Automatic Hyperparameter Tuning in Sparse Matrix Factorization

Ryota Kawasumi* Koujin Takeda[†]
4th January, 2023

Abstract

We study the problem of hyperparameter tuning in sparse matrix factorization under Bayesian framework. In the prior work, an analytical solution of sparse matrix factorization with Laplace prior was obtained by variational Bayes method under several approximations. Based on this solution, we propose a novel numerical method of hyperparameter tuning by evaluating the zero point of normalization factor in sparse matrix prior. We also verify that our method shows excellent performance for ground-truth sparse matrix reconstruction by comparing it with the widely-used algorithm of sparse principal component analysis.

Keywords: matrix factorization, sparsity, hyperparameter tuning, variational Bayes analysis

1 Introduction

Among machine learning problems, matrix factorization (MF) is significant because MF appears in many applications such as recommendation system, signal processing, etc. We restrict ourselves to sparse MF problem in this article, where either factorized matrix must be sparse. This is originally discussed as sparse coding in neuroscience [1, 2], and recognized as a significant problem in neuronal information processing in the brain. It also appears in sparse modeling in information science such as dictionary learning [3, 4] or sparse principal component analysis (sparse PCA) [5, 6].

Many attempts have been made so far for understanding theoretical aspects of MF, and analytical tools for random systems in statistical physics are found to be useful, e.g. Markov chain Monte Carlo method [7], replica analysis [8, 9, 10, 11, 12], and message passing [9, 10, 11, 12, 13, 14], where some works are not limited to sparse matrix case. The authors of this article also analyzed sparse MF under Laplace prior (or ℓ_1 regularizer) by variational

^{*}Department of Mathematics, Graduate School of Science and Engineering, Chuo University, e-mail: rykawasumi@gmail.com

[†]Department of Mechanical Systems Engineering, Graduate School of Science and Engineering, Ibaraki University, e-mail: koujin.takeda.kt@vc.ibaraki.ac.jp

Bayes (VB) method for Kullback-Leibler (KL) divergence, or equivalently variational calculus for free energy from the viewpoint of statistical physics [15]. They obtained the analytical solution of sparse MF under several approximations, which serves as sparse MF algorithm. They also experimentally found that ground-truth sparse MF solution can be reconstructed by this algorithm under appropriate hyperparameter value in Laplace prior. However, hyperparameter tuning is generally difficult in machine learning. Several methods based on risk estimate, information criterion, or cross validation have been proposed in well-known Lasso [16, 17, 18, 19], some of which are based on replica analysis or message passing [20, 21, 22]. On the other hand, such method has not been sufficiently discussed for sparse MF, especially under ℓ_1 regularizer.

Here we propose a novel numerical approach for hyperparameter tuning in Laplace prior in sparse MF solution [15]. The essential idea is the zero point of normalization factor. In sparse MF solution, the contribution from Laplace prior is expressed by the correction terms to MF solution under uniform prior. All these terms include normalization factor of probability distribution for sparse matrix in their denominators. Hence, their contribution becomes dominant in the vicinity of the zero point of normalization factor. Based on this idea, we construct an MF algorithm including evaluation of the zero point for hyperparameter tuning. However, sparse MF solution was obtained under several approximations, and our numerical approach does not necessarily lead to sparse MF solution. Nevertheless, our experiment shows that ground-truth sparse MF solution can be reconstructed with high accuracy. We also compare the performance of our algorithm with widely-used sparse PCA algorithm [23]. Consequently, in the case of sparser ground-truth matrix, we find that our method shows better performance than sparse PCA algorithm under hyperparameter tuning by hand.

2 Outline of VB analysis for sparse MF

VB analysis for sparse MF solution is outlined here. In MF, observed matrix $\mathbf{V} \in \mathbb{R}^{L \times M}$ is represented by $\mathbf{V} = \mathbf{A}\mathbf{B} + \mathbf{E}$, where $\mathbf{A} \in \mathbb{R}^{L \times H}$, $\mathbf{B} \in \mathbb{R}^{H \times M}$, and noise matrix $\mathbf{E} \in \mathbb{R}^{L \times M}$. Here matrix is denoted by boldface letter. The task is to find factorized matrices \mathbf{A}, \mathbf{B} from \mathbf{V} , and Bayesian approach is taken for the purpose. We assume multivariate Gaussian prior (or ℓ_2 regularizer) for dense matrix \mathbf{A} and Laplace prior (or ℓ_1 regularizer) for sparse matrix \mathbf{B} ,

