Analysis of Variational Bayesian Matrix Factorization

Shinichi Nakajima[1] and Masashi Sugiyama[2]

[1] Nikon Corporation, 1-6-3 Nishi-Ohi, Shinagawa-ku, Tokyo 140-8601, Japan
[2] Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
nakajima.s@nikon.co.jp sugi@cs.titech.ac.jp
http://watanabe-www.pi.titech.ac.jp/~nkj23/index.html

Abstract. Recently, the variational Bayesian approximation was applied to probabilistic matrix factorization and 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 theoretically elucidate properties of a variational Bayesian matrix factorization method. In particular, its mechanism of avoiding overfitting is analyzed. Our analysis relies on the key fact that the matrix factorization model induces non-identifiability, i.e., the mapping between factorized matrices and the original matrix is not one-to-one. The positive-part James-Stein shrinkage operator and the Marcenko-Pastur law—the limiting distribution of eigenvalues of the central Wishart distribution—play important roles in our analysis.

1 Introduction

The problem of estimating a matrix that describes a linear relation between two vectors has been extensively studied by the name of multivariate linear regression with multiple responses, canonical correlation analysis, or reduced rank regression [1]. On the other hand, a recent focus of matrix estimation includes imputation of missing entries of a single matrix, e.g., in the context of microarray data analysis [2] and recommender systems[3] [3, 4]. In this paper, we consider the problem of interpolating missing entries of a matrix.

The paper [5] proposed the weighted low-rank approximation (WLRA) method, which based on the expectation-maximization (EM) algorithm: a matrix is fitted to the data without a rank constraint in the E-step and it is projected back to the set of low-rank matrices by the singular value decomposition (SVD) in the M-step. The web article [4] proposed the regularized SVD method, which minimizes the loss function combined with the Frobenius-norm penalty by gradient descent. If the *trace-norm* penalty is used instead of the Frobenius-norm penalty, a low-rank solution can be obtained without having an explicit low-rank constraint; when the trace-norm penalty is combined with the hinge-loss, a semi-definite programming formulation is obtained [6] (see also [7]

[3] A recommender system is aimed at predicting a preference score of users based on the preference scores of other users. If we consider a matrix where each row corresponds to each user and each column corresponds to each item, the task can be regarded as completing missing entries. This formulation is often referred to as *collaborative filtering*.

for a gradient method with smooth approximation). When the trace-norm penalty and the squared-loss are used, a computationally efficient algorithm is obtained [8].

The above methods, minimizing a loss function combined with a regularizer, could be viewed as Bayesian MAP estimation. On the other hand, it is said that full-Bayesian estimation (considering the entire posterior distribution) is often more accurate than only taking into account the mode of the posterior distribution [9]. However, working with the entire posterior distribution is often computationally intractable, and the variational approximation [10] is known to be a useful approach to coping with this problem. Following this idea, the papers [11, 12] proposed variational Bayesian (VB) approaches to matrix factorization and experimentally showed their superior performance.

In this paper, we try to give a theoretical insight into the experimental facts that the VB approach often has better performance; more specifically, we investigate how the VB method avoids overfitting and why a low-rank solution tends to be produced. We first show that the VB solution can be regarded as a type of the *positive-part James-Stein shrinkage estimator* [13], which is known to *dominate* the least-squares estimator under some conditions. Our analysis strongly relies on the fact that the matrix factorization model induces *non-identifiability*, i.e., the decomposition is redundant and therefore the mapping between factorized matrices and the original matrix is not one-to-one [14, 15]. We then analyze the generalization performance of the VB solution based on the *Marcenko-Pastur law* [16, 17], which elucidates the limiting distribution of eigenvalues of the *central Wishart distribution*.

2 Probabilistic Approach to Matrix Factorization

In this section, we first give a probabilistic formulation of the matrix factorization problem and then review existing approaches.

2.1 Formulation

Let us consider the problem of estimating a target matrix X ($\in \mathbb{R}^{L\times M}$) from its observation Y ($\in \mathbb{R}^{L\times M}$). In the case of recommender systems, the entry $X_{l,m}$ represents the preference score (e.g., one to five stars) given by the l-th user for the m-th item. Assume that the observed matrix Y is subject to the following additive-noise model:

$$Y = X + \mathcal{E}, \tag{1}$$

where $\mathcal{E}$ ($\in \mathbb{R}^{L\times M}$) is a noise matrix. Assuming that each entry of $\mathcal{E}$ is independently subject to the normal distribution with mean zero and variance σ^2, we have the following likelihood:

$$p(Y|X) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|Y - X\|_{\mathrm{Fro}}^2\right), \tag{2}$$

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

If there are missing entries in the observation Y, the likelihood is expressed as

$$p(Y|X) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|W \bullet (Y - X)\|_{\mathrm{Fro}}^2\right), \tag{3}$$

where $\bullet$ denotes the Hadamard product (or the element-wise product) and W is the $L \times M$ matrix with $W_{l,m} = 0$ if $Y_{l,m}$ is a missing entry; otherwise $W_{l,m} = 1$.

