Implicit Regularization in Variational Bayesian Matrix Factorization

Shinichi Nakajima NAKAJIMA.S@NIKON.CO.JP
Nikon Corporation, Tokyo 140-8601, Japan

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Tokyo Institute of Technology and JST PRESTO, Tokyo 152-8552, Japan

Abstract

Matrix factorization into the product of low-rank matrices induces *non-identifiability*, i.e., the mapping between the target matrix and factorized matrices is not one-to-one. In this paper, we theoretically investigate the influence of non-identifiability on Bayesian matrix factorization. More specifically, we show that a variational Bayesian method involves regularization effect even when the prior is non-informative, which is intrinsically different from the maximum a posteriori approach. We also extend our analysis to empirical Bayes scenarios where hyperparameters are also learned from data.

1. Introduction

The goal of *matrix factorization* (MF) is to find a low-rank expression of a target matrix. MF has been used for learning linear relation between vectors such as *reduced rank regression* (Baldi & Hornik, 1995; Reinsel & Velu, 1998), *canonical correlation analysis* (Rao, 1965; Anderson, 1984), and *partial least-squares* (Rosipal & Krämer, 2006). More recently, MF is applied to *collaborative filtering* (CF) in the context of recommender systems (Konstan et al., 1997; Funk, 2006) and microarray data analysis (Baldi & Brunak, 1998). For this reason, MF has attracted considerable attention these days.

Recently, the *variational Bayesian* (VB) approach (Attias, 1999) has been applied to MF (Lim & Teh, 2007; Raiko et al., 2007). The VBMF method was shown to perform very well in experiments. However, its good performance was not completely understood beyond its experimental success. The purpose of this paper is to provide new insight into Bayesian MF.

A key characteristic of MF models is *non-identifiability* (Watanabe, 2009), where the mapping between parameters and functions is not one-to-one—in the context of MF, the mapping between the target matrix and the factorized matrices is not one-to-one. Previous theoretical studies on non-identifiable models showed that, when combined with *full-Baysian* (FB) estimation, regularization effect is significantly stronger than the MAP method (Watanabe, 2001; Yamazaki & Watanabe, 2003). Since a single point in the function space corresponds to a set of points in the (redundant) parameter space in non-identifiable models, simple distributions such as the Gaussian distribution in the function space produce highly complicated *multimodal* distributions in the parameter space. This causes the MAP and FB solutions to be significantly different. Thus the behavior of non-identifiable models is substantially different from that of identifiable models.

Theoretical properties of VB has been investigated for Gaussian mixture models (Watanabe & Watanabe, 2006) and linear neural networks (Nakajima & Watanabe, 2007). In this paper, we extend these results and investigate the behavior of the VBMF estimator. More specifically, we show that VBMF consists of two shrinkage factors, the *positive-part James-Stein* (PJS) shrinkage (James & Stein, 1961; Efron & Morris, 1973) and the *trace-norm* shrinkage (Srebro et al., 2005), operating on each singular component separately for producing low-rank solutions. The trace-norm shrinkage is simply induced by non-flat prior information, as in the MAP approach (Salakhutdinov & Mnih, 2008). Thus, no trace-norm shrinkage remains when priors are non-informative. On the other hand, we show a counter-intuitive fact that the PJS shrinkage factor still remains even with uniform priors. This allows the VBMF method to avoid overfitting (or in some

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

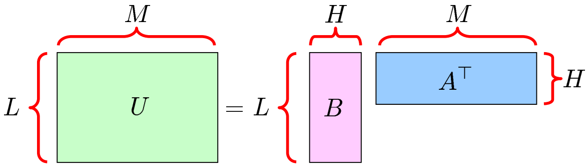

Figure 1. Matrix factorization model ($H \le L \le M$).

cases this might cause underfitting) even when non-informative priors are provided.

We further extend the above analysis to *empirical VBMF* (EVBMF) scenarios, where hyperparameters in prior distributions are also learned based on the *VB free energy*. We derive bounds of the EVBMF estimator, and show that the effect of PJS shrinkage is at least doubled compared with the uniform prior cases.

2. Bayesian Approaches to Matrix Factorization

In this section, we give a probabilistic formulation of the *matrix factorization* (MF) problem and review its Bayesian methods.

2.1. Formulation

The goal of the MF problem is to estimate a target matrix U $(\in \mathbb{R}^{L\times M})$ from its n observations

$$\mathcal{V}^n = \{V^{(i)} \in \mathbb{R}^{L\times M} \mid i = 1, \ldots, n\}.$$

Throughout the paper, we assume $L \le M$. If $L > M$, we may simply re-define the transpose $U^\top$ as U so that $L \le M$ holds. Thus this does not impose any restriction.

A key assumption of MF is that U is a low-rank matrix. Let H $(\le L)$ be the rank of U. Then the matrix U can be decomposed into the product of $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_H) \in \mathbb{R}^{M\times H}$ and $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_H) \in \mathbb{R}^{L\times H}$ as follows (see Figure 1):

$$U = BA^\top.$$

Assume that the observed matrix V is subject to the additive-noise model

$$V = U + \mathcal{E},$$

where $\mathcal{E}$ $(\in \mathbb{R}^{L\times M})$ is a noise matrix. Each entry of $\mathcal{E}$ is assumed to independently follow the Gaussian distribution with mean zero and variance σ^2. Then, the likelihood $p(\mathcal{V}^n|A, B)$ is given by

$$p(\mathcal{V}^n|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \|V^{(i)} - BA^\top\|^2_{\mathrm{Fro}}\right), \quad (1)$$

where $\|\cdot\|^2_{\mathrm{Fro}}$ denotes the *Frobenius norm* of a matrix.

