Analysis of Empirical MAP and Empirical Partially Bayes: Can They be Alternatives to Variational Bayes?

Shinichi Nakajima
Nikon Corporation

Masashi Sugiyama
Tokyo Institute of Technology

Abstract

Variational Bayesian (VB) learning is known to be a promising approximation to Bayesian learning with computational efficiency. However, in some applications, e.g., large-scale collaborative filtering and tensor factorization, VB is still computationally too costly. In such cases, looser approximations such as MAP estimation and partially Bayesian (PB) learning, where a part of the parameters are point-estimated, seem attractive. In this paper, we theoretically investigate the behavior of the MAP and the PB solutions of matrix factorization. A notable finding is that the global solutions of MAP and PB in the empirical Bayesian scenario, where the hyperparameters are also estimated from observation, are trivial and useless, while their local solutions behave similarly to the global solution of VB. This suggests that empirical MAP and empirical PB with *local search* can be alternatives to empirical VB equipped with the useful automatic relevance determination property. Experiments support our theory.

1 INTRODUCTION

In probabilistic models where Bayesian learning is computationally intractable, variational Bayesian (VB) approximation (Attias, 1999) is a promising alternative equipped with the useful automatic relevance determination (ARD) property (Neal, 1996). VB has experimentally shown its good performance in many applications (Bishop, 1999; Ghahramani and Beal, 2001; Jaakkola and Jordan, 2000; Barber and Chiappa, 2006; Lim and Teh, 2007; Ilin and Raiko, 2010), and its model selection accuracy has been theoretically guaranteed (Nakajima et al., 2012) in fully-observed Bayesian matrix factorization (MF) (Salakhutdinov and Mnih, 2008) or probabilistic PCA (Tipping and Bishop, 1999; Roweis and Ghahramani, 1999).

Generally, VB is solved with efficient iterative local search algorithms. However, in some applications where even VB is computationally too costly, looser approximations, where all or a part of the parameters are point-estimated, with less computation costs are attractive alternatives. For example, Chu and Ghahramani (2009) applied partially Bayesian (PB) learning, where the core tensor is integrated out and the factor matrices are point-estimated, to Tucker factorization (Kolda and Bader, 2009; Carroll and Chang, 1970; Harshman, 1970; Tucker, 1996). Mørup and Hansen (2009) applied the MAP estimation to Tucker factorization with the empirical Bayesian procedure, i.e., the hyperparameters are also estimated from observation. The empirical MAP estimation, with the same order of computation costs as the ordinary alternating least squares algorithm (Kolda and Bader, 2009), showed its model selection capability through the ARD property.

On the other hand, it was shown that, in fully-observed MF, the objective function for empirical PB and empirical MAP is lower-unbounded at the origin, which implies that their global solutions are trivial and useless (Nakajima and Sugiyama, 2011; Nakajima et al., 2011). Here, a question arises: Is there any essential difference between those factorization models?

This paper answers to the question. We theoretically investigate the behavior of the local solutions of empirical PB and empirical MAP in fully-observed MF. More specifically, we obtain an analytic-form of the local solutions. A notable finding is that, although the global solutions of empirical PB and empirical MAP are useless, the local solutions behave similarly to the global solution of empirical VB. Experiments support our theory.

We also investigate empirical PB and empirical MAP in collaborative filtering (or partially observed MF)

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

and tensor factorization. We theoretically show that their global solutions are also trivial and useless, but experimentally show that, with *local search*, they work similarly to empirical VB.

2 BACKGROUND

In this section, we formulate the matrix factorization model, and introduce the free energy minimization framework, which contains VB, PB, and MAP.

2.1 Probabilistic Matrix Factorization

Assume that an observed matrix $V \in \mathbb{R}^{L\times M}$ consists of 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 *matrix factorization* (MF), the target matrix is assumed to be low rank, and can be factorized as

$$U = BA^\top,$$

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

In this paper, we consider the probabilistic MF model (Salakhutdinov and Mnih, 2008), where the observation noise $\mathcal{E}$ and the priors of A and B are assumed to be Gaussian:

$$p(V|A,B) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|V - BA^\top\|_{\mathrm{Fro}}^2\right), \quad (1)$$
$$p(A) \propto \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(AC_A^{-1}A^\top\right)\right), \quad (2)$$
$$p(B) \propto \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(BC_B^{-1}B^\top\right)\right). \quad (3)$$

Here, we denote by $\top$ the transpose of a matrix or vector, by $\|\cdot\|_{\mathrm{Fro}}$ the Frobenius norm, and by $\mathrm{tr}(\cdot)$ the trace of a matrix. We assume that the prior covariance matrices C_A and C_B are diagonal and positive definite, i.e.,

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

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_AC_B are arranged in the non-increasing order, i.e., $c_{a_h}c_{b_h} \geq c_{a_{h'}}c_{b_{h'}}$ 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 Free Energy Minimization Framework for Approximate Bayesian Inference

The Bayes posterior is given by

$$p(A, B|V) = \tfrac{p(V|A,B)p(A)p(B)}{p(V)}, \quad (4)$$

where $p(V) = \langle p(V|A,B)\rangle_{p(A)p(B)}$. Here, $\langle\cdot\rangle_p$ denotes the expectation over the distribution p. Since this expectation is hard to compute, many approximation methods have been proposed, including sampling methods (Chen et al., 2001) and deterministic methods (Attias, 1999; Minka, 2001). This paper focuses on deterministic approximation methods.