Let[4] $H = \min(L, M)$, and let us decompose the matrix X into the product of $A \in \mathbb{R}^{M \times H}$ and $B \in \mathbb{R}^{L \times H}$: $X = BA^\top$, where $\top$ denotes the transpose of a matrix/vector. Then, the likelihood (3) is written as

$$p(Y|A, B) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|W \bullet (Y - BA^\top)\|_{\mathrm{Fro}}^2\right). \tag{4}$$

2.2 Maximum A Posteriori (MAP) Estimation

An estimate of X can be obtained by a procedure similar to the expectation-maximization (EM) algorithm in latent variable models[5]. Let Z ($\in \mathbb{R}^{L \times M}$) be a (latent) complete observation matrix, and let $\phi(X)$ be the prior distribution of X. Then the maximum a posteriori (MAP) solution can be obtained by the following EM algorithm:

E-step: $Z^{(t)} = W \bullet Y + (1 - W) \bullet X^{(t)}$,
M-step: $X^{(t+1)} = \mathrm{argmax}_X\, p(Z^{(t)}|X)\phi(X)$,

where $t = 0, 1, \ldots$ is the iteration number. The M-step corresponds to MAP estimation given a fully observed matrix $Z^{(t)}$, while the E-step updates the latent variable $Z^{(t)}$.

Weighted Low-rank Approximation (WLRA): To avoid overfitting, the paper [5] proposed the WLRA method, which approximates the matrix X with a given rank H' ($\leq H$). This can be regarded as the ℓ_0-norm constraint on the singular values. The WLRA method can be obtained from the following prior distribution on X: $\phi(X) \propto \exp\left(-\frac{1}{c^2}\sum_{h=1}^{H}\theta(\widehat{\gamma}_h > 0)\right)$, where $\theta(\cdot)$ denotes the indicator function, $\widehat{\gamma}_h$ is the h-th largest singular value of X, and c^2 is a constant determined by H'. Then the M-step yields

M-step: $X^{(t+1)} = \sum_{h=1}^{H'} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top$,

where γ_h is the h-th largest singular value of Z, and $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ are the corresponding right and the left singular vectors, respectively. Thus the WLRA algorithm sharply cuts off irrelevant singular values for avoiding overfitting.

Matrix Estimation with Trace-norm Regularization (METR): Another possibility of avoiding overfitting would be regularization—the METR method employs the *trace-norm regularization*, which imposes the ℓ_1-norm constraint on the singular values [8]. METR can be obtained from the following prior distribution on X: $\phi(X) \propto \exp\left(-\frac{1}{c^2}\sum_{h=1}^{H}\widehat{\gamma}_h\right)$, where c^2 is a hyperparameter. The M-step yields

[4] Although we can analyze the case that $H < \min(L, M)$ in a similar way, we assume $H = \min(L, M)$ for the sake of simplicity.

[5] Note that there are other computationally efficient approaches to obtaining a MAP solution. However, the purpose of the review here is not to discuss computational issues, but to compare regularization schemes. For this reason, we focus on the EM algorithm.

M-step: $X^{(t+1)} = \sum_{h=1}^{H} \max\left(0, \gamma_h - \frac{\sigma^2}{c^2}\right) \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top$.

Note that the METR method can also be obtained as MAP estimation when Gaussian priors are assumed on A and B as Eq.(10) [6].

Matrix Estimation with Frobenius Regularization (MEFR): Another regularization approach is to use the *Frobenius regularization*. The MEFR method imposes the ℓ_2-norm constraint on the singular values. MEFR is obtained from the following prior distribution on X: $\phi(X) \propto \exp\left(-\frac{1}{2c^2}\sum_{h=1}^{H} \widehat{\gamma}_h^2\right)$. The M-step yields

M-step: $X^{(t+1)} = \sum_{h=1}^{H} \left(1 - \frac{\sigma^2}{\sigma^2 + c^2}\right) \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top$.

However, the MEFR method is not useful in missing entry completion as it is since it only proportionally shrinks the original matrix and therefore missing values are always zero. Thus the MEFR method should be combined with a low-rank constraint [4].

Maximum-margin Matrix Factorization (MMMF): The paper [6] proposed a matrix factorization method called MMMF, which involves the trace-norm regularization similar to the METR method, but employs the *hinge-loss* inspired by the large-margin principle of support vector machines. For the binary observation Y ($\in \{\pm 1\}^{L \times M}$), the MMMF optimization problem is expressed as

$$\min_X \left[\sum_{h=1}^{H} \widehat{\gamma}_h + \lambda \sum_{l,m} \max(0, 1 - Y_{l,m} X_{l,m})\right],$$

where λ is a regularization parameter and $\sum_{l,m}$ goes over all non-missing entries of Y.

