Robust Bayesian Matrix Factorisation

Balaji Lakshminarayanan
Yandex Labs*
balaji@yandex-team.ru

Guillaume Bouchard
Xerox Research Centre Europe
guillaume.bouchard@xrce.xerox.com

Cedric Archambeau
Xerox Research Centre Europe
cedric.archambeau@xrce.xerox.com

Abstract

We analyse the noise arising in collaborative filtering when formalised as a probabilistic matrix factorisation problem. We show empirically that modelling row- and column-specific variances is important, the noise being in general non-Gaussian and heteroscedastic. We also advocate for the use of a Student-t prior for the latent features as the standard Gaussian is included as a special case. We derive several variational inference algorithms and estimate the hyperparameters by type-II maximum likelihood. Experiments on real data show that the predictive performance is significantly improved.

1 INTRODUCTION

Techniques to factorise large, partially observed matrices are recognised as reference methods for large-scale collaborative filtering tasks [HP99, BL07] and data imputation [KMBM10]. The simplest method to perform this task is the singular value decomposition (SVD), which finds a low-rank bilinear approximation of the observation matrix. Entries of the matrix are modelled by an inner product between two low-dimensional feature vectors, which are usually estimated by minimising a squared loss. In the fully observed case, this problem can be solved by classical SVD. It is known to have a unique solution, which can be obtained by a series of rank-1 approximations applied on the residuals, i.e. the difference between the observations and the model predictions. In collaborative filtering, however, most of the entries are missing. The corresponding optimisation problem is a weighted version of SVD, which is substantially more difficult to solve. Simple EM-like algorithms estimating the distribution of missing entries are typically slow to converge when the data are sparse [SJ03]. Most existing methods use block-coordinate gradient descent algorithms based on alternating regressions, sometimes called criss-cross regression [GZ79]. Several authors considered robust alternating regression to add robustness against outliers and to stabilise the algorithm (i.e. to avoid local minima) [CFPR03, MY08]. They showed that robust estimators for the weighted SVD problem provide significant improvements in terms of parameter estimation and predictive performances. While these approaches can be applied to collaborative filtering, they are adhoc in the sense that they were designed to alleviate the inherent difficulties of the estimation problem, but are not justified in terms of probabilistic modelling.

Recently, several probabilistic interpretations using a Gaussian noise model, commonly denoted by the term probabilistic matrix factorisation (PMF), were shown to lead to significant improvements over the standard SVD [LT07, RIK07, SM08b], essentially due to the implicit smoothing included in the model [NS10]. While assuming Gaussian noise makes sense when the observations are continuous, it is less justified when they are on an ordinal scale. Still, it was reported that PMF performed well on this kind of data. A potential explanation is that a continuous noise model accounts for the user's occasional mood swings, leading to feedback of varying quality. However, there is no reason to think that other continuous models would not perform as well or better than a Gaussian model.

The full Bayesian treatment of PMF, known as Bayesian matrix factorisation (BMF) [SM08a], further extends the probabilistic model by imposing Gaussian-Wishart priors over the low-rank decomposition. Again, there is no reason to think a priori that these priors should be Gaussian-Wishart in practice. A study of the posteriors obtained by BMF on standard movie recommendation benchmarks show that the tails are significantly stronger than the tails of the

*Most of the work was done during an internship at Xerox Research Centre Europe

Appearing in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

Gaussian distribution (see e.g. the histograms in Fig. 3 in [SM08a]).

Robustness is key to the success of practical supervised and unsupervised learning techniques (see e.g. [MHN07] for an application of M-estimators to collaborative filtering). A reduced sensitivity to model misspecifications is necessary to handle outliers and atypical observations, which are very common in real data. A natural way to obtain robust probabilistic models is to replace the Gaussian distributions by heavy-tailed distributions such as the Student-t. This approach was adopted in robust principal component analysis [ADV06], robust linear models for regression [TL05] and robust clustering [AV07]. In this work robustness is incorporated in two different ways. First, we replace the Gaussian noise model used in BMF by a Gaussian scale mixture-based noise model. This model is closely related to a Student-t one, which is obtained by integrating out a Gamma distributed scale variable in a Gaussian scale mixture (see Appendix A). Instead, we consider *two* rescaling factors per observation (rating, one due to the row feature (user) and one due to the column feature (item). We show empirically that modelling row- and column-specific variances in bilinear forms significantly improves the predictive performance. Second, we consider Student-t priors on the latent features. We advocate that in practice the distribution of these features is non-Gaussian and show that considering heavy-tailed priors further improves the quality of the predictions. The paper is organised as follows. Section 2 motivates and introduces a heteroscedastic noise model for BMF. Section 3 proposes non-Gaussian prior distributions for the features matched to the noise. Sections 4 and 5 detail the inference and learning algorithms. They are followed by implementation details, experiments and conclusion in Sections 6, 7 and 8.

2 NOISE MODELS FOR BMF

Let N be the number of users and M the number of rated items. The aim is to compute a factorised form of the rating matrix $\mathbf{R} \in \mathbb{R}^{N\times M}$. This low-rank approximation will have the effect of pooling users and items together, which in turn will help us predict unobserved entries. We further let $\mathbf{W} \in \mathbb{R}^{N\times M}$ be the indicator matrix of the observed elements of $\mathbf{R}$, that is w_{nm} equals 1 if r_{nm} has been observed and 0 otherwise. The collaborative filtering problem can then be written as a weighted SVD problem [SJ03]:

$$\min_{\mathbf{\Phi},\mathbf{\Omega}} \|\mathbf{W} * (\mathbf{\Phi}\mathbf{\Omega}^\top - \mathbf{R})\|_F^2 \qquad (1)$$

where $\|.\|_F$ is the Frobenius norm and $*$ the Hadamard product. The tall matrices $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N)^\top \in \mathbb{R}^{N\times K}$ and $\mathbf{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_M)^\top \in \mathbb{R}^{M\times K}$ denote respectively the user-specific and item-specific feature matrices. In our notation, the rows of $\mathbf{\Phi}$ correspond to the latent row features and the rows of $\mathbf{\Omega}$ to the latent column features. Solving the weighted SVD problem is more involved than standard SVD as the typically used sequence of rank-1 approximations of the residuals is not guaranteed to converge to a solution of (1).