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

$$F(r) = \left\langle \log \tfrac{r(A,B)}{p(V|A,B)p(A)p(B)} \right\rangle_{r(A,B)} \quad (5)$$
$$= \left\langle \log \tfrac{r(A,B)}{p(A,B|V)} \right\rangle_{r(A,B)} - \log p(V).$$

In the last equation, the first term 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 a distribution closest to the Bayes posterior in the sense of the KL distance. A general approach to approximate Bayesian inference is to find the minimizer of the free energy (5) with respect to r in some restricted function space. In the following subsections, we review three types of approximation methods with different restricted function spaces.

Let $\widehat{r}$ be such a minimizer. We define the MF solution by the mean of the target matrix U:

$$\widehat{U} = \left\langle BA^\top \right\rangle_{\widehat{r}(A,B)}. \quad (6)$$

The hyperparameters (C_A, C_B) can also be estimated by minimizing the free energy:

$$(\widehat{C}_A, \widehat{C}_B) = \mathrm{argmin}_{C_A,C_B}\left(\min_r F(r; C_A, C_B)\right).$$

This approach is referred to as the *empirical Bayesian* procedure, on which this paper focuses.

Let

$$V = \textstyle\sum_{h=1}^H \gamma_h \boldsymbol{\omega}_{b_h}\boldsymbol{\omega}_{a_h}^\top \quad (7)$$

be the singular value decomposition (SVD) of V. In the three approximate Bayesian methods discussed in this paper, the MF solution can be written as truncated shrinkage SVD, i.e,

$$\widehat{U} = \sum_{h=1}^H \widehat{\gamma}_h \boldsymbol{\omega}_{b_h}\boldsymbol{\omega}_{a_h}^\top, \text{ where } \widehat{\gamma}_h = \begin{cases} \breve{\gamma}_h & \text{if } \gamma_h \geq \underline{\gamma}_h, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Here, $\underline{\gamma}_h$ is a truncation threshold and $\breve{\gamma}_h$ is a shrinkage estimator, respectively, both of which depend on the approximation method.

2.3 Empirical Variational Bayes

In the VB approximation, the independence between the entangled parameter matrices A and B is assumed:

$$r^{\mathrm{VB}}(A,B) = r_{\mathrm{A}}^{\mathrm{VB}}(A) r_{\mathrm{B}}^{\mathrm{VB}}(B). \tag{9}$$

Under this constraint, a tractable local search algorithm for minimizing the free energy (5) was derived (Bishop, 1999; Lim and Teh, 2007). Furthermore, Nakajima et al. (2012) have recently derived an analytic-form of the global solution when the observed matrix V has no missing entry:

Proposition 1 *(Nakajima et al., 2012) Let*

$$\underline{K} = \min(L,M), \quad \overline{K} = \max(L,M), \quad \alpha = \underline{K}/\overline{K}.$$

Let $\underline{\kappa} = \underline{\kappa}(\alpha)$ (> 1) be the zero-cross point of the following decreasing function:

$$\Xi(\kappa;\alpha) = \Phi(\sqrt{\alpha}\kappa) + \Phi\left(\tfrac{\kappa}{\sqrt{\alpha}}\right),$$
$$\text{where } \Phi(x) = \tfrac{\log(x+1)}{x} - \tfrac{1}{2}.$$

Then, the empirical VB solution is given by Eq.(8) *with the following truncation threshold and shrinkage estimator:*

$$\underline{\gamma}_h^{\mathrm{EVB}} = \sigma\sqrt{M+L+\sqrt{LM}\left(\underline{\kappa}+\tfrac{1}{\underline{\kappa}}\right)}, \tag{10}$$

$$\breve{\gamma}_h^{\mathrm{EVB}} = \tfrac{\gamma_h}{2}\left(1-\tfrac{(M+L)\sigma^2}{\gamma_h^2}+\sqrt{\left(1-\tfrac{(M+L)\sigma^2}{\gamma_h^2}\right)^2-\tfrac{4LM\sigma^4}{\gamma_h^4}}\right). \tag{11}$$