The MMMF method could also be regarded as MAP estimation with the same prior as METR; but the noise model is different from Eq.(2).

2.3 Variational Bayes (VB) Estimation

The papers [11, 12] proposed matrix factorization algorithms based on the VB approach [10] to approximating the posterior $p(A, B|Y)$.

Let $\phi(A)$ and $\phi(B)$ be priors on the factors A and B. Then the posterior distribution of A and B is written as follows:

$$p(A, B|Y) = \frac{\phi(A)\phi(B)p(Y|A,B)}{\int \phi(A)\phi(B)p(Y|A,B)dAdB}. \tag{5}$$

This is the minimizer of the following functional, called the free energy, with respect to $r(A, B)$:

$$F(r|Y) = \int r(A, B) \log \frac{r(A,B)}{p(Y|A,B)\phi(A)\phi(B)} dAdB. \tag{6}$$

The VB approach approximates the posterior $p(A, B|Y)$ within a function class where A and B are independent of each other:

$$r(A, B) = r(A)r(B). \tag{7}$$

Then, using the variational method to minimize Eq.(6), we obtain the following conditions:

$$r(A) \propto \phi(A) \exp\left(\langle \log p(Y|A,B) \rangle_{r(B)}\right), \tag{8}$$

$$r(B) \propto \phi(B) \exp\left(\langle \log p(Y|A,B) \rangle_{r(A)}\right), \tag{9}$$

where $\langle \cdot \rangle_p$ is the expectation over a distribution p. Since $p(Y|A,B)$ is bilinear with respect to A and B (see Eq.(4)), the expectations in Eqs.(8) and (9) can be calculated simply by using the Gaussian integration.

Let us assume the Gaussian priors on the factors A and B:

$$\phi(A) \propto \exp\left(-\tfrac{1}{2c_a^2}\|A\|_{\mathrm{Fro}}^2\right) \quad \text{and} \quad \phi(B) \propto \exp\left(-\tfrac{1}{2c_b^2}\|B\|_{\mathrm{Fro}}^2\right), \tag{10}$$

where c_a^2 and c_b^2 are hyperparameters corresponding to the prior variance of the elements of A and B, respectively. Then the conditions (8) and (9) show that the posterior is also Gaussian. Based on this property, the papers [11, 12] proposed algorithms that iteratively update the mean and the covariance of A and B by Eqs.(8) and (9), respectively. Then the posterior *mean* of $BA^\top$, i.e., $\langle BA^\top \rangle_{r(A,B)}$, is outputted[6] as an estimate of X.

3 Analysis of the VB Approach

VB estimation in general is shown to be a useful alternative to MAP estimation [10, 9], and its good performance has been theoretically investigated in the light of model *non-identifiability*—a statistical model is said to be non-identifiable if the mapping between a parameter value and a probability distribution is not one-to-one [14, 15].

The VB-based matrix factorization methods reviewed in Section 2.3 are shown to work well in experiments [11, 12]. However, their good performance was not completely understood beyond their experimental success. In this section, we theoretically investigate properties of a VB-based matrix factorization method. Note that the factorized matrix model (4) is also non-identifiable since the mapping between (A, B) and X is not one-to-one.

In order to make the analysis feasible, let us consider a variant of VB-based matrix factorization which consists of the following VBEM iterations:

VBE-step: $Z^{(t)} = W \bullet Y + (1 - W) \bullet X^{(t)}$,
VBM-step: $X^{(t+1)} = \langle BA^\top \rangle_{r(A,B|Z^{(t)})}$

3.1 Regularization Properties of VBEM

Here, we investigate the regularization properties of the above VBEM algorithm. Unlike other MAP estimation methods, the VBM-step is not explicitly given. We first show an analytic form of the VBM-step, and then elucidate the regularization mechanism of VBEM.

Note that our analysis below can be regarded as an extension of the paper [15], which analyzes properties of reduced rank regression in *asymptotic* settings. In the current setting of matrix factorization, on the other hand, we need *non-asymptotic* analysis since only one observation matrix is available.

[6] A method to estimate the hyperparameters c_a^2 and c_b^2 has also been proposed.