2.1 Gaussian Noise

A probabilistic interpretation of (1) is obtained by assuming that given the latent features $\boldsymbol{\phi}_n$ and $\boldsymbol{\omega}_m$, the observation r_{nm} is isotropic Gaussian with mean $\boldsymbol{\phi}_n^\top \boldsymbol{\omega}_m$ and constant precision $\tau > 0$. The log-likelihood of the observed ratings is then given by:

$$\ln p(\mathbf{R}|\mathbf{\Phi}, \mathbf{\Omega}, \tau) = \sum_\ell \ln \mathcal{N}(r_\ell | \boldsymbol{\phi}_{n_\ell}^\top \boldsymbol{\omega}_{m_\ell}, \tau), \qquad (2)$$

where ℓ now indexes the observed ratings.[1] A solution to (1) is obtained by maximising (2) with respect to the features. To prevent overfitting one can impose isotropic Gaussian priors on them, leading to maximum a posteriori estimators. This approach was adopted in PMF [SM08b]. The many Bayesian extensions [LT07, RIK07, SM08a] consider Gaussian priors of various flavours.

All those models make strong assumptions about the noise: not only it is considered to be normally distributed, but it is also assumed to be homoscedastic, i.e. having a constant noise level across rows and columns of $\mathbf{R}$. This can be counter-intuitive in many situations. For example, in movie rating applications, some users might have the tendency to give nearly constant ratings, while others might have more diverse opinions, using the full scale of ratings. A similar observation can be made about the movies (or items in general): conventional movies might have a significantly smaller variance in their ratings than avant-garde films which tend to polarize the users' appreciation. To account for this variability, simple normalisation can be used (e.g. standardising user ratings by subtracting the mean and dividing by the standard deviation), but this pre-processing is subject to estimation error (e.g. when there are few observations per items). More importantly, normalisation is applicable, either to the rows, or the columns of $\mathbf{R}$, but not to the rows *and* the columns simultaneously, which is precisely the situation that concerns collaborative filtering problems.

As an illustration, we computed the PMF solution on one third of the ratings of the One Million MovieLens

[1] n_ℓ (m_ℓ) refers to the user index (item index) of the ℓth observed rating. The notation ℓ_n (ℓ_m) refers to all observed ratings for user n (item m).

data set, used another third to compute the per-user and per-item unbiased variance estimates, and the last third was used as the test set to analyse the residuals, which are defined as $r_{nm} - \boldsymbol{\phi}_n^\top \boldsymbol{\omega}_m$. For each user, we computed the unbiased empirical variance of the residuals of their predicted test ratings and plotted the histogram in Fig. 1 (top). We then computed the distribution of these variances by sampling a large number of ratings under the PMF model assumption, i.e. Gaussian homoscedastic noise model, for the same features and noise as before. The resulting *predicted* distribution corresponds to the red curve in Fig. 1 (top). We see that there is a significant mismatch between the model predictions and the actual ratings assuming a constant variance. In other words, if the homoscedastic Gaussian distribution assumption was true, the range of the observed variances would have been much smaller. We also computed the predicted variance distribution under the per-user heteroscedastic noise model (green thick curve), where the per-user variances were estimated with one third of the data. It can be observed that with this more flexible model, the per-user heteroscedastic model predicts the distribution of variances much more accurately. Fig. 1 (bottom) corresponds to the analysis of the residuals per movie (item). Again, there is a model mismatch in terms of spread the homoscedastic model. By contrast, the per-movie heteroscedastic model leads to a more accurate fit of the noise variance distribution.

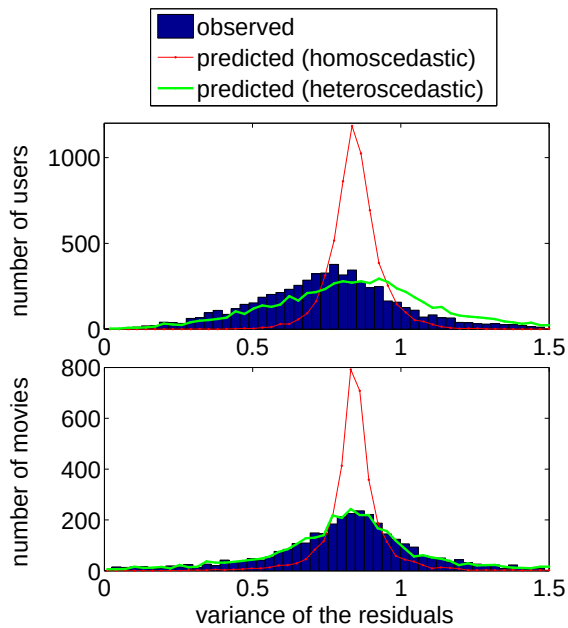

Figure 1: Histogram of empirical variances of the residuals per user (top) or per movie (bottom) under PMF. Line plot of the predicted distributions of the residual variances under PMF with a homoscedastic and a heteroscedastic noise model. Only users or movies with at least 100 ratings in the 1 Million MovieLens data set were considered.

2.2 Gaussian Scale Mixture Noise

A natural noise structure to account for the heterogeneity in the user- and item-specific variances is the following heteroscedastic Gaussian scale mixture noise model:

$$\ln p(\mathbf{R}|\boldsymbol{\Phi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) = \sum_{\ell} \ln \mathcal{N}(r_\ell | \boldsymbol{\phi}_{n_\ell}^\top \boldsymbol{\omega}_{m_\ell}, \tau \alpha_{n_\ell} \beta_{m_\ell}).$$

The key difference with (2) is that the precision of rating r_ℓ is now reweighted by the user-specific scale α_{n_ℓ} and the item-specific scale β_{m_ℓ}. The prior on the scale parameters are given by

$$p(\alpha_n) = \mathcal{G}a(\alpha_n | \tfrac{a_0}{2}, \tfrac{b_0}{2}), \tag{3}$$

$$p(\beta_m) = \mathcal{G}a(\beta_m | \tfrac{c_0}{2}, \tfrac{d_0}{2}), \tag{4}$$

for all n and m. The parameters a_0 and b_0 are shared by all users, while the parameters c_0 and d_0 by all items.