2.4 Empirical Partially Bayes

Point-estimation amounts to approximating the distribution with the delta function. PB-A learning point-estimates B, while PB-B learning point-estimates A, respectively, i.e.,

$$r^{\text{PB-A}}(A,B) = r_{\mathrm{A}}^{\text{PB-A}}(A)\delta(B;\widehat{B}), \tag{12}$$

$$r^{\text{PB-B}}(A,B) = \delta(A;\widehat{A}) r_{\mathrm{B}}^{\text{PB-B}}(B), \tag{13}$$

where $\delta(B;\widehat{B})$ denotes the (*pseudo*-)Dirac delta function of B located at $B=\widehat{B}$.[1] PB learning chooses the one, giving a lower free energy, of the PB-A and the PB-B solutions (Nakajima et al., 2011).

Analyzing the free energy, Nakajima et al. (2011) showed that the global solution of empirical PB is useless:

[1] By the *pseudo*-Dirac delta function, we mean an extremely localized density function, e.g., $\delta(B;\widehat{B}) \propto \exp\left(-\frac{\|B-\widehat{B}\|_{\mathrm{Fro}}^2}{2\varepsilon}\right)$ with a very small but strictly positive variance $\varepsilon > 0$, such that its tail effect can be neglected, while its negative entropy $\chi_B = \langle \log\delta(B;\widehat{B})\rangle_{\delta(B;\widehat{B})}$ remains finite.

Proposition 2 *(Nakajima et al., 2011) The empirical PB solution is $\widehat{\gamma}_h^{EPB} = 0$, regardless of observation.*

2.5 Empirical MAP

In the maximum a posteriori (MAP) estimation, all the parameters are point-estimated, i.e.,

$$r^{\mathrm{MAP}}(A,B) = \delta(A;\widehat{A})\delta(B;\widehat{B}). \tag{14}$$

The global solution of empirical MAP is also useless:

Proposition 3 *(Nakajima and Sugiyama, 2011) The empirical MAP solution is $\widehat{\gamma}_h^{EMAP} = 0$, regardless of observation.*

3 THEORETICAL ANALYSIS

As explained in Section 2, previous theoretical work proved that the global solutions of empirical PB and empirical MAP are trivial in fully-observed MF. In this section, we however show that empirical PB and empirical MAP have non-trivial *local* solutions, which behave like the global solution of empirical VB.

3.1 Analytic-forms of Non-trivial Local Solutions

Due to the diagonality, proven by Nakajima et al. (2013), of the VB posterior covariances, the VB posterior can be expressed as

$$r(A,B) \propto \textstyle\prod_{h=1}^{H} \exp\left(-\frac{\|\boldsymbol{a}_h-\widehat{\boldsymbol{a}}_h\|^2}{2\sigma_{a_h}^2} - \frac{\|\boldsymbol{b}_h-\widehat{\boldsymbol{b}}_h\|^2}{2\sigma_{b_h}^2}\right), \tag{15}$$

with $\widehat{\boldsymbol{a}}_h = a_h\boldsymbol{\omega}_{a_h}$, and $\widehat{\boldsymbol{b}}_h = b_h\boldsymbol{\omega}_{b_h}$. Then, the free energy (5) can be explicitly written as follows:

$$F = \tfrac{1}{2}\left(LM\log(2\pi\sigma^2) + \tfrac{\|V\|_{\mathrm{Fro}}^2}{\sigma^2} + \textstyle\sum_{h=1}^{H} 2F_h\right), \tag{16}$$

where

$$2F_h = M\log\tfrac{c_{a_h}^2}{\sigma_{a_h}^2} + L\log\tfrac{c_{b_h}^2}{\sigma_{b_h}^2} + \tfrac{a_h^2+M\sigma_{a_h}^2}{c_{a_h}^2} + \tfrac{b_h^2+L\sigma_{b_h}^2}{c_{b_h}^2}$$
$$-(L+M) + \tfrac{-2a_hb_h\gamma_h+\left(a_h^2+M\sigma_{a_h}^2\right)\left(b_h^2+L\sigma_{b_h}^2\right)}{\sigma^2}. \tag{17}$$

The free energy of PB and MAP can also be expressed by Eq.(16) with some factors in Eq.(17) smashed: $\sigma_{b_h}^2 \to \varepsilon$ for a very small constant $\varepsilon > 0$ in PB-A and MAP, and $\sigma_{a_h}^2 \to \varepsilon$ in PB-B and MAP. Below, we neglect the large constants related to the entropy, i.e., $-L\log\sigma_{b_h}^2$ in PB-A and MAP, and $-M\log\sigma_{a_h}^2$ in PB-B and MAP. We fix the ratio between the hyper-parameters to $c_{a_h}/c_{b_h} = 1$, since the ratio does not affect the estimation of the other parameters.

