Journal of Machine Learning Research 14 (2013) 1-37 Submitted 6/12; Revised 11/12; Published 1/13

Global Analytic Solution of Fully-observed Variational Bayesian Matrix Factorization*

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

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
Tokyo 152-8552, Japan

S. Derin Babacan DBABACAN@ILLINOIS.EDU
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA

Ryota Tomioka TOMIOKA@MIST.I.U-TOKYO.AC.JP
Department of Mathematical Informatics
The University of Tokyo
Tokyo 113-8656, Japan

Editor: Manfred Opper

Abstract

The variational Bayesian (VB) approximation is known to be a promising approach to Bayesian estimation, when the rigorous calculation of the Bayes posterior is intractable. The VB approximation has been successfully applied to *matrix factorization* (MF), offering automatic dimensionality selection for principal component analysis. Generally, finding the VB solution is a non-convex problem, and most methods rely on a local search algorithm derived through a standard procedure for the VB approximation. In this paper, we show that a better option is available for fully-observed VBMF—the global solution can be *analytically* computed. More specifically, the global solution is a reweighted SVD of the observed matrix, and each weight can be obtained by solving a quartic equation with its coefficients being functions of the observed singular value. We further show that the global optimal solution of *empirical* VBMF (where hyperparameters are also learned from data) can also be analytically computed. We illustrate the usefulness of our results through experiments in multi-variate analysis.

Keywords: variational Bayes, matrix factorization, empirical Bayes, model-induced regularization, probabilistic PCA

1. Introduction

The problem of finding a low-rank approximation of a target matrix through *matrix factorization* (MF) recently attracted considerable attention. In this paper, we consider *fully-observed MF* where

[*] This paper is a combined and extended version of our earlier conference papers (Nakajima et al., 2010, 2011).

the observed matrix has no missing entry.[1] This formulation includes multivariate analysis techniques such as *principal component analysis* (Hotelling, 1933) and *reduced rank regression* (Reinsel and Velu, 1998). *Canonical correlation analysis* (Hotelling, 1936; Anderson, 1984; Hardoon et al., 2004) and *partial least-squares* (Worsley et al., 1997; Rosipal and Krämer, 2006) are also closely related to MF.

Singular value decomposition (SVD) is a classical method for MF, which gives the optimal low-rank approximation to the target matrix in terms of the squared error. Regularized variants of SVD have been studied for the *Frobenius-norm* penalty (i.e., singular values are regularized by the ℓ_2-penalty) (Paterek, 2007) or the *trace-norm* penalty (i.e., singular values are regularized by the ℓ_1-penalty) (Srebro et al., 2005). Since the Frobenius-norm penalty does not automatically produce a low-rank solution, it should be combined with an explicit low-rank constraint, which is non-convex. In contrast, the trace-norm penalty tends to produce sparse solutions, so a low-rank solution can be obtained without explicit rank constraints. This implies that the optimization problem of trace-norm MF is still convex, and thus the global optimal solution can be obtained. Recently, optimization techniques for trace-norm MF have been extensively studied (Rennie and Srebro, 2005; Cai et al., 2010; Ji and Ye, 2009; Tomioka et al., 2010).

Bayesian approaches to MF have also been actively explored. A *maximum a posteriori* (MAP) estimation, which computes the mode of the posterior distributions, was shown (Srebro et al., 2005) to be equivalent to the ℓ_1-MF when Gaussian priors are imposed on factorized matrices (Salakhutdinov and Mnih, 2008). The *variational Bayesian* (VB) method (Attias, 1999; Bishop, 2006), which approximates the posterior distributions by decomposable distributions, has also been applied to MF (Bishop, 1999; Lim and Teh, 2007; Ilin and Raiko, 2010). The VB-based MF method (VBMF) was shown to perform well in experiments, and its theoretical properties have been investigated (Nakajima and Sugiyama, 2011).

However, the optimization problem of VBMF is non-convex. In practice, the VBMF solution is computed by the *iterated conditional modes* (ICM) algorithm (Besag, 1986; Bishop, 2006), where the mean and the covariance of the posterior distributions are iteratively updated until convergence (Lim and Teh, 2007; Ilin and Raiko, 2010). One may obtain a local optimal solution by the ICM algorithm, but many restarts would be necessary to find a good local optimum.

In this paper, we show that, despite the non-convexity of the optimization problem, the global optimal solution of VBMF can be *analytically* computed. More specifically, the global solution is a reweighted SVD of the observed matrix, and each weight can be obtained by solving a quartic equation with its coefficients being functions of the observed singular value. This is highly advantageous over the standard ICM algorithm since the global optimum can be found without any iterations and restarts. We also consider an *empirical* VB scenario where the hyperparameters (prior variances) are also learned from data. Again, the optimization problem of empirical VBMF is non-convex, but we show that the global optimal solution of empirical VBMF can still be analytically computed. The usefulness of our results is demonstrated through experiments.

Our analysis can be seen as an extension of Nakajima and Sugiyama (2011). The major progress is twofold:

1. Weakened decomposability assumption.

[1] This excludes the *collaborative filtering* setup, which is aimed at imputing missing entries of an observed matrix (Konstan et al., 1997; Funk, 2006).

Nakajima and Sugiyama (2011) analyzed the behavior of VBMF under the *column-wise* independence assumption (Ilin and Raiko, 2010), that is, the columns of the factorized matrices are forced to be mutually independent in the VB posterior. This was one of the limitations of the previous work, since the weaker *matrix-wise* independence assumption (Lim and Teh, 2007) is rather standard, and sufficient to derive the ICM algorithm. It was not clear how these different assumptions affect the approximation accuracy to the Bayes posterior. In this paper, we show that the VB solution under the matrix-wise independence assumption is column-wise independent, meaning that the stronger column-wise independence assumption does not degrade the quality of approximation accuracy.

2. Exact analysis for rectangular cases.

   Nakajima and Sugiyama (2011) derived bounds of the VBMF solution (more specifically, bounds of the weights for the reweighed SVD). Those bounds are tight enough to give the exact analytic solution only when the observed matrix is square. In this paper, we conduct a more precise analysis, which results in a quartic equation with its coefficients depending on the observed singular value. Satisfying this quartic equation is a necessary condition for the weight, and further consideration specifies which of the four solutions is the VBMF solution.

In summary, we derive the exact global analytic solution for general rectangular cases under the standard matrix-wise independence assumption.

The rest of this paper is organized as follows. We first introduce the framework of Bayesian matrix factorization and the variational Bayesian approximation in Section 2. Then, we analyze the VB free energy, and derive the global analytic solution in Section 3. Section 4 is devoted to explaining the relation between MF and multivariate analysis techniques. In Section 5, we show practical usefulness of our analytic-form solutions through experiments. In Section 6, we derive simple analytic-form solutions for special cases, discuss the relation between model pruning and spontaneous symmetry breaking, and consider the possibility of extending our results to more general problems. Finally, we conclude in Section 7.

2. Formulation

In this section, we first formulate the problem of probabilistic MF (Section 2.1). Then, we introduce the VB approximation (Section 2.2) and its empirical variant (Section 2.3). We also introduce a simplified variant (Section 2.4), which was analyzed in Nakajima and Sugiyama (2011) and will be shown to be equivalent to the (non-simple) VB approximation in the subsequent section.

2.1 Probabilistic Matrix Factorization

Assume that we have an observation matrix $V \in \mathbb{R}^{L \times M}$, which is the sum of a target matrix $U \in \mathbb{R}^{L \times M}$ and a noise matrix $\mathcal{E} \in \mathbb{R}^{L \times M}$:

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

In the *matrix factorization* model, the target matrix is assumed to be low rank, and expressed in the following factorized form:

$$U = BA^{\top},$$

where $A \in \mathbb{R}^{M\times H}$ and $B \in \mathbb{R}^{L\times H}$. Here, $\top$ denotes the transpose of a matrix or vector. Thus, the rank of U is upper-bounded by $H \leq \min(L,M)$.

We consider the Gaussian probabilistic MF model (Salakhutdinov and Mnih, 2008), given as follows:

$$p(V|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}\|V - BA^\top\|^2_{\mathrm{Fro}}\right), \tag{1}$$

$$p(A) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(AC_A^{-1}A^\top\right)\right), \tag{2}$$

$$p(B) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(BC_B^{-1}B^\top\right)\right), \tag{3}$$

where σ^2 is the noise variance. Here, we denote by $\|\cdot\|_{\mathrm{Fro}}$ the Frobenius norm, and by $\mathrm{tr}(\cdot)$ the trace of a matrix. We assume that $L \leq M$. If $L > M$, we may simply re-define the transpose $V^\top$ as V so that $L \leq M$ holds. Thus this does not impose any restriction. We assume that the prior covariance matrices C_A and C_B are diagonal and positive definite, that is,

$$C_A = \mathrm{diag}(c^2_{a_1}, \ldots, c^2_{a_H}),$$
$$C_B = \mathrm{diag}(c^2_{b_1}, \ldots, c^2_{b_H}),$$

for $c_{a_h}, c_{b_h} > 0, h = 1, \ldots, H$. Without loss of generality, we assume that the diagonal entries of the product $C_A C_B$ are arranged in the non-increasing order, that is, $c_{a_h}c_{b_h} \geq c_{a_{h'}}c_{b_{h'}}$ for any pair $h < h'$.

Throughout the paper, we denote a column vector of a matrix by a bold small letter, and a row vector by a bold small letter with a tilde, namely,

$$A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_H) = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M)^\top \in \mathbb{R}^{M\times H},$$
$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_H) = \left(\widetilde{\boldsymbol{b}}_1, \ldots, \widetilde{\boldsymbol{b}}_L\right)^\top \in \mathbb{R}^{L\times H}.$$

2.2 Variational Bayesian Approximation

The Bayes posterior is written as

$$p(A,B|V) = \frac{p(V|A,B)p(A)p(B)}{p(V)}, \tag{4}$$

where $p(V) = \langle p(V|A,B)\rangle_{p(A)p(B)}$ is the marginal likelihood. Here, $\langle\cdot\rangle_p$ denotes the expectation over the distribution p. Since the Bayes posterior (4) is computationally intractable, the VB approximation was proposed (Bishop, 1999; Lim and Teh, 2007; Ilin and Raiko, 2010).

Let $r(A,B)$, or r for short, be a trial distribution. The following functional with respect to r is called the free energy:

$$\begin{aligned} F(r|V) &= \left\langle \log \frac{r(A,B)}{p(V|A,B)p(A)p(B)} \right\rangle_{r(A,B)} \\ &= \left\langle \log \frac{r(A,B)}{p(A,B|V)} \right\rangle_{r(A,B)} - \log p(V). \end{aligned} \tag{5}$$

The first term in Equation (5) is the Kullback-Leibler (KL) distance from the trial distribution to the Bayes posterior, and the second term is a constant. Therefore, minimizing the free energy (5) amounts to finding the distribution closest to the Bayes posterior in the sense of the KL distance. In the VB approximation, the free energy (5) is minimized over some restricted function space.

A standard constraint for the MF model is *matrix-wise* independence (Bishop, 1999; Lim and Teh, 2007), that is,

$$r^{\mathrm{VB}}(A,B) = r_{\mathrm{A}}^{\mathrm{VB}}(A) r_{\mathrm{B}}^{\mathrm{VB}}(B). \tag{6}$$

This constraint breaks the entanglement between the parameter matrices A and B, and leads to a computationally-tractable iterative algorithm, called the *iterated conditional modes* (ICM) algorithm (Besag, 1986; Bishop, 2006). The resulting distribution is called the *VB posterior*.

Using the variational method, we can show that the VB posterior minimizing the free energy (5) under the constraint (6) can be written as

$$r^{\mathrm{VB}}(A,B) = \prod_{m=1}^{M} \mathcal{N}_H(\widetilde{\boldsymbol{a}}_m; \widehat{\widetilde{\boldsymbol{a}}}_m, \Sigma_A) \prod_{l=1}^{L} \mathcal{N}_H(\widetilde{\boldsymbol{b}}_l; \widehat{\widetilde{\boldsymbol{b}}}_l, \Sigma_B), \tag{7}$$

where the parameters satisfy

$$\widehat{A} = \left(\widehat{\widetilde{\boldsymbol{a}}}_1, \ldots, \widehat{\widetilde{\boldsymbol{a}}}_M\right)^{\top} = V^{\top} \widehat{B} \frac{\Sigma_A}{\sigma^2}, \tag{8}$$

$$\widehat{B} = \left(\widehat{\widetilde{\boldsymbol{b}}}_1, \ldots, \widehat{\widetilde{\boldsymbol{b}}}_L\right)^{\top} = V \widehat{A} \frac{\Sigma_B}{\sigma^2}, \tag{9}$$

$$\Sigma_A = \sigma^2 \left(\widehat{B}^{\top} \widehat{B} + L\Sigma_B + \sigma^2 C_A^{-1}\right)^{-1}, \tag{10}$$

$$\Sigma_B = \sigma^2 \left(\widehat{A}^{\top} \widehat{A} + M\Sigma_A + \sigma^2 C_B^{-1}\right)^{-1}. \tag{11}$$

Here, $\mathcal{N}_d(\cdot; \boldsymbol{\mu}, \Sigma)$ denotes the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Note that, in the VB posterior (7), the rows $\{\widetilde{\boldsymbol{a}}_m\}$ ($\{\widetilde{\boldsymbol{b}}_l\}$) of A (B) are independent of each other, and share a common covariance Σ_A (Σ_B) (Bishop, 1999).

The ICM for VBMF iteratively updates the parameters $(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B)$ by Equations (8)–(11) until convergence, allowing one to obtain a local minimum of the free energy (5). Finally, the VB estimator of U is computed as

$$\widehat{U}^{\mathrm{VB}} = \widehat{B}\widehat{A}^{\top}.$$

2.3 Empirical VB Approximation

The free energy minimization principle also allows us to estimate the hyperparameters C_A and C_B from data. This is called the *empirical* Bayesian scenario. In this scenario, C_A and C_B are updated in each iteration by the following formulas:

$$c_{a_h}^2 = \|\widehat{\boldsymbol{a}}_h\|^2/M + (\Sigma_A)_{hh}, \tag{12}$$

$$c_{b_h}^2 = \|\widehat{\boldsymbol{b}}_h\|^2/L + (\Sigma_B)_{hh}. \tag{13}$$

When the noise variance σ^2 is unknown, it can also be estimated based on the free energy minimization. The update rule for σ^2 is given by

$$\sigma^2 = \frac{\|V\|_{\mathrm{Fro}}^2 - \mathrm{tr}(2V^\top \widehat{B}\widehat{A}^\top) + \mathrm{tr}\left((\widehat{A}^\top \widehat{A} + M\Sigma_A)(\widehat{B}^\top \widehat{B} + L\Sigma_B)\right)}{LM}, \tag{14}$$

which should be applied in each iteration of the ICM algorithm.