2.2. Bayesian Matrix Factorization

We use the Gaussian priors on the parameters A, B:

$$\begin{aligned} \phi(U) &= \phi_A(A)\phi_B(B), \\ \phi_A(A) &\propto \exp\left(-\textstyle\sum_{h=1}^{H} \|\boldsymbol{a}_h\|^2/(2c^2_{a_h})\right), \\ \phi_B(B) &\propto \exp\left(-\textstyle\sum_{h=1}^{H} \|\boldsymbol{b}_h\|^2/(2c^2_{b_h})\right). \end{aligned}$$

$c^2_{a_h}$ and $c^2_{b_h}$ are hyperparameters corresponding to the prior variance of those vectors. Without loss of generality, we assume that the product $c_{a_h}c_{b_h}$ is non-increasing with respect to h.

The *Bayes posterior* $p(A, B|\mathcal{V}^n)$ can be written as

$$p(A,B|\mathcal{V}^n) = \frac{p(\mathcal{V}^n|A,B)\phi_A(A)\phi_B(B)}{\langle p(\mathcal{V}^n|A,B)\rangle_{\phi_A(A)\phi_B(B)}}, \quad (2)$$

where $\langle\cdot\rangle_p$ denotes the expectation over p. The *full-Bayesian* (FB) solution is given by the *Bayes posterior mean*:

$$\widehat{U}^{\mathrm{FB}} = \langle BA^\top\rangle_{p(A,B|\mathcal{V}^n)}. \quad (3)$$

The hyperparameters c_{a_h} and c_{b_h} may be determined so that the *Bayes free energy* $F(\mathcal{V}^n)$ is minimized.

$$F(\mathcal{V}^n) = -\log\langle p(\mathcal{V}^n|A,B)\rangle_{\phi_A(A)\phi_B(B)}. \quad (4)$$

This method is called the *empirical Bayes* method (Bishop, 2006).

2.3. Maximum A Posteriori Matrix Factorization (MAPMF)

When computing the Bayes posterior (2), the expectation in the denominator of Eq.(2) is often intractable due to high dimensionality of the parameters A and B. A simple approach to mitigating this problem is to use the *maximum a posteriori* (MAP) approximation. The MAP solution $\widehat{U}^{\mathrm{MAP}}$ is given by

$$\widehat{U}^{\mathrm{MAP}} = \widehat{B}^{\mathrm{MAP}}\widehat{A}^{\mathrm{MAP}\top},$$

where $(\widehat{A}^{\mathrm{MAP}}, \widehat{B}^{\mathrm{MAP}}) = \mathrm{argmax}_{A,B}\, p(A,B|\mathcal{V}^n)$.

2.4. Variational Bayesian Matrix Factorization (VBMF)

Another approach to avoiding computational intractability of the FB method is to use the VB approximation (Attias, 1999; Bishop, 2006). Here, we review the VBMF method proposed by Lim and Teh (2007) and Raiko et al. (2007).

For a *trial* distribution

$$r(A, B|\mathcal{V}^n) = \prod_{h=1}^{H} r_{a_h}(\boldsymbol{a}_h|\mathcal{V}^n) r_{b_h}(\boldsymbol{b}_h|\mathcal{V}^n),$$

the *VB free energy* is defined as

$$F(r|\mathcal{V}^n) = \left\langle \log \frac{r(A, B|\mathcal{V}^n)}{p(\mathcal{V}^n, A, B)} \right\rangle_{r(A,B|\mathcal{V}^n)}.$$

The VB approach minimizes the VB free energy with respect to the trial distribution $r(A, B|\mathcal{V}^n)$. The resulting distribution is called the *VB posterior*. The VB solution $\widehat{U}^{\mathrm{VB}}$ is given by the *VB posterior mean*:

$$\widehat{U}^{\mathrm{VB}} = \langle BA^\top \rangle_{r(A,B|\mathcal{V}^n)}.$$

Applying the variational method to the VB free energy shows that the VB posterior satisfies

$$r_{a_h}(\boldsymbol{a}_h|\mathcal{V}^n) \propto \phi_{a_h}(\boldsymbol{a}_h) \exp\left(\langle \log p(\mathcal{V}^n|A, B)\rangle_{r(\backslash \boldsymbol{a}_h|\mathcal{V}^n)}\right),$$
$$r_{b_h}(\boldsymbol{b}_h|\mathcal{V}^n) \propto \phi_{b_h}(\boldsymbol{b}_h) \exp\left(\langle \log p(\mathcal{V}^n|A, B)\rangle_{r(\backslash \boldsymbol{b}_h|\mathcal{V}^n)}\right),$$

where $r(\backslash \boldsymbol{a}_h|\mathcal{V}^n)$ denotes the VB posterior except $\boldsymbol{a}_h$.