Analytic Solution of VBM-step: Let

$$A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_H) \quad \text{and} \quad B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_H).$$

Below, we assume as in [12] that $\{\boldsymbol{a}_h\}_{h=1}^H$ and $\{\boldsymbol{b}_h\}_{h=1}^H$ are independent of each other, i.e., we restrict our function class used for approximating the posterior $p(A, B|Z)$ to

$$r(A, B) = \textstyle\prod_{h=1}^H r(\boldsymbol{a}_h) r(\boldsymbol{b}_h). \tag{11}$$

Then, we have a simpler update rule than Eqs.(8) and (9) as follows:

$$r(\boldsymbol{a}_h) \propto \phi(\boldsymbol{a}_h) \exp\left(\langle \log p(Z|A, B)\rangle_{r(A,B)/r(\boldsymbol{a}_h)}\right), \tag{12}$$

$$r(\boldsymbol{b}_h) \propto \phi(\boldsymbol{b}_h) \exp\left(\langle \log p(Z|A, B)\rangle_{r(A,B)/r(\boldsymbol{b}_h)}\right). \tag{13}$$

Substituting Eqs.(2) (with $Y = Z$ and $X = BA^\top$) and (10) into Eqs.(12) and (13), we can express the VB posterior as

$$r(A, B) = \textstyle\prod_{h=1}^H \mathcal{N}_M(\boldsymbol{a}_h; \boldsymbol{\mu}_{a_h}, \Sigma_{a_h}) \cdot \mathcal{N}_L(\boldsymbol{b}_h; \boldsymbol{\mu}_{b_h}, \Sigma_{b_h}), \tag{14}$$

where $\mathcal{N}_d(\cdot; \mu, \Sigma)$ denotes the density of the d-dimensional normal distribution with mean μ and covariance matrix Σ. Note that $\boldsymbol{\mu}_{a_h}$, $\boldsymbol{\mu}_{b_h}$, Σ_{a_h}, and Σ_{b_h} satisfy

$$\boldsymbol{\mu}_{a_h} = \sigma^{-2}\Sigma_{a_h} Z \boldsymbol{\mu}_{b_h}, \qquad \boldsymbol{\mu}_{b_h} = \sigma^{-2}\Sigma_{b_h} Z \boldsymbol{\mu}_{a_h}, \tag{15}$$

$$\Sigma_{a_h} = \sigma^2 \left[(\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})) + \sigma^2 c_a^{-2}\right]^{-1} I_M, \tag{16}$$

$$\Sigma_{b_h} = \sigma^2 \left[(\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h})) + \sigma^2 c_b^{-2}\right]^{-1} I_L, \tag{17}$$

where I_d denotes the d-dimensional identity matrix. Solving the system of equations (15)–(17), we have the following theorem (its proof is omitted due to lack of space):

Theorem 1. *Let $\widehat{X}$ be the VB posterior mean of X, i.e., $\widehat{X} = \langle BA^\top\rangle_{r(A,B)}$. Let*

$$K = \max(L, M).$$

Let γ_h be the h-th largest singular value of Z and let $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ be the corresponding right and the left singular vectors. Then $\widehat{X}$ is analytically given by

$$\widehat{X} = \textstyle\sum_{h=1}^H \widehat{\gamma}_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \quad \textit{where} \quad \widehat{\gamma}_h = \max\left\{0, \left(1 - \frac{K\sigma^2}{\gamma_h^2}\right)\gamma_h - \Delta_h\right\}. \tag{18}$$

Δ_h (≥ 0) in the above is bounded as follows:

$$\Delta_h = \tfrac{\sigma^2}{c_a c_b} \qquad (L = M), \tag{19}$$

$$0 \leq \Delta_h \leq \tfrac{\sigma^2}{c_a c_b}\left(\sqrt{\tfrac{K}{H}} + \tfrac{K}{c_a c_b (K-H) H}\gamma_h\right) \qquad (L \neq M). \tag{20}$$

Furthermore, when $L = M$, the VB posterior is explicitly given by Eq.(14) *with*

$$\boldsymbol{\mu}_{a_h} = \sqrt{\tfrac{c_a}{c_b}\widehat{\gamma}_h}\boldsymbol{\omega}_{a_h}, \qquad \boldsymbol{\mu}_{b_h} = \sqrt{\tfrac{c_b}{c_a}\widehat{\gamma}_h}\boldsymbol{\omega}_{b_h}, \tag{21}$$

$$\Sigma_{a_h} = \tfrac{c_a}{2c_b K}\left\{\sqrt{\left(\widehat{\gamma}_h + \tfrac{\sigma^2}{c_a c_b}\right)^2 + 4\sigma^2 K} - \left(\widehat{\gamma}_h + \tfrac{\sigma^2}{c_a c_b}\right)\right\} I_K, \tag{22}$$

$$\Sigma_{b_h} = \tfrac{c_b}{2c_a K}\left\{\sqrt{\left(\widehat{\gamma}_h + \tfrac{\sigma^2}{c_a c_b}\right)^2 + 4\sigma^2 K} - \left(\widehat{\gamma}_h + \tfrac{\sigma^2}{c_a c_b}\right)\right\} I_K. \tag{23}$$

Regularization Mechanism of VBEM: From Theorem 1, we have the following interpretation.

