0[V] + \frac{\lambda^2}{2} E_0[f(x)V^2] \\
&\quad - \frac{\lambda^2}{2} E_0[f(x)] E_0[V^2] - \lambda^2 E_0[f(x)V] E_0[V] + \lambda^2 E_0[f(x)] E_0[V]^2 \pm \dots \\
&= E_0[f(x)] - \lambda \mathrm{Cov}_0[f(x), V(x)] - \lambda^2 E_0[V(x)] \mathrm{Cov}_0[f(x), V(x)] \\
&\quad + \frac{\lambda^2}{2} \mathrm{Cov}_0[f(x), V^2(x)] \pm \dots \qquad (5)
\end{aligned}
$$

3 Variational Matrix Factorization

As toy example, we factorize a scalar "matrix" $r = xy + \epsilon$ with $\epsilon \sim \mathcal{N}(\epsilon; 0, \beta^{-1})$. In the parlance of recommender systems (as we'll later consider sparsely observed matrices R) a user (modelled by a latent scalar x) rates an item (modelled by y). One rating

$$
r \sim \mathcal{N}(r; xy, \beta^{-1})
$$

[1] Note $\frac{1}{1-z} = 1 + z + z^2 + \dots$ only converges for $|z| < 1$.

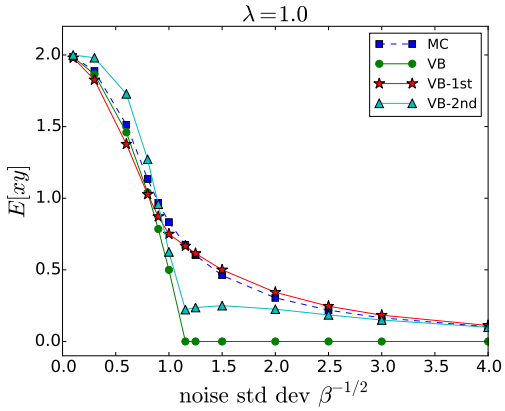

Figure 1: The expected affinity $E[xy]$ under $p(x, y|r, \beta)$ as a Monte Carlo ground truth (MC), and its approximations. We plot the results with $r = 2$ and $\alpha_x = \alpha_y = 1$. With these parameters the VB estimate is zero for $\beta^{-\frac{1}{2}} > 2/\sqrt{3}$ (see Fig. 2 and analysis in [2]), but is accurately corrected through the first order term in (7). Under less noise symmetry is broken ($\beta^{-\frac{1}{2}} < 2/\sqrt{3}$), and the first order correction (6) is remains accurate.

is observed. Assume priors $p(x) = \mathcal{N}(x; 0, \alpha_x^{-1})$ and $p(y) = \mathcal{N}(y; 0, \alpha_y^{-1})$. Let $q(x, y) = q(x)q(y)$ be a factorized approximation with $q(x) = \mathcal{N}(x; \mu_x, \gamma_x^{-1})$ and $q(y) = \mathcal{N}(y; \mu_y, \gamma_y^{-1})$. We will investigate how the variational prediction of a rating $E_0[xy]$ and its further terms in (5) change as the observation noise standard deviation $\beta^{-\frac{1}{2}}$ increases.

Motivation. This toy example is the building block for computing the corrections for real recommender system models, i.e. matrix factorization models with sparsely observed ratings matrices and K latent dimension for both user and item vectors. As shown in the Supplementary Material, the corrections for the predicted rating from user i to item j, $E[\mathbf{x}_i^T \mathbf{y}_j]$, can be shown to be the sum of $K(N_i + M_j)$ corrections like the ones presented in this section, when q fully factorizes. N_i denotes the number of items rated by user i and M_j the number of users that have rated item j. We expect the correction to be largest when N_i and M_j are small.

Before commencing to technical details, consider Fig. 1, which is accompanied by Fig. 2. When there is little observation noise, with $\beta^{-\frac{1}{2}}$ being small, the variational Bayes (VB) solution $E_0[xy]$ closely matches $E[xy]$, obtained from a Monte Carlo (MC) estimate. As the noise is increased, the VB solution stops breaking symmetry and snaps to zero (see Fig. 2, and Nakajima *et al.*'s analysis [3]), after which $E_0[xy] = 0$ will always be predicted. The first order correction gives a substantial improvement towards the ground truth, although with no guarantee that (5) is a convergent series, the inclusion of a second order term incorporates a degradation.