2.5. Empirical Variational Bayesian Matrix Factorization (EVBMF)

The VB free energy also allows us to determine the hyperparameters $c_{a_h}^2$ and $c_{b_h}^2$ in a computationally tractable way. That is, instead of the Bayes free energy $F(\mathcal{V}^n)$, the VB free energy $F(r|\mathcal{V}^n)$ is minimized with respect to $c_{a_h}^2$ and $c_{b_h}^2$. We call this method *empirical VBMF* (EVBMF).

3. Analysis of MAPMF, VBMF, and EVBMF

In this section, we theoretically analyze the behavior of MAPMF, VBMF and EVBMF solutions. More specifically, we derive analytic-form expression of the MAPMF solution (Section 3.1), semi-analytic expressions of the VBMF solution (Section 3.2) and the EVBMF solution (Section 3.3), and we elucidate their regularization mechanism.

3.1. MAPMF

Let γ_h (≥ 0) be the h-th largest singular value of $\overline{V}$:

$$\overline{V} = \frac{1}{n} \sum_{i=1}^{n} V^{(i)}.$$

Let $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ be the associated right and left singular vectors:

$$\overline{V} = \sum_{h=1}^{L} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top.$$

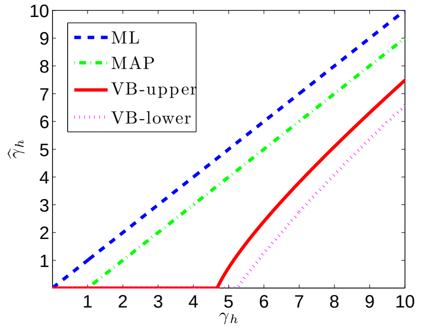

Figure 2. Shrinkage of the ML estimator (6), the MAP estimator (5), and the VB estimator (8) when $n = 1$, $\sigma^2 = 0.1$, $c_{a_h} c_{b_h} = 0.1$, $L = 100$, and $M = 200$.

Then we have the following theorem.

Theorem 1 *The MAP estimator $\widehat{U}^{MAP}$ is given by*

$$\widehat{U}^{MAP} = \sum_{h=1}^{H} \widehat{\gamma}_h^{MAP} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top,$$

where, for $[c]^+ = \max(0, c)$,

$$\widehat{\gamma}_h^{MAP} = \left[\gamma_h - \frac{\sigma^2}{n c_{a_h} c_{b_h}}\right]^+. \tag{5}$$

The theorem implies that the MAP solution cuts off the singular values less than $\sigma^2/(n c_{a_h} c_{b_h})$; otherwise it reduces the singular values by $\sigma^2/(n c_{a_h} c_{b_h})$ (see Figure 2). This shrinkage effect allows the MAPMF method to avoid overfitting.

Similarly to Theorem 1, we can show that the *maximum likelihood* (ML) estimator is given by

$$\widehat{U}^{\mathrm{ML}} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{ML}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top,$$

where

$$\widehat{\gamma}_h^{\mathrm{ML}} = \gamma_h \text{ for all } h. \tag{6}$$

Thus the ML solution is reduced to $\overline{V}$ when $H = L$ (see Figure 2):

$$\widehat{U}^{\mathrm{ML}} = \sum_{h=1}^{L} \widehat{\gamma}_h^{\mathrm{ML}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top = \overline{V}.$$

A parametric model is said to be *identifiable* if the mapping between parameters and functions is one-to-one; otherwise the model is said to be *non-identifiable* (Watanabe, 2001). Since the decomposition $U = BA^\top$

is redundant, the MF model is non-identifiable. For identifiable models, the MAP estimator with the uniform prior is reduced to the ML estimator (Bishop, 2006). On the other hand, the MF model is non-identifiable because of the redundancy of the decomposition $U = BA^\top$. This implies that a single point in the space of U corresponds to a set of points in the joint space of A and B. For this reason, the uniform priors on A and B do not produce the uniform prior on U. Nevertheless, Eqs.(5) and (6) imply that MAP is reduced to ML when the priors on A and B are uniform (i.e., $c_{a_h}, c_{b_h} \to \infty$).

More precisely, Eqs.(5) and (6) show that $c_{a_h}c_{b_h} \to \infty$ is sufficient for MAP to be reduced to ML, which is weaker than $c_{a_h}, c_{b_h} \to \infty$. This implies that both priors on A and B do not have to be uniform; only the condition that one of the priors is uniform is sufficient for MAP to be reduced to ML in the MF model. This phenomenon is distinctively different from the case of identifiable models.

When the prior is uniform and the likelihood is Gaussian, the posterior is Gaussian. Thus the mean and mode of the posterior agrees with each other due to the symmetry of the Gaussian density. For identifiable models, this fact implies that the FB and MAP solutions agree with each other. However, the FB and MAP solutions are generally different in non-identifiable models since the symmetry of the Gaussian density in the space of U is no longer kept in the joint space of A and B.

In Section 4.1, we further investigate these distinctive features of the MF model using illustrative examples.

3.2. VBMF

Next, let us analyze the behavior of the VBMF estimator. We have the following theorem.