If $c_a c_b \to \infty$, Δ_h vanishes and the VB estimator is expressed as

$$\widehat{X} = \sum_{h=1}^{H} \max\left\{0, \left(1 - \frac{K\sigma^2}{\gamma_h^2}\right)\gamma_h\right\} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}. \tag{24}$$

Thus, the *positive-part James-Stein (PJS) shrinkage operator* [13] is applied to singular values in a component-wise manner. The PJS estimator has a regularization effect that the estimation variance is reduced by shrinking the estimator (but the bias is increased in turn). It has been proved that the PJS estimator *dominates* the least-squares estimator under some conditions.

If $L = M$ and γ_h is large enough, the VB estimator is expressed as

$$\widehat{X} \approx \sum_{h=1}^{H} \left(\gamma_h - \frac{\sigma^2}{c_a c_b}\right) \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}. \tag{25}$$

Thus, the singular value γ_h is decreased by a constant $\sigma^2/(c_a c_b)$. This may be regarded as a similar effect to the *trace-norm regularization* (i.e., the ℓ_1-norm regularization of singular values; see Section 2.2).

If γ_h is large enough and $\Delta_h \approx c\gamma_h$ $(0 \leq c \leq 1)$, the VB estimator is expressed as

$$\widehat{X} \approx \sum_{h=1}^{H} (1-c)\gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}. \tag{26}$$

Thus, the singular value γ_h is *shrunk* proportionally. This may be regarded as a similar effect to the *Frobenius-norm regularization* (i.e., the ℓ_2-norm regularization of singular values; see Section 2.2).

Thus, VBEM regularizes the solution based on the combination of PJS shrinkage, trace-norm regularization, and (possibly) Frobenius-norm regularization.

Posterior Mode, Posterior Mean, and Model Non-identifiability: When the uniform prior (i.e., $c_a^2, c_b^2 \to \infty$) is used and the rank of $\widehat{X}$ is not reduced, one may intuitively think that no reguralization mechanism is involved. This intuition is true when MAP estimation is used—MAP estimation merely results in maximum likelihood (ML) estimation which has no regularization effect. However, in VBEM, the solution involves the PJS-type regularization (see Eq.(24)) and therefore overfitting can be avoided even when the uniform prior is used without rank constraints. This argument is in good agreement with the exprerimental results reported in the paper [12].

Based on Theorem 1, we explain the reason for the significant difference between MAP and VB. For illustration purposes, let us start from the simplest case where $L = M = 1$ (i.e., X is a scalar) and the noise variance is $\sigma^2 = 1$. The top graphs in Fig.1 shows the contours of the Bayes posterior with the uniform prior on A and B when $Z = 0, 1, 2$ is observed (the horizontal and vertical axes correspond to A and B, respectively); the MAP estimators are indicated by the dashed curves (all points on the curves are the MAP estimators, which give the same solution $\widehat{X}$). In the bottom graphs of Fig.1, the VB posteriors—which are independent Gaussians—are plotted for $Z = 0, 1, 2$. The asterisks indicate their expectations, i.e., the VB estimators. When $Z = 0$, the MAP and the VB estimators both give the same value $\widehat{X} = \widehat{B}\widehat{A} = 0$. When

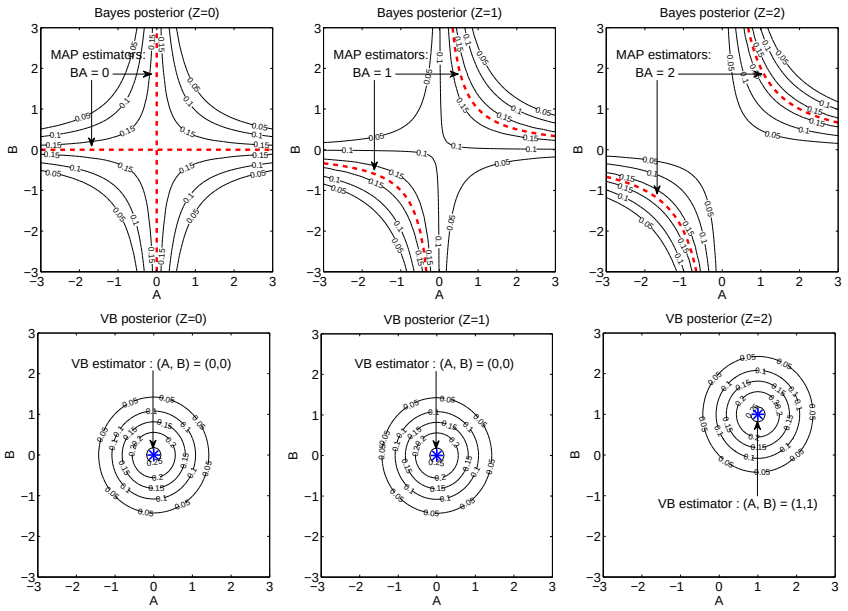


Fig. 1. MAP and VB solutions when $L = M = 1$ (the matrices X, A, and B are scalars).