2.4 SimpleVB Approximation

A simplified variant, called the SimpleVB approximation, assumes *column-wise* independence of each matrix (Ilin and Raiko, 2010; Nakajima and Sugiyama, 2011), that is,

$$r^{\mathrm{SVB}}(A,B) = \prod_{h=1}^{H} r_{a_h}^{\mathrm{SVB}}(\boldsymbol{a}_h) \prod_{h=1}^{H} r_{b_h}^{\mathrm{SVB}}(\boldsymbol{b}_h). \tag{15}$$

Note that the *column-wise* independence constraint (15) is stronger than the *matrix-wise* independence constraint (6), that is, any column-wise independent distribution is matrix-wise independent.

The SimpleVB posterior can be written as

$$r^{\mathrm{SVB}}(A,B) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \widehat{\boldsymbol{a}}_h^{\mathrm{SVB}}, \sigma_{a_h}^{2\,\mathrm{SVB}} I_M) \prod_{h=1}^{H} \mathcal{N}_L(\boldsymbol{b}_h; \widehat{\boldsymbol{b}}_h^{\mathrm{SVB}}, \sigma_{b_h}^{2\,\mathrm{SVB}} I_L),$$

where the parameters satisfy

$$\widehat{\boldsymbol{a}}_h^{\mathrm{SVB}} = \frac{\sigma_{a_h}^{2\,\mathrm{SVB}}}{\sigma^2} \left(V - \sum_{h' \neq h} \widehat{\boldsymbol{b}}_{h'}^{\mathrm{SVB}} \widehat{\boldsymbol{a}}_{h'}^{\mathrm{SVB}\top} \right)^\top \widehat{\boldsymbol{b}}_h^{\mathrm{SVB}}, \tag{16}$$

$$\widehat{\boldsymbol{b}}_h^{\mathrm{SVB}} = \frac{\sigma_{b_h}^{2\,\mathrm{SVB}}}{\sigma^2} \left(V - \sum_{h' \neq h} \widehat{\boldsymbol{b}}_{h'}^{\mathrm{SVB}} \widehat{\boldsymbol{a}}_{h'}^{\mathrm{SVB}\top} \right) \widehat{\boldsymbol{a}}_h^{\mathrm{SVB}}, \tag{17}$$

$$\sigma_{a_h}^{2\,\mathrm{SVB}} = \sigma^2 \left(\|\widehat{\boldsymbol{b}}_h^{\mathrm{SVB}}\|^2 + L\sigma_{b_h}^{2\,\mathrm{SVB}} + \sigma^2 c_{a_h}^{-2} \right)^{-1}, \tag{18}$$

$$\sigma_{b_h}^{2\,\mathrm{SVB}} = \sigma^2 \left(\|\widehat{\boldsymbol{a}}_h^{\mathrm{SVB}}\|^2 + M\sigma_{a_h}^{2\,\mathrm{SVB}} + \sigma^2 c_{b_h}^{-2} \right)^{-1}. \tag{19}$$

Here, I_d denotes the d-dimensional identity matrix. The constraint (15) restricts the covariances Σ_A and Σ_B in Equation (7) to be diagonal, and thus reduces necessary memory storage and computational cost (Ilin and Raiko, 2010).

Iterating Equations (16)–(19) until convergence, we can obtain a local minimum of the free energy. Equations (14), (12), and (13) are similarly applied if the noise variance σ^2 is unknown and in the empirical Bayesian scenario, respectively.

The column-wise independence (15) also simplifies theoretical analysis. Thanks to this simplification, Nakajima and Sugiyama (2011) showed that the SimpleVBMF solution is a reweighted SVD, and successfully derived theoretical bounds of the weights. Their analysis revealed interesting properties of VBMF, called *model-induced regularization*. However, it has not been clear how restrictive the column-wise independence assumption is. In Section 3, we theoretically show that the column-wise independence assumption has actually no effect, before deriving the exact global analytic solution.

3. Theoretical Analysis

In this section, we first prove the equivalence between VBMF and SimpleVBMF (Section 3.1). After that, starting from a proposition given in Nakajima and Sugiyama (2011), we derive the global analytic solution for VBMF (Section 3.2). Finally, we derive the global analytic solution for the empirical VBMF (Section 3.3).

3.1 Equivalence between VBMF and SimpleVBMF

Under the *matrix-wise* independence constraint (6), the free energy (5) can be written as

$$\begin{aligned}F^{\mathrm{VB}} &= \langle \log r_A(A) + \log r_B(B) - \log p(V|A,B)p(A)p(B)\rangle_{r(A)r(B)} \\ &= \frac{\|V\|_{\mathrm{Fro}}^2}{2\sigma^2} + \frac{LM}{2}\log\sigma^2 + \frac{M}{2}\log\frac{|C_A|}{|\Sigma_A|} + \frac{L}{2}\log\frac{|C_B|}{|\Sigma_B|} \\ &\quad + \frac{1}{2}\mathrm{tr}\Big\{C_A^{-1}\left(\widehat{A}^\top\widehat{A} + M\Sigma_A\right) + C_B^{-1}\left(\widehat{B}^\top\widehat{B} + L\Sigma_B\right) \\ &\qquad + \sigma^{-2}\left(-2\widehat{A}^\top V^\top\widehat{B} + \left(\widehat{A}^\top\widehat{A} + M\Sigma_A\right)\left(\widehat{B}^\top\widehat{B} + L\Sigma_B\right)\right)\Big\} + \mathrm{const.}, \qquad (20)\end{aligned}$$

where $|\cdot|$ denotes the determinant of a matrix. Note that Equations (8)–(11) together form the stationarity condition of Equation (20) with respect to $(\widehat{A},\widehat{B},\Sigma_A,\Sigma_B)$.

We say that two points $(\widehat{A},\widehat{B},\Sigma_A,\Sigma_B)$ and $(\widehat{A}',\widehat{B}',\Sigma_A',\Sigma_B')$ are *equivalent* if both give the same free energy and $\widehat{B}\widehat{A}^\top = \widehat{B}'\widehat{A}'^\top$ holds. We obtain the following theorem (its proof is given in Appendix A):

Theorem 1 *When C_AC_B is non-degenerate (i.e., $c_{a_h}c_{b_h} > c_{a_{h'}}c_{b_{h'}}$ for any pair $h < h'$), any solution minimizing the free energy* (20) *has diagonal Σ_A and Σ_B. When C_AC_B is degenerate, any solution has an* equivalent *solution with diagonal Σ_A and Σ_B.*

The result that Σ_A and Σ_B become diagonal would be natural because we assumed the independent Gaussian priors on A and B: the fact that any V can be decomposed into orthogonal singular components may imply that the observation V cannot convey any preference for singular-component-wise correlation. Note, however, that Theorem 1 does not necessarily hold when the observed matrix has missing entries.

Obviously, any VBMF solution (minimizer of the free energy (20)) with diagonal covariances is a SimpleVBMF solution (minimizer of the free energy (20) under the constraint that the covariances are diagonal). Theorem 1 states that, if C_AC_B is non-degenerate, the set of VBMF solutions and the set of SimpleVBMF solutions are identical. In the case when C_AC_B is degenerate, the set of VBMF solutions is the union of the set of SimpleVBMF solutions and the set of their *equivalent* solutions with non-diagonal covariances. Actually, any VBMF solution can be obtained by rotating its *equivalent* SimpleVBMF solution (VBMF solution with diagonal covariances) (see Appendix A.4). In practice, it is however sufficient to focus on the SimpleVBMF solutions, since *equivalent* solutions share the same free energy F^{VB} and the same mean prediction $\widehat{B}\widehat{A}^\top$. In this sense, we can conclude that the stronger *column-wise* independence constraint (15) does not degrade approximation accuracy, and the VBMF solution under the *matrix-wise* independence (6) *essentially* agrees with the SimpleVBMF solution.

Since we have shown the equivalence between VBMF and SimpleVBMF, we can use the results obtained in Nakajima and Sugiyama (2011), where SimpleVBMF was analyzed, for pursuing the global analytic solution for (non-simple) VBMF.

3.2 Global Analytic Solution for VBMF

Here, we derive an analytic-form expression of the VBMF global solution. We denote by $\mathbb{R}^d_{++}$ the set of the d-dimensional vectors with positive elements, and by $\mathbb{S}^d_{++}$ the set of $d \times d$ symmetric positive-definite matrices. We solve the following problem:

$$\begin{aligned} \text{Given} \quad & (c^2_{a_h}, c^2_{b_h}) \in \mathbb{R}^2_{++} \ (\forall h = 1, \ldots, H),\ \sigma^2 \in \mathbb{R}_{++}, \\ \min \quad & F^{\mathrm{VB}}(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B) \\ \text{s.t.} \quad & \widehat{A} \in \mathbb{R}^{M \times H},\ \widehat{B} \in \mathbb{R}^{L \times H},\ \Sigma_A \in \mathbb{S}^H_{++},\ \Sigma_B \in \mathbb{S}^H_{++}, \end{aligned}$$

where $F^{\mathrm{VB}}(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B)$ is the free energy given by Equation (20). This is a non-convex optimization problem, but we show that the global optimal solution can still be analytically obtained.

We start from the following proposition, which is obtained by summarizing Lemma 11, Lemma 13, Lemma 14, Lemma 15, and Lemma 17 in Nakajima and Sugiyama (2011):

Proposition 2 *(Nakajima and Sugiyama, 2011) Let γ_h (≥ 0) be the h-th largest singular value of V, and let $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ be the associated right and left singular vectors:*

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

Then, the global SimpleVB solution (under the column-wise independence (15)*) can be expressed as*

$$\widehat{U}^{\mathrm{SVB}} \equiv \langle BA^\top \rangle_{r^{\mathrm{SVB}}(A,B)} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{SVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top.$$

Let

$$\widetilde{\gamma}_h = \sqrt{\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c^2_{a_h}c^2_{b_h}} + \sqrt{\left(\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c^2_{a_h}c^2_{b_h}}\right)^2 - LM\sigma^4}}.$$

When

$$\gamma_h \leq \widetilde{\gamma}_h,$$

the SimpleVB solution for the h-th component is $\widehat{\gamma}_h^{\mathrm{SVB}} = 0$. When

$$\gamma_h > \widetilde{\gamma}_h, \tag{21}$$

$\widehat{\gamma}_h^{\mathrm{SVB}}$ is given as a positive real solution of

$$\widehat{\gamma}_h^2 + q_1(\widehat{\gamma}_h) \cdot \widehat{\gamma}_h + q_0 = 0, \tag{22}$$

where

$$\begin{aligned} q_1(\widehat{\gamma}_h) &= \frac{-(M-L)^2(\gamma_h - \widehat{\gamma}_h) + (L+M)\sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h)^2 + \frac{4\sigma^4 LM}{c^2_{a_h}c^2_{b_h}}}}{2LM}, \\ q_0 &= \frac{\sigma^4}{c^2_{a_h}c^2_{b_h}} - \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2. \end{aligned}$$

When Inequality (21) *holds, Equation* (22) *has only one positive real solution, which lies in*

$$0 < \widehat{\gamma}_h < \gamma_h.$$

In Nakajima and Sugiyama (2011), it was shown that any SimpleVBMF solution is a stationary point, and Equation (22) was derived from the stationarity condition (16)–(19). Bounds of $\widehat{\gamma}_h^{\mathrm{SVB}}$ were obtained by approximating Equation (22) with a quadratic equation (more specifically, by bounding $q_1(\widehat{\gamma}_h)$ by constants). This analysis revealed interesting properties of VBMF, including the model-induced regularization effect and the sparsity induction mechanism. Thanks to Theorem 1, almost the same statements as Proposition 2 hold for VBMF (Lemma 8 in Appendix B).

In this paper, our purpose is to obtain the exact solution, and therefore, we should treat Equation (22) more precisely. If $q_1(\widehat{\gamma}_h)$ were a constant, Equation (22) would be quadratic with respect to $\widehat{\gamma}_h$, and its solutions could be easily obtained. However, Equation (22) is not even polynomial, because $q_1(\widehat{\gamma}_h)$ depends on the square root of $\widehat{\gamma}_h$. With some algebra, we can convert Equation (22) to a quartic equation, which has four solutions in general. By examining which solution corresponds to the positive solution of Equation (22), we obtain the following theorem (the proof is given in Appendix B):

Theorem 3 *Let* $\widehat{\gamma}_h^{\mathrm{second}}$ *be the* second *largest real solution of the following quartic equation with respect to* $\widehat{\gamma}_h$*:*

$$f(\widehat{\gamma}_h) := \widehat{\gamma}_h^4 + \xi_3\widehat{\gamma}_h^3 + \xi_2\widehat{\gamma}_h^2 + \xi_1\widehat{\gamma}_h + \xi_0 = 0, \tag{23}$$

where the coefficients are defined by

$$\begin{aligned}
\xi_3 &= \frac{(L-M)^2\gamma_h}{LM}, \\
\xi_2 &= -\left(\xi_3\gamma_h + \frac{(L^2+M^2)\eta_h^2}{LM} + \frac{2\sigma^4}{c_{a_h}^2 c_{b_h}^2}\right), \\
\xi_1 &= \xi_3\sqrt{\xi_0}, \\
\xi_0 &= \left(\eta_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}\right)^2, \\
\eta_h^2 &= \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2.
\end{aligned}$$

Then, the global VB solution can be expressed as

$$\widehat{U}^{\mathrm{VB}} \equiv \langle BA^\top\rangle_{r^{\mathrm{VB}}(A,B)} = \widehat{BA}^\top = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h}\boldsymbol{\omega}_{a_h}^\top,$$

where

$$\widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \widehat{\gamma}_h^{\mathrm{second}} & \textit{if } \gamma_h > \widetilde{\gamma}_h, \\ 0 & \textit{otherwise}. \end{cases}$$

The coefficients of the quartic equation (23) are analytic, so $\widehat{\gamma}_h^{\text{second}}$ can also be obtained analytically, for example, by *Ferrari's method* (Hazewinkel, 2002).[2] Therefore, the global VB solution can be analytically computed.[3] This is a strong advantage over the standard ICM algorithm since many iterations and restarts would be necessary to find a good solution by ICM.