Theorem 2 *$\widehat{U}^{VB}$ is expressed as*

$$\widehat{U}^{VB} = \sum_{h=1}^{H} \widehat{\gamma}_h^{VB} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top. \tag{7}$$

When $\gamma_h \leq \sqrt{M\sigma^2/n}$, $\widehat{\gamma}_h^{VB} = 0$. When $\gamma_h > \sqrt{M\sigma^2/n}$, $\widehat{\gamma}_h^{VB}$ is bounded as

$$\left[\left(1-\frac{M\sigma^2}{n\gamma_h^2}\right)\gamma_h - \frac{\sigma^2\sqrt{M/L}}{nc_{a_h}c_{b_h}}\right]^+ \leq \widehat{\gamma}_h^{VB} < \left(1-\frac{M\sigma^2}{n\gamma_h^2}\right)\gamma_h. \tag{8}$$

The upper- and lower-bounds in Eq.(8) are illustrated in Figure 2. In the limit when $c_{a_h}c_{b_h} \to \infty$, the lower-bound agrees with the upper-bound, and we have

$$\lim_{c_{a_h}c_{b_h}\to\infty} \widehat{\gamma}_h^{\mathrm{VB}} = \left[\left(1 - \frac{M\sigma^2}{n\gamma_h^2}\right)\gamma_h\right]^+ \tag{9}$$

if $\gamma_h > 0$; otherwise $\widehat{\gamma}_h^{\mathrm{VB}} = 0$. This is the same form as the *positive-part James-Stein (PJS) shrinkage estimator* (James & Stein, 1961; Efron & Morris, 1973). The factor $M\sigma^2/n$ is the expected contribution of the noise to γ_h^2—when the target matrix is $U = 0$, the expectation of γ_h^2 over all h is given by $M\sigma^2/n$. When $\gamma_h^2 \leq M\sigma^2/n$, Eq.(9) implies that $\widehat{\gamma}_h^{\mathrm{VB}} = 0$. Thus, the PJS estimator cuts off the singular components dominated by noise. As γ_h^2 increases, the PJS shrinkage factor $M\sigma^2/(n\gamma_h^2)$ tends to 0, and thus the estimated singular value $\widehat{\gamma}_h^{\mathrm{VB}}$ becomes close to the original singular value γ_h.

Let us compare the behavior of the VB solution (9) with that of the MAP solution (5) when $c_{a_h}c_{b_h} \to \infty$. In this case, the MAP solution merely results in the ML solution where no regularization is incorporated. In contrast, VB offers PJS-type regularization even when $c_{a_h}c_{b_h} \to \infty$; thus VB can still mitigate overfitting. This fact is in good agreement with the experimental results reported in Raiko et al. (2007), where no overfitting is observed when $c_{a_h}^2 = 1$ and $c_{b_h}^2$ is set to large values. This counter-intuitive fact stems again from the non-identifiability of the MF model—the Gaussian noise $\mathcal{E}$ imposed in the space of U possesses a very complex surface in the joint space of A and B, in particular, *multimodal* structure. This causes the MAP solution to be distinctively different from the VB solution. In Section 4.2, we investigate the above phenomena in more detail using illustrative examples.

We can derive another upper-bound of $\widehat{\gamma}_h^{\mathrm{VB}}$, which depends on hyperparameters c_{a_h} and c_{b_h}.

Theorem 3 *When $\gamma_h > \sqrt{M\sigma^2/n}$, $\widehat{\gamma}_h^{VB}$ is upper-bounded as*

$$\widehat{\gamma}_h^{VB} \leq \left[\sqrt{\left(1 - \frac{L\sigma^2}{n\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{n\gamma_h^2}\right)} \cdot \gamma_h - \frac{\sigma^2}{nc_{a_h}c_{b_h}}\right]^+.$$

When $L = M$ and $\gamma_h > \sqrt{M\sigma^2/n}$, the above upper-bound agrees with the lower-bound in Eq.(8), and thus we have

$$\widehat{\gamma}_h^{\mathrm{VB}} = \left[\left(1 - \frac{M\sigma^2}{n\gamma_h^2}\right)\gamma_h - \frac{\sigma^2}{nc_{a_h}c_{b_h}}\right]^+ \tag{10}$$

if $\gamma_h > 0$; otherwise $\widehat{\gamma}_h^{\mathrm{VB}} = 0$. Then the complete VB posterior can be obtained analytically.

Corollary 1 *When $L = M$, the VB posteriors are given by*

$r_A(A|\mathcal{V}^n) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \boldsymbol{\mu}_{a_h}, \Sigma_{a_h})$,

$r_B(B|\mathcal{V}^n) = \prod_{h=1}^{H} \mathcal{N}_L(\boldsymbol{b}_h; \boldsymbol{\mu}_{b_h}, \Sigma_{b_h})$,

$$\boldsymbol{\mu}_{a_h} = \pm\sqrt{\tfrac{c_{a_h}}{c_{b_h}}\widehat{\gamma}_h^{VB}} \cdot \boldsymbol{\omega}_{a_h}, \quad \Sigma_{a_h} = \tfrac{c_{a_h}}{2Mc_{b_h}}\kappa I_M,$$

$$\boldsymbol{\mu}_{b_h} = \pm\sqrt{\frac{c_{b_h}}{c_{a_h}}\widehat{\gamma}_h^{VB}} \cdot \boldsymbol{\omega}_{b_h}, \quad \Sigma_{b_h} = \frac{c_{b_h}}{2Mc_{a_h}}\kappa I_M,$$

$$\kappa = \sqrt{\left(\widehat{\gamma}_h^{VB} + \tfrac{\sigma^2}{nc_{a_h}c_{b_h}}\right)^2 + \tfrac{4\sigma^2 M}{n}} - \left(\widehat{\gamma}_h^{VB} + \tfrac{\sigma^2}{nc_{a_h}c_{b_h}}\right),$$

where $\mathcal{N}_d(\cdot; \boldsymbol{\mu}, \Sigma)$ denotes the d-dimensional Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and $\widehat{\gamma}_h^{VB}$ is given by Eq.(10).