3 PRIORS FOR BMF

The latent features in PMF [SM08b] and BMF [LT07, RIK07, SM08a] are assumed to be independent Gaussians. However, there is no reason in practice to believe the Gaussian assumption is always justified.

A popular measure of "non-Gaussianity" is the kurtosis, which quantifies the peakedness of a distribution. A high kurtosis is related to the distribution tails as more of the variance is due to infrequent, possibly relatively extreme deviations. Student-t and other Gaussian scale mixtures have a kurtosis typically higher than that of the Gaussian. Recently, it was shown in [ADV06] that probabilistic principal component analysis (PCA) with Student-t features leads to improved models of large dimensional data. In the context of collaborative filtering, where PMF is known to be the equivalent of probabilistic PCA with missing information, a similar improvement is to be expected when imposing Gaussian scale mixtures priors on the features:

$$p(\boldsymbol{\Phi}) = \prod\nolimits_n \int \mathcal{N}(\boldsymbol{\phi}_n | \mathbf{0}, \alpha_n \boldsymbol{\Lambda}_\phi)\, p(\alpha_n)\, d\alpha_n, \tag{5}$$

$$p(\boldsymbol{\Omega}) = \prod\nolimits_m \int \mathcal{N}(\boldsymbol{\omega}_m | \mathbf{0}, \beta_m \boldsymbol{\Lambda}_\omega)\, p(\beta_n)\, d\beta_n, \tag{6}$$

where $p(\alpha_n)$ and $p(\beta_n)$ are given by (3) and (4).

Following [LT07], we restrict $\boldsymbol{\Lambda}_\phi$ and $\boldsymbol{\Lambda}_\omega$ to be diagonal precision matrices, say $\boldsymbol{\Lambda}_\phi^{-1} = \operatorname{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\}$,

$\boldsymbol{\Lambda}_\omega^{-1} = \operatorname{diag}\{\rho_1^2, \rho_2^2, \ldots, \rho_K^2\}$. As shown in Appendix A, the integrals in (5) and (6) are analytically tractable leading to multivariate Student-t distributions.

4 VARIATIONAL INFERENCE

We are interested in a deterministic approximation of the full Bayesian matrix factorisation problem and follow therefore a variational EM approach [NH98, Att00, Bea03]. In this approach, the log-marginal likelihood is lower bounded by the negative variational free energy:

$$\ln p(\mathbf{R}) = -\mathcal{F}(q, \boldsymbol{\theta}) + \mathcal{KL}\left(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{R}, \boldsymbol{\theta})\right) \geq -\mathcal{F}(q, \boldsymbol{\theta}),$$

where $\mathbf{z} = \{\boldsymbol{\Phi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are the latent variables, $\boldsymbol{\theta} = \{\tau, \boldsymbol{\Lambda}_\phi, \boldsymbol{\Lambda}_\omega, a_0, b_0, c_0, d_0\}$ are the parameters and the variational free energy $\mathcal{F}$ is defined as follows:

$$\mathcal{F}(q, \boldsymbol{\theta}) = -\langle \ln p(\mathbf{R}, \mathbf{z}|\boldsymbol{\theta})\rangle_q - \mathrm{H}[q], \tag{7}$$

In order to find a tractable solution, it is in general assumed that the variational approximate posterior q fully factorises, that is $q(\mathbf{z}) = q(\boldsymbol{\Phi})q(\boldsymbol{\Omega})q(\boldsymbol{\alpha})q(\boldsymbol{\beta})$. This approach is known as mean field (see e.g. [Bis06]). In this work we also consider a structured variational approximation [Wie00] of the form $q(\mathbf{z}) = q(\boldsymbol{\Phi}, \boldsymbol{\alpha})q(\boldsymbol{\Omega}, \boldsymbol{\beta})$. Structured variational approximations have been shown to be beneficial in practice as the correlations between latent variables are not unnecessarily ignored and they reduce the bound gap (see e.g. [AV07]).

4.1 Gaussian Scale Mixture Noise with Gaussian Priors on the Features

The mean field approximation leads to the following Gaussian variational posteriors for the latent features:

$$q(\boldsymbol{\Phi}) = \prod\nolimits_n \mathcal{N}(\boldsymbol{\phi}_n|\bar{\boldsymbol{\phi}}_n, \bar{\mathbf{S}}_n),$$
$$q(\boldsymbol{\Omega}) = \prod\nolimits_m \mathcal{N}(\boldsymbol{\omega}_m|\bar{\boldsymbol{\omega}}_m, \bar{\mathbf{R}}_m),$$

where

$$\bar{\boldsymbol{\phi}}_n = \tau\bar{\alpha}_n\bar{\mathbf{S}}_n^{-1}\sum\nolimits_{\ell_n}\bar{\beta}_{m_\ell}\bar{\boldsymbol{\omega}}_{m_\ell}r_\ell,$$
$$\bar{\mathbf{S}}_n = \tau\bar{\alpha}_n\sum\nolimits_{\ell_n}\langle\beta_{m_\ell}\boldsymbol{\omega}_{m_\ell}\boldsymbol{\omega}_{m_\ell}^\top\rangle + \boldsymbol{\Lambda}_\phi,$$
$$\bar{\boldsymbol{\omega}}_m = \tau\bar{\beta}_m\bar{\mathbf{R}}_m^{-1}\sum\nolimits_{\ell_m}\bar{\alpha}_{n_\ell}\bar{\boldsymbol{\phi}}_{n_\ell}r_\ell,$$
$$\bar{\mathbf{R}}_m = \tau\bar{\beta}_m\sum\nolimits_{\ell_m}\langle\alpha_{n_\ell}\boldsymbol{\phi}_{n_\ell}\boldsymbol{\phi}_{n_\ell}^\top\rangle + \boldsymbol{\Lambda}_\omega.$$

The variational posteriors for the scale parameters are $q(\boldsymbol{\alpha}) = \prod_n \mathcal{G}a(\alpha_n|\frac{a_n}{2}, \frac{b_n}{2})$ and $q(\boldsymbol{\beta}) = \prod_m \mathcal{G}a(\beta_m|\frac{c_m}{2}, \frac{d_m}{2})$.