Based on the above result, the complete VB posterior can be obtained analytically as follows (the proof is also given in Appendix B):

Theorem 4 *The VB posterior is given by*

$$r^{\text{VB}}(A,B) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \widehat{\boldsymbol{a}}_h, \sigma^2_{a_h} I_M) \prod_{h=1}^{H} \mathcal{N}_L(\boldsymbol{b}_h; \widehat{\boldsymbol{b}}_h, \sigma^2_{b_h} I_L),$$

where, for $\widehat{\gamma}_h^{\text{VB}}$ being the solution given by Theorem 3,

$$\widehat{\boldsymbol{a}}_h = \pm\sqrt{\widehat{\gamma}_h^{\text{VB}} \widehat{\delta}_h} \cdot \boldsymbol{\omega}_{a_h},$$

$$\widehat{\boldsymbol{b}}_h = \pm\sqrt{\widehat{\gamma}_h^{\text{VB}} \widehat{\delta}_h^{-1}} \cdot \boldsymbol{\omega}_{b_h},$$

$$\sigma^2_{a_h} = \frac{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 - \sigma^2(M-L))^2 + 4M\sigma^2\widehat{\eta}_h^2}}{2M(\widehat{\gamma}_h^{\text{VB}}\widehat{\delta}_h^{-1} + \sigma^2 c_{a_h}^{-2})},$$

$$\sigma^2_{b_h} = \frac{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(M-L))^2 + 4L\sigma^2\widehat{\eta}_h^2}}{2L(\widehat{\gamma}_h^{\text{VB}}\widehat{\delta}_h + \sigma^2 c_{b_h}^{-2})},$$

$$\widehat{\delta}_h = \frac{(M-L)(\gamma_h - \widehat{\gamma}_h^{\text{VB}}) + \sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h^{\text{VB}})^2 + \frac{4\sigma^4 LM}{c_{a_h}^2 c_{b_h}^2}}}{2\sigma^2 M c_{a_h}^{-2}},$$

$$\widehat{\eta}_h^2 = \begin{cases} \eta_h^2 & \textit{if } \gamma_h > \widetilde{\gamma}_h, \\ \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} & \textit{otherwise.} \end{cases}$$

3.3 Global Analytic Solution for Empirical VBMF

Solving the following problem gives the empirical VBMF solution:

$$\begin{aligned} \text{Given} \quad & \sigma^2 \in \mathbb{R}_{++}, \\ \min \quad & F^{\text{VB}}(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B, \{c_{a_h}^2, c_{b_h}^2; h = 1, \ldots, H\}), \\ \text{s.t.} \quad & \widehat{A} \in \mathbb{R}^{M \times H},\ \widehat{B} \in \mathbb{R}^{L \times H},\ \Sigma_A \in \mathbb{S}_{++}^{H},\ \Sigma_B \in \mathbb{S}_{++}^{H}, \\ & (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}_{++}^2\ (\forall h = 1, \ldots, H), \end{aligned}$$

where $F^{\text{VB}}(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B, \{c_{a_h}^2, c_{b_h}^2; h = 1, \ldots, H\})$ is the free energy given by Equation (20). Although this is again a non-convex optimization problem, the global optimal solution can be obtained analytically. As discussed in Nakajima and Sugiyama (2011), the ratio c_{a_h}/c_{b_h} is arbitrary in empirical VB. Accordingly, we fix the ratio to $c_{a_h}/c_{b_h} = 1$ without loss of generality.

[2] In practice, one may solve the quartic equation numerically, for example, by the '`roots`' function in MATLAB®.

[3] In our latest work on performance analysis of VBMF, we have derived a simpler-form solution, which does not require to solve a quartic equation (Nakajima et al., 2012b).

Nakajima and Sugiyama (2011) obtained a closed form solution of the optimal hyperparameter value $\widehat{c}_{a_h}\widehat{c}_{b_h}$ for SimpleVBMF. Therefore, we can easily obtain the global analytic solution for empirical VBMF. We have the following theorem (the proof is given in Appendix C):

Theorem 5 *The global empirical VB solution is given by*

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

where

$$\widehat{\gamma}_h^{\mathrm{EVB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{VB}} & \textit{if } \gamma_h > \underline{\gamma}_h \textit{ and } \Delta_h \leq 0, \\ 0 & \textit{otherwise.} \end{cases}$$

Here,

$$\underline{\gamma}_h = (\sqrt{L}+\sqrt{M})\sigma, \tag{24}$$

$$\breve{c}_{a_h}^2 \breve{c}_{b_h}^2 = \frac{1}{2LM}\left(\gamma_h^2 - (L+M)\sigma^2 + \sqrt{\left(\gamma_h^2 - (L+M)\sigma^2\right)^2 - 4LM\sigma^4}\right), \tag{25}$$

$$\Delta_h = M\log\left(\frac{\gamma_h}{M\sigma^2}\breve{\gamma}_h^{\mathrm{VB}}+1\right) + L\log\left(\frac{\gamma_h}{L\sigma^2}\breve{\gamma}_h^{\mathrm{VB}}+1\right) + \frac{1}{\sigma^2}\left(-2\gamma_h\breve{\gamma}_h^{\mathrm{VB}} + LM\breve{c}_{a_h}^2\breve{c}_{b_h}^2\right), \tag{26}$$

and $\breve{\gamma}_h^{\mathrm{VB}}$ is the VB solution for $c_{a_h}c_{b_h} = \breve{c}_{a_h}\breve{c}_{b_h}$.

By using Theorem 3 and Theorem 5, the global empirical VB solution can be computed analytically. This is again a strong advantage over the standard ICM algorithm since ICM would require many iterations and restarts to find a good local minimum. The calculation procedure for the empirical VB solution is as follows: After obtaining $\{\gamma_h\}$ by singular value decomposition of V, we first check if $\gamma_h > \underline{\gamma}_h$ holds for each h, by using Equation (24). If it holds, we compute $\breve{\gamma}_h^{\mathrm{VB}}$ by using Equation (25) and Theorem 3. Otherwise, $\widehat{\gamma}_h^{\mathrm{EVB}} = 0$. Finally, we check if $\Delta_h \leq 0$ holds by using Equation (26).

When the noise variance σ^2 is unknown, it may be estimated by minimizing the VB free energy with respect to σ^2. In practice, this single-parameter minimization may be carried out numerically based on Equation (20) and Theorem 4.

4. Matrix Factorization for Multivariate Analysis

In this section, we explicitly describe the relation between MF and multivariate analysis techniques.

4.1 Probabilistic PCA

The relation to principal component analysis (PCA) (Hotelling, 1933) is straightforward. In probabilistic PCA (Tipping and Bishop, 1999), the observation $\boldsymbol{v} \in \mathbb{R}^L$ is assumed to be driven by a latent vector $\widetilde{\boldsymbol{a}} \in \mathbb{R}^H$ in the following form:

$$\boldsymbol{v} = B\widetilde{\boldsymbol{a}} + \boldsymbol{\varepsilon}.$$

Here, $B \in \mathbb{R}^{L\times H}$ specifies the linear relationship between $\widetilde{\boldsymbol{a}}$ and $\boldsymbol{v}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^L$ is a Gaussian noise subject to $\mathcal{N}_L(\boldsymbol{0}, \sigma^2 I_L)$. Suppose that we are given M observed samples $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M)$ generated

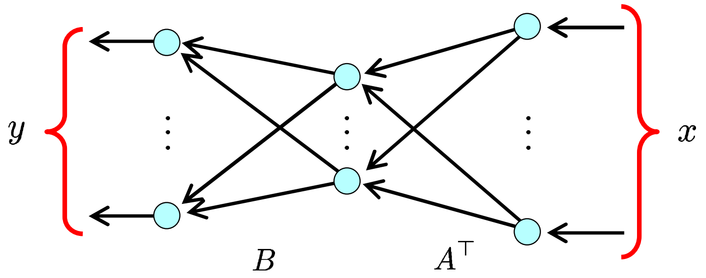


Figure 1: Linear neural network.

from the latent vectors $A^\top = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M)$, and each latent vector is subject to $\widetilde{\boldsymbol{a}} \sim \mathcal{N}_H(\mathbf{0}, I_H)$. Then, the probabilistic PCA model is written as Equations (1) and (2) with $C_A = I_H$.

If we apply Bayesian inference, the intrinsic dimension H is automatically selected without predetermination (Bishop, 1999). This useful property is called *automatic dimensionality selection* (ADS). It was shown that ADS originates from the *model-induced regularization* effect (Nakajima and Sugiyama, 2011).

4.2 Reduced Rank Regression

Reduced rank regression (RRR) (Baldi and Hornik, 1995; Reinsel and Velu, 1998) is aimed at learning a relation between an input vector $\boldsymbol{x} \in \mathbb{R}^M$ and an output vector $\boldsymbol{y} \in \mathbb{R}^L$ by using the following linear model:

$$\boldsymbol{y} = BA^\top \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{27}$$

where $A \in \mathbb{R}^{M \times H}$ and $B \in \mathbb{R}^{L \times H}$ are parameter matrices, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_L(\mathbf{0}, \sigma'^2 I_L)$ is a Gaussian noise vector. This can be expressed as a linear neural network (Figure 1). Thus, we can interpret this model as first projecting the input vector $\boldsymbol{x}$ onto a lower-dimensional latent subspace by $A^\top$ and then performing linear prediction by B.

Suppose we are given n pairs of input and output vectors:

$$\mathcal{V}^n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid \boldsymbol{x}_i \in \mathbb{R}^M, \boldsymbol{y}_i \in \mathbb{R}^L, i = 1, \ldots, n\}. \tag{28}$$

Then, the likelihood of the RRR model (27) is expressed as

$$p(\mathcal{V}^n | A, B) \propto \exp\left(-\frac{1}{2\sigma'^2} \sum_{i=1}^n \|\boldsymbol{y}_i - BA^\top \boldsymbol{x}_i\|^2 \right). \tag{29}$$

Let us assume that the samples are centered:

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = \mathbf{0} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i = \mathbf{0}.$$

Furthermore, let us assume that the input samples are *pre-whitened* (Hyvärinen et al., 2001), that is, they satisfy

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top = I_M.$$

Let

$$V = \Sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{x}_i^\top \tag{30}$$

be the sample *cross-covariance* matrix, and

$$\sigma^2 = \frac{\sigma'^2}{n} \tag{31}$$

be a rescaled noise variance. Then the likelihood (29) can be written as

$$p(\mathcal{V}^n|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}\|V - BA^\top\|^2_{\mathrm{Fro}}\right)\exp\left(-\frac{1}{2\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{y}_i\|^2 - \|V\|^2_{\mathrm{Fro}}\right)\right). \tag{32}$$

The first factor in Equation (32) coincides with the likelihood of the MF model (1), and the second factor is constant with respect to A and B. Thus, RRR is reduced to MF.

However, the second factor depends on the rescaled noise variance σ^2, and therefore, should be considered when σ^2 is estimated based on the free energy minimization principle. Furthermore, the normalization constant of the likelihood (29) is slightly different from that of the MF model. Taking into account of these differences, the VB free energy of the RRR model (29) with the priors (2) and (3) is given by

$$\begin{aligned}
F^{\mathrm{VB-RRR}} &= \left\langle \log r_A(A) + \log r_B(B) - \log p(\mathcal{V}^n|A,B)p(A)p(B)\right\rangle_{r(A)r(B)} \\
&= \frac{\sum_{i=1}^{n}\|\boldsymbol{y}_i\|^2}{2n\sigma^2} + \frac{nL}{2}\log\sigma^2 + \frac{M}{2}\log\frac{|C_A|}{|\Sigma_A|} + \frac{L}{2}\log\frac{|C_B|}{|\Sigma_B|} \\
&\quad + \frac{1}{2}\mathrm{tr}\left\{C_A^{-1}\left(\widehat{A}^\top\widehat{A} + M\Sigma_A\right) + C_B^{-1}\left(\widehat{B}^\top\widehat{B} + L\Sigma_B\right)\right. \\
&\qquad \left. + \sigma^{-2}\left(-2\widehat{A}^\top V^\top \widehat{B} + \left(\widehat{A}^\top\widehat{A} + M\Sigma_A\right)\left(\widehat{B}^\top\widehat{B} + L\Sigma_B\right)\right)\right\} + \mathrm{const.}
\end{aligned} \tag{33}$$

Note that the difference from Equation (20) exists only in the first two terms. Minimizing Equation (33), instead of Equation (20), gives an estimator for the rescaled noise variance. For the standard ICM algorithm, the following update rule should be substituted for Equation (14):

$$(\sigma^2)^{\mathrm{RRR}} = \frac{n^{-1}\sum_{i=1}^{n}\|\boldsymbol{y}_i\|^2 - \mathrm{tr}(2V^\top\widehat{B}\widehat{A}^\top) + \mathrm{tr}\left((\widehat{A}^\top\widehat{A} + M\Sigma_A)(\widehat{B}^\top\widehat{B} + L\Sigma_B)\right)}{nL}. \tag{34}$$

Once the rescaled noise variance σ^2 is estimated, Equation (31) gives the original noise variance σ'^2 of the RRR model (29).

4.3 Partial Least-Squares

Partial least-squares (PLS) (Worsley et al., 1997; Rosipal and Krämer, 2006) is similar to RRR. In PLS, the parameters A and B are learned so that the squared Frobenius norm of the difference from the sample *cross-covariance matrix* (30) is minimized:

$$(A,B) := \underset{A,B}{\mathrm{argmin}}\,\|\Sigma_{XY} - BA^\top\|^2_{\mathrm{Fro}}. \tag{35}$$

Clearly, PLS can be seen as the maximum likelihood estimation of the MF model (1).

4.4 Canonical Correlation Analysis

For paired samples (28), the goal of *canonical correlation analysis* (CCA) (Hotelling, 1936; Anderson, 1984) is to seek vectors $\boldsymbol{a} \in \mathbb{R}^M$ and $\boldsymbol{b} \in \mathbb{R}^L$ such that the correlation between $\boldsymbol{a}^\top \boldsymbol{x}$ and $\boldsymbol{b}^\top \boldsymbol{y}$ is maximized. $\boldsymbol{a}$ and $\boldsymbol{b}$ are called *canonical vectors*.

More formally, given the first $(h-1)$ canonical vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{h-1}$ and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{h-1}$, the h-th canonical vectors are defined as

$$(\boldsymbol{a}_h, \boldsymbol{b}_h) := \underset{\boldsymbol{a}, \boldsymbol{b}}{\mathrm{argmax}} \frac{\boldsymbol{a}^\top \Sigma_{XY} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_{XX} \boldsymbol{a}} \sqrt{\boldsymbol{b}^\top \Sigma_{YY} \boldsymbol{b}}},$$
$$\text{s.t. } \boldsymbol{a}^\top \Sigma_{XX} \boldsymbol{a}_{h'} = 0 \text{ and } \boldsymbol{b}^\top \Sigma_{YY} \boldsymbol{b}_{h'} = 0 \text{ for } h' = 1, \ldots, h-1,$$

where Σ_{XX} and Σ_{YY} are the sample covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and Σ_{XY} is the sample cross-covariance matrix, defined in Equation (30), of $\boldsymbol{x}$ and $\boldsymbol{y}$. The entire solution $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_H)$ and $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_H)$ are given as the H largest singular vectors of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$.