Technical details. The terms that constitute $V(x, y) = H(x, y) - H_q(x, y)$ are

$$H(x, y) = \frac{1}{2}\left[\beta(r - xy)^2 + \alpha_x x^2 + \alpha_y y^2\right] \qquad H_q(x, y) = \frac{1}{2}\left[\gamma_x(x - \mu_x)^2 + \gamma_y(y - \mu_y)^2\right] .$$

Let's look at the first order of the expansion in (5),

$$E[xy] = E_q[xy] - \lambda E_q[xyV] + \lambda E_q[xy]E_q[V] ,$$

bearing in mind that q is the minimizer of $F[q] = -\log Z_q + E_q[V]$. The terms that we require in the expansion are

$$\begin{aligned} E_q[xy] &= \mu_x\mu_y \\ E_q[xyV] - E_q[xy]E_q[V] &= -\frac{1}{2}\left[\mu_x\mu_y\left\{-2\frac{\alpha_x}{\gamma_x} - 2\frac{\alpha_y}{\gamma_y} - 8\frac{\beta}{\gamma_x\gamma_y}\right\} + \mu_x^2\left\{\frac{2\beta r}{\gamma_y}\right\} + \mu_y^2\left\{\frac{2\beta r}{\gamma_x}\right\}\right. \\ &\qquad \left. + \frac{2\beta r}{\gamma_x\gamma_y} - \mu_x\mu_y^3\left\{\frac{2\beta}{\gamma_x}\right\} - \mu_x^3\mu_y\left\{\frac{2\beta}{\gamma_y}\right\}\right] , \end{aligned} \tag{6}$$

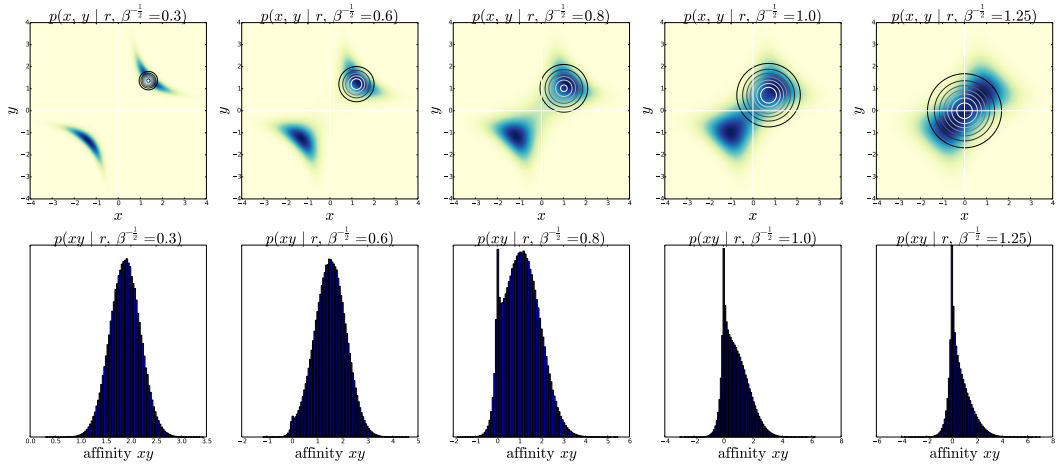

Figure 2: The posterior $p(x, y|r, \beta^{-\frac{1}{2}})$, and the accompanying fully factorized VB solution (circular contours), for $r = 2$. The approximation breaks symmetry at $\beta^{-\frac{1}{2}} = \frac{2}{\sqrt{3}}$. Fig. 1 presents corrections to the mean of $p(xy|r, \beta^{-\frac{1}{2}})$.

with the latter being $\mathrm{Cov}_q[xy, V]$. To illustrate the effect of the correction, consider high observation noise $\beta^{-\frac{1}{2}} > \frac{2}{\sqrt{3}}$ in Fig. 1. The VB approximation locks on to a zero mean [3], and does not break symmetry [1]. With $\mu_u = \mu_v = 0$ for example, the only remaining term to first order is (using $\lambda = 1$)

$$E[xy] \approx \frac{\beta}{\gamma_x \gamma_y} r \,, \tag{7}$$

and the correction is verifiably in the direction of r.