Their parameters are given by

$$a_n = a_0 + 1,$$
$$b_n = b_0 + \tau\sum\nolimits_{l_n}\langle\beta_{m_\ell}(r_\ell - \boldsymbol{\phi}_n^\top\boldsymbol{\omega}_{m_\ell})^2\rangle,$$
$$c_m = c_0 + 1,$$
$$d_m = d_0 + \tau\sum\nolimits_{\ell_m}\langle\alpha_{n_\ell}(r_\ell - \boldsymbol{\phi}_{n_\ell}^\top\boldsymbol{\omega}_m)^2\rangle.$$

4.2 Gaussian Noise with Student Priors on the Features

We consider a structured variational approximation by restricting the variational posteriors for $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$ to be of the same form as the priors, that is Gaussians with scaled covariances:

$$q(\boldsymbol{\Phi}|\boldsymbol{\alpha}) = \prod\nolimits_n \mathcal{N}(\boldsymbol{\phi}_n|\bar{\boldsymbol{\phi}}_n, \alpha_n\bar{\mathbf{S}}_n), \tag{8}$$
$$q(\boldsymbol{\Omega}|\boldsymbol{\beta}) = \prod\nolimits_m \mathcal{N}(\boldsymbol{\omega}_m|\bar{\boldsymbol{\omega}}_m, \beta_m\bar{\mathbf{R}}_m), \tag{9}$$

Direct maximisation of the bound $-\mathcal{F}$ wrt $\bar{\boldsymbol{\phi}}_n$ and $\bar{\mathbf{S}}_n$, as well as $\bar{\boldsymbol{\omega}}_m$ and $\bar{\mathbf{R}}_m$ leads to

$$\bar{\boldsymbol{\phi}}_n = \tau\widetilde{\mathbf{S}}_n^{-1}\sum\nolimits_{\ell_n}\bar{\boldsymbol{\omega}}_{m_\ell}r_\ell,$$
$$\bar{\mathbf{S}}_n = \tau\langle\alpha_n^{-1}\rangle\sum\nolimits_{\ell_n}\langle\boldsymbol{\omega}_{m_\ell}\boldsymbol{\omega}_{m_\ell}^\top\rangle + \boldsymbol{\Lambda}_\phi,$$
$$\bar{\boldsymbol{\omega}}_m = \tau\widetilde{\mathbf{R}}_m^{-1}\sum\nolimits_{\ell_m}\bar{\boldsymbol{\phi}}_{n_\ell}r_\ell,$$
$$\bar{\mathbf{R}}_m = \tau\langle\beta_m^{-1}\rangle\sum\nolimits_{\ell_m}\langle\boldsymbol{\phi}_{n_\ell}\boldsymbol{\phi}_{n_\ell}^\top\rangle + \boldsymbol{\Lambda}_\omega,$$

where $\widetilde{\mathbf{S}}_n = \tau\sum_{\ell_n}\langle\boldsymbol{\omega}_{m_\ell}\boldsymbol{\omega}_{m_\ell}^\top\rangle + \bar{\alpha}_n\boldsymbol{\Lambda}_\phi$ and $\widetilde{\mathbf{R}}_m = \tau\sum_{\ell_m}\langle\boldsymbol{\phi}_{n_\ell}\boldsymbol{\phi}_{n_\ell}^\top\rangle + \bar{\beta}_m\boldsymbol{\Lambda}_\omega$.

The posterior for the scale parameters are obtained by using the result discussed in Appendix B, which holds for the structured variational approximation that we consider here, namely $q(\boldsymbol{\Phi}, \boldsymbol{\alpha}) = \prod_n q(\boldsymbol{\phi}_n, \alpha_n)$ and $q(\boldsymbol{\Omega}, \boldsymbol{\beta}) = \prod_m q(\boldsymbol{\omega}_m, \beta_m)$.

While the variational posteriors of the scale variables are not Gamma distributions in this case, they can be recognised as products of generalised inverse Gaussians: $q(\boldsymbol{\alpha}) = \prod_n \mathcal{N}^{-1}(\alpha_n|\frac{\nu_n}{2}, \chi_n, \phi_n)$ and $q(\boldsymbol{\beta}) = \prod_m \mathcal{N}^{-1}(\beta_m|\frac{\nu_m}{2}, \chi_m, \phi_m)$, where the parameters are defined as follows:

$$\nu_n = a_0, \qquad \nu_m = c_0,$$
$$\chi_n = \tau\sum\nolimits_{\ell_n}\langle\boldsymbol{\omega}_{m_\ell}^\top\bar{\mathbf{S}}_n^{-1}\boldsymbol{\omega}_{m_\ell}\rangle, \qquad \chi_m = \tau\sum\nolimits_{\ell_m}\langle\boldsymbol{\phi}_{n_\ell}^\top\bar{\mathbf{R}}_m^{-1}\boldsymbol{\phi}_{n_\ell}\rangle,$$
$$\varphi_n = b_0 + \bar{\boldsymbol{\phi}}_n^\top\boldsymbol{\Lambda}_\phi\bar{\boldsymbol{\phi}}_n, \qquad \varphi_m = d_0 + \bar{\boldsymbol{\omega}}_m^\top\boldsymbol{\Lambda}_\omega\bar{\boldsymbol{\omega}}_m.$$

The generalised inverse Gaussian distribution is defined in Appendix C.

4.3 Gaussian Scale Mixture Noise with Student Priors on the Features

We consider a structured variational approximation for the case where both the noise and the priors are Gaus-

sian scale mixtures. It is easy to show that the form of the variational posterior that arises is given by (8) and (9) with parameters now defined by

$$\begin{aligned}\bar{\boldsymbol{\phi}}_n &= \tau \bar{\mathbf{S}}_n^{-1} \sum\nolimits_{\ell_n} \bar{\beta}_{m\ell} \bar{\boldsymbol{\omega}}_{m\ell} r_\ell,\\ \bar{\mathbf{S}}_n &= \tau \sum\nolimits_{\ell_n} \langle \beta_{m\ell} \boldsymbol{\omega}_{m\ell} \boldsymbol{\omega}_{m\ell}^\top \rangle + \boldsymbol{\Lambda}_\phi,\\ \bar{\boldsymbol{\omega}}_m &= \tau \bar{\mathbf{R}}_m^{-1} \sum\nolimits_{\ell_m} \bar{\alpha}_{n\ell} \bar{\boldsymbol{\phi}}_{n\ell} r_\ell,\\ \bar{\mathbf{R}}_m &= \tau \sum\nolimits_{\ell_m} \langle \alpha_{n\ell} \boldsymbol{\phi}_{n\ell} \boldsymbol{\phi}_{n\ell}^\top \rangle + \boldsymbol{\Lambda}_\omega.\end{aligned}$$