Let us assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are both pre-whitened, that is, $\Sigma_{XX} = I_M$ and $\Sigma_{YY} = I_L$. Then the solutions A and B are given as the singular vectors of Σ_{XY} associated with the H largest singular values. Since the solutions of Equation (35) are also given by the H dominant singular vectors of Σ_{XY} (Stewart, 1993), CCA is reduced to the maximum likelihood estimation of the MF model (1).

5. Experimental Results

In this section, we show experimental results on artificial and benchmark data sets, which illustrate practical usefulness of our analytic solution.

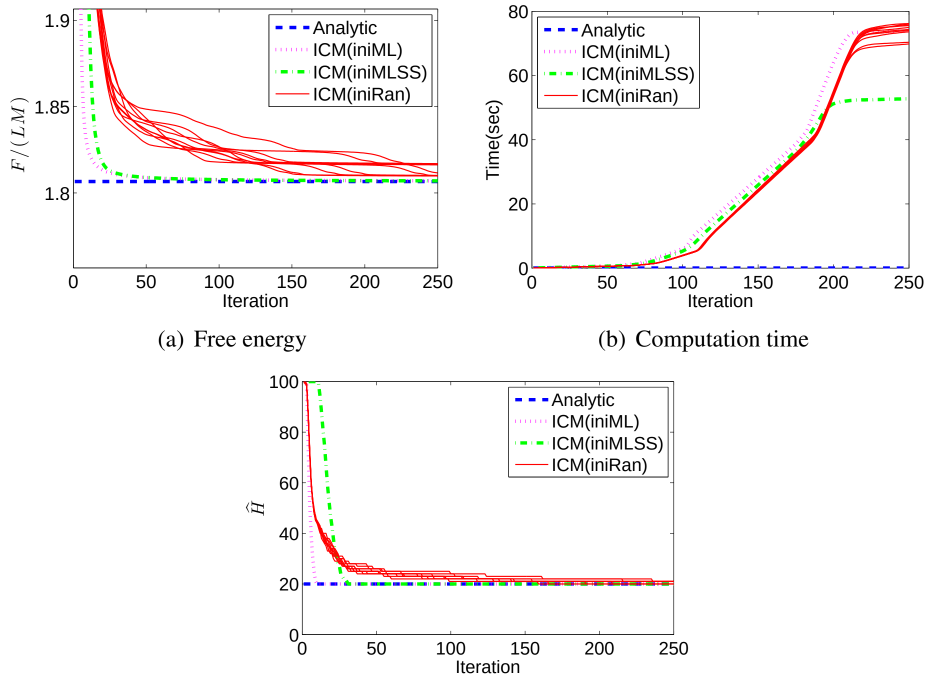
5.1 Experiment on Artificial Data

We compare the standard *ICM* algorithm and the *analytic* solution in the *empirical* VB scenario with unknown noise variance, that is, the hyperparameters (C_A, C_B) and the noise variance σ^2 are also estimated from observation. We use the full-rank model (i.e., $H = \min(L, M)$), and expect the ADS effect to automatically find the true rank H^*.

Figure 2 shows the free energy, the computation time, and the estimated rank over iterations for an artificial (*Artificial1*) data set with the data matrix size $L = 100$ and $M = 300$, and the true rank $H^* = 20$. We randomly created *true* matrices $A^* \in \mathbb{R}^{M \times H^*}$ and $B^* \in \mathbb{R}^{L \times H^*}$ so that each entry of A^* and B^* follows $\mathcal{N}_1(0,1)$. An observed matrix V was created by adding a noise subject to $\mathcal{N}_1(0,1)$ to each entry of $B^* A^{*\top}$.

The standard ICM algorithm consists of the update rules (8)–(14). Initial values were set in the following way: $\widehat{A}$ and $\widehat{B}$ are randomly created so that each entry follows $\mathcal{N}_1(0,1)$. Other variables are set to $\Sigma_A = \Sigma_B = C_A = C_B = I_H$ and $\sigma^2 = 1$. Note that we rescale V so that $\|V\|_{\mathrm{Fro}}^2/(LM) = 1$, before starting iterations. We ran the standard ICM algorithm 10 times, starting from different initial points, and each trial is plotted by a solid line (labeled as 'ICM(iniRan)') in Figure 2. The analytic solution consists of applying Theorem 5 combined with a naive 1-dimensional search for the estimation of noise variance σ^2. The analytic solution is plotted by the dashed line (labeled as 'Analytic'). We see that the analytic solution estimates the true rank $\widehat{H} = H^* = 20$ immediately (~ 0.1 sec on average over 10 trials), while the ICM algorithm does not converge in 60 sec.

Figure 3 shows experimental results on another artificial data set (*Artificial2*) where $L = 70$, $M = 300$, and $H^* = 40$. In this case, all the 10 trials of the ICM algorithm are trapped at local



(a) Free energy

(b) Computation time

(c) Estimated rank

Figure 2: Experimental results for *Artificial1* data set, where the data matrix size is $L = 100$ and $M = 300$, and the true rank is $H^* = 20$.

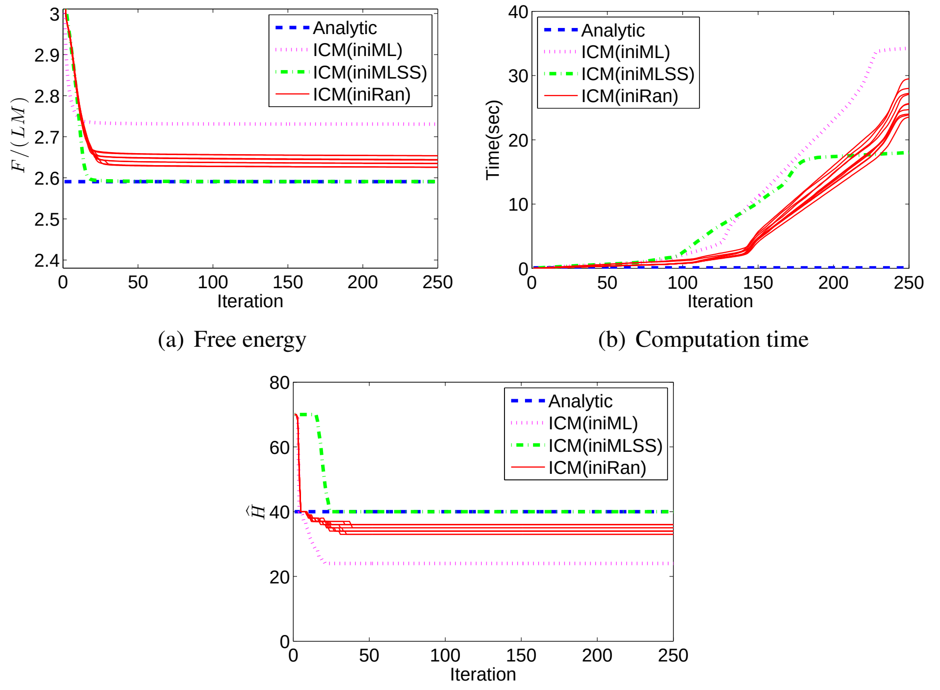
minima. We empirically observed that the local minima problem tends to be more critical, when H^* is large (close to H).

We also evaluated the ICM algorithm with different initialization schemes. The line labeled as 'ICM(iniML)' indicates the ICM algorithm starting from the maximum likelihood (ML) solution: $(\widehat{\boldsymbol{a}}_h, \widehat{\boldsymbol{b}}_h) = (\sqrt{\gamma_h}\boldsymbol{\omega}_{a_h}, \sqrt{\gamma_h}\boldsymbol{\omega}_{b_h})$. The initial values for other variables are the same as the random initialization. Figures 2 and 3 show that the ML initialization generally makes convergence faster than the random initialization, but suffers from the local minima problem more often.

We observed that starting from a small noise variance tends to alleviate the local minima problem at the expense of slightly slower convergence. The line labeled as 'ICM(iniMLSS)' indicates the ICM algorithm starting from the ML solution with a small noise variance $\sigma^2 = 0.0001$. We see in Figures 2 and 3 that this initialization improves quality of solutions, and successfully finds the true rank for these artificial data sets. However, we will show in Section 5.2 that this scheme still suffers from the local minima problem on benchmark data sets.

5.2 Experiment on Benchmark Data

Figures 4–6 show the PCA results on the *Glass*, the *Satimage*, and the *Spectf* data sets available from the UCI repository (Asuncion and Newman, 2007). Similar tendency to the artificial data experiment (Figures 2 and 3) is observed: 'ICM(iniRan)' converges slowly, and is often trapped



(a) Free energy

(b) Computation time

(c) Estimated rank

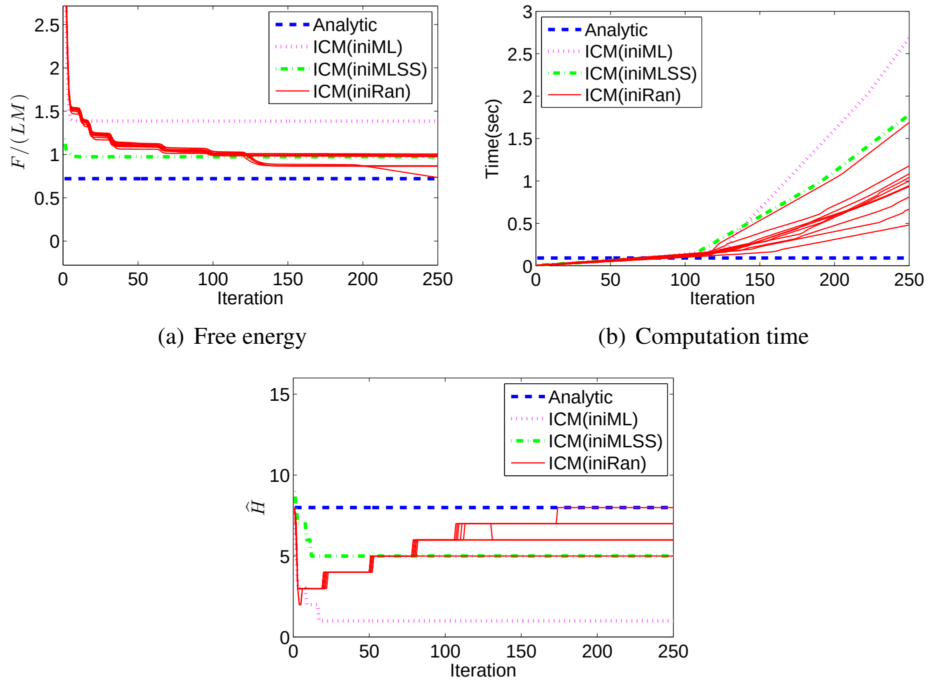
Figure 3: Experimental results for *Artificial2* data set ($L = 70$, $M = 300$, and $H^* = 40$).

at local minima with *wrong* ranks;[4] ‘ICM(iniML)’ converges slightly faster but to worse local minima; ‘ICM(iniMLSS)’ tends to give better solutions. Unlike in the artificial data experiment, ‘ICM(iniMLSS)’ fails to find the *correct* rank with these benchmark data sets. We also conducted experiments on other benchmark data sets, and found that the ICM algorithm generally converges slowly, and sometimes suffers from the local minima problem, while our analytic-form gives the global solution immediately.

Finally, we applied VBMF to a *reduced rank regression* (RRR) (Reinsel and Velu, 1998) task, and show the results in Figure 7. We centered the $L = 3$-dimensional outputs and the $M = 7$-dimensional inputs of the *Concrete Slump Test* data set, and pre-whitened the inputs. We also standardized the outputs so that the variance of each element is equal to one. Note that we have to minimize Equation (33), instead of Equation (20), for estimating the noise variance in our proposed method with the analytic solution, and use Equation (34), instead of Equation (14), for updating the noise variance in the standard ICM algorithm.

Overall, the proposed global analytic solution is shown to be a useful alternative to the standard ICM algorithm.

[4] Since the *true* ranks of the benchmark data sets are unknown, we mean by a *wrong* rank a rank different from the one given by the global ’Analytic’ solution.



(a) Free energy

(b) Computation time

(c) Estimated rank

Figure 4: PCA results for the *Glass* data set ($L = 9, M = 214$).

6. Discussion

In this section, we first derive simple analytic-form solutions for special cases, where the *model-induced regularization* and the *prior-induced* regularization can be clearly distinguished (Section 6.1). Then, we discuss the relation between model pruning by VB and spontaneous symmetry breaking (Section 6.2). Finally, we consider possibilities of extending our results to more general cases (Section 6.3).

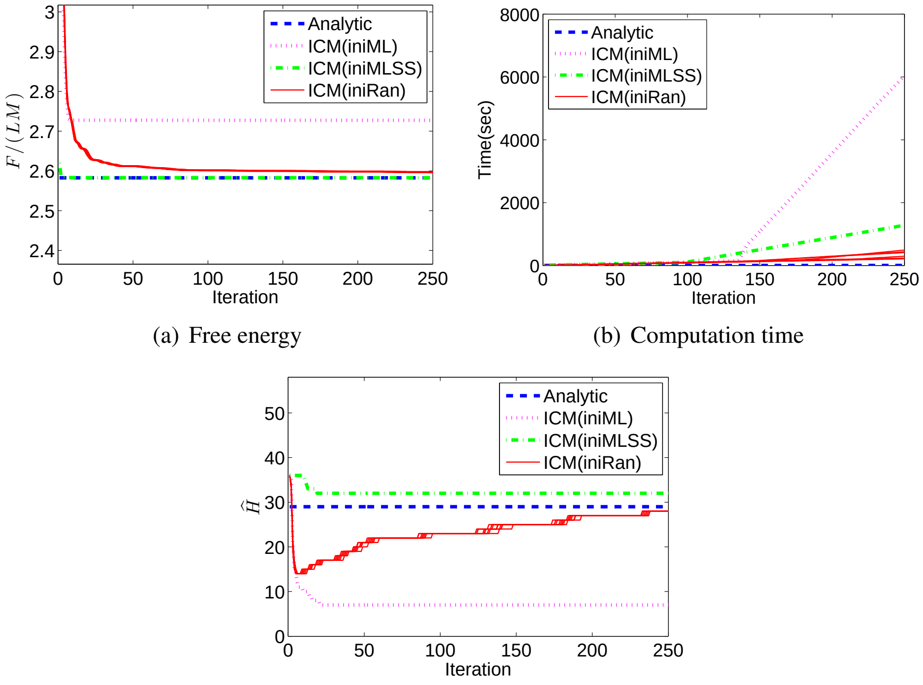
6.1 Special Cases

Here, we consider two special cases, where simple-form solutions are obtained.