4 Summary

We've motivated, by means of a toy example, the benefit of and framework for a perturbation theoretical view on variational inference. Outside the scope of the workshop, current and future work encompass the application to matrix factorization (see Supplementary Material), variational inference of stochastic differential equations, Gaussian Process (GP) classification, and the variational approach for sparse GP regression.

References

[1] D. J. C. MacKay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Technical report, University of Cambridge Cavendish Laboratory, 2001.

[2] S. Nakajima and M. Sugiyama. Theoretical analysis of Bayesian matrix factorization. *Journal of Machine Learning Research*, 12(Nov):2583–2648, 2011.

[3] S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *Journal of Machine Learning Research*, 14(Jan):1–37, 2013.

[4] M. Opper, U. Paquet, and O. Winther. Perturbative corrections for approximate inference in Gaussian latent variable models. *Journal of Machine Learning Research*, 14(Sep):2857–2898, 2013.

[5] M. Opper and O. Winther. Variational linear response. In *Advances in Neural Information Processing Systems 16*, pages 1157–1164. MIT Press, 2004.

[6] U. Paquet, O. Winther, and M. Opper. Perturbation corrections in approximate inference: Mixture modelling applications. *Journal of Machine Learning Research*, 10(Jun):1263–1304, 2009.

Supplementary Material

Matrix factorization - general case

We now consider a recommender system with N users, M items and K latent dimensions. An $N \times M$ sparse rating matrix $\mathbf{R}$ can be constructed using each user's list of rated items in the catalogue. An element r_{ij} in $\mathbf{R}$ contains the rating that user i has given to item j. Matrix factorization techniques factorize the rating matrix as $\mathbf{R} = \mathbf{X}^T\mathbf{Y} + \mathbf{E}$, where $\mathbf{X}$ is an $K \times N$ user matrix, the item matrix $\mathbf{Y}$ is $K \times M$ and the residual term $\mathbf{E}$ is $N \times M$. We denote the set of observed entries in $\mathbf{R}$ by $\mathbb{R}$.

We use a Gaussian likelihood for the ratings, i.e. $r_{ij} \sim \mathcal{N}(r_{ij}; \mathbf{x}_i^T\mathbf{y}_j, \beta^{-1})$, and isotropic Gaussian priors for both user and item vectors:

$$p(\mathbf{x}_i) = \prod_{k=1}^{K} \mathcal{N}(x_{ik}; 0, \alpha_x^{-1}) \qquad p(\mathbf{y}_j) = \prod_{k=1}^{K} \mathcal{N}(y_{jk}; 0, \alpha_y^{-1}) \,. \tag{8}$$

It is common to take take a fully factorized approximation for user i's and item j's vectors over dimensions $k = 1, \ldots, K$:[2]

$$q(\mathbf{x}_i) = \prod_k q(x_{ik}) = \prod_k \mathcal{N}(x_{ik}; \mu_{ik}, \gamma_{ik}^{-1}) \qquad q(\mathbf{y}_j) = \prod_k q(y_{jk}) = \prod_k \mathcal{N}(y_{jk}; \eta_{jk}, \zeta_{jk}^{-1}) \,.$$

The main quantity of interest in a recommender system is the predicted rating $E[\mathbf{x}_i^T\mathbf{y}_j]$, whose first order correction requires the computation of $\mathrm{Cov}_0[\mathbf{x}_i^T\mathbf{y}_j, V(\mathbf{X}, \mathbf{Y})]$. With our assumptions, the functions H and H_q in $V(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}, \mathbf{Y}) - H_q(\mathbf{X}, \mathbf{Y})$ are

$$H(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\left[\beta \sum_{(i,j)} (r_{ij} - \mathbf{x}_i^T\mathbf{y}_j)^2 + \alpha_x \sum_i \mathbf{x}_i^T\mathbf{x}_i + \alpha_y \sum_j \mathbf{y}_j^T\mathbf{y}_j\right] \,,$$

$$H_q(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\left[\sum_{i,k} \gamma_{ik}(x_{ik} - \mu_{ik})^2 + \sum_{j,k} \zeta_{jk}(y_{jk} - \eta_{jk})^2\right] \,.$$