$Z = 1$, the MAP estimator gives $\widehat{X} = 1$, while the VB estimator is still $\widehat{X} = 0$. When $Z = 2$, the VB estimator is off the origin ($\widehat{X} = 1$), but is still closer to the origin (i.e., strongly regularized) than the MAP solution $\widehat{X} = 2$. More generally, as the observed value Z is increased, the VB estimator approaches to the MAP estimator. However, the VB solution is always closer to the origin than the MAP solution. Note that $(-1, -1)$ is another VB solution when $Z = 2$, although only one VB solution at $(1, 1)$ is depicted in the figure for clear visibility.

The above analysis shows that even with the same uniform prior on A and B, MAP and VB give different solutions—the VB solution tends to be more strongly regularized than the MAP solution. We focused on $L = M = 1$ and $c_a^2, c_b^2 \to \infty$ in the above analysis for illustration purposes. But from Eqs.(18)–(20) we see that the amplitude of each component of the VB estimator is no larger than that of the PJS estimator (24) for any L, M, and $c_a^2, c_b^2 \geq 0$. This means that the VB solution always tends to be more strongly regularized than the MAP solution.

3.2 Generalization Properties of VBEM

Here, we investigate the generalization properties of the VBEM algorithm. First the generalization error of an estimated matrix is defined and it is decomposed into the 'necessary' part and the 'redundant' part. We then elucidate properties of the redundant part, in particular, sparseness of the solution and the generalization performance.

Generalization Error of VBEM: Our analysis is based on the assumption that the fully observed matrix Z is subject to the true distribution $p(Z|X^*)$, which is of the

form (2) with the true matrix X^* and Z substituted for Y. Let H^* be the rank of X^* and assume

$$H^* \ll H.$$

This would be acceptable in, e.g., collaborative filtering since H is typically very large.

Let us measure the generalization error of the VB solution $\widehat{X}$ by the average Kullback-Leibler divergence from the true distribution to the estimated distribution:

$$G = \left\langle \log \frac{p(Z|X^*)}{p(Z|\widehat{X})} \right\rangle_{p(Z|X^*)} = \frac{1}{2\sigma^2} \left\langle \|\widehat{X} - X^*\|_{\mathrm{Fro}}^2 \right\rangle_{p(Z|X^*)}. \tag{27}$$

Let $\mathcal{W}_d(m, \Sigma, \Lambda)$ be the d-dimensional Wishart distribution with m degrees of freedom, scale matrix Σ, and non-centrality matrix Λ. Then, it is easy to show that $ZZ^\top$ follows the non-central Wishart distribution:

$$ZZ^\top \sim \mathcal{W}_H(K, \sigma^2 I_H, X^* X^{*\top}) \quad \text{if } L \leq M.$$

If $L > M$, we may simply re-define $X^\top$ as X so that $L \leq M$ holds.

By assumption, X^* consists of only H^* singular components. Let us decompose $\widehat{X}$ into the component projected onto the space spanned by X^* (the 'necessary' part) and its complement (the 'redundant' part):

$$\widehat{X} = \widehat{X}_{\mathrm{nec}} + \widehat{X}_{\mathrm{red}}.$$

Then, Eq.(27) implies that the generalization error can be decomposed as

$$G = G_{\mathrm{nec}} + G_{\mathrm{red}},$$
$$\text{where} \quad G_{\mathrm{nec}} = \left\langle \|\widehat{X}_{\mathrm{nec}} - X^*\|_{\mathrm{Fro}}^2 \right\rangle_{p(Z|X^*)} \quad \text{and} \quad G_{\mathrm{red}} = \left\langle \|\widehat{X}_{\mathrm{red}}\|_{\mathrm{Fro}}^2 \right\rangle_{p(Z|X^*)}.$$

Since $H^* \ll H$ by assumption, the contribution of the necessary components would be negligibly small compared with the contribution of the redundant components. Based on this reasoning, we focus on G_{red} in the following analysis.

Analysis of Eigenvalue Distribution of Redundant Components: Since the Gaussian noise is invariant under rotation, G_{red} can be expressed without loss of generality as

$$G_{\mathrm{red}} = \left\langle \|\widehat{X}(R)\|_{\mathrm{Fro}}^2 \right\rangle_{\mathcal{N}(R)}, \tag{28}$$

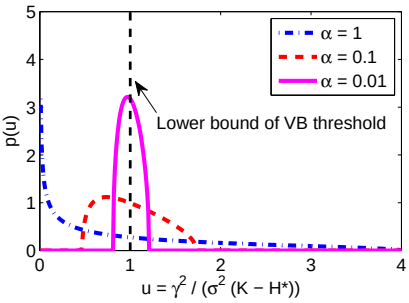
where $\widehat{X}(R)$ denotes the VB estimator given observation R, which is a $(H - H^*) \times (K - H^*)$ random matrix with entries independently subject to the normal distribution (denoted by $\mathcal{N}(R)$) with mean zero and variance σ^2. $RR^\top$ follows the central Wishart distribution:

$$RR^\top \sim \mathcal{W}_{H-H^*}(K - H^*, \sigma^2 I_{H-H^*}).$$

Let $u_1, u_2, \ldots, u_{H-H^*}$ be the eigenvalues of $\frac{1}{\sigma^2(K-H^*)} RR^\top$, and define the empirical distribution of the eigenvalues by

$$p(u) = \frac{\delta(u_1) + \delta(u_2) + \cdots + \delta(u_{H-H^*})}{H - H^*},$$