6.1.1 FLAT PRIOR

When $c_{a_h}c_{b_h} \to \infty$ (i.e., the prior is *almost* flat), a simple-form exact solution for SimpleVBMF has been obtained in Nakajima and Sugiyama (2011). Thanks to Theorem 1, the same applies to VBMF under the standard *matrix-wise* independence assumption. This solution can be obtained also by factorizing the quartic equation (23) as follows:

$$\lim_{c_{a_h}c_{b_h}\to\infty} f(\widehat{\gamma}_h) = \left(\widehat{\gamma}_h + \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\gamma_h\right)\left(\widehat{\gamma}_h + \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h\right)$$
$$\cdot\left(\widehat{\gamma}_h - \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h\right)\left(\widehat{\gamma}_h - \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\gamma_h\right) = 0.$$

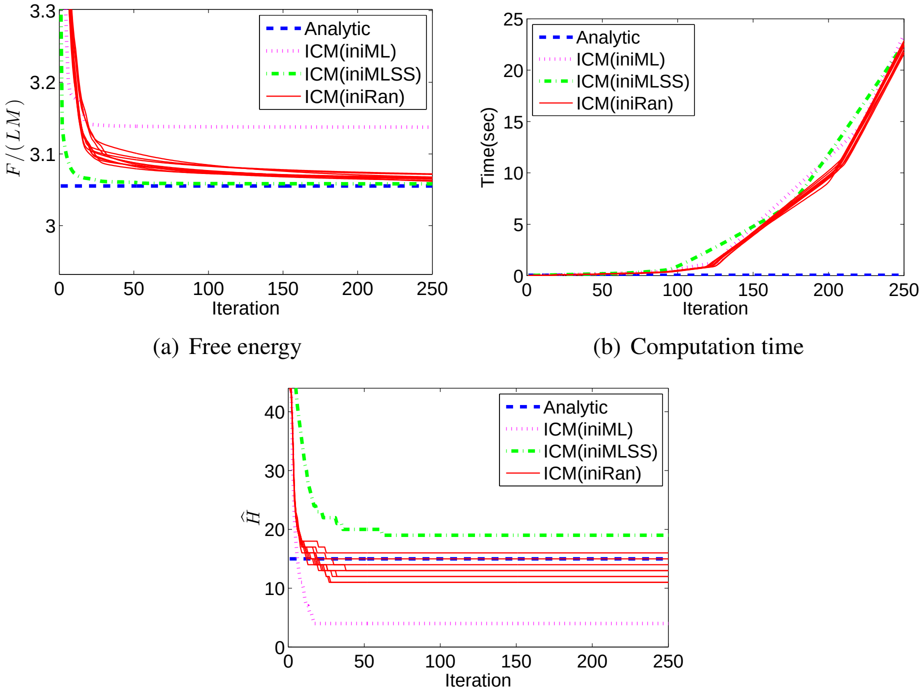


(a) Free energy

(b) Computation time

(c) Estimated rank

Figure 5: PCA results for the *Satimage* data set ($L = 36, M = 6435$).

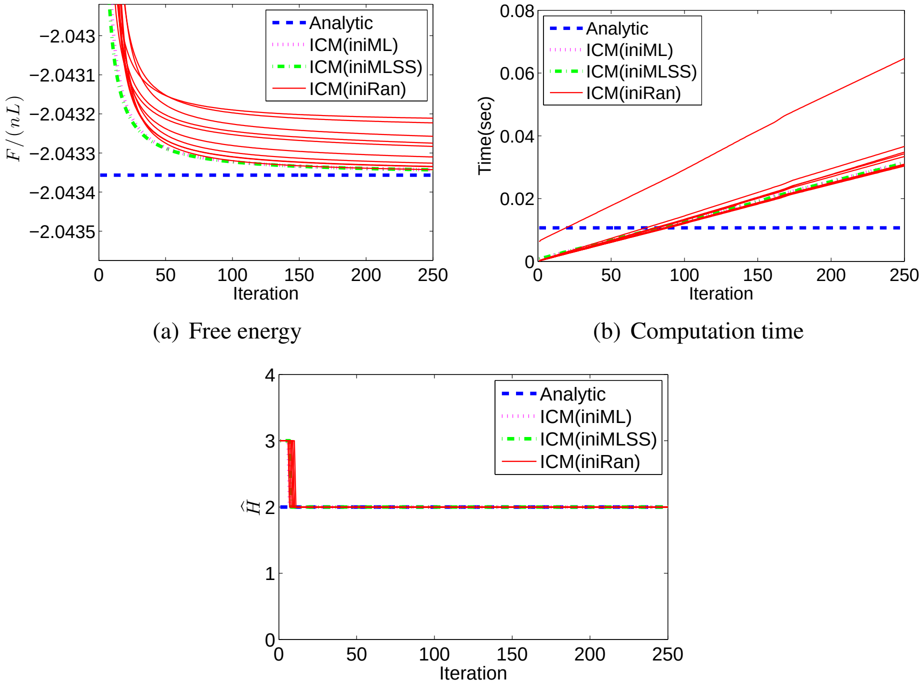


(a) Free energy

(b) Computation time

(c) Estimated rank

Figure 6: PCA results for the *Spectf* data set ($L = 44, M = 267$).



(a) Free energy

(b) Computation time

(c) Estimated rank

Figure 7: RRR results for the *Concrete Slump Test* data set ($L = 3, M = 7$).

Since Theorem 3 states that its *second* largest solution gives the VB estimator for $\gamma_h > \lim_{c_{a_h} c_{b_h} \to \infty} \widetilde{\gamma}_h = \sqrt{M\sigma^2}$, we have the following corollary:

Corollary 1 *The global VB solution with the almost flat prior (i.e., $c_{a_h} c_{b_h} \to \infty$) is given by*

$$\lim_{c_{a_h} c_{b_h} \to \infty} \widehat{\gamma}_h^{\mathrm{VB}} = \widehat{\gamma}_h^{\mathrm{PJS}} = \begin{cases} \max\left\{0, \left(1 - \dfrac{M\sigma^2}{\gamma_h^2}\right)\gamma_h\right\} & \mathit{if}\ \gamma_h > 0, \\ 0 & \mathit{otherwise}. \end{cases} \tag{36}$$

Equation (36) is the *positive-part James-Stein* (PJS) shrinkage estimator (James and Stein, 1961), operated on each singular component separately. This is actually the upper-bound of the VB solution for arbitrary $c_{a_h} c_{b_h} > 0$. The counter-intuitive fact—a shrinkage is observed even in the limit of flat priors—can be explained by strong non-uniformity of the *volume element of the Fisher metric*, that is, the *Jeffreys* prior (Jeffreys, 1946), in the parameter space. This effect is called *model-induced regularization* (MIR), because it is induced not by priors but by the structure of the model likelihood function (Nakajima and Sugiyama, 2011). MIR was shown to generally appear in Bayesian estimation when the model is *non-identifiable* (i.e., the mapping between parameters and distribution functions is not one-to-one) (Watanabe, 2009). The mechanism how non-identifiability causes MIR and ADS in VBMF was explicitly illustrated in Nakajima and Sugiyama (2011).

6.1.2 SQUARE MATRIX

Also when $L = M$ (i.e., the observation matrix V is square), a simple-form solution can be obtained. Since $\xi_3 = \xi_1 = 0$ (see Theorem 3) in this case, the quartic equation (23) can be solved as a quadratic equation with respect to $\widehat{\gamma}_h^2$ (Nakajima and Sugiyama, 2011). We can also find the solution by factorizing the quartic equation (23) for $\gamma_h > \sqrt{M\sigma^2}$ as follows:

$$f^{\mathrm{square}}(\widehat{\gamma}_h) = \left(\widehat{\gamma}_h + \widehat{\gamma}_h^{\mathrm{PJS}} + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\left(\widehat{\gamma}_h + \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)$$
$$\cdot\left(\widehat{\gamma}_h - \widehat{\gamma}_h^{\mathrm{PJS}} + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\left(\widehat{\gamma}_h - \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right) = 0.$$

Using Theorem 3, we have the following corollary:

Corollary 2 *When $L = M$, the global VB solution is given by*

$$\widehat{\gamma}_h^{\mathrm{VB-square}} = \max\left\{0, \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right\}. \tag{37}$$

Equation (37) shows that, in this case, MIR and *prior-induced regularization* (PIR) can be completely decomposed—the estimator is equipped with the *model-induced* PJS shrinkage ($\widehat{\gamma}_h^{\mathrm{PJS}}$) and the *prior-induced* trace-norm shrinkage ($-\sigma^2/(c_{a_h}c_{b_h})$).

The empirical VB solution is also simplified in this case. The following corollary is obtained simply by combining Theorem 1 in this paper and Corollary 2 in Nakajima and Sugiyama (2011):

Corollary 3 *When $L = M$, the global empirical VB solution is given by*

$$\widehat{\gamma}_h^{\mathrm{EVB}} = \begin{cases} \left(1 - \frac{M\sigma^2}{\gamma_h^2} - \rho_-\right)\gamma_h & \textit{if } \gamma_h > \underline{\gamma}_h \textit{ and } \Delta_h' \le 0, \\ 0 & \textit{otherwise}, \end{cases}$$

where

$$\underline{\gamma}_h = 2\sqrt{M}\sigma,$$
$$\Delta_h' = \log\left(\frac{\gamma_h^2}{M\sigma^2}(1-\rho_-)\right) - \frac{\gamma_h^2}{M\sigma^2}(1-\rho_-) + \left(1 + \frac{\gamma_h^2}{2M\sigma^2}\rho_+^2\right),$$
$$\rho_\pm = \sqrt{\frac{1}{2}\left(1 - \frac{2M\sigma^2}{\gamma_h^2} \pm \sqrt{1 - \frac{4M\sigma^2}{\gamma_h^2}}\right)}.$$

By using Corollary 2 and Corollary 3, respectively, we can easily compute the VB and the empirical VB solutions in this case without a quartic solver.

6.2 Model Pruning and Spontaneous Symmetry Breaking

Mackay (2001) pointed out that there are cases when VB prunes model components *inappropriately*, giving a toy example of a mixture of Gaussians. There, *appropriateness* is measured in terms of the similarity to the rigorous Bayesian estimation. He plotted the free energy of the mixture of Gaussians as a function of hidden *responsibility* variables—the probabilities that each sample belongs to each Gaussian component—and argued that VB sometimes favors simpler models too much. In this case, degrees of freedom are pruned when spontaneous symmetry breaking occurs.

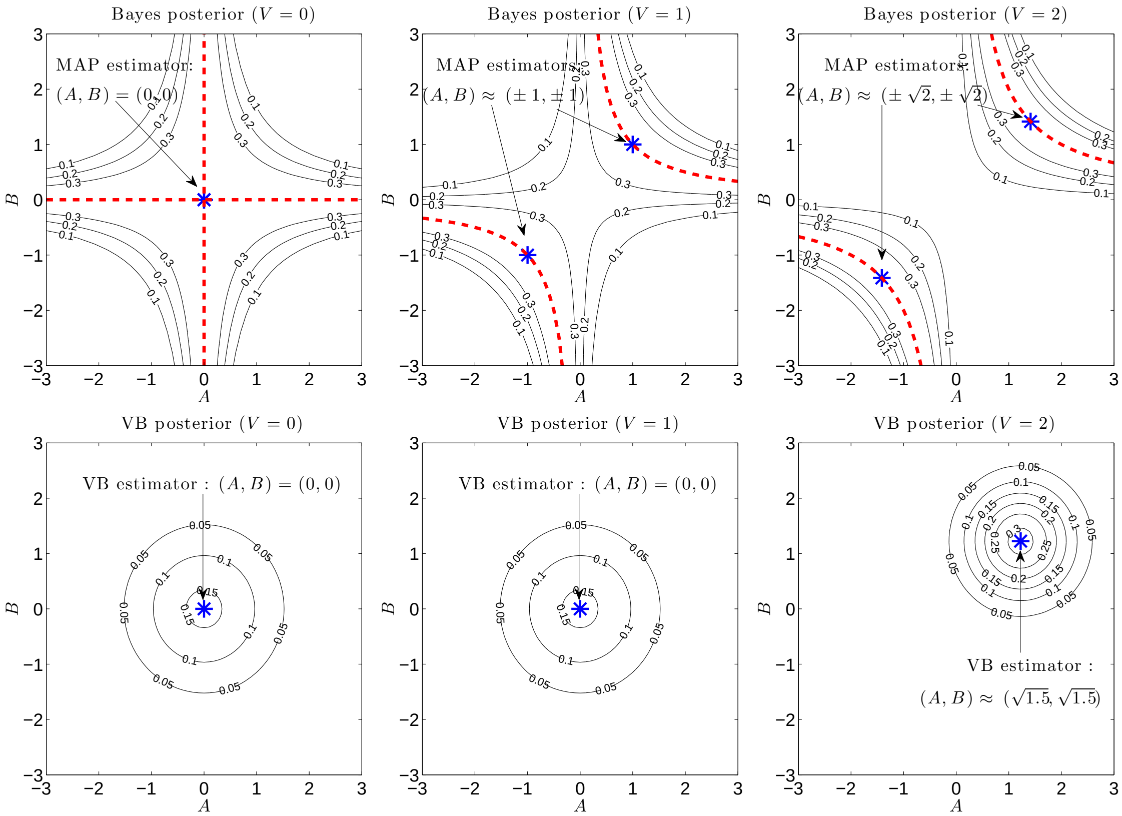


Figure 8: Bayes posteriors (top row) and the VB posteriors (bottom row) of a *scalar factorization* model (i.e., a MF model for $L = M = H = 1$) with $\sigma^2 = 1$ and $c_a = c_b = 100$ (almost flat priors), when the observed values are $V = 0$ (left), $V = 1$ (middle), and $V = 2$ (right), respectively. In the top row, the asterisks indicate the MAP estimators, and the dashed lines the ML estimators (the modes of the contour). In the bottom row, the asterisks indicate the VB estimators. All graphs are quoted from Nakajima and Sugiyama (2011).

In VBMF, degrees of freedom are pruned when spontaneous symmetry breaking does *not* occur. Figure 8 shows the Bayes posteriors (top row) and the VB posteriors (bottom row) of a *scalar factorization* model (i.e., a MF model for $L = M = H = 1$) with $\sigma^2 = 1$ and $c_a = c_b = 100$ (almost flat priors). As we can see in the top row, the Bayes posterior has two modes unless $V = 0$, and the distance between the two modes increases as $|V|$ increases. Since the VB posterior tries to approximate the Bayes posterior with a single uncorrelated distribution, it stays at the origin when $|V|$ is not sufficiently large. When $|V|$ is large enough, the VB posterior approximates one of the modes, as seen in the graphs in the right column (for the case when $V = 2$) of Figure 8 (note that there also exists an *equivalent* VB solution located at $(A, B) \approx (-\sqrt{1.5}, -\sqrt{1.5})$).

Equation (36) implies that symmetry breaking occurs when $V > \widetilde{\gamma}_h \sim \sqrt{M\sigma^2} = 1$, which coincides with the average contribution of noise to the observed singular values over all singular components. In this way, VBMF discards singular components dominated by noise. EVBMF has a different transition point, and tends to give a sparser solution (see Section 4 in Nakajima and Sugiyama (2011) for further discussion).