Counterintuitively, we see that while we are only interested in correcting $E_0[\mathbf{x}_i^T\mathbf{y}_j]$, both H and H_q depend also on terms relative to items not seen by user i and users that have not rated item j. As we will see shortly, however, these terms do not contribute in $\mathrm{Cov}_0[\mathbf{x}_i^T\mathbf{y}_j, V(\mathbf{X}, \mathbf{Y})]$. In its unsimplified form,

$$\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, V(\mathbf{X}, \mathbf{Y})\right] = \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \frac{1}{2}\left\{\beta \sum_{(n,m)\in\mathbb{R}} (r_{nm} - \mathbf{x}_n^T\mathbf{y}_m)^2 + \alpha_x \sum_{n=1}^{N} \mathbf{x}_n^T\mathbf{x}_n + \alpha_y \sum_{m=1}^{M} \mathbf{y}_m^T\mathbf{y}_m \right.\right.$$
$$\left.\left. - \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}(x_{nk} - \mu_{nk})^2 - \sum_{m=1}^{M}\sum_{k=1}^{K} \zeta_{mk}(y_{mk} - \eta_{mk})^2 \right\}\right]$$

In the computation of $\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, V(\mathbf{X}, \mathbf{Y})\right]$, all the terms in $V(\mathbf{X}, \mathbf{Y})$ that do not depend on $\mathbf{x}_i$ or $\mathbf{y}_j$ are constants and leave the covariance unchanged: considering the user vectors we have for example

$$\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \alpha_x \sum_{n=1}^{N} \mathbf{x}_n^T\mathbf{x}_n\right] = \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \alpha_x \mathbf{x}_i^T\mathbf{x}_i\right] \,.$$

In a similar way, among the likelihood terms $\sum_{(n,m)\in\mathbb{R}} (r_{nm} - \mathbf{x}_n^T\mathbf{y}_m)^2$ we need only to keep the ones corresponding to the items rated by user i and the users that rated item j (i.e. only the i-th row and j-th column of the ratings matrix). Defining $\mathcal{M}(i)$ as the set containing the indices of the items rated by user i and $\mathcal{N}(j)$ the set containing the indices of the users that rated item j, we have

$$\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{(n,m)\in\mathbb{R}} (r_{nm} - \mathbf{x}_n^T\mathbf{y}_m)^2\right] = \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{m\in\mathcal{M}(i)} (r_{im} - \mathbf{x}_i^T\mathbf{y}_m)^2 + \sum_{\substack{n\in\mathcal{N}(j) \\ n\neq i}} (r_{nj} - \mathbf{x}_n^T\mathbf{y}_j)^2\right] \,.$$

[2] Notice that we now use η and ζ for item mean and precisions, to avoid subscript spaghetti.

After removing all the constant terms the simplified covariance is given by

$$\begin{aligned}
\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, V\right] &= \mathrm{Cov}_0\Bigg[\mathbf{x}_i^T\mathbf{y}_j, \frac{\beta}{2}\sum_{m\in\mathcal{M}(i)}(r_{im}-\mathbf{x}_i^T\mathbf{y}_m)^2 + \frac{\beta}{2}\sum_{\substack{n\in\mathcal{N}(j)\\ n\neq i}}(r_{nj}-\mathbf{x}_n^T\mathbf{y}_j)^2 + \\
&\qquad + \frac{\alpha_x}{2}\mathbf{x}_i^T\mathbf{x}_i + \frac{\alpha_y}{2}\mathbf{y}_j^T\mathbf{y}_j - \frac{1}{2}\sum_{k=1}^K\gamma_{ik}(x_{ik}-\mu_{ik})^2 - \frac{1}{2}\sum_{k=1}^K\zeta_{jk}(y_{jk}-\eta_{jk})^2\Bigg] \\
&= \frac{\beta}{2}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{m\in\mathcal{M}(i)}(r_{im}-\mathbf{x}_i^T\mathbf{y}_m)^2\right] + \frac{\beta}{2}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{\substack{n\in\mathcal{N}(j)\\ n\neq i}}(r_{nj}-\mathbf{x}_n^T\mathbf{y}_j)^2\right] + \\
&\quad + \frac{1}{2}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \alpha_x\mathbf{x}_i^T\mathbf{x}_i + \alpha_y\mathbf{y}_j^T\mathbf{y}_j - \sum_{k=1}^K\gamma_{ik}(x_{ik}-\mu_{ik})^2 - \sum_{k=1}^K\zeta_{jk}(y_{jk}-\eta_{jk})^2\right] \\
&= \frac{\beta}{2}f_{\mathbf{x}}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{Y}_i) + \frac{\beta}{2}f_{\mathbf{y}}(\mathbf{x}_i,\mathbf{y}_j,\mathbf{X}_j) + \frac{1}{2}g(\mathbf{x}_i,\mathbf{y}_j)\ .
\end{aligned}$$