The variational posteriors for the scale parameters are Gamma distributions with parameters given by

$$\begin{aligned}a_n &= a_0 + 1,\\ b_n &= b_0 + \tau \sum\nolimits_{\ell_n} \langle \beta_{m\ell} (r_\ell - \bar{\boldsymbol{\phi}}_n^\top \boldsymbol{\omega}_{m\ell})^2 \rangle + \bar{\boldsymbol{\phi}}_n^\top \boldsymbol{\Lambda}_\phi \bar{\boldsymbol{\phi}}_n,\\ c_m &= c_0 + 1,\\ d_m &= d_0 + \tau \sum\nolimits_{\ell_m} \langle \alpha_{n\ell} (r_\ell - \boldsymbol{\phi}_{n\ell}^\top \bar{\boldsymbol{\omega}}_m)^2 \rangle + \bar{\boldsymbol{\omega}}_m^\top \boldsymbol{\Lambda}_\omega \bar{\boldsymbol{\omega}}_m.\end{aligned}$$

5 TYPE II ML ESTIMATION

The parameters are estimated by type II maximum likelihood (or empirical Bayes). The updates are obtained by direct maximisation of $-\mathcal{F}$.

When the noise is a Gaussian scale mixture and the priors on the latent features are Student-t distributions, the updates for the parameters are given by

$$\begin{aligned}\tau^{-1} &\leftarrow \frac{\sum_\ell \langle \alpha_{n\ell} \beta_{m\ell} (r_\ell - \boldsymbol{\phi}_{n\ell}^\top \boldsymbol{\omega}_{m\ell})^2 \rangle}{L},\\ \sigma_k^2 &\leftarrow \frac{\sum_n \langle \alpha_n \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\top \rangle_{kk}}{N}, \qquad b_0 \leftarrow \frac{a_0 N}{\sum_n \bar{\alpha}_n},\\ \rho_k^2 &\leftarrow \frac{\sum_m \langle \beta_m \boldsymbol{\omega}_m \boldsymbol{\omega}_m^\top \rangle_{kk}}{M}, \qquad d_0 \leftarrow \frac{c_0 M}{\sum_m \bar{\beta}_m}.\end{aligned}$$

Parameter a_0 is found by solving the nonlinear equation $\sum_n \left\{ \ln \frac{b_0}{2} + \langle \ln \alpha_n \rangle - \psi(\frac{a_0}{2}) \right\} = 0$ by line search. Parameter c_0 is updated in the same manner.

When the noise is Gaussian, the update for the precision is replaced by

$$\tau^{-1} \leftarrow \frac{\sum_\ell \langle (r_\ell - \boldsymbol{\phi}_{n\ell}^\top \boldsymbol{\omega}_{m\ell})^2 \rangle}{L}.$$

When the priors on the latent features are Gaussians, the updates for the diagonal elements of $\boldsymbol{\Lambda}_\phi$ and $\boldsymbol{\Lambda}_\omega$ are replaced by

$$\sigma_k^2 \leftarrow \frac{\sum_n \langle \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\top \rangle_{kk}}{N}, \qquad \rho_k^2 \leftarrow \frac{\sum_m \langle \boldsymbol{\omega}_m \boldsymbol{\omega}_m^\top \rangle_{kk}}{M}.$$

6 IMPLEMENTATION DETAILS

In practice, it is important to model the user (or item) specific offsets. These can easily be incorporated into the model by appending a '1' to the latent item feature $\boldsymbol{\omega}_m$ (or the latent user feature $\boldsymbol{\phi}_n$). Also, the model is only identifiable up to a rescaling of $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$. Thus, we can fix $\boldsymbol{\Lambda}_\omega$ to the identity matrix $\mathbf{I}_K$ and only optimise with respect to $\boldsymbol{\Lambda}_\phi$.

In our experiments, we used full covariance matrices for $\bar{\mathbf{S}}_n$ and $\bar{\mathbf{R}}_m$. However, to scale up the model to a large K, one might restrict $\bar{\mathbf{S}}_n$ and $\bar{\mathbf{R}}_m$ to be diagonal matrices. In this case, it is only required to store K variance parameters per user and per item, and the computational cost of inverting these matrices is reduced from $\mathcal{O}(K^3)$ to $\mathcal{O}(K)$.

7 EXPERIMENTS

In this section, we first describe the data sets and the performance measures we used to validate the models. Next, we discuss the results.

7.1 Data Sets

MovieLens[2]: The data set contains movie ratings by MovieLens users. The ratings are ordinal values on the scale 1 to 5. Users have rated at least 20 movies. We considered the *MovieLens 100K* (approximately $100,000$ ratings by 943 users on 1682 movies), the *MovieLens 1 Million* ($1,000,209$ ratings by $6,040$ users on approximately $3,900$ movies), and the *MovieLens 10 Million* (10000054 ratings by 71567 users on 10681 movies) rating data sets.

Jester Joke [GRGP01]: The *Jester-1* data set contains 4.1 million ratings of 100 jokes from $73,421$ users. We will use *Jester-1-3*, which is a subset of *Jester-1*, containing ratings of $24,938$ users who have rated between 15 and 35 jokes. These ratings are on a continuous scale from -10.0 to 10.0.

7.2 Experimental Setup

We randomly choose 70% of the ratings for training and use the remaining ratings as test data. Every movie and every user appears at least once in the training set. In the MovieLens data sets, we include only movies that have been rated at least thrice.