Fig. 2. Normalized eigenvalue (squared singular value of R) distribution of a large-scale Wishart matrix. Singular values smaller than the threshold are eliminated by the PJS operator in VBEM. We can show that the median of γ^2 is less than $\sigma^2(K-H^*)$ for any α. On the other hand, the VB threshold is no less than $\sigma^2 K$, which is always larger than the median. Thus, at least 50% of singular values are zero in VBEM (see Theorem 2 and Fig.3 for detail).



where $\delta(u)$ denotes the Dirac measure at u. Let α be the 'squareness' index of the target matrix X defined by

$$\alpha = \frac{H-H^*}{K-H^*},$$

which satisfies $0 < \alpha \leq 1$. Then, the following proposition is known regarding the distribution of eigenvalues of the central Wishart distribution.

Proposition 1 (Marcenko-Pastur law). *[16, 17] In the large-scale limit where K, H, and H^* go to infinity in the same order, the probability measure of the empirical distribution of the eigenvalue u of $\frac{1}{\sigma^2(K-H^*)}RR^\top$ converges almost surely to*

$$p(u)du = \frac{\sqrt{(u-\underline{u})(\overline{u}-u)}}{2\pi\alpha u}\theta(\underline{u} < u < \overline{u})du, \tag{29}$$

where $\underline{u} = (\sqrt{\alpha}-1)^2$, $\overline{u} = (\sqrt{\alpha}+1)^2$, and $\theta(\cdot)$ denotes the indicator function.

Fig.2 depicts the eigenvalue distribution of a large-scale Wishart matrix for $\alpha = 0.01, 0.1, 1$, where the eigenvalues (or the squared singular values of R) are normalized by $\sigma^2(K-H^*)$ in the graph for better comparison.

Remember that the VB estimator $\widehat{X}(R)$ eliminates the singular values (of R) smaller than a certain positive value, which we call the *VB threshold*. When $c_a c_b \to \infty$, $\gamma^2 = K\sigma^2$ is the VB threshold (see Eq.(24)). Since $H^* \geq 0$, we have

$$\sigma^2 K \geq \sigma^2(K-H^*),$$

which corresponds to a lower bound of the VB threshold for any $H^*, c_a^2, c_b^2 \geq 0$ (see Eq.(18)). In Fig.2, eigenvalues smaller than this threshold (which is normalized to one in the figure) are discarded.

Analysis of Sparseness of Redundant Components: We can evaluate the proportion of the singular values larger than the VB threshold as follows. Let

$$J_k(u_0) = 2\pi\alpha \int_{u_0}^{\overline{u}} u^k p(u)du. \tag{30}$$

Note that $(2\pi\alpha)^{-1}J_k(\underline{u})$ corresponds to the k-th moment of the Marcenko-Pastur distribution. $J_k(u_0)$ for $k = -1, 0, 1$ has analytic forms as follows:

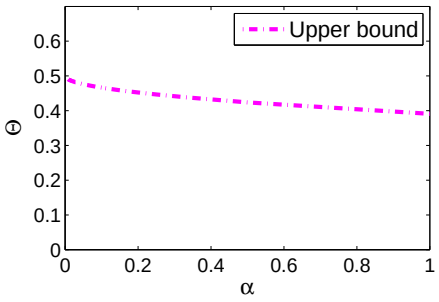


Fig. 3. The proportion Θ of non-zero singular values. Θ is always below 0.5 and it converges to 0.5 as $\alpha \to 0, c_a c_b \to \infty$ and $H^* = 0$.

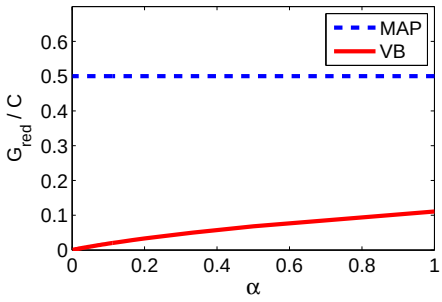


Fig. 4. The behaviors of $G_{\rm red}$ for MAP estimation and VB estimation, when $c_a c_b \to \infty$ and $H^* = 0$ (the values for VB becomes smaller when $H^* > 0$).