Given that the rigorous Bayesian estimator in MF is not sparse (see Figure 10 in Nakajima and Sugiyama, 2011), one might argue that the sparsity of VBMF is *inappropriate*. On the other hand, given that model pruning by VB has been acknowledged as a practically useful property, one might also argue that *appropriateness* should be measured in terms of performance. Motivated by the latter idea, we have conducted performance analysis of EVBMF in our latest work (Nakajima et al., 2012b), and shown that model pruning by EVBMF works perfectly under some condition. Conducting performance analysis in other models would be our future work.

6.3 Extensions

In this paper, we derived the global analytic solution of VBMF, by fully making use of the assumptions that the likelihood and priors are both spherical Gaussian, and that the observed matrix has no missing entry. They were necessary to solve the free energy minimization problem as a reweighted SVD. In this subsection, we discuss possibilities to extend our results to more general problems.

6.3.1 Robust PCA

VBMF gives a low-rank solution, which can be seen as a singular-component-wise sparse solution. We can extend our analysis so that a wider variety of sparsity can be handled.

Robust PCA (Candes et al., 2009) has recently gathered a great deal of attention. Equipped with an element-wise sparse term in addition to a low-rank term, robust PCA separates the low dimensional data structure from spiky noise. Its VB variant has also been proposed (Babacan et al., 2012). To obtain the VB solution of robust PCA, we have proposed a novel algorithm where the analytic VBMF solution is applied to partial problems (Nakajima et al., 2012a). Although the global optimality is not guaranteed, this algorithm has been experimentally shown to give a better solution than the standard ICM algorithm. In addition, our proposed algorithm can handle a variety of sparse terms beyond robust PCA.

6.3.2 Tensor Factorization

We have shown that the VB solution under *matrix-wise* independence essentially agrees with the SimpleVB solution under *column-wise* independence. We expect that similar *redundancy* would be found also in other models, for example, *tensor factorization* (Kolda and Bader, 2009; Carroll and Chang, 1970; Harshman, 1970; Tucker, 1996). In our preliminary study so far, we saw that the analytic VB solution for tensor factorization is not attainable, at least in the same way as MF. However, we have found that the optimal solution has diagonal covariances for the core tensor in Tucker decomposition (Nakajima, 2012), which would allow us to greatly simplify inference algorithms and reduce necessary memory storage and computational costs.

6.3.3 Correlated Priors

Our analysis assumed uncorrelated priors. With correlated priors, the posterior is no longer uncorrelated and thus it is not straightforward in general to obtain the global solution from the results obtained in this paper. One exception is the following: Suppose there exists an $H \times H$ non-singular matrix T such that both of $C'_A = TC_AT^\top$ and $C'_B = (T^{-1})^\top C_B T^{-1}$ are diagonal. We can show that the free energy (20) is invariant under the following transformation for any T:

$$\begin{aligned}
&A \to AT^\top, && \Sigma_A \to T\Sigma_A T^\top, && C_A \to TC_A T^\top, \\
&B \to BT^{-1}, && \Sigma_B \to (T^{-1})^T \Sigma_B T^{-1}, && C_B \to (T^{-1})^\top C_B T^{-1}.
\end{aligned}$$

Accordingly, the following procedure gives the global solution analytically: the analytic solution given the diagonal (C'_A, C'_B) is first computed, and the above transformation is then applied.

Handling priors correlated over *rows* of A and B is more challenging and remains as future work. Such a prior allows model correlations in the observation space, and can capture useful characteristics of data, for example, short-term correlation in time-series data and correlation among neighboring pixels in image data.

6.3.4 MISSING ENTRIES PREDICTION

Missing entries prediction is another prototypical application of MF (Konstan et al., 1997; Funk, 2006; Lim and Teh, 2007; Ilin and Raiko, 2010; Salakhutdinov and Mnih, 2008), where finding the global VBMF solution seems a very hard problem. However, one may use our analytic solution as a subroutine, for example, in the *soft-thresholding* step of SOFT-IMPUTE (Mazumder et al., 2010). Along this line, Seeger and Bouchard (2012) have recently proposed an algorithm, which tends to give a better local solution than the standard ICM algorithm for missing entries prediction. They also proposed a way to cope with non-Gaussian likelihood functions.

7. Conclusion

Overcoming the non-convexity of VB methods has been one of the important challenges in the Bayesian machine learning community, since it sometimes prevented us from applying the VB methods to highly complex real-world problems. In this paper, we focused on the matrix factorization (MF) problem with no missing entry, and showed that this weakness could be overcome by *analytically* computing the global optimal solution. We further derived the global optimal solution analytically for the empirical VBMF method, where hyperparameters are also optimized based on data samples. Since no hand-tuning parameter remains in empirical VBMF, our analytic-form solution is practically useful and computationally highly efficient. Numerical experiments showed that the proposed approach is promising.

We discussed the possibility that our analytic solution can be used as a building block of novel algorithms for more general problems. Tackling such possible extensions and conducting performance analysis of those methods are our future work.

Acknowledgments

The authors thank anonymous reviewers for their helpful comments, which significantly improved the paper. Shinichi Nakajima and Masashi Sugiyama thank the support from Grant-in-Aid for Scientific Research on Innovative Areas: Prediction and Decision Making, 23120004. S. Derin Babacan was supported by a Beckman Postdoctoral Fellowship. Ryota Tomioka was supported by Grant-in-Aid for Young Scientists (B) 22700138.

Appendix A. Proof of Theorem 1

In the same way as in the analysis for the SimpleVB approximation (see the proof of Lemma 10 in Nakajima and Sugiyama, 2011), we can show that any minimizer of the free energy (20) is a stationary point. Therefore, Equations (8)–(11) hold for any solution.

We consider the following three cases:

Case 1 When no pair of diagonal entries of $C_A C_B$ coincide.

Case 2 When all diagonal entries of $C_A C_B$ coincide.

Case 3 When (not all but) some pairs of diagonal entries of $C_A C_B$ coincide.

In the following, we prove that, in Case 1, Σ_A and Σ_B are diagonal for any solution $(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B)$, and that, in other cases, any solution has its *equivalent* solution with diagonal Σ_A and Σ_B.

Our proof relies on a technique related to the following proposition:

Proposition 6 *(Ruhe, 1970) Let $\lambda_h(\Phi), \lambda_h(\Psi)$ be the h-th largest eigenvalues of positive-definite symmetric matrices $\Phi, \Psi \in \mathbb{R}^{H \times H}$, respectively. Then, it holds that*

$$tr\{\Phi^{-1}\Psi\} \geq \sum_{h=1}^{H} \frac{\lambda_h(\Psi)}{\lambda_h(\Phi)}.$$

We use the following lemma (its proof is given in Appendix D.1):

Lemma 7 *Let $\Gamma, \Omega, \Phi \in \mathbb{R}^{H \times H}$ be a non-degenerate diagonal matrix, an orthogonal matrix, and a symmetric matrix, respectively. Let $\{\Lambda^{(k)}, \Lambda'^{(k)} \in \mathbb{R}^{H \times H}; k = 1, \ldots, K\}$ be arbitrary diagonal matrices. If*

$$G(\Omega) = tr\left\{\Gamma\Omega\Phi\Omega^\top + \sum_{k=1}^{K} \Lambda^{(k)}\Omega\Lambda'^{(k)}\Omega^\top\right\} \tag{38}$$

is minimized (as a function of Ω, given $\Gamma, \Phi, \{\Lambda^{(k)}, \Lambda'^{(k)}\}$) when $\Omega = I_H$, then Φ is diagonal. Here, K can be any natural number including $K = 0$ (when only the first term exists).

A.1 Proof for Case 1

Here, we consider the case when $c_{a_h} c_{b_h} > c_{a_{h'}} c_{b_{h'}}$ for any pair $h < h'$. We will show that any minimizer has diagonal covariances in this case.

Assume that $(A^*, B^*, \Sigma_A^*, \Sigma_B^*)$ is a minimizer of the free energy (20), and consider the following variation from it with respect to an arbitrary $H \times H$ orthogonal matrix Ω:

$$\widehat{A} = A^* C_A^{-1/2} \Omega^\top C_A^{1/2}, \tag{39}$$

$$\widehat{B} = B^* C_A^{1/2} \Omega^\top C_A^{-1/2}, \tag{40}$$

$$\Sigma_A = C_A^{1/2} \Omega C_A^{-1/2} \Sigma_A^* C_A^{-1/2} \Omega^\top C_A^{1/2}, \tag{41}$$

$$\Sigma_B = C_A^{-1/2} \Omega C_A^{1/2} \Sigma_B^* C_A^{1/2} \Omega^\top C_A^{-1/2}. \tag{42}$$

Note that this variation does not change $\widehat{B}\widehat{A}^\top$, and that $(\widehat{A},\widehat{B},\Sigma_A,\Sigma_B) = (A^*,B^*,\Sigma_A^*,\Sigma_B^*)$ holds if $\Omega = I_H$. Then, the free energy (20) can be written as a function of Ω:

$$F^{\mathrm{VB}}(\Omega) = \frac{1}{2}\mathrm{tr}\left\{C_A^{-1}C_B^{-1}\Omega C_A^{1/2}\left(B^{*\top}B^* + L\Sigma_B^*\right)C_A^{1/2}\Omega^\top\right\} + \mathrm{const.} \tag{43}$$

(the terms except the second term in the curly braces in Equation (20) are constant).

We define

$$\Phi = C_A^{1/2}\left(B^{*\top}B^* + L\Sigma_B^*\right)C_A^{1/2},$$

and rewrite Equation (43) as

$$F^{\mathrm{VB}}(\Omega) = \frac{1}{2}\mathrm{tr}\left\{C_A^{-1}C_B^{-1}\Omega\Phi\Omega^\top\right\} + \mathrm{const.} \tag{44}$$

The assumption that $(A^*,B^*,\Sigma_A^*,\Sigma_B^*)$ is a minimizer requires that Equation (44) is minimized when $\Omega = I_H$. Then, Lemma 7 (for $K = 0$) implies that Φ is diagonal.[5] Therefore,

$$C_A^{-1/2}\Phi C_A^{-1/2} (= \Phi C_A^{-1}) = B^{*\top}B^* + L\Sigma_B^*$$

is also diagonal. Consequently, Equation (10) implies that Σ_A^* is diagonal.

Next, consider the following variation with respect to an arbitrary $H \times H$ orthogonal matrix Ω',

$$\begin{aligned}
\widehat{A} &= A^* C_B^{1/2}\Omega'^\top C_B^{-1/2},\\
\widehat{B} &= B^* C_B^{-1/2}\Omega'^\top C_B^{1/2},\\
\Sigma_A &= C_B^{-1/2}\Omega' C_B^{1/2}\Sigma_A^* C_B^{1/2}\Omega'^\top C_B^{-1/2},\\
\Sigma_B &= C_B^{1/2}\Omega' C_B^{-1/2}\Sigma_B^* C_B^{-1/2}\Omega'^\top C_B^{1/2}.
\end{aligned}$$

Then, the free energy as a function of Ω' is given by

$$F^{\mathrm{VB}}(\Omega') = \frac{1}{2}\mathrm{tr}\left\{C_A^{-1}C_B^{-1}\Omega' C_B^{1/2}\left(A^{*\top}A^* + M\Sigma_A^*\right)C_B^{1/2}\Omega'^\top\right\} + \mathrm{const.}$$

From this, we can similarly prove that Σ_B^* is also diagonal, which completes the proof for Case 1. ■

A.2 Proof for Case 2

Here, we consider the case when $C_A C_B = cI_H$ for some positive $c \in \mathbb{R}$. In this case, there exist solutions with non-diagonal covariances. However, any of them belongs to an *equivalent* class involving a solution with diagonal covariances.

We can easily show that the free energy (20) is invariant of Ω under the transformation (39)–(42). This arbitrariness forms an *equivalent* class of solutions. Since there exists Ω that diagonalizes any given Σ_A^* through Equation (41), each *equivalent* class involves a solution with diagonal Σ_A. In the following, we will prove that any solution with diagonal Σ_A has diagonal Σ_B.

[5] Proposition 6 implies that the diagonal entries of Φ are arranged in non-increasing order.

Assume that $(A^*, B^*, \Sigma_A^*, \Sigma_B^*)$ is a solution with diagonal Σ_A^*, and consider the following variation from it with respect to an arbitrary $H \times H$ orthogonal matrix Ω:

$$\begin{aligned}
\widehat{A} &= A^* C_A^{-1/2} \Gamma^{-1/2} \Omega^\top \Gamma^{1/2} C_A^{1/2}, \\
\widehat{B} &= B^* C_A^{1/2} \Gamma^{1/2} \Omega^\top \Gamma^{-1/2} C_A^{-1/2}, \\
\Sigma_A &= C_A^{1/2} \Gamma^{1/2} \Omega \Gamma^{-1/2} C_A^{-1/2} \Sigma_A^* C_A^{-1/2} \Gamma^{-1/2} \Omega^\top \Gamma^{1/2} C_A^{1/2}, \\
\Sigma_B &= C_A^{-1/2} \Gamma^{-1/2} \Omega \Gamma^{1/2} C_A^{1/2} \Sigma_B^* C_A^{1/2} \Gamma^{1/2} \Omega^\top \Gamma^{-1/2} C_A^{-1/2}.
\end{aligned}$$

Here, $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_H)$ is an arbitrary non-degenerate ($\gamma_h \neq \gamma_{h'}$ for $h \neq h'$) positive-definite diagonal matrix. Then, the free energy can be written as a function of Ω:

$$\begin{aligned}
F^{\mathrm{VB}}(\Omega) = \frac{1}{2} \mathrm{tr} \Big\{ & \Gamma \Omega \Gamma^{-1/2} C_A^{-1/2} \left(A^{*\top} A^* + M \Sigma_A^* \right) C_A^{-1/2} \Gamma^{-1/2} \Omega^\top \\
& + c^{-1} \Gamma^{-1} \Omega \Gamma^{1/2} C_A^{1/2} \left(B^{*\top} B^* + L \Sigma_B^* \right) C_A^{1/2} \Gamma^{1/2} \Omega^\top \Big\}. \qquad (45)
\end{aligned}$$

We define

$$\begin{aligned}
\Phi_A &= \Gamma^{-1/2} C_A^{-1/2} \left(A^{*\top} A^* + M \Sigma_A^* \right) C_A^{-1/2} \Gamma^{-1/2}, \\
\Phi_B &= c^{-1} \Gamma^{1/2} C_A^{1/2} \left(B^{*\top} B^* + L \Sigma_B^* \right) C_A^{1/2} \Gamma^{1/2},
\end{aligned}$$

and rewrite Equation (45) as

$$F^{\mathrm{VB}}(\Omega) = \frac{1}{2} \mathrm{tr} \left\{ \Gamma \Omega \Phi_A \Omega^\top + \Gamma^{-1} \Omega \Phi_B \Omega^\top \right\}. \qquad (46)$$

Since Σ_A^* is diagonal, Equation (10) implies that Φ_B is diagonal. The assumption that $(A^*, B^*, \Sigma_A^*, \Sigma_B^*)$ is a solution requires that Equation (46) is minimized when $\Omega = I_H$. Accordingly, Lemma 7 implies that Φ_A is diagonal. Consequently, Equation (11) implies that Σ_B^* is diagonal.