We are interested in evaluating the effect of the noise model and of the priors imposed on the latent factors. To this effect, we compare the following models on the

[2] www.grouplens.org/node/73

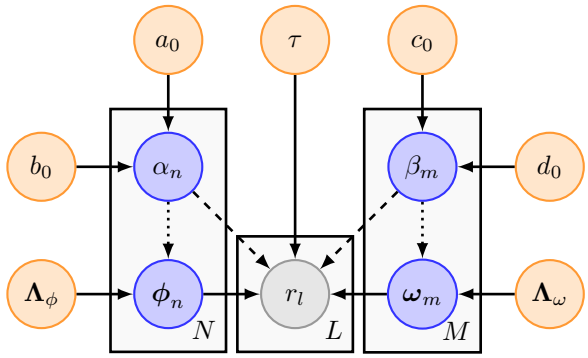

Figure 2: Graphical models for the different matrix factorisations. GG is obtained when only considering the solid arrows. GR is obtained when considering the solid and the dotted arrows. RG is obtained when considering the solid and the dashed arrows. Finally, RR is obtained when considering all the arrows.

rating prediction task: (see Fig. 2 for the graphical representation of the models):

- GG: Homoscedastic Gaussian noise, Gaussian priors [LT07]
- RG: Heteroscedastic Gaussian noise, Gaussian priors (Section 4.1)
- GR: Homoscedastic Gaussian noise, Student-t priors (Section 4.2)
- RR: Heteroscedastic Gaussian noise, Student-t priors (Section 4.3)

where 'G' indicates *Gaussian* and 'R' indicates a *robust* heavy-tailed distribution. Additionally, we present the results for the mean field approximation of the GR model described in Section 4.2, referred to as GR-mf in the following. Comparing the performance of GR and GR-mf will help us understand the gains that can be obtained through the use of a structured variational inference algorithm in place of a standard mean field algorithm.

Prediction is done by using point estimates of both $\mathbf{\Phi}$ and $\mathbf{\Omega}$. We compare the models in terms of predictive performance based on the following metrics:

- *Root mean squared error* (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum\nolimits_t (r_t - \bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t})^2},$$

  where t indexes the test ratings. This metric is closely related to the Gaussian noise model.

- *Mean absolute error* (MAE):

$$\text{MAE} = \frac{1}{T}\sum\nolimits_t |r_t - \bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t}|.$$

  This metric is related to the heteroscedastic Gaussian noise model.

- *Ordinal log-likelihood* (OLL):

$$\text{OLL} = \sum\nolimits_t \log P(r_t),$$

  where $P(r_t)$ is given by (10) in the case of the homoscedastic Gaussian noise and by (11) in the case of the heteroscedastic Gaussian noise.

The OLL metric is only used in the experiments with the MovieLens data. Since the true ratings are ordinal, the predicted ratings need to be rounded off to predict a value in the support of the original ratings. It is possible to compute the likelihood of the predicted value lying within the boundaries corresponding to the true rating. We compute this log-likelihood in the same fashion as described in [CG05].

Let us denote the boundaries defining the ordinal scale by $\{b_0, b_1, \cdots, b_5\}$. In our experiments we do not attempt to optimise these values, but used $b_0 = -\infty, b_i = i + 0.5, b_5 = +\infty$. The ordinal likelihood for the Gaussian noise model is given by

$$P(r_t = i) \approx \int_{b_{i-1}-\bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t}}^{b_i-\bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t}} \mathcal{N}(\epsilon_t|0, \langle\tau\rangle)\, d\epsilon_t, \qquad (10)$$

where ϵ_t is the prediction error (or residual) and $i \in \{1, \ldots, 5\}$. For the Gaussian scale mixture noise model, we use again point estimates, leading to the following approximation:

$$P(r_t = i) \approx \int_{b_{i-1}-\bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t}}^{b_i-\bar{\boldsymbol{\phi}}_{n_t}^\top \bar{\boldsymbol{\omega}}_{m_t}} \mathcal{N}(\epsilon_t|0, \bar{\alpha}_{n_t}\bar{\beta}_{m_t}\langle\tau\rangle)\, d\epsilon_t. \qquad (11)$$

7.3 Results

The RMSE, MAE and OLL values obtained for MovieLens 100k data are shown in Table 1. The dimension of the low-rank approximation is set to $K = 30$. Considering GG as the reference method, we see that the performance improves on all metrics when a heteroscedastic noise model is considered (RG and RR). Interestingly, assuming Student-t priors does not improve the results: RG and GG perform similarly as respectively RR and GR. This suggests that the use of Student-t prior is not supported by the data. Checking the shape parameters a_0 and c_0 confirms this finding as they are relatively large (> 40), meaning that the

Table 1: Test results on the MovieLens 100k data using $K = 30$.

Model	RMSE	MAE	OLL
RR	0.900	0.705	-37638
RG	0.901	0.708	-37054
GR	0.906	0.710	-38193
GR-mf	0.907	0.710	-38312
GG	0.906	0.710	-38234

Table 2: Test results on the MovieLens Million rating data using $K = 20$.

Model	RMSE	MAE	OLL
RR	0.845	0.661	-354585
RG	0.846	0.663	-351667
GR	0.853	0.666	-365136
GR-mf	0.853	0.666	-365232
GG	0.851	0.665	-364319

priors are close to Gaussians. It should be noted that mean field performs similarly to the structured variational approximation for these data sets, except in terms of OLL.

To analyse the significance of the performance improvement of RR and RG over GG, we evaluated the methods on 10 different train/test splits and recorded the values of RMSE, MAE and OLL. Next, we performed a one-tailed paired-t test checking whether the models RR and RG significantly performed better than GG. We found that the improvement was significant for the three metrics at level < 0.001.

The results for the MovieLens 1 Million data exhibit a similar trend (see Table 2). We obtain better performances in terms of RMSE and similar ones in terms of MAE as the ones reported in [DV08], who consider generalised bilinear forms. The number of latent factors was not optimised for performance, but all the models are nearly optimum for $K = 20$. The best results for GG were RMSE $= 0.849$, MAE $= 0.664$ and OLL $= -363690$ (for $K = 30$). While better than the results shown in Table 2, these are still worse than the ones obtained for RG and RR.

We further evaluate the heteroscedastic noise model at larger scale by running our algorithms on the 10 Million rating data set with $K = 15$. The RMSE, MAE and OLL are respectively equal to 0.786, 0.610 and -3310589 for RR, and 0.789, 0.612 and -3433667 for GG. Hence, RR outperforms GG again on all the metrics.

Table 3: Test results on Jester-1-3 for $K = 15$.