where $\mathbf{X}_j$ is the restriction of $\mathbf{X}$ to the columns indexed by $\mathcal{N}(j)$, and $\mathbf{Y}_i$ is the restriction of $\mathbf{Y}$ to the columns indexed by $\mathcal{M}(i)$.

Computation of $g(\mathbf{x}_i,\mathbf{y}_j)$. Thanks to our choice of a fully factorized VB distribution and the bi-linearity of the covariance operator we have

$$\begin{aligned}
g(\mathbf{x}_i,\mathbf{y}_j) &= \mathrm{Cov}_0\left[\sum_{k=1}^K x_{ik}y_{ik}, \sum_{k=1}^K\left(\alpha_x x_{ik}^2 + \alpha_y y_{jk}^2 - \gamma_{ik}(x_{ik}-\mu_{ik})^2 - \zeta_{jk}(y_{jk}-\eta_{jk})^2\right)\right] \\
&= \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{ik}, \alpha_x x_{ik}^2 + \alpha_y y_{jk}^2 - \gamma_{ik}(x_{ik}-\mu_{ik})^2 - \zeta_{jk}(y_{jk}-\eta_{jk})^2\right]\ .
\end{aligned}$$

Computation of $f_{\mathbf{x}}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{Y}_i)$ and $f_{\mathbf{y}}(\mathbf{x}_i,\mathbf{y}_j,\mathbf{X}_j)$. We now focus below only on the computation of $f_{\mathbf{x}}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{Y}_i)$, as the results for $f_{\mathbf{y}}(\mathbf{x}_i,\mathbf{y}_j,\mathbf{X}_j)$ are analogous.
We rewrite the term $f_{\mathbf{x}}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{Y}_i)$ as

$$\begin{aligned}
f_{\mathbf{x}}(\mathbf{x}_i,\mathbf{y}_i,\mathbf{Y}_i) &= \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{m\in\mathcal{M}(i)}(r_{im}-\mathbf{x}_i^T\mathbf{y}_m)^2\right] \\
&= \sum_{m\in\mathcal{M}(i)}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, (r_{im}-\mathbf{x}_i^T\mathbf{y}_m)^2\right] \\
&= \sum_{m\in\mathcal{M}(i)}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, r_{im}^2 - 2r_{im}\mathbf{x}_i^T\mathbf{y}_m + (\mathbf{x}_i^T\mathbf{y}_m)^2\right] \\
&= \sum_{m\in\mathcal{M}(i)}\left\{-2r_{im}\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \mathbf{x}_i^T\mathbf{y}_m\right] + \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, (\mathbf{x}_i^T\mathbf{y}_m)^2\right]\right\}
\end{aligned}$$

and analyse separately each of the covariance terms. Thanks to the full factorization of the VB posterior we have

$$\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \mathbf{x}_i^T\mathbf{y}_m\right] = \mathrm{Cov}_0\left[\sum_{k=1}^K x_{ik}y_{jk}, \sum_{k=1}^K x_{ik}y_{mk}\right] = \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right]\ .$$