$$P(\mathbf{A}) \propto \prod_{l,h} \exp\left(-\frac{a_{lh}^2}{2(C_{\mathbf{A}})_{hh}}\right),$$
 (1)

$$P(\boldsymbol{B}) \propto \prod_{h,m} \exp\left(-\frac{|b_{hm}|}{k}\right),$$
 (2)

where C_A is covariance matrix for matrix A, and k is hyperparameter in Laplace prior. The element in noise matrix E is drawn from Gaussian distribution $\mathcal{N}(0, \sigma^2)$. From Gaussianity of noise, the likelihood for sparse MF model is given by

$$P(\mathbf{V}|\mathbf{A}, \mathbf{B}) \propto \prod_{l,m} \exp\left\{-\frac{1}{2\sigma^2} \left(v_{lm} - \sum_{h} a_{lh} b_{hm}\right)^2\right\}.$$
 (3)

With these priors and the likelihood, the matrices $\boldsymbol{A}, \boldsymbol{B}$ can be estimated from Bayes formula by maximization of posterior, which is however computationally infeasible. In reference [15], VB analysis is used to tackle this problem. In VB analysis, the minimum of KL divergence between trial functions and true posterior for the matrices $\boldsymbol{A}, \boldsymbol{B}$ is evaluated. However, unlike the case of Gaussian prior for both matrices [24, 25], the analytical solution of KL minimum cannot be obtained without approximations. Hence, several approximations are used: mean field approximation, 1/k expansion of $P(\boldsymbol{B})$ up to the first order, and neglect of covariance.

By KL divergence minimization with VB analysis, the expressions of means \bar{a}_{lh} , \bar{b}_{hm} and variances $(\Sigma_{Al})_{hh}$, $(\Sigma_{Bm})_{hh}$ for trial functions of \boldsymbol{A} , \boldsymbol{B} are obtained as below. More precisely, the variables \bar{a}_{lh} , $(\Sigma_{Al})_{hh}$ represent the mean and the variance of lh-element in \boldsymbol{A} , respectively. Similarly, the variables \bar{b}_{hm} , $(\Sigma_{Bm})_{hh}$ denote the mean and the variance of hm-element in \boldsymbol{B} , respectively. The sums $\sum_{l}(\Sigma_{Al})_{hh}$, $\sum_{m}(\Sigma_{Bm})_{hh}$ are the diagonal hh-element of covariance matrix $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{B}\boldsymbol{B}^T$, respectively. See reference [15] for the details of the analysis.

$$\bar{a}_{lh} = \sum_{m,h'} (\hat{\Sigma}_{Al}^{-1})_{hh'} v_{lm} \bar{b}_{h'm},$$
 (4)

$$(\Sigma_{\mathbf{A}l})_{hh} = \sigma^2(\hat{\Sigma}_{\mathbf{A}l}^{-1})_{hh}, \tag{5}$$

$$\bar{b}_{hm} = \sum_{h',l} (\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{hh'} v_{lm} \bar{a}_{lh'} - \sum_{h'} \frac{\sigma^2(\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{h'h}}{kZ_B} \operatorname{erf}(\omega_{h'm}), \tag{6}$$

$$(\Sigma_{\mathbf{B}m})_{hh} = \sigma^{2}(\hat{\Sigma}_{\mathbf{B}m}^{-1})_{hh} - \sum_{h'} \sqrt{\frac{2}{\pi \sigma^{2}(\hat{\Sigma}_{\mathbf{B}m}^{-1})_{h'h'}}} \frac{\{\sigma^{2}(\hat{\Sigma}_{\mathbf{B}m}^{-1})_{h'h}\}^{2}}{kZ_{B}} e^{-\omega_{h'm}^{2}} - \left\{\sum_{h'} \frac{\sigma^{2}(\hat{\Sigma}_{\mathbf{B}m}^{-1})_{h'h}}{kZ_{B}} \operatorname{erf}(\omega_{h'm})\right\}^{2},$$

$$(7)$$

where $\operatorname{erf}(x) = (2/\sqrt{\pi}) \int_x^\infty \exp(-t^2) dt$ and the superscript -1 means inverse matrix. The matrices $\hat{\Sigma}_{Al}$, $\hat{\Sigma}_{Bm} \in \mathbb{R}^{H^2}$ are obtained as

$$(\hat{\Sigma}_{Al})_{hh'} = \frac{\sigma^2}{(C_A)_{hh}} \delta_{h,h'} + \sum_{m} \left((\Sigma_{Bm})_{hh} \delta_{h,h'} + \bar{b}_{hm} \bar{b}_{h'm} \right), \tag{8}$$

$$(\hat{\Sigma}_{\boldsymbol{B}m})_{hh'} = \sum_{l} \left((\Sigma_{\boldsymbol{A}l})_{hh} \delta_{h,h'} + \bar{a}_{lh} \bar{a}_{lh'} \right). \tag{9}$$