Proposition 2. *[15] $J_k(u_0)$ has the following analytic forms for $k = -1, 0, 1$.*

$$J_{-1}(u_0) = \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1} s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0<\alpha<1), \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1), \end{cases}$$

$$J_0(u_0) = -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1} s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)},$$

$$J_1(u_0) = 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s), \quad \textit{where} \quad s = \frac{u_0-(1+\alpha)}{2\sqrt{\alpha}}.$$

This proposition enables us to calculate the proportion of nonzero redundant components as shown in the following theorem:

Theorem 2. *Let Θ be the proportion of nonzero redundant components in the large scale limit. Then, its upper bound is given by*

$$\Theta \leq (2\pi\alpha)^{-1} J_0(\kappa), \quad \textit{where} \quad \kappa = \frac{K}{K-H^*}. \tag{31}$$

The equality holds when $c_a c_b \to \infty$.

This theorem implies that VBEM gives a low-rank solution without explicit rank restriction. The curve in Fig.3 shows the value of $(2\pi\alpha)^{-1}J_0(1)$, which is the upper bound of Θ for any H^*, c_a^2, c_b^2. This value is always below 0.5, which means that at least 50% of singular values always become zero in VBEM; in practice the solution would be even more sparser.

Analysis of Redundant-component Generalization Error: Next, we obtain the following theorem which enables us to evaluate the value of $G_{\rm red}$:

Theorem 3. *The upper bound of the contribution of the redundant components to the generalization error in the large-scale limit is given by*

$$G_{\rm red} \leq \frac{C\{J_1(\kappa) - 2\kappa J_0(\kappa) + \kappa^2 J_{-1}(\kappa)\}}{4\pi\alpha}, \quad \textit{where} \quad C = (K-H^*)(H-H^*). \tag{32}$$

The equality holds when $c_a c_b \to \infty$.

Based on the above theorem and Proposition 2, we can compute the value of $G_{\rm red}$ analytically. In Fig.4, $G_{\rm red}/C$ for VBEM estimation and MAP estimation (which is equivalent to ML estimation due to the flat prior) are depicted. $G_{\rm red}/C$ for MAP is independent of α and is equal to 0.5. On the other hand, $G_{\rm red}/C$ for VBEM is increasing with respect to α, but is always much smaller than that of MAP. This implies that VBEM is highly robust against large observation noise.

4 Conclusions

In this paper, we have analyzed a variational Bayesian expectation-maximization (VBEM) method of matrix factorization. In particular, we elucidated the mechanism of inducing a low-rank solution and avoiding overfitting, where the principle of the positive-part James-Stein shrinkage operator and the Marcenko-Pastur law played important roles in the analysis.

Future work is to explicitly treat the missing values in the VBEM procedure, and to directly analyze the generalization error including the 'necessary' part (see Section 3.2).

References

1. Baldi, P.F., Hornik, K.: Learning in Linear Neural Networks: a Survey. IEEE Trans. on Neural Networks **6** (1995) 837–858
2. Baldi, P., Brunak, S.: Bioinformatics. MIT Press (2001)
3. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: Grouplens: applying collaborative filtering to usenet news. Commun. ACM **40** (1997) 77–87
4. Funk, S.: Try this at home. http://sifter.org/˜simon/journal/20061211.html (2006)
5. Srebro, N., Jaakkola, T.: Weighted Low Rank Approximation. In: Proc. of ICML. (2003)
6. Srebro, N., Rennie, J., Jaakkola, T.: Maximum Margin Matrix Factorization. In: Advances in NIPS. Volume 17. (2005)
7. Rennie, J.D.M., Srebro, N.: Fast Maximum Margin Matrix Factorization for Collaborative Prediction. In: Proc. of ICML. (2005)
8. Salakhutdinov, R., Mnih, A.: Probabilistic Matrix Factorization. In: Advances in NIPS. Volume 20. (2008)
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
10. Attias, H.: Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In: Proc. of UAI. (1999)
11. Lim, Y.J., Teh, T.W.: Variational Bayesian Approach to Movie Rating Prediction. In: Proc. of KDD Cup and Workshop. (2007)
12. Raiko, T., Ilin, A., Karhunen, J.: Principal Component Analysis for Large Sale Problems with Lots of Missing Values. In: Proc. of ECML. Volume LNAI 4701. (2007) 691–698
13. James, W., Stein, C.: Estimation with Quadratic Loss. In: Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob. (1961) 361–379
14. Watanabe, S.: Algebraic Analysis for Nonidentifiable Learning Machines. Neural Computation **13** (2001) 899–933
15. Nakajima, S., Watanabe, S.: Variational Bayes Solution of Linear Neural Networks and its Generalization Performance. Neural Computation **19** (2007) 1112–1153
16. Marcenko, V.A., Pastur, L.A.: Distribution of Eigenvalues for Some Sets of Random Matrices. Mathematics of the USSR-Sbornik **1** (1967) 457–483
17. Wachter, K.W.: The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements. Annals of Probability **6** (1978) 1–18