3.3. EVBMF

Finally, we analyze the behavior of EVBMF, where hyperparameters c_{a_h} and c_{b_h} are also estimated from data. We have the following theorem.

Theorem 4 *The EVB estimator is given by the following form.*

$$\widehat{U}^{EVB} = \sum_{h=1}^{H} \widehat{\gamma}_h^{EVB} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top. \tag{11}$$

When $\gamma_h < (\sqrt{L} + \sqrt{M})\sigma/\sqrt{n}$, $\widehat{\gamma}_h^{EVB} = 0$. When $\gamma_h \geq (\sqrt{L} + \sqrt{M})\sigma/\sqrt{n}$, $\widehat{\gamma}_h^{EVB}$ is upper-bounded as

$$\widehat{\gamma}_h^{EVB} < \left(1 - \frac{M\sigma^2}{n\gamma_h^2}\right)\gamma_h. \tag{12}$$

When $\gamma_h \geq \sqrt{7M}\sigma/\sqrt{n}$ $(> (\sqrt{L} + \sqrt{M})\sigma/\sqrt{n})$, $\widehat{\gamma}_h^{EVB}$ is lower-bounded as

$$\widehat{\gamma}_h^{EVB} > \left[\left(1 - \frac{2M\sigma^2}{n\gamma_h^2 - \sqrt{n\gamma_h^2(L + M + \sqrt{LM})\sigma^2}}\right)\gamma_h\right]^+. \tag{13}$$

Note that the inequality in Eq.(13) is strict.

As pointed out by Raiko et al. (2007), if c_{a_h}, c_{b_h}, A, and B are all estimated so that the Bayes posterior is maximized (i.e., '*empirical MAP*'; EMAP), the solution is trivial (i.e., $\widehat{\gamma}^{\mathrm{EMAP}} = 0$) irrespective of the observation. In contrast, Theorem 4 implies that EVB gives a non-trivial solution (i.e., $\widehat{\gamma}_h^{\mathrm{EVB}} > 0$) when $\gamma_h \geq \sqrt{7M}\sigma/\sqrt{n}$. It is also note worthy that the upper-bound in Eq.(12) is the same as that in Eq.(8). Thus, even when the hyperparameters c_{a_h} and c_{b_h} are learned from data, the same upper-bound as the fixed-hyperparameter case holds.

Another upper-bound of $\widehat{\gamma}_h^{\mathrm{EVB}}$ is given as follows.

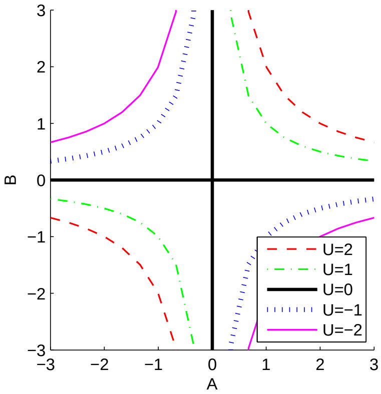

Figure 3. Equivalence class. Any A and B such that their product is unchanged give the same matrix U.

Theorem 5 *When $\gamma_h \geq (\sqrt{L} + \sqrt{M})\sigma/\sqrt{n}$, $\widehat{\gamma}_h^{EVB}$ is upper-bounded as*

$$\widehat{\gamma}_h^{EVB} < \sqrt{\left(1 - \frac{L\sigma^2}{n\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{n\gamma_h^2}\right)}\gamma_h - \frac{\sqrt{LM}\sigma^2}{n\gamma_h}.$$

When $L = M$, the above upper-bound is sharper than that in Eq.(12), resulting in

$$\widehat{\gamma}_h^{\mathrm{EVB}} < \left(1 - \frac{2M\sigma^2}{n\gamma_h^2}\right)\gamma_h. \tag{14}$$

Thus, the PJS shrinkage factor of the upper-bound (14) of EVBMF is $2M\sigma^2/(n\gamma_h^2)$. On the other hand, as shown in Eq.(9), the PJS shrinkage factor of plain VBMF with uniform priors on A and B (i.e., $c_a, c_b \to \infty$) is $M\sigma^2/(n\gamma_h^2)$, which is *less than a half* of EVBMF. Thus, EVBMF provides substantially stronger regularization effect than plain VBMF with uniform priors.

Furthermore, from Eq.(10), we can confirm that the upper-bound (14) is equivalent to the VB solution when $c_{a_h}c_{b_h} = \gamma_h/M$.

4. Illustration of Influence of Non-identifiability

In order to understand the regularization mechanism of MAPMF, VBMF, and EVBMF more intuitively, let us illustrate the influence of non-identifiability when $L = M = H = 1$ (i.e., U, V, A, and B are merely scalars). In this case, any A and B such that their product is unchanged form an *equivalence class* and give the same value U (see Figure 3). When $U = 0$, the equivalence class is a cross shape on the A- and B-axes; otherwise, it forms a hyperbolic curve.