For $\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, (\mathbf{x}_i^T\mathbf{y}_m)^2\right]$ we get instead

$$\begin{aligned}
\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, (\mathbf{x}_i^T\mathbf{y}_m)^2\right] &= \mathrm{Cov}_0\left[\sum_{k=1}^K x_{ik}y_{jk}, \sum_{k=1}^K (x_{ik}y_{mk})^2 + 2\sum_{k=1}^K\sum_{p<k} x_{ik}y_{mk}x_{ip}y_{mp}\right] \\
&= \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right] + \mathrm{Cov}_0\left[\sum_{k=1}^K x_{ik}y_{jk}, 2\sum_{k=1}^K\sum_{p<k} x_{ik}y_{mk}x_{ip}y_{mp}\right] \\
&= \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right] + \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, 2\left(\sum_{c\neq k} x_{ic}y_{mc}\right)x_{ik}y_{mk}\right] \\
&= \sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right] + \sum_{k=1}^K 2\left(\sum_{c\neq k} E_0[x_{ic}y_{mc}]\right)\mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right]
\end{aligned}$$

Combining these results we obtain

$$\begin{aligned}
f_{\mathbf{x}}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{Y}_i) = \sum_{m\in\mathcal{M}(i)}\sum_{k=1}^K \Bigg\{ &-2r_{im}\mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right] + \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right] + \\
&+2\left(\sum_{c\neq k} E_0[x_{ic}y_{mc}]\right)\mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right]\Bigg\} \\
= \sum_{m\in\mathcal{M}(i)}\sum_{k=1}^K \left\{-2\left(r_{im} - \sum_{c\neq k} E_0[x_{ic}y_{mc}]\right)\mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right] + \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right]\right\}
\end{aligned}$$

If we then define $r_{im}^{(k)}$ as the expected contribution to the rating r_{im} from component k, i.e.

$$r_{im}^{(k)} = r_{im} - \sum_{c\neq k} E_0[x_{ic}y_{mc}] = r_{im} - \sum_{c\neq k}\mu_{ic}\eta_{mc}\ ,$$

we can rewrite $f_{\mathbf{x}}(\mathbf{x}_i, \mathbf{y}_i, Y_i)$ as

$$\begin{aligned}
f_{\mathbf{x}}(\mathbf{x}_i, \mathbf{y}_i, Y_i) &= \mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, \sum_{m\in\mathcal{M}(i)}(r_{im} - \mathbf{x}_i^T\mathbf{y}_m)^2\right] \\
&= \sum_{m\in\mathcal{M}(i)}\sum_{k=1}^K \left\{-2r_{im}^{(k)}\mathrm{Cov}_0\left[x_{ik}y_{jk}, x_{ik}y_{mk}\right] + \mathrm{Cov}_0\left[x_{ik}y_{jk}, (x_{ik}y_{mk})^2\right]\right\} \\
&= \sum_{m\in\mathcal{M}(i)}\sum_{k=1}^K \mathrm{Cov}_0\left[x_{ik}y_{jk}, (r_{im}^{(k)} - x_{ik}y_{mk})^2\right]\ .
\end{aligned}$$

For each of the $|\mathcal{M}(i)|$ points we will then only need the correction for K univariate problems (that are far simpler to compute). Notice in particular that this result is only possible thanks to our assumption of fully factorized priors and VB posteriors, as there are no correlations among variables of the same user/item vectors to be taken into account when computing the covariances.

Putting all together: corrections for the predicted ratings

Combining the results above, the first order correction for the required expectation of this matrix factorization problem is given by