The factors ω_{hm} and Z_B are defined by

$$\omega_{hm} = \frac{\sum_{h',l} (\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{hh'} v_{lm} \bar{a}_{lh'}}{\sqrt{2\sigma^2 (\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{hh}}},$$
(10)

$$Z_B = 1 - \frac{1}{k} \sum_{m,h} \left\{ \sqrt{\frac{2\sigma^2(\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{hh}}{\pi}} e^{-\omega_{hm}^2} + \left(\sum_{h',l} (\hat{\Sigma}_{\boldsymbol{B}m}^{-1})_{hh'} v_{lm} \bar{a}_{lh'} \right) \operatorname{erf}\left(\omega_{hm}\right) \right\}.$$

The factor Z_B serves as the normalization factor for the probability $P(\mathbf{B})$. Note that Z_B may become negative or ill-defined for very small k due to 1/k expansion.

By evaluating the means and the matrices in equations (4)-(7) in conjunction with equations (8)-(11) numerically, we can obtain the inferred factorized matrices \mathbf{A}, \mathbf{B} from $\bar{a}_{lh}, \bar{b}_{hm}$. As suggested in reference [15], the performance of sparse matrix reconstruction depends on the value of k, and appropriate value of k will exist.

3 The idea for hyperparameter tuning

In our approach for hyperparameter tuning, we focus on the correction terms in sparse MF solution equations (4)-(7). The key is the zero point of normalization factor Z_B . In numerical experiment in reference [15], the zero point of Z_B is close to optimal k for sparse matrix reconstruction. The reason will be as follows. In equations (6) and (7), all k-dependent terms originated from Laplace prior have $1/kZ_B$ factor. Hence, these terms will become dominant in the vicinity of the zero point of Z_B , whereas they disappear in the limit of $k \to \infty$ or uniform prior case for B. From equation (11), the value of k at the zero point of Z_B is given by

$$k = \sum_{m,h} \left\{ \sqrt{\frac{2\sigma^2(\widehat{\Sigma}_{Bm}^{-1})_{hh}}{\pi}} e^{-\omega_{hm}^2} + \left(\sum_{h',l} v_{lm}(\widehat{\Sigma}_{Bm}^{-1})_{hh'} \bar{a}_{lh'} \right) \operatorname{erf}(\omega_{hm}) \right\}.$$
(12)

Note that the variables $\bar{a}_{lh'}$, $(\widehat{\Sigma}_{Bm}^{-1})_{hh'}$, ω_{hm} on r.h.s. depend on k. Therefore, we need to solve this equation numerically to evaluate the zero point of Z_B .

The simplest idea for the zero point is to iteratively update k by computing r.h.s of equation (12). However, simple iteration for k leads to instability. We should introduce partial update of k for convergence,

$$k^{(t+1)} = (1 - \epsilon)k^{(t)} + \epsilon \sum_{m,h} \left\{ \sqrt{\frac{2\sigma^2(\widehat{\Sigma}_{Bm}^{-1})_{hh}}{\pi}} e^{-\omega_{hm}^2} + \left(\sum_{h',l} v_{lm}(\widehat{\Sigma}_{Bm}^{-1})_{hh'} \bar{a}_{lh'} \right) \operatorname{erf}(\omega_{hm}) \right\},$$
(13)

where ϵ is an arbitrary small constant for partial update, and the superscript (t) denotes iteration step.

By considering equations (4)-(11) and (13) as an iterative algorithm, we can construct a sparse MF algorithm as in algorithm 1. For the termination condition of the algorithm, we use the threshold Z_{Bthres} for normalization factor Z_B .