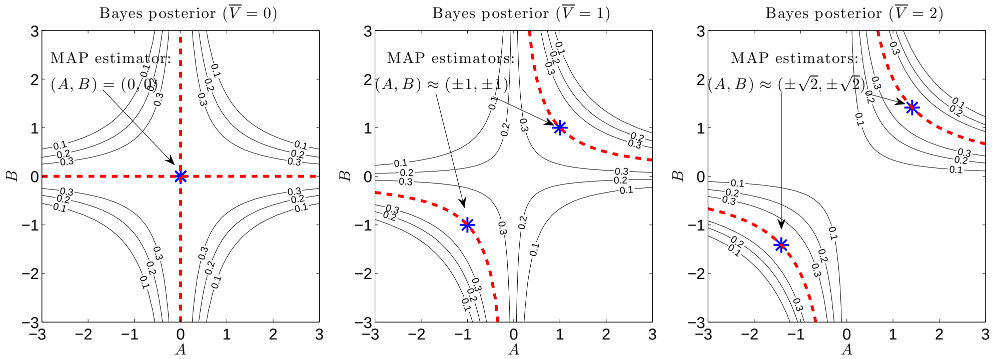

Figure 4. Bayes posteriors with $c_a = c_b = 100$ (i.e., almost flat priors). The asterisks are the MAP solutions, and the dashed lines indicate the modes when $c_a, c_b \to \infty$.

4.1. MAPMF

When $L = M = H = 1$, the Bayes posterior $p(A, B|\mathcal{V}^n)$ is proportional to

$$\exp\left(-\frac{n}{2\sigma^2}(\overline{V} - BA)^2 - \frac{A^2}{2c_a^2} - \frac{B^2}{2c_b^2}\right). \tag{15}$$

Figure 4 shows the contour of the above Bayes posterior when $\overline{V} = 0, 1, 2$ are observed, where the number of samples is $n = 1$, the noise variance is $\sigma^2 = 1$, and the hyperparameters are $c_a = c_b = 100$ (i.e., almost flat priors). When $\overline{V} = 0$, the surface has a cross shape and its maximum is at the origin. When $\overline{V} > 0$, the surface is divided into the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$), and the two 'modes' get farther as $\overline{V} > 0$ increases.

For finite c_a and c_b, the MAP solution can be expressed as

$$\widehat{A}^{\mathrm{MAP}} = \pm\sqrt{\frac{c_a}{c_b}\left[|\overline{V}| - \frac{\sigma^2}{nc_ac_b}\right]^+},$$

$$\widehat{B}^{\mathrm{MAP}} = \pm\mathrm{sign}(\overline{V})\sqrt{\frac{c_b}{c_a}\left[|\overline{V}| - \frac{\sigma^2}{nc_ac_b}\right]^+},$$

where sign(·) denotes the sign of a scalar. In Figure 4, the MAP estimators are indicated by the asterisks; the dashed lines indicate the modes of the contour of Eq.(15) when $c_a, c_b \to \infty$. When $\overline{V} = 0$, the Bayes posterior takes the maximum value on the A- and B-axes, which results in $\widehat{U}^{\mathrm{MAP}} = 0$. When $\overline{V} = 1$, the profile of the peaks of the Bayes posterior is hyperbolic and the maximum value is achieved on the hyperbolic curves in the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$); in either case, $\widehat{U}^{\mathrm{MAP}} \approx 1$. When $\overline{V} = 2$, a similar multimodal structure is observed. From these plots, we can visually confirm that the MAP solution with almost flat priors ($c_a = c_b = 100$) approximately agrees with the ML solution: $\widehat{U}^{\mathrm{MAP}} \approx \widehat{U}^{\mathrm{ML}} = \overline{V}$.

Furthermore, these graphs explain the reason why $c_ac_b \to \infty$ is sufficient for MAP to agree with ML in the MF setup (see Section 3). Suppose c_a is kept small, say $c_a = 1$, in Figure 4. Then the Gaussian 'decay' remains along the horizontal axis in the profile of the Bayes posterior. However, the MAP solution $\widehat{U}^{\mathrm{MAP}}$ does not change since the mode of the Bayes posterior is kept lying on the dashed line (equivalence class). Thus, MAP agrees with ML if either of c_a and c_b tends to infinity.

If $\overline{V} = 0, 1, 2$ are observed, the FB solutions (see Eq.(3)) are given by $0, 0.92, 1.93$, respectively (which were numerically computed). Since the corresponding MAP solutions are $0, 1, 2$, FB and MAP were shown to produce different solutions. This happened because the Gaussian density in the space of U is no longer symmetric in the joint space of A and B (see Figure 4 again), and thus the posterior mean and mode are different.

We can further show that the prior proportional to $\sqrt{A^2 + B^2}$ (which is *improper*) corresponds to the *Jeffreys non-informative prior* (Jeffreys, 1946) for the MF model.