Model	RMSE	MAE
RR	4.454	3.439
RG	4.456	3.468
GR	4.463	3.451
GR-mf	4.482	3.460
GG	4.406	3.480

Next, we tested the algorithm on the Jester-1-3 data sets; the results are shown in Table 3. We note that GG performs best in terms of RMSE, but worst in terms of MAE. The RMSE corresponds to the loss that is minimised in Gaussian PMF; this means that using a Gaussian model is asymptotically optimal. Since only users with at least 15 ratings were selected in this data set, the variance reduction due to the use of a robust estimator is not large enough to improve the test RMSE. Moreover, RMSE is the most commonly used metric in collaborative filtering, but it is not a robust measure for assessing the model predictive performance,especially when the distribution of the data is peaked and even if they take values on a finite domain. Finally, note that GR-mf performs worse than GR for all metrics.

8 CONCLUSION

Adequate noise models are often the key to good predictive performance. Perhaps, a good example are autoregressive conditional heteroscedastic (ARCH) models proposed in the late eighties and which are now commonly used to model financial time series with varying volatility. Recently, several works, ranging from regression [TL05] to clustering [AV07], indicated that constructing probabilistic models based on heavy-tailed distributions is very useful in practice. Real data is in general not only partially observed, but it is often also corrupted by atypical errors due to faulty measurement tools, human mistakes or even malicious intent. Heavy-tailed distributions are natural candidates to handle these outliers.

In this work, however, we show that heavy-tailed distributions, which exhibit a high kurtosis, are also useful to incorporate robustness in probabilistic models of data with a bounded range, whether the scale is continuous or ordinal. In the already well studied task of collaborative filtering, we showed that performance improves when considering noise models that take the high variability of ratings into account. It is expected that these results will carry over to most data imputation tasks with a mix of continuous and/or ordinal data.

Acknowledgements

We would like thank Nicolas Delannay for initial discussions about extensions of PMF and BMF, and the anonymous reviewers for helpful feedback.

References

[ADV06] C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In *23rd International Conference on Machine Learning (ICML)*, pages 33–40, 2006.

[Att00] H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12 (NIPS)*, 2000.

[AV07] C. Archambeau and M. Verleysen. Robust Bayesian clustering, 2007.

[Bea03] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BL07] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup and Workshop*, 2007.

[CFPR03] C. Croux, P. Filzmoser, G. Pison, and PJ Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13(1):23–36, 2003.

[CG05] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6(1):1019–1041, 2005.

[DV08] N. Delannay and M. Verleysen. Collaborative filtering with interlaced generalized linear models. *Neurocomputing*, 71:1300–1310, 2008.

[GRGP01] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[GZ79] K.R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, 1979.

[HP99] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *International Joint Conference in Artificial Intelligence (IJCAI)*, 1999.

[Jør82] B. Jørgensen. *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Springer-Verlag, 1982.

[KMBM10] M. E. Khan, B. Marlin, G. Bouchard, and K. Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems 23 (NIPS)*, 2010.

[LR95] C. Liu and D. B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

[LT07] Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *KDD Cup and Workshop*, 2007.

[MHN07] B. Mehta, T. Hofmann, and W. Nejdl. Robust collaborative filtering. In *ACM Conference On Recommender Systems*, pages 49 – 56, 2007.

[MY08] R.A. Maronna and V.J. Yohai. Robust Low-Rank Approximation of Data Matrices With Elementwise Contamination. *Technometrics*, 50(3):295–304, 2008.

[NH98] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, 89:355–368, 1998.

[NS10] S. Nakajima and M.i Sugiyama. Implicit regularization in variational bayesian matrix factorization. In *27th International Conference on Machine Learning (ICML)*, 2010.

[RIK07] T. Raiko, E. Ilin, and J. Karhunen. Principal component analysis for sparse high-dimensional data. In *14th International Conference on Neural Information Processing (ICONIP)*, pages 566–575, 2007.

[SJ03] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *In 20th International Conference on Machine Learning*, pages 720–727. AAAI Press, 2003.

[SM08a] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *25th International Conference on Machine Learning (ICML)*, 2008.

[SM08b] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2008.

[TL05] M. E. Tipping and N. D. Lawrence. Variational inference for Student-t models: robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69:123–141, 2005.

[Wie00] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Uncertainty in Artificial Intelligence (UAI)*, pages 626–633, 2000.

A LATENT VARIABLE VIEW OF THE STUDENT-t

Let $\mathbf{w}$ be a D-dimensional Gaussian random vector, such that $p(\mathbf{w}|\tau_{\mathbf{w}}) = \mathcal{N}(\mathbf{w}|\tau_{\mathbf{w}}\boldsymbol{\Lambda}_{\mathbf{w}})$. Assuming a Gamma prior $\mathcal{G}a(\tau_{\mathbf{w}}|a,b)$ leads to

$$\begin{aligned} p(\mathbf{w}) &= \int_{\tau_{\mathbf{w}}} p(\mathbf{w}|\tau_{\mathbf{w}})\, p(\tau_{\mathbf{w}})\, d\tau_{\mathbf{w}} \\ &= \frac{|\boldsymbol{\Lambda}_{\mathbf{w}}|^{1/2}}{(2\pi)^{D/2}} \frac{b^a}{\Gamma(a)} \int_{\tau_{\mathbf{w}}} \tau_{\mathbf{w}}^{a+\frac{D}{2}-1} e^{-(\frac{1}{2}\mathbf{w}^\top \boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{w}+b)\tau_{\mathbf{w}}}\, d\tau_{\mathbf{w}} \\ &= \frac{|\boldsymbol{\Lambda}_{\mathbf{w}}|^{1/2}}{(2\pi)^{D/2}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(\frac{D}{2}+a)}{(b+\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{w})^{\frac{D}{2}+a}} \\ &= \mathcal{S}t(0, \tfrac{a}{b}\boldsymbol{\Lambda}_{\mathbf{w}}, 2a), \end{aligned}$$

where the zero-mean multivariate Student-t probability density function is defined as follows

$$\mathcal{S}t(\mathbf{x}|0,\boldsymbol{\Lambda},\nu) = \frac{\Gamma(\frac{D+\nu}{2})|\boldsymbol{\Lambda}|^{1/2}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{D/2}} \left(1+\frac{1}{\nu}\mathbf{x}^\top\boldsymbol{\Lambda}\mathbf{x}\right)^{-\frac{D+\nu}{2}}.$$

Vector $\mathbf{x}$ is D-dimensional, $\boldsymbol{\Lambda}$ is the precision matrix and ν is the shape parameter (or the number of degrees of freedom when an integer). Note that this is a different hierarchical construction than the one proposed in [LR95].