We should remember that sparse MF solution equations (4)-(7) is obtained by the first order approximation of 1/k, which may lose the advantage of ℓ_1 regularizer for sparsity. Therefore, there is no guarantee that our approach will lead to sparse MF solution. We should verify the validity of our approach carefully.

```
Algorithm 1 iterative MF algorithm with hyperparameter tuning for sparse prior
```

$$\begin{array}{l} \text{set initial } k^{(0)} \text{ and } Z_B^{(0)} \text{ very large} \\ \text{initialize } \bar{a}_{lh}^{(0)}, \bar{b}_{hm}^{(0)} \text{ randomly } \forall l, h, m \\ \text{initialize } \Sigma_{Bm}^{(0)} \text{ as identity matrix } \forall m \\ \\ \textbf{while } Z_B^{(t)} > Z_{B\text{thres}} \text{ do} \\ (\hat{\Sigma}_{Al})_{hh'}^{(t+1)} &= \frac{\sigma^2}{(C_A)_{hh}} \delta_{h,h'} + \sum_m \left((\Sigma_{Bm})_{hh}^{(t)} \delta_{h,h'} + \bar{b}_{hm}^{(t)} \bar{b}_{h'm}^{(t)} \right), \ \forall h, h' \\ (\Sigma_{Al})_{hh'}^{(t+1)} &= \sigma^2 (\hat{\Sigma}_{Al}^{-1})_{hh}^{(t+1)}, \ \forall h \\ \bar{a}_{lh}^{(t+1)} &= \sum_{m,h'} (\hat{\Sigma}_{Al}^{-1})_{hh'}^{(t+1)}, \ \forall l, h \\ (\hat{\Sigma}_{Bm})_{hh'}^{(t+1)} &= \sum_{l} \left((\Sigma_{Al})_{hh}^{(t+1)} \delta_{h,h'} + \bar{a}_{lh}^{(t+1)} \bar{a}_{lh'}^{(t+1)} \right), \ \forall h, h' \\ \\ \omega_{hm}^{(t+1)} &= \sum_{l} \left((\Sigma_{Al})_{hh}^{(t+1)} \delta_{h,h'} + \bar{a}_{lh}^{(t+1)} \bar{a}_{lh'}^{(t+1)} \right), \ \forall h, m \\ \\ k^{(t+1)} &= (1-\epsilon)k^{(t)} + \epsilon \left\{ \sum_{m,h} \left(\sqrt{\frac{2\sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)}}{2}} e^{-(\omega_{hm}^{(t+1)})^2} \right) \\ &+ \left(\sum_{h',l} (\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)} v_{lm} \bar{a}_{lh'}^{(t+1)} \right) \exp \left((\omega_{hm}^{(t+1)}) \right) \right\} \\ Z_B^{(t+1)} &= 1 - \frac{1}{k^{(t+1)}} \sum_{m,h} \left\{ \sqrt{\frac{2\sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)}}{\pi}} e^{-(\omega_{hm}^{(t+1)})^2} \\ &+ \left(\sum_{h',l} (\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)} v_{lm} \bar{a}_{lh'}^{(t+1)} \right) \exp \left((\omega_{hm}^{(t+1)}) \right) \right\} \\ (\Sigma_{Bm})_{hh}^{(t+1)} &= \sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh}^{(t+1)} - \sum_{h'} \sqrt{\frac{2\sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)}}{\pi}} e^{-(\omega_{hm}^{(t+1)})^2} \\ &- \left\{ \sum_{h'} \frac{\sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)}}{k^{(t+1)}Z_B^{(t+1)}} \exp \left((\omega_{h'm}^{(t+1)}) \right)^2, \ \forall h \\ \bar{b}_{hm}^{(t+1)} &= \sum_{h',l} (\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)} v_{lm} \bar{a}_{lh'}^{(t)} - \sum_{h'} \frac{\sigma^2(\hat{\Sigma}_{Bm}^{-1})_{hh'}^{(t+1)}}{k^{(t+1)}Z_B^{(t+1)}} \exp \left((\omega_{h'm}^{(t+1)}), \ \forall h, m \\ t \longrightarrow t+1 \end{array} \right.$$

end while

The result of MF is obtained from $\bar{a}_{lh}^{(t)}, \bar{b}_{hm}^{(t)} \, \forall l, h, m$.

4 Numerical experiment

For evaluation of sparse MF performance, we numerically reconstruct ground-truth factorized matrix A^* , B^* by algorithm 1. The setup of our numerical experiment for synthetic data is as follows. Each element in A^* , B^* is drawn from standard Gaussian distribution $\mathcal{N}(0,1)$ and Bernoulli-Gaussian distribution $P(B_{hm}^*) = (1-\rho)\delta(B_{hm}^*) + \rho \exp(-(B_{hm}^*)^2/2)/\sqrt{2\pi} \ \forall h, m$, respectively. The parameter ρ describes sparsity of B^* . The covariance matrix C_A is set to be identity. In the experiments of synthetic data we set $\epsilon = 0.1$ and L = M = 500. We set $Z_{B\text{thres}} = 10^{-5}$ and the result is averaged over 20 trials excepting figure 1. Noise magnitude σ is set to be 0.05 in figures 1, 2, and 4.