4.2. VBMF

Next, we illustrate the behavior of the VB estimator, where the Bayes posterior is approximated by a spherical Gaussian. In the current one-dimensional setup, Corollary 1 implies that the VB posteriors $r_A(A|\mathcal{V}^n)$ and $r_B(B|\mathcal{V}^n)$ are expressed as

$$r_A(A|\mathcal{V}^n) = \mathcal{N}\left(A; \pm\sqrt{\widehat{\gamma}^{\mathrm{VB}}\frac{c_a}{c_b}}, \zeta\frac{c_a}{c_b}\right),$$

$$r_B(B|\mathcal{V}^n) = \mathcal{N}\left(B; \pm\mathrm{sign}(\overline{V})\sqrt{\widehat{\gamma}^{\mathrm{VB}}\frac{c_b}{c_a}}, \zeta\frac{c_b}{c_a}\right),$$

$$\zeta = \sqrt{\left(\frac{\widehat{\gamma}^{\mathrm{VB}}}{2} + \frac{\sigma^2}{2nc_ac_b}\right)^2 + \frac{\sigma^2}{n}} - \left(\frac{\widehat{\gamma}^{\mathrm{VB}}}{2} + \frac{\sigma^2}{2nc_ac_b}\right),$$

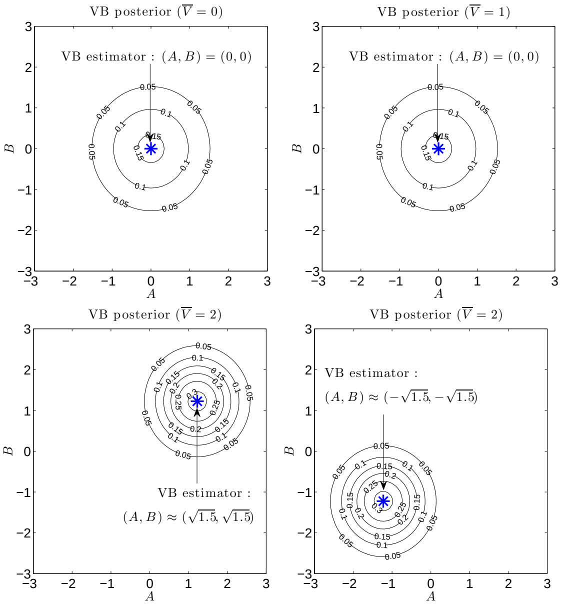

Figure 5. VB posteriors and VB solutions when $L = M = 1$ (i.e., the matrices $\overline{V}$, U, A, and B are scalars). When $\overline{V} = 2$, VB gives either one of the two solutions shown in the bottom row.

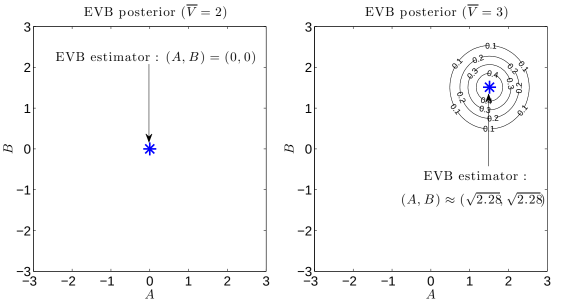

Figure 6. EVB posteriors and EVB solutions when $L = M = 1$. Left: When $\overline{V} = 2$, the EVB posterior is the delta function located at the origin. Right: The solution is detached from the origin when $\overline{V} = 3$.

$$\widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \left[\left(1 - \frac{\sigma^2}{n\overline{V}^2}\right) |\overline{V}| - \frac{\sigma^2}{nc_a c_b} \right]^+ & \text{if } \overline{V} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 5 shows the contour of the VB posteriors $r_A(A|\mathcal{V}^n)$ and $r_B(B|\mathcal{V}^n)$ when $\overline{V} = 0, 1, 2$ are observed, where the number of samples is $n = 1$, the noise variance is $\sigma^2 = 1$, and the hyperparameters are $c_a = c_b = 100$ (i.e., almost flat priors).

When $\overline{V} = 0$, the cross-shaped contour of the Bayes posterior (see Figure 4) is approximated by a spherical Gaussian function located at the origin. Thus, the VB estimator is $\widehat{U}^{\mathrm{VB}} = 0$, which is equivalent to the MAP solution. When $\overline{V} = 1$, two hyperbolic 'modes' of the Bayes posterior are approximated again by a spherical Gaussian function located at the origin. Thus, the VB estimator is still $\widehat{U}^{\mathrm{VB}} = 0$, which is different from the MAP solution.

$\overline{V} = \sqrt{M\sigma^2/n} = 1$ is actually a transition point of the behavior of the VB estimator. When $\overline{V}$ is not larger than the threshold $\sqrt{M\sigma^2/n}$, the VB method tries to approximate the two 'modes' of the Bayes posterior by a single Gaussian located at the origin. When $\overline{V}$ goes beyond the threshold, the 'distance' between two hyperbolic 'modes' of the Bayes posterior becomes so large that the VB method chooses to approximate one of the two modes in the positive and negative orthants. As such, the symmetry is broken spontaneously and the VB solution is detached from the origin. Note that, as discussed in Section 3, $M\sigma^2/n$ amounts to the expected contribution of noise $\mathcal{E}$ to the squared singular value γ^2 ($= \overline{V}^2$ in the current setup).

The bottom row of Figure 5 shows the contour of two possible VB posteriors when $\overline{V} = 2$. Note that, in either case, the VB solution is the same: $\widehat{U}^{\mathrm{VB}} \approx 3/2$. The VB solution is closer to the origin than the MAP solution $\widehat{U}^{\mathrm{MAP}} = 2$, and the difference between the VB and MAP solutions tends to shrink as $\overline{V}$ increases.