$$\begin{aligned}
\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, V\right] &= \frac{\beta}{2}\sum_{m\in\mathcal{M}(i)}\sum_{k=1}^{K}\mathrm{Cov}_0\left[x_{ik}y_{jk}, (r_{im}^{(k)} - x_{ik}y_{mk})^2\right] + \\
&+ \frac{\beta}{2}\sum_{\substack{n\in\mathcal{N}(j)\\ n\neq i}}\sum_{k=1}^{K}\mathrm{Cov}_0\left[x_{ik}y_{jk}, (r_{nj}^{(k)} - x_{nk}y_{jk})^2\right] + \\
&+ \frac{1}{2}\sum_{k=1}^{K}\mathrm{Cov}_0\left[x_{ik}y_{ik}, \alpha_x x_{ik}^2 + \alpha_y y_{jk}^2 - \gamma_{ik}(x_{ik}-\mu_{ik})^2 - \zeta_{jk}(y_{jk}-\eta_{jk})^2\right] \\
&= \sum_{k=1}^{K}\left\{\frac{\beta}{2}\sum_{m\in\mathcal{M}(i)}\mathrm{Cov}_0\left[x_{ik}y_{jk}, (r_{im}^{(k)} - x_{ik}y_{mk})^2\right] + \right. \\
&\quad + \frac{\beta}{2}\sum_{\substack{n\in\mathcal{N}(j)\\ n\neq i}}\mathrm{Cov}_0\left[x_{ik}y_{jk}, (r_{nj}^{(k)} - x_{nk}y_{jk})^2\right] + \\
&\quad \left. + \frac{1}{2}\mathrm{Cov}_0\left[x_{ik}y_{ik}, \alpha_x x_{ik}^2 + \alpha_y y_{jk}^2 - \gamma_{ik}(x_{ik}-\mu_{ik})^2 - \zeta_{jk}(y_{jk}-\eta_{jk})^2\right]\right\}
\end{aligned}$$

Due to the fully factorized VB approximation, the computation of these covariances does not require vector operations and can be easily done manually or using symbolic mathematics software such as Sympy[3]. The required non-central Gaussian moments can be computed using Wick's theorem. Assuming that r_{ij} is an unobserved rating that we want to predict[4], the final expression for the covariance term in the first order correction of $E[\mathbf{x}_i^T\mathbf{y}_j]$ is

$$\begin{aligned}
\mathrm{Cov}_0\left[\mathbf{x}_i^T\mathbf{y}_j, V\right] = \sum_{k=1}^{K}\Bigg\{ &\sum_{m\in\mathcal{M}(i)}\left[\frac{\beta}{\gamma_{ik}}\mu_{ik}\eta_{jk}\eta_{mk}^2 - \frac{\beta r_{im}^{(k)}}{\gamma_{ik}}\eta_{jk}\eta_{mk} + \frac{\beta}{\gamma_{ik}\zeta_{mk}}\mu_{ik}\eta_{jk}\right] + \\
&+ \sum_{\substack{n\in\mathcal{N}(j)\\ n\neq i}}\left[\frac{\beta}{\zeta_{jk}}\mu_{ik}\eta_{jk}\mu_{nk}^2 - \frac{\beta r_{nj}^{(k)}}{\zeta_{jk}}\mu_{ik}\mu_{nk} + \frac{\beta}{\gamma_{nk}\zeta_{jk}}\mu_{ik}\eta_{jk}\right] + \\
&+ \left[\left(\frac{\alpha_x}{\gamma_{ik}} + \frac{\alpha_y}{\zeta_{jk}}\right)\mu_{ik}\eta_{jk}\right]\Bigg\}
\end{aligned}$$

To illustrate the effect of the correction we can compute its value in the case where symmetries in the user vectors $\mathbf{u}_i$ are not broken, i.e. $\boldsymbol{\mu}_i = \mathbf{0}$ (this may happen for "noisy" users with only few rated items). In this case, the expected rating with a first order correction is (with $\lambda = 1$)

$$E[\mathbf{x}_i^T\mathbf{y}_j] \approx \sum_{k=1}^{K}\left\{\sum_{m\in\mathcal{M}(i)}\left[\frac{\beta r_{im}^{(k)}}{\gamma_{ik}}\eta_{jk}\eta_{mk}\right]\right\}.$$

We see that the expectation depends on the rating given by user i to other items in the catalogue weighted by how similar their vectors are to $\mathbf{y}_j$ (so that we have a positive contribution from $r_{im}^{(k)}$ to

[3] http://www.sympy.org

[4] This means that the likelihood term in $V(\mathbf{X}, \mathbf{Y})$ does not depend on $\mathbf{x}_i^T\mathbf{y}_j$. In the toy example presented in Section 3 instead we were computing a correction on an observed rating.

the predicted rating only if the k-th components of the means of the j-th and m-th item vectors have the same direction).