We should note that sparse MF has degenerate solutions: V is invariant under permutation of h in A_{lh}^*, B_{hm}^* and scaling $\{A_{lh}^*, B_{hm}^*\} \rightarrow \{c_h A_{lh}^*, B_{hm}^*/c_h\} \ \forall l, m$, where c_h is an arbitrary constant. Therefore, for appropriate measure of ground-truth matrix reconstruction, we define rooted mean squared error (RMSE) between ground-truth matrices and reconstructed ones as follows.

$$RMSE_{\mathbf{A}} = \sqrt{\frac{1}{LH} \sum_{h} \left(\min_{h', s_h} \sum_{l} \left(\mathbf{A}_{lh}^* - \frac{s_h}{N_{h'}} \bar{a}_{lh'} \right)^2 \right)}, \tag{14}$$

$$RMSE_{\boldsymbol{B}} = \sqrt{\frac{1}{HM}} \sum_{h} \left(\min_{h'} \sum_{m} \left(\boldsymbol{B}_{hm}^* - s_h' N_{h'} \bar{b}_{h'm} \right)^2 \right), \tag{15}$$

where

$$N_{h'} = \sqrt{\frac{\sum_{l'} (\bar{a}_{l'h'})^2}{L}}, \tag{16}$$

$$s_h' = \underset{s_h}{\operatorname{argmin}} \left(\min_{h'} \sum_{l} \left(\boldsymbol{A}_{lh}^* - \frac{s_h}{N_{h'}} \bar{a}_{lh'} \right)^2 \right). \tag{17}$$

The normalization factor $N_{h'}$ is necessary for comparison between reconstructed and ground-truth matrix elements. Remember that ℓ_2 -norm of column vector in \mathbf{A}^* is approximately \sqrt{L} for $L \to \infty$. The factor $s_h \in \{\pm 1\}$ is for removal of sign ambiguity from scale invariance. Minimization with respect to h' is for correct one-to-one correspondence between vectors in ground-truth/reconstructed matrices, because MF has permutation invariance as stated above. Similarly, for sparsity measure, element in reconstructed \mathbf{B} is regarded as sparse if its absolute value is smaller than 10^{-2} after scaling the matrix as $\{\bar{a}_{lh}, \bar{b}_{hm}\} \to \{\bar{a}_{lh}/N'_h, N'_h \bar{b}_{hm}\}$ $\forall l, m$ for normalizing \mathbf{A} . Then the sparsity of \mathbf{B} is measured by the fraction of sparse element in \mathbf{B} .

In addition, we also define rooted mean squared error for multiplied matrix for validity of sparse MF solution,

$$RMSE_{\mathbf{V}} = \sqrt{\frac{1}{LM} \sum_{l,m} \left(\mathbf{V}_{lm} - \sum_{h} \bar{a}_{lh} \bar{b}_{hm} \right)^2}.$$
 (18)

 $RMSE_{\mathbf{V}}$ will be equal to noise magnitude σ when MF is successful.

First, a typical dynamical behavior of our algorithm under $\rho = 0.8, H = 20$ is depicted in figure 1. As Z_B gets close to zero, RMSE_A and RMSE_B decrease, and sparsity of B

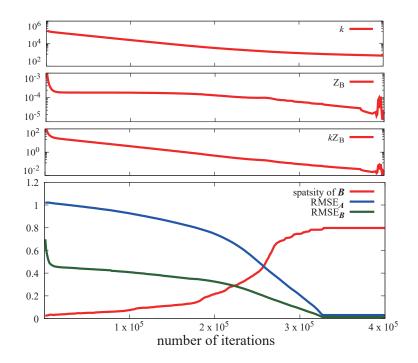


Figure 1: A typical dynamical behavior of the proposed sparse MF algorithm.

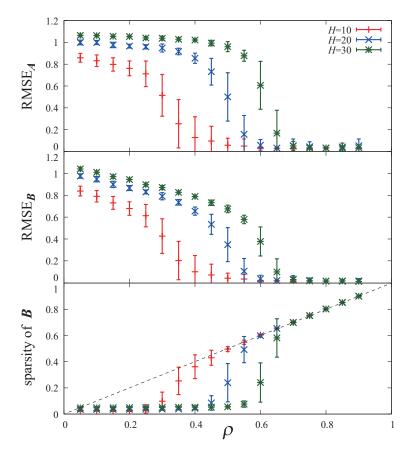


Figure 2: The result of sparse matrix reconstruction: $RMSE_{\boldsymbol{A}}$ (top), $RMSE_{\boldsymbol{B}}$ (middle), and sparsity of reconstructed \boldsymbol{B} (bottom) are shown.