Thus, we have proved that any solution has its *equivalent* solution with diagonal covariances, which completes the proof for Case 2. ■

A.3 Proof for Case 3

Finally, we consider the case when $c_{a_h} c_{b_h} = c_{a_h'} c_{b_{h'}}$ for (not all but) some pairs $h \neq h'$. First, in the same way as for Case 1, we can prove that Σ_A and Σ_B are block diagonal where the blocks correspond to the groups sharing the same $c_{a_h} c_{b_h}$. Next, we can apply the proof for Case 2 to each block, and show that any solution has its *equivalent* solution with diagonal Σ_A and Σ_B. Combining these results completes the proof of Theorem 1. ■

A.4 General Expression

In summary, for any minimizer of Equation (20), the covariances can be written in the following form:

$$\Sigma_A = C_A^{1/2} \Theta C_A^{-1/2} \Gamma_{\Sigma_A} C_A^{-1/2} \Theta^\top C_A^{1/2} (= C_B^{-1/2} \Theta C_B^{1/2} \Gamma_{\Sigma_A} C_B^{1/2} \Theta^\top C_B^{-1/2}), \qquad (47)$$

$$\Sigma_B = C_A^{-1/2} \Theta C_A^{1/2} \Gamma_{\Sigma_B} C_A^{1/2} \Theta^\top C_A^{-1/2} (= C_B^{1/2} \Theta C_B^{-1/2} \Gamma_{\Sigma_B} C_B^{-1/2} \Theta^\top C_B^{1/2}). \qquad (48)$$

Here, Γ_{Σ_A} and Γ_{Σ_B} are positive-definite diagonal matrices, and Θ is a block diagonal matrix such that the blocks correspond to the groups sharing the same $c_{a_h} c_{b_h}$, and each block consists of an orthogonal matrix. Furthermore, if there exists a solution with (Σ_A, Σ_B) written in the form of Equations (47) and (48) with a certain set of $(\Gamma_{\Sigma_A}, \Gamma_{\Sigma_B}, \Theta)$, then there also exist its *equivalent* solutions with the same $(\Gamma_{\Sigma_A}, \Gamma_{\Sigma_B})$ for *any* Θ. Focusing on the solution with $\Theta = I_H$ as the representative of each *equivalent* class, we can assume that Σ_A and Σ_B are diagonal without loss of generality.

Appendix B. Proof of Theorem 3 and Theorem 4

Combining Theorem 1 and Proposition 2, we have the following lemma:

Lemma 8 *Let* γ_h (≥ 0) *be the h-th largest singular value of* V*, and let* $\boldsymbol{\omega}_{a_h}$ *and* $\boldsymbol{\omega}_{b_h}$ *be the associated right and left singular vectors:*

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

Then, the global VB solution (under the matrix-wise independence (6)*) can be expressed as*

$$\widehat{U}^{\mathrm{VB}} \equiv \langle BA^{\top} \rangle_{r^{\mathrm{VB}}(A,B)} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}.$$

Let

$$\widetilde{\gamma}_h = \sqrt{\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM\sigma^4}}. \tag{49}$$

When

$$\gamma_h \leq \widetilde{\gamma}_h,$$

the VB solution for the h-th component is $\widehat{\gamma}_h^{\mathrm{VB}} = 0$*. When*

$$\gamma_h > \widetilde{\gamma}_h, \tag{50}$$

$\widehat{\gamma}_h^{\mathrm{VB}}$ *is given as a positive real solution of*

$$\widehat{\gamma}_h^2 + q_1(\widehat{\gamma}_h) \cdot \widehat{\gamma}_h + q_0 = 0, \tag{51}$$

where

$$q_1(\widehat{\gamma}_h) = \frac{-(M-L)^2(\gamma_h - \widehat{\gamma}_h) + (L+M)\sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h)^2 + \frac{4\sigma^4 LM}{c_{a_h}^2 c_{b_h}^2}}}{2LM}, \tag{52}$$

$$q_0 = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} - \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2. \tag{53}$$

When Inequality (50) *holds, Equation* (51) *has only one positive real solution, which lies in*

$$0 < \widehat{\gamma}_h < \gamma_h. \tag{54}$$

To obtain an analytic-form solution, we will find the positive solution of Equation (51) for $\gamma_h > \widetilde{\gamma}_h$. Because $q_1(\widehat{\gamma}_h)$ depends on the square root of $\widehat{\gamma}_h$, Equation (51) is not polynomial. However, since it has only one non-polynomial term, we can easily convert it to a polynomial form in the following way.

Substituting Equations (52) and (53), we can rewrite Equation (51) as

$$-\frac{(M^2+L^2)}{2LM}\widehat{\gamma}_h^2 + \frac{(M-L)^2\gamma_h}{2LM}\widehat{\gamma}_h + \sqrt{\xi_0} = \left(\frac{(M+L)\sqrt{\{(M-L)(\gamma_h - \widehat{\gamma}_h)\}^2 + \frac{4\sigma^4 ML}{c_{a_h}^2 c_{b_h}^2}}}{2LM} \right)\widehat{\gamma}_h. \tag{55}$$

Squaring both sides of Equation (55) removes the square root in the right-hand side, and leads to the quartic equation (23),

$$f(\widehat{\gamma}_h) := \widehat{\gamma}_h^4 + \xi_3\widehat{\gamma}_h^3 + \xi_2\widehat{\gamma}_h^2 + \xi_1\widehat{\gamma}_h + \xi_0 = 0, \tag{23}$$

where

$$\xi_3 = \frac{(L-M)^2\gamma_h}{LM}, \tag{56}$$

$$\xi_2 = -\left(\xi_3\gamma_h + \frac{(L^2+M^2)\eta_h^2}{LM} + \frac{2\sigma^4}{c_{a_h}^2 c_{b_h}^2}\right), \tag{57}$$

$$\xi_1 = \xi_3\sqrt{\xi_0}, \tag{58}$$

$$\xi_0 = \left(\eta_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}\right)^2, \tag{59}$$

$$\eta_h^2 = \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2. \tag{60}$$

Since we derived Equation (23) from Equation (51), any solution satisfying Equation (51) satisfies Equation (23). However, the converse does not necessarily hold, because squaring both sides of Equation (55) can create solutions that do not satisfy the original equation (51). By examining the possible range of the positive solution of Equation (51), we obtain the following lemma (the proof is given in Appendix D.2):

Lemma 9 *Assume that Inequality* (50) *holds. Any positive solution of Equation* (51) *lying in the range* (54) *satisfies the quartic equation* (23)*, and lies in the following range:*

$$0 < \widehat{\gamma}_h < \xi_0^{1/4}. \tag{61}$$

Conversely, any positive solution of the quartic equation (23) *lying in the range* (61) *satisfies Equation* (51)*, and lies in the range* (54).

Lemma 8 and Lemma 9 imply that finding the VB solution is achieved by finding a positive solution of the quartic equation (23) lying in the range (61). Investigating the shape of the quartic function $f(\widehat{\gamma}_h)$, defined in Equation (23), we have the following lemma (the proof is given in Appendix D.3):

Lemma 10 *Assume that Inequality* (50) *holds. The quartic equation* (23) *has two positive real solutions. The smaller one lies in the range* (61)*, and the larger one lies outside the range.*

Combining Lemma 8, Lemma 9, and Lemma 10 completes the proof of Theorem 3.

The following lemma is obtained by summarizing Lemma 11, Lemma 12, Lemma 14, Lemma 15, and Lemma 17 in Nakajima and Sugiyama (2011), and then combining with Theorem 1 in this paper:[6]

Lemma 11 *Let*

$$
\begin{aligned}
(\widehat{\eta}_h^2)^{\mathrm{null}} &= \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}, \\
(\sigma_{a_h}^2)^{\mathrm{null}} &= \frac{-\left((\widehat{\eta}_h^2)^{\mathrm{null}} - \sigma^2(M-L)\right) + \sqrt{((\widehat{\eta}_h^2)^{\mathrm{null}} - \sigma^2(M-L))^2 + 4M\sigma^2(\widehat{\eta}_h^2)^{\mathrm{null}}}}{2M\sigma^2 c_{a_h}^{-2}}, \\
(\sigma_{b_h}^2)^{\mathrm{null}} &= \frac{-\left((\widehat{\eta}_h^2)^{\mathrm{null}} + \sigma^2(M-L)\right) + \sqrt{((\widehat{\eta}_h^2)^{\mathrm{null}} + \sigma^2(M-L))^2 + 4L\sigma^2(\widehat{\eta}_h^2)^{\mathrm{null}}}}{2L(\widehat{\gamma}_h\widehat{\delta}_h + \sigma^2 c_{b_h}^{-2})}.
\end{aligned}
$$

When $\gamma_h \leq \widetilde{\gamma}_h$, the means and the variances of the VB posterior for the h-th component are given by

$$
(\widehat{\boldsymbol{a}}_h, \widehat{\boldsymbol{b}}_h, (\Sigma_A)_{h,h}, (\Sigma_B)_{h,h}) = \left(\mathbf{0}, \mathbf{0}, (\sigma_{a_h}^2)^{\mathrm{null}}, (\sigma_{b_h}^2)^{\mathrm{null}}\right).
$$

For $\gamma_h > \widetilde{\gamma}_h$, let

$$
\begin{aligned}
\widehat{\delta}_h &= \frac{(M-L)(\gamma_h - \widehat{\gamma}_h) + \sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h)^2 + \frac{4\sigma^4 LM}{c_{a_h}^2 c_{b_h}^2}}}{2\sigma^2 M c_{a_h}^{-2}}, \qquad (62) \\
\sigma_{a_h}^2 &= \frac{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 - \sigma^2(M-L))^2 + 4M\sigma^2\widehat{\eta}_h^2}}{2M(\widehat{\gamma}_h\widehat{\delta}_h^{-1} + \sigma^2 c_{a_h}^{-2})}, \\
\sigma_{b_h}^2 &= \frac{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(M-L))^2 + 4L\sigma^2\widehat{\eta}_h^2}}{2L(\widehat{\gamma}_h\widehat{\delta}_h + \sigma^2 c_{b_h}^{-2})}.
\end{aligned}
$$

When $\gamma_h > \widetilde{\gamma}_h$, the means and the variances of the VB posterior for the h-th component are given by

$$
(\widehat{\boldsymbol{a}}_h, \widehat{\boldsymbol{b}}_h, (\Sigma_A)_{h,h}, (\Sigma_B)_{h,h}) = (\pm\sqrt{\widehat{\gamma}_h\widehat{\delta}_h}\boldsymbol{\omega}_{a_h}, \pm\sqrt{\widehat{\gamma}_h\widehat{\delta}_h^{-1}}\boldsymbol{\omega}_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2).
$$

Combining Theorem 3 and Lemma 11 completes the proof of Theorem 4. ■

Appendix C. Proof of Theorem 5

Summarizing Lemma 22, Lemma 23, and Lemma 24 in Nakajima and Sugiyama (2011), and then combining with Theorem 1 in this paper, we have the following lemma:

[6] We also used Equation (147) in Nakajima and Sugiyama (2011), which is identical to Equation (62) in this paper.

Lemma 12 *If $\gamma_h \geq \underline{\gamma}_h$, the VB free energy* (20) *can have two local minima, i.e., $c_{a_h}c_{b_h} \to 0$ and $c_{a_h}c_{b_h} = \breve{c}_{a_h}\breve{c}_{b_h}$. Otherwise, $c_{a_h}c_{b_h} \to 0$ is the only local minimum.*

It was also shown in Nakajima and Sugiyama (2011) that the (scaled) free energy difference between the two local minima is given by Δ_h (the *positive* local minimum with $c_{a_h}c_{b_h} = \breve{c}_{a_h}\breve{c}_{b_h}$ gives lower free energy than the *null* local minimum with $c_{a_h}c_{b_h} \to 0$ if and only if $\Delta_h \leq 0$).[7] Thus, we have the following lemma:

Lemma 13 *The hyperparameter $\widehat{c}_{a_h}\widehat{c}_{b_h}$ that globally minimizes the VB free energy* (20) *is given by $\widehat{c}_{a_h}\widehat{c}_{b_h} = \breve{c}_{a_h}\breve{c}_{b_h}$ if $\gamma_h > \underline{\gamma}_h$ and $\Delta_h \leq 0$. Otherwise $\widehat{c}_{a_h}\widehat{c}_{b_h} \to 0$.*

Combining Lemma 13 and Theorem 3 completes the proof of Theorem 5. ■

Appendix D. Proof of Lemmas

In this appendix, we prove lemmas used in the previous appendices.

D.1 Proof of Lemma 7

Let

$$\Phi = \Omega'\Gamma'\Omega'^{\top} \tag{63}$$

be the eigenvalue decomposition of Φ. Let $\boldsymbol{\gamma}, \boldsymbol{\gamma}', \{\boldsymbol{\lambda}^{(k)}\}, \{\boldsymbol{\lambda}'^{(k)}\}$ be the vectors consist of the diagonal entries of $\Gamma, \Gamma', \{\Lambda^{(k)}\}, \{\Lambda'^{(k)}\}$, respectively, i.e,

$$\Gamma = \mathrm{diag}(\boldsymbol{\gamma}), \qquad \Gamma' = \mathrm{diag}(\boldsymbol{\gamma}'), \qquad \Lambda^{(k)} = \mathrm{diag}(\boldsymbol{\lambda}^{(k)}), \qquad \Lambda'^{(k)} = \mathrm{diag}(\boldsymbol{\lambda}'^{(k)}).$$

Then, Equation (38) can be written as

$$G(\Omega) = \mathrm{tr}\left\{\Gamma\Omega\Phi\Omega^{\top} + \sum_{k=1}^{K}\Lambda^{(k)}\Omega\Lambda'^{(k)}\Omega^{\top}\right\} = \boldsymbol{\gamma}^{\top}Q\boldsymbol{\gamma}' + \sum_{k=1}^{K}\boldsymbol{\lambda}^{(k)\top}R\boldsymbol{\lambda}'^{(k)}, \tag{64}$$

where

$$Q = (\Omega\Omega') * (\Omega\Omega'), \qquad R = \Omega * \Omega.$$

Here, $*$ denotes the Hadamard product.[8]

Using this expression, we will prove that Φ is diagonal if $\Omega = I_H$ minimizes Equation (64). Let us consider a bilateral perturbation $\Omega = \Delta$ such that the 2×2 matrix $\Delta_{(h,h')}$ consisting of the h-th and the h'-th columns and rows form an 2×2 orthogonal matrix

$$\Delta_{(h,h')} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

[7] Equation (26) was obtained as Equation (172) in Nakajima and Sugiyama (2011).