4.3. EVBMF

Finally, we illustrate the behavior of the EVB estimator. When $L = M$, the EVB estimators of c_{a_h} and c_{b_h} can be analytically expressed (the details are omitted due to lack of space). Combing the analytic expression and Corollary 1, we can explicitly plot the EVB posterior (Figure 6).

When $\overline{V} = 2 \leq (\sqrt{L} + \sqrt{M})\sigma/\sqrt{n} = 2$, the infimum of the free energy is attained at $\Sigma_a, \Sigma_b, c_a, c_b \to 0$ under $\Sigma_a/c_a = 1$ and $\Sigma_b/c_b = 1$. Thus, the EVB posterior is the delta function located at the origin, and the EVB estimator is $(\widehat{A}^{\mathrm{EVB}}, \widehat{B}^{\mathrm{EVB}}) = (0, 0)$ (see the left graph). On the other hand, when $\overline{V} = 3 \geq \sqrt{7M}\sigma/\sqrt{n} = \sqrt{7} \approx 2.65$, the solution $(\widehat{A}^{\mathrm{EVB}}, \widehat{B}^{\mathrm{EVB}})$ is detached from the origin (see the right graph). Note that the EVB solution is not unique in terms of (A, B) in this case, but is unique in terms of $U = BA$.

The graphs exhibit the stronger shrinkage effect of EVB than VB with the almost flat priors.

5. Conclusion

In this paper, we theoretically analyzed the behavior of Bayesian matrix factorization methods. More specifically, in Section 3, we derived *non-asymptotic* bounds of the *maximum a posteriori matrix factorization* (MAPMF) estimator, the *variational Bayesian* *matrix factorization* (VBMF) estimator, and the *empirical VBMF* (EVBMF) estimator. Then we showed that MAPMF consists of the *trace-norm* shrinkage alone, while VBMF consists of the *positive-part James-Stein* (PJS) shrinkage and the trace-norm shrinkage. An interesting finding was that, while the trace-norm shrinkage does not take effect when the priors are flat, the PJS shrinkage remains activated even with flat priors. We also showed that in EVBMF, the 'strength' of the PJS shrinkage is more than doubled compared with VBMF with the flat prior. We illustrated these facts using one-dimensional examples in Section 4.

The fact that the PJS shrinkage remains activated even with flat priors is induced by the *non-identifiability* of the MF models, where parameters form equivalent classes. Thus, flat priors in the space of factorized matrices are no longer flat in the space of the target (composite) matrix. Furthermore, simple distributions such as the Gaussian distribution in the space of target matrix produce highly complicated *multimodal* distributions in the space of factorized matrices. As Gelman (2004) pointed out, re-parameterization involving modification of conjugate priors affects the behavior of statistical models. Although such re-parameterization is often introduced solely for computational purposes, its role is essential in matrix factorization.

Acknowledgments: The authors thank anonymous reviewers for their helpful comments. MS thanks the support from the FIRST program.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* New York: Wiley. Second edition.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of the Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)* (pp. 21–30).

Baldi, P., & Brunak, S. (1998). *Bioinformatics: The machine learning approach.* Cambridge, MA, USA: MIT Press.

Baldi, P. F., & Hornik, K. (1995). Learning in Linear Neural Networks: A Survey. *IEEE Transactions on Neural Networks*, *6*, 837–858.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York, NY, USA: Springer.

Efron, B., & Morris, C. (1973). Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association*, *68*, 117–130.

Funk, S. (2006). Try this at home. http://sifter.org/~simon/journal/20061211.html.

Gelman, A. (2004). Parameterization and Bayesian Modeling. *Journal of the American Statistical Association*, *99*, 537–545.

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 361–379). Berkeley, CA, USA: University of California Press.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* (pp. 453–461).

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, *40*, 77–87.

Lim, Y. J., & Teh, T. W. (2007). Variational Bayesian Approach to Movie Rating Prediction. *Proceedings of KDD Cup and Workshop.*

Nakajima, S., & Watanabe, S. (2007). Variational Bayes Solution of Linear Neural Networks and its Generalization Performance. *Neural Computation*, *19*, 1112–1153.

Raiko, T., Ilin, A., & Karhunen, J. (2007). Principal component analysis for large sale problems with lots of missing values. *Proceedings of the 18th European conference on Machine Learning* (pp. 691–698).

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York, NY, USA: Wiley.

Reinsel, G. R., & Velu, R. P. (1998). *Multivariate reduced-rank regression: Theory and applications.* New York, NY, USA: Springer.

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection Techniques* (pp. 34–51). Berlin, Germany: Springer.

Salakhutdinov, R., & Mnih, A. (2008). Probabilistic matrix factorization. *Advances in Neural Information Processing Systems 20* (pp. 1257–1264). Cambridge, MA, USA: MIT Press.

Srebro, N., Rennie, J., & Jaakkola, T. (2005). Maximum Margin Matrix Factorization. *Advances in Neural Information Processing Systems 17.*

Watanabe, K., & Watanabe, S. (2006). Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation. *Journal of Machine Learning Research*, *7*, 625–644.

Watanabe, S. (2001). Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, *13*, 899–933.

Watanabe, S. (2009). *Algebraic geometry and statistical learning.* Cambridge, UK: Cambridge University Press.

Yamazaki, K., & Watanabe, S. (2003). Singularities in Mixture Models and Upper Bounds of Stochastic Complexity. *Neural Networks*, *16*, 1029–1038.