approaches the original value of ρ (= 0.8). This means good reconstruction performance of our algorithm. At late stage of iteration, unstable behavior of Z_B is observed, therefore our algorithm should be terminated before such instability occurs. For comparison, we also conduct the experiment under fixed hyperparameter k like in the prior work [15], where the update of k in algorithm 1 is removed. However, we find that reconstruction performance is poor. The MF experiment under fixed k indicates that the sparsity of B in MF solution remains almost constant against the change of k even after one million of iterations, whose value is much smaller than the ground-truth sparsity ρ . This implies that update of k is essential for ground-truth matrix reconstruction with high accuracy.

Next, we conduct an experiment with sparsity ρ and dimension H being varied. The result in figure 2 indicates the threshold of ρ for crossover to nearly-perfect reconstructable region. For larger H, many sparse MF solutions other than the ground-truth will exist due to larger number of factorized matrix elements, which will make ground-truth matrix reconstruction more difficult and lead to larger threshold value. Note that transition between perfect/imperfect reconstruction phases under noiseless case and behavior of RMSE are analyzed in prior works [8, 9, 10, 11, 12, 14], however not under ℓ_1 regularizer. For robustness to noise, we also observe σ dependence under $\rho = 0.8, H = 20$. In this case, standard deviation of element in matrix A^*B^* is evaluated as 2.01 ± 0.04 for comparison with σ or for signal-to-noise ratio. The result in figure 3 means that ground-truth solution can be found with high accuracy even in relatively noisy case.

We also compare the performance of our algorithm with widely-used sparse PCA algorithm [23], where Lasso for sparse \boldsymbol{B} and dictionary estimation for \boldsymbol{A} are performed alternately. This is implemented in scikit-learn library in Python and easily accessible. In this experiment Python version 2.7 is used. For sparse PCA, we conduct 10^4 iterations and vary the coefficient of ℓ_1 regularizer, denoted by α in figure 4. Default values are used for other hyperparameters. We evaluate RMSE_{\boldsymbol{A}}, RMSE_{\boldsymbol{B}}, and RMSE_{\boldsymbol{V}} for both algorithms under H=20 as shown in figure 4. The result shows that reconstruction performance of our algorithm is better than sparse PCA for larger ρ . Another advantage of our algorithm is that RMSE_{\boldsymbol{V}} is kept constant and almost equal to σ (0.05 in the present case) regardless of the value ρ , while RMSE_{\boldsymbol{V}} increases for smaller ρ in sparse PCA. In our algorithm, we assume that the value of noise magnitude σ is known in advance, which will lead to constant RMSE_{\boldsymbol{V}}. However, it should be emphasized that such constant behavior of RMSE_{\boldsymbol{V}} is nontrivial because VB solution for sparse MF was obtained under several approximations.

Finally, for verifying practical utility of our algorithm, we also conduct an experiment for extraction of dictionary in real image. In this experiment, we use four monochrome images of 256×256 pixels (L = M = 256) in Volume 3 in The USC-SIPI Image Database [26], namely Tree (image # 4.1.06), Moon Surface (5.1.09), Aerial (5.1.10), and Clock (5.1.12). Before applying MF, the values of each pixel are normalized to have zero mean and one variance, which is regarded as the observed matrix \mathbf{V} in our formulation. For comparison, we apply two MF methods, our algorithm and sparse PCA algorithm. In applying our algorithm to real image we set $\epsilon = 10^{-3}$, $Z_{B thres} = 10^{-5}$ and the result is averaged over 5 trials by changing initial random factorized matrices. For sparse PCA, we conduct 10^4 iterations and vary the ℓ_1 regularizer coefficient α . The results of RMSE $_{\mathbf{V}}$ and sparsity of \mathbf{B} are evaluated for both algorithms.

The result is shown in figure 5. We mainly conduct the experiment under H=40 and

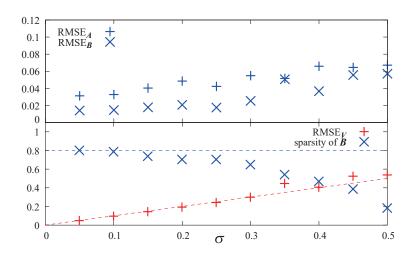


Figure 3: The dependence on noise magnitude σ .