[8] Note that Q as well as R is the Hadamard square of an orthogonal matrix, which is known to be doubly stochastic (i.e., any of the columns and the rows sums up to one) (Marshall et al., 2009). Therefore, it can be seen that Q reassigns the components of $\boldsymbol{\gamma}$ to those of $\boldsymbol{\gamma}'$ when calculating the element-wise product in the first term of Equation (64). The same applies to R and $\{\boldsymbol{\lambda}^{(k)}, \boldsymbol{\lambda}'^{(k)}\}$ in the second term. Naturally, rearranging the components of $\boldsymbol{\gamma}$ in non-decreasing order and the components of $\boldsymbol{\gamma}'$ in non-increasing order minimizes $\boldsymbol{\gamma}^{\top}Q\boldsymbol{\gamma}'$, which proves Proposition 6 (Ruhe, 1970; Marshall et al., 2009).

and the rest entries coincide with those of the identity matrix. Then, the elements of Q become

$$Q_{i,j} = \begin{cases} (\Omega'_{h,j}\cos\theta - \Omega'_{h',j}\sin\theta)^2 & \text{if } i = h, \\ (\Omega'_{h,j}\sin\theta + \Omega'_{h',j}\cos\theta)^2 & \text{if } i = h', \\ \Omega'^2_{i,j} & \text{otherwise,} \end{cases}$$

and Equation (64) can be written as a function of θ:

$$G(\theta) = \sum_{j=1}^{H} \left\{\gamma_h(\Omega'_{h,j}\cos\theta - \Omega'_{h',j}\sin\theta)^2 + \gamma_{h'}(\Omega'_{h,j}\sin\theta + \Omega'_{h',j}\cos\theta)^2\right\}\gamma'_j$$
$$+ \sum_{k=1}^{K} \begin{pmatrix} \lambda_h^{(k)} & \lambda_{h'}^{(k)} \end{pmatrix} \begin{pmatrix} \cos^2\theta & \sin^2\theta \\ \sin^2\theta & \cos^2\theta \end{pmatrix} \begin{pmatrix} \lambda_h^{(k)} \\ \lambda_{h'}^{(k)} \end{pmatrix} + \text{const.}. \tag{65}$$

Since Equation (65) is differentiable at $\theta = 0$, our assumption that Equation (64) is minimized when $\Omega = I_H$ requires that $\theta = 0$ is a stationary point of Equation (65) for any $h \neq h'$. Therefore, it holds that

$$0 = \left.\frac{\partial G}{\partial\theta}\right|_{\theta=0} = 2\,(\gamma_{h'} - \gamma_h)\sum_j \Omega'_{h,j}\gamma'_j\Omega'_{h',j} = 2\,(\gamma_{h'} - \gamma_h)\,\Phi_{h,h'}. \tag{66}$$

In the last equation, we used Equation (63). Since we assume that Γ is non-degenerate ($\gamma_h \neq \gamma_{h'}$ for $h \neq h'$), Equation (66) implies that Φ is diagonal, which completes the proof of Lemma 7. ■

D.2 Proof of Lemma 9

Assume that Inequality (50) holds, i.e.,

$$\gamma_h > \widetilde{\gamma}_h. \tag{50}$$

By using Equation (60), we have

$$\eta_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = \gamma_h^{-2}\left(\gamma_h^2 - \widetilde{\gamma}_h^2\right)\left(\gamma_h^2 - \acute{\gamma}_h^2\right), \tag{67}$$

where

$$\acute{\gamma}_h = \sqrt{\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2} - \sqrt{\left(\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM\sigma^4}}. \tag{68}$$

Comparing Equations (49) and (68) leads to

$$\widetilde{\gamma}_h > \acute{\gamma}_h,$$

and therefore, Equation (67) is positive, i.e.,

$$\eta_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} > 0. \tag{69}$$

Combining Equations (59) and (60), and Inequality (69) leads to

$$0 < \xi_0^{1/4} < \eta_h < \gamma_h. \tag{70}$$

Combining Equations (53), (59), and (60) leads to

$$q_0 = -\sqrt{\xi_0}. \tag{71}$$

Let us first assume that we have a positive solution of Equation (51) lying in the range (54),

$$0 < \widehat{\gamma}_h < \gamma_h. \tag{54}$$

Since Equation (23) was derived from Equation (51), this solution naturally satisfies Equation (23). For the solution, Equation (52) implies that

$$q_1(\widehat{\gamma}_h) > 0.$$

Inequalities (70) and Equation (71) imply that

$$q_0 < 0.$$

Therefore, by ignoring the positive second term in the left-hand side of Equation (51), we find that the solution lies in the range (61),

$$0 < \widehat{\gamma}_h < \sqrt{-q_0} = \xi_0^{1/4}. \tag{61}$$

Here, we used Equation (71) in the last equality.

Conversely, assume that we have a positive solution of Equation (23) lying in the range (61). Since Equation (23) was derived by squaring both sides of Equation (55), the solution satisfies Equation (55) if the both sides of Equation (55) have the same sign. Clearly, the right-hand side of Equation (55) is positive. We will show that the left-hand side of Equation (55),

$$g(\widehat{\gamma}_h) = -\frac{(M^2+L^2)}{2LM}\widehat{\gamma}_h^2 + \frac{(M-L)^2\gamma_h}{2LM}\widehat{\gamma}_h + \sqrt{\xi_0},$$

is also positive.

Note that $g(\widehat{\gamma}_h)$ is strictly concave because it is a quadratic function with a negative coefficient of the quadratic term. Since we are assuming that the solution lies in the range (61), the following holds:

$$\begin{aligned} g(\widehat{\gamma}_h) &> \min\left\{g(0), g(\xi_0^{1/4})\right\} \\ &> \min\left\{\sqrt{\xi_0}, \frac{(M-L)^2\gamma_h}{2LM}\xi_0^{1/4}(\gamma_h - \xi_0^{1/4})\right\} \\ &> 0. \end{aligned}$$

We used Inequalities (70) in the last inequality. Thus, we have shown that the left-hand side, $g(\widehat{\gamma}_h)$, of Equation (55) is also positive, and therefore, the solution satisfies Equation (55). This means that the solution also satisfies its equivalent equation (51). Since Inequalities (70) imply that the range (61) is included in the range (54), the solution trivially lies in the range (54), which completes the proof of Lemma 9. ■

D.3 Proof of Lemma 10

We will investigate the shape of the quartic function (23),

$$f(\widehat{\gamma}_h) := \widehat{\gamma}_h^4 + \xi_3\widehat{\gamma}_h^3 + \xi_2\widehat{\gamma}_h^2 + \xi_1\widehat{\gamma}_h + \xi_0. \tag{23}$$

Since the coefficient of the quartic term is positive (equal to one), $f(\widehat{\gamma}_h)$ goes to infinity as $\widehat{\gamma}_h \to -\infty$ or $\widehat{\gamma}_h \to \infty$. Since Equation (59) implies that $\xi_0 > 0$, it holds that $f(0) > 0$.

By using Equation (58), we have

$$\begin{aligned} f(\xi_0^{1/4}) &= \xi_0 + \xi_3\xi_0^{3/4} + \xi_2\sqrt{\xi_0} + \xi_1\xi_0^{1/4} + \xi_0 \\ &= \xi_0 + \xi_3\xi_0^{3/4} + \xi_2\sqrt{\xi_0} + \xi_3\xi_0^{3/4} + \xi_0 \\ &= \sqrt{\xi_0}\left(2\sqrt{\xi_0} + 2\xi_3\xi_0^{1/4} + \xi_2\right). \end{aligned}$$

By using Inequalities (70) and Equation (57), this can be bounded as

$$\begin{aligned} f(\xi_0^{1/4}) &< \sqrt{\xi_0}\left(2\eta_h^2 + 2\xi_3\eta_h - \xi_3\gamma_h - \frac{(L^2+M^2)\eta_h^2}{LM}\right) \\ &= \sqrt{\xi_0}\left(2\xi_3\eta_h - \xi_3\gamma_h - \frac{(L-M)^2\eta_h^2}{LM}\right) \\ &= \frac{\sqrt{\xi_0}\xi_3}{\gamma_h}\left(2\eta_h\gamma_h - \gamma_h^2 - \eta_h^2\right) \\ &= -\frac{\sqrt{\xi_0}\xi_3}{\gamma_h}\left(\gamma_h - \eta_h\right)^2 \\ &< 0. \end{aligned}$$

Here, we used Equation (56) in the third equality, and Inequalities (70) in the last inequality.

In summary, we have the following:

$$\begin{aligned} \lim_{\widehat{\gamma}_h \to -\infty} f(\widehat{\gamma}_h) &= \infty, \\ f(0) &> 0, && (72) \\ f(\xi_0^{1/4}) &< 0, && (73) \\ \lim_{\widehat{\gamma}_h \to \infty} f(\widehat{\gamma}_h) &= \infty. && (74) \end{aligned}$$

Furthermore, since Equation (57) implies that $\xi_2 < 0$, $f(\widehat{\gamma}_h)$ has a negative curvature at the origin, i.e., $(\partial^2 f/\partial^2\widehat{\gamma}_h)(0) < 0$. This means that $f(\widehat{\gamma}_h)$ has one inflection point each in the positive region $\widehat{\gamma}_h > 0$ and in the negative region $\widehat{\gamma}_h < 0$. The shape of the quartic function $f(\widehat{\gamma}_h)$ is shown in Figure 9. Note that the points at which $f(\widehat{\gamma}_h)$ crosses the horizontal axis are the solutions of the quartic equation (23).

Inequality (73) and Equation (74) imply that at least one solution exists in the region

$$\widehat{\gamma}_h > \xi_0^{1/4}.$$

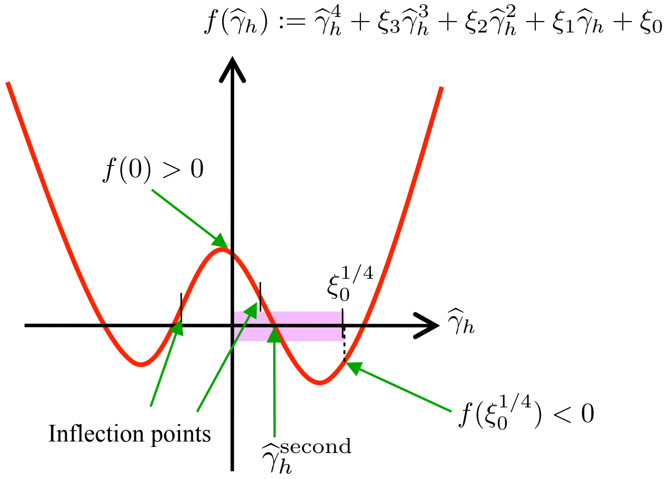


Figure 9: The shape of a quartic function $f(\widehat{\gamma}_h) := \widehat{\gamma}_h^4 + \xi_3\widehat{\gamma}_h^3 + \xi_2\widehat{\gamma}_h^2 + \xi_1\widehat{\gamma}_h + \xi_0$, where $\xi_2 < 0$, $\xi_0 (= f(0)) > 0$, and $f(\xi_0^{1/4}) < 0$. The range $0 < \widehat{\gamma}_h < \xi_0^{1/4}$, where the second largest positive real solution $\widehat{\gamma}_h^{\mathrm{second}}$ exists, is highlighted.

Inequalities (72) and (73) imply that at least one solution exists in the region

$$0 < \widehat{\gamma}_h < \xi_0^{1/4}.$$

Since $f(\widehat{\gamma}_h)$ has only one inflection point in the positive region, it has no more solution in the positive region without contradiction with Inequality (72) (see Figure 9), which completes the proof of Lemma 10. ■

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, second edition, 1984.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Trans. on Signal Processing*, 60(8):3964–3977, 2012.

P. F. Baldi and K. Hornik. Learning in linear neural networks: A survey. *IEEE Transactions on Neural Networks*, 6(4):837–858, 1995.

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48: 259–302, 1986.

C. M. Bishop. Variational principal components. In *Proc. of ICANN*, volume 1, pages 514–509, 1999.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009.

J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 35:283–319, 1970.

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

D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

M. Hazewinkel, editor. *Encyclopaedia of Mathematics*. Springer, 2002.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *JMLR*, 11:1957–2000, 2010.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, Berkeley, CA., USA, 1961. University of California Press.

H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, volume 186, pages 453–461, 1946.

S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of International Conference on Machine Learning*, pages 457–464, 2009.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

Y. J. Lim and T. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, 2007.

D. J. C. Mackay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Available from `http://www.inference.phy.cam.ac.uk/mackay/minima.pdf`. 2001.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications, Second Edition*. Springer, 2009.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

S. Nakajima. Variational Bayesian algorithm for relational tensor factorization. *Under Preparation*, 2012.

S. Nakajima and M. Sugiyama. Theoretical analysis of Bayesian matrix factorization. *Journal of Machine Learning Research*, 12:2579–2644, 2011.

S. Nakajima, M. Sugiyama, and R. Tomioka. Global analytic solution for variational Bayesian matrix factorization. In *Advances in Neural Information Processing Systems 23*, pages 1759–1767, 2010.

S. Nakajima, M. Sugiyama, and S. D. Babacan. Global solution of fully-observed variational Bayesian matrix factorization is column-wise independent. In *Advances in Neural Information Processing Systems 24*, 2011.

S. Nakajima, M. Sugiyama, and S. D. Babacan. Sparse additive matrix factorization for robust PCA and its generalization. In *Proceedings of Fourth Asian Conference on Machine Learning*, 2012a.

S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Perfect dimensionality recovery by variational Bayesian PCA. In *Advances in Neural Information Processing Systems 25*, 2012b.

A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, 2007.

G. R. Reinsel and R. P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York, 1998.

J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine learning*, pages 713–719, 2005.

R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51, Berlin, 2006. Springer.

A. Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10:343–354, 1970.

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1257–1264, Cambridge, MA, 2008. MIT Press.

M. Seeger and G. Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proc. of AISTATS*, La Palma, Spain, 2012.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4): 551–556, 1993.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61:611–622, 1999.

R. Tomioka, T. Suzuki, M. Sugiyama, and H. Kashima. An efficient and general augmented Lagrangian algorithm for learning low-rank matrices. In *Proceedings of International Conference on Machine Learning*, 2010.

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1996.

S. Watanabe. *Algebraic Geometry and Statistical Learning*. Cambridge University Press, Cambridge, UK, 2009.

K. J. Worsley, J-B. Poline, K. J. Friston, and A. C. Evanss. Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage*, 6(4):305–319, 1997.