B Specific form of $q(\alpha_n)$ and $q(\beta_m)$

Consider the observed variable $\mathbf{x}$ and the set of latent variables $\mathbf{y} = \{\mathbf{y}_k\}_k$ and $\mathbf{z} = \{\mathbf{z}_k\}_k$. Next we show how to easily compute the posterior $q(\mathbf{z}_k)$ when considering a structured variational approximation with joint posterior $q(\mathbf{y},\mathbf{z}) = \prod_k q(\mathbf{y}_k,\mathbf{z}_k)$.

Assume the marginal $p(\mathbf{x})$ is analytically intractable. The variational lower bound is defined as

$$-\mathcal{F}(q) = \langle \ln p(\mathbf{x},\mathbf{y},\mathbf{z})\rangle_{q(\mathbf{y},\mathbf{z})} + \mathrm{H}\left[q(\mathbf{y},\mathbf{z})\right],$$

where $q(\mathbf{y},\mathbf{z}) = \prod_k q(\mathbf{y}_k,\mathbf{z}_k)$. Let us denote the incomplete product $\prod_{k'\neq k} q(\mathbf{y}_{k'},\mathbf{z}_{k'})$ by $q(\mathbf{y}_{\setminus k},\mathbf{z}_{\setminus k})$. Exploiting the specific form of the variational posterior and ignoring terms independent of $\mathbf{z}_k$, we obtain

$$\begin{aligned} -\mathcal{F}(q) &= \int \langle \ln p(\mathbf{x},\mathbf{y},\mathbf{z})\rangle_{q(\mathbf{y}_k|\mathbf{z}_k)q(\mathbf{y}_{\setminus k},\mathbf{z}_{\setminus k})} q(\mathbf{z}_k) d\mathbf{z}_k \\ &\quad + \int \mathrm{H}\left[q(\mathbf{y}_k|\mathbf{z}_k)\right] q(\mathbf{z}_k) d\mathbf{z}_k + \mathrm{H}\left[q(\mathbf{z}_k)\right] + const \\ &= \int \ln e^{\langle \ln p(\mathbf{x},\mathbf{y},\mathbf{z})\rangle_{q(\mathbf{y}_k|\mathbf{z}_k)q(\mathbf{y}_{\setminus k},\mathbf{z}_{\setminus k})} + \mathrm{H}[q(\mathbf{y}_k|\mathbf{z}_k)]} q(\mathbf{z}_k) d\mathbf{z}_k \\ &\quad + \mathrm{H}\left[q(\mathbf{z}_k)\right] + const\ . \end{aligned}$$

Up to a normalising constant Z we have that $-\mathcal{F}(q)$ is equal to

$$-\mathcal{KL}\left(q(\mathbf{z}_k)\|\tfrac{1}{Z}e^{\langle \ln p(\mathbf{x},\mathbf{y},\mathbf{z})\rangle_{q(\mathbf{y}_k|\mathbf{z}_k)q(\mathbf{y}_{\setminus k},\mathbf{z}_{\setminus k})} + \mathrm{H}[q(\mathbf{y}_k|\mathbf{z}_k)]}\right).$$

Hence, when assuming $q(\boldsymbol{\Phi},\boldsymbol{\alpha}) = \prod_n q(\boldsymbol{\phi}_n,\alpha_n)$, the bound (7) is maximal if

$$\begin{aligned} q(\alpha_n) \propto\ & e^{\langle \ln p(\mathbf{R},\boldsymbol{\Phi},\boldsymbol{\Omega},\boldsymbol{\alpha},\boldsymbol{\beta})\rangle_{q(\boldsymbol{\phi}_n|\alpha_n)q(\boldsymbol{\Phi}_{\setminus n},\boldsymbol{\alpha}_{\setminus n})q(\boldsymbol{\Omega},\boldsymbol{\beta})}} \\ & \times e^{\mathrm{H}[q(\boldsymbol{\phi}_n|\alpha_n)]}, \end{aligned}$$

where $q(\boldsymbol{\Phi}_{\setminus n},\boldsymbol{\alpha}_{\setminus n}) = \prod_{n'\neq n} q(\boldsymbol{\phi}_{n'},\alpha_{n'})$. The variational posterior $q(\beta_m)$ is obtained in the same manner.

C GENERALISED INVERSE GAUSSIAN

The generalised inverse Gaussian distribution [Jør82] is given by

$$\mathcal{N}^{-1}(x|\nu,\chi,\varphi) = \frac{1}{\mathcal{Z}(\nu,\chi,\varphi)} x^{\nu-1} e^{-\frac{1}{2}(\chi x^{-1}+\varphi x)},$$

where $\mathcal{Z}(\nu,\chi,\varphi) = 2\left(\chi/\phi\right)^{\nu/2} K_\nu(\sqrt{\chi\varphi})$ with $K_\nu(\cdot)$ denoting the modified Bessel function of the second kind with index $\nu \in \mathbb{R}$. The following expectations are useful: $\langle x\rangle = \sqrt{\chi/\phi}R_\nu(\sqrt{\chi\phi})$, $\langle x^{-1}\rangle = \sqrt{\phi/\chi}R_{-\nu}(\sqrt{\chi\phi})$ and $\langle \ln x\rangle = \ln\sqrt{\frac{\chi}{\phi}} + \frac{d\ln K_\nu(\sqrt{\chi\phi})}{d\nu}$, where we defined $R_\nu(\cdot) = K_{\nu+1}(\cdot)/K_\nu(\cdot)$.

When $\chi = 0$ and $\nu > 0$, the generalised inverse Gaussian reduces to the Gamma:

$$\mathcal{G}a(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx},$$

where $a,b > 0$ and $\Gamma(\cdot)$ is the (complete) gamma function. It is straightforward to verify this result by setting $a = \nu$ and $b = \varphi/2$, and noting that $\lim_{z\to 0} K_\nu(z) = \Gamma(\nu)2^{\nu-1}z^{-\nu}$ for $\nu > 0$.