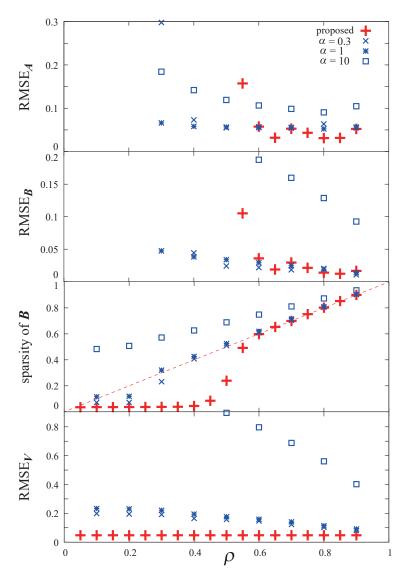


Figure 4: Comparison of performance with sparse PCA algorithm in [23]: From top to bottom, $RMSE_{A}$, $RMSE_{B}$, sparsity of reconstructed B, and $RMSE_{V}$ are shown.

without noise. In the application to the noiseless case, we cannot set the noise parameter $\sigma=0$ in our algorithm, then we set $\sigma=0.03$ for the MF solution. In sparse PCA, the results for all four images indicate that RMSE decreases for smaller α and has almost a constant minimum value below a certain value of α , whereas the sparsity of \boldsymbol{B} constantly decreases. In contrast, almost minimum RMSE $_{\boldsymbol{V}}$ is obtained by our algorithm without tuning hyperparameter. Furthermore, it appears that the sparsity of \boldsymbol{B} by our algorithm is close to the largest value within the region of almost-constant minimum RMSE $_{\boldsymbol{V}}$ by sparse PCA. Such behaviors do not change even under different H (H=20) or the noisy case ($\sigma=0.1$ in \boldsymbol{E} , where we also set $\sigma=0.1$ in our algorithm).

This result implies that our algorithm might be able to find the MF solution having the nearly sparsest B within the region of nearly minimum $RMSE_V$. This is the significant advantage of our algorithm because we do not need to tune hyperparameter for sparse MF, namely the coefficient of ℓ_1 regularizer. In addition, this result also suggests that our algorithm excellently works even for real data. This fact is nontrivial because the elements in factorized matrices A, B should be correlated in real image, whereas we assume i.i.d prior for A, B in our formulation.

5 Summary

We proposed a sparse MF algorithm including hyperparameter tuning in Laplace prior. Surprisingly, although VB solution for sparse MF is derived under several approximations, our algorithm is successful for finding ground-truth sparse MF solution with high accuracy. We also found that our algorithm shows the excellent performance for extracting dictionary in real image.

Here we only numerically verified the performance of our algorithm, and several problems remain unresolved. For example, we need to analyze reconstruction performance and convergence condition theoretically. However it will be difficult due to the complex expression of VB solution. The analysis of VB solution under Gaussian prior [24, 25], which is much simpler than the present case, will help us to understand performance and dynamics of our algorithm. Next problem is partial update of hyperparameter for convergence. In general, very small update parameter ϵ leads to stable convergence, while it makes the algorithm very slow. Therefore, we need to find the strategy for faster convergence. Finally, our method for hyperparameter tuning is completely novel, and future application may give new insights to other sparse modeling problems.

Acknowledgments

We appreciate comments from Tomoki Tamai. This work is supported by KAKENHI Nos. 18K11175, 19K12178, 20H05774, and 20H05776.

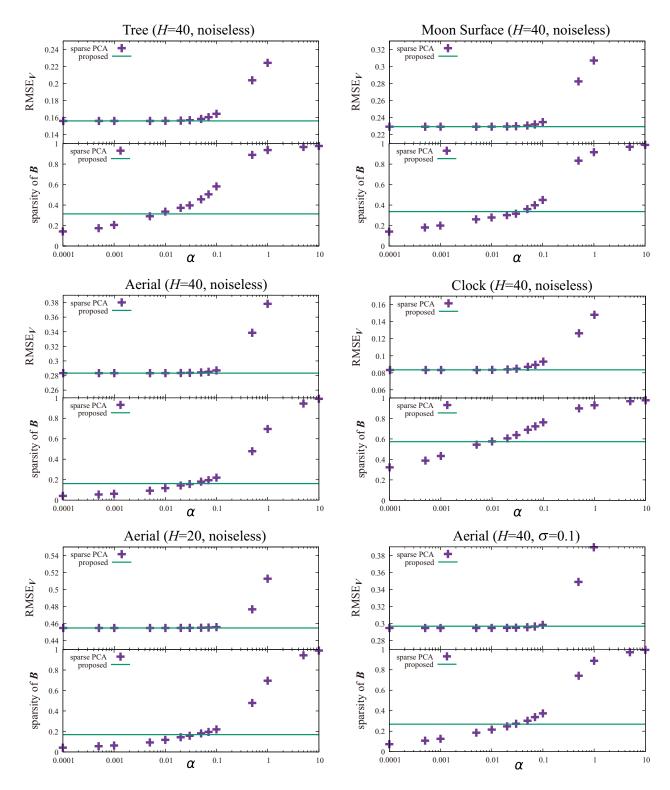


Figure 5: Comparison of performance with sparse PCA algorithm for real image: The results for Tree (top left, H=40, noiseless), Moon Surface (top right, H=40, noiseless), Aerial (middle left, H=40, noiseless), Clock (middle right, H=40, noiseless), Aerial under smaller H (bottom left, H=20, noiseless), and noisy Aerial (bottom right, H=40, $\sigma=0.1$) are shown.

References

- [1] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607 609, 1996.
- [2] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? Vis. Res., 37(23):3311 3325, 1997.
- [3] K. Engan, S. O. Aase, and J. H. Husoy. Method of optimal directions for frame design. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2443 – 2446, March 1999.
- [4] M. Aharon, M. Elad, and A. M. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311 4322, November 2006.
- [5] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. J. Comput. Graph. Stat., 15:265 – 286, 2006.
- [6] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert R G Lanckriet. A direct formulation for sparse pca using semidefinite programming. SIAM Rev., 49:434 – 448, 2007.
- [7] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proc. of the International Conference on Machine Learning*, pages 880 887, 2008.
- [8] A. Sakata and Y. Kabashima. Statistical mechanics of dictionary learning. *EPL* 103:28008, 2013.
- [9] Florent Krzakala, Marc Mézard, and Lenka Zdeborovà. Phase diagram and approximate message passing for blind calibration and dictionary learning. In *Proc. of IEEE International Symposium on Information Theory*, pages 659 663, 2013.
- [10] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborovà. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Trans. Inf. Theory*, 62(7):4228 – 4265, July 2016.
- [11] Christophe Schulke, Philip Schniter, and Lenka Zdeborovà. Phase diagram of matrix compressed sensing. *Phys. Rev. E*, 94:062136, December 2016.
- [12] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborovà. Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *J. Stat. Mech.*, page 073403, 2017.
- [13] Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In *Proc. of Advances in Neural Information Processing Systems*, pages 917 925, 2013.

- [14] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborovà. Phase transitions in sparse pca. In *Proc. of IEEE International Symposium on Information Theory*, pages 1635 1639, 2015.
- [15] R. Kawasumi and K. Takeda. Approximate method of variational bayesian matrix factorization/completion with sparse prior. J. Stat. Mech., page 053404, 2018.
- [16] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.*, 35:2173 2192, 2007.
- [17] Charles Dossal, Maher Kachour, Jalal M. Fadili, Gabriel Peyré, and Christophe Chesneau. The degrees of freedom of the lasso for general design matrix. Stat. Sin., 23:809 828, 2013.
- [18] Samuel Vaiter, Charles-Alban Deledalle, Jalal M. Fadili, Gabriel Peyré, and Charles Dossal. The degrees of freedom of partly smooth regularizers. *Ann. Inst. Stat. Math.*, 69:791 832, 2017.
- [19] Shuaiwen Wang, Wenda Zhou, Haihao Lu, Arian Maleki, and Vahab Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In Proc. of the International Conference on Machine Learning, pages 5228 5237, 2018.
- [20] Mohsen Bayati, Murat A. Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. In *Proc. of Advances in Neural Information Processing Systems*, pages 944 – 952, 2013.
- [21] Tomoyuki Obuchi and Yoshiyuki Kabashima. Cross validation in lasso and its acceleration. J. Stat. Mech., page 053304, 2016.
- [22] Ali Mousavi, Arian Maleki, and Richard G. Baraniuk. Consistent parameter estimation for lasso and approximate message passing. *Ann. Stat.*, 45:2427 2454, 2017.
- [23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. of the International Conference on Machine Learning*, pages 689 – 696, 2009.
- [24] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res., 11:1957 2000, 2010.
- [25] S. Nakajima and M. Sugiyama. Theoretical analysis of bayesian matrix factorization. *J. Mach. Learn. Res.*, 12:2583 2648, 2011.
- [26] Signal and Image Processing Institute. The USC-SIPI Image Database: Version 6, 2018.