We can give an intuition of Proposition 2 and Proposition 3, i.e., why the global solutions of empirical PB and empirical MAP are useless. Consider the empirical PB-A solution, which minimizes the PB-A free energy:

$$2F_h^{\text{PB-A}} = M \log \frac{c_{a_h}^2}{\sigma_{a_h}^2} + L \log c_{b_h}^2 + \frac{a_h^2 + M\sigma_{a_h}^2}{c_{a_h}^2} + \frac{b_h^2}{c_{b_h}^2} + \frac{-2a_h b_h \gamma_h + \left(a_h^2 + M\sigma_{a_h}^2\right) b_h^2}{\sigma^2} + \text{const.} \quad (18)$$

Clearly, Eq.(18) diverges to $-\infty$ as $c_{a_h} c_{b_h} \to 0$ with $a_h = b_h = 0, \sigma_{a_h}^2 = c_{a_h} c_{b_h}$, and therefore,

$$(\widehat{c}_{a_h} \widehat{c}_{b_h})^{\text{EPB-A}} \to 0 \quad (19)$$

is the global solution. The same applies to empirical PB-B and empirical MAP, which leads to the useless trivial estimators, i.e.,

$$\widehat{\gamma}_h^{\text{EPB}} = \widehat{\gamma}_h^{\text{EMAP}} = 0. \quad (20)$$

Note that this phenomenon is caused by smashing a part of the posterior, which makes the infinitely large entropy constant: In empirical VB without smashing, the first four terms in Eq.(17) balance the prior and the posterior variances, and Eq.(17) is lower-bounded.[2]

Nevertheless, we will show that empirical PB and empirical MAP with *local search* can work similarly to empirical VB. We obtain the following theorems:

Theorem 1 *The PB free energy has a non-trivial local minimum such that*

$$a_h b_h = \breve{\gamma}_h^{\textit{local-EPB}} \textit{ if and only if } \gamma_h > \underline{\gamma}_h^{\textit{local-EPB}}, \quad (21)$$

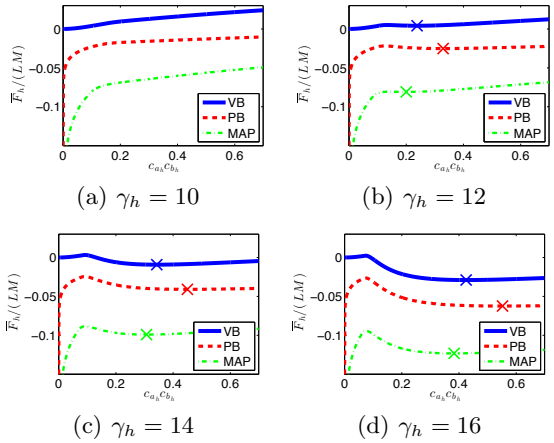
where

$$\underline{\gamma}_h^{\textit{local-EPB}} = \sigma \sqrt{L + M + \sqrt{2LM + \underline{K}^2}}, \quad (22)$$

$$\breve{\gamma}_h^{\textit{local-EPB}} = \frac{\gamma_h}{2} \left(1 + \frac{-\overline{K}\sigma^2 + \sqrt{\gamma_h^4 - 2(L+M)\sigma^2\gamma_h^2 + \overline{K}^2\sigma^4}}{\gamma_h^2} \right). \quad (23)$$

(Sketch of proof) The PB-A solution given $c_{a_h}^2$ and $c_{b_h}^2$ has been obtained by Nakajima et al. (2011). Substituting it into the free energy (18), we can write the PB-A free energy as a function of $c_{a_h} c_{b_h}$. By taking the first derivative, we obtain the stationary points. By checking the sign of the second derivative, we pick up the non-trivial local minimum from those stationary points. We can analyze the PB-B free energy in the same way, and obtain its local minimum. Choosing the one of PB-A and PB-B giving a lower free energy leads to the solution (21), which completes the proof. □

[2] Hyperprior on $c_{a_h}^2$ and $c_{b_h}^2$ can make the free energy in PB and MAP lower-bounded. However, for the purpose of ARD, it should be almost non-informative. With such an almost non-informative hyperprior, e.g., the inverge-Gamma, $p(c_{a_h}^2, c_{b_h}^2) \propto (c_{a_h}^2 c_{b_h}^2)^{1.001} + 0.001/(c_{a_h}^2 c_{b_h}^2)$, used in Bishop (1999), a deep valley exists close to the origin, which makes the global estimator uselessly conservative.



(a) $\gamma_h = 10$ (b) $\gamma_h = 12$ (c) $\gamma_h = 14$ (d) $\gamma_h = 16$

Figure 1: Free Energy dependence on $c_{a_h} c_{b_h}$, where $L = 20, M = 50$. A cross indicates a non-trivial local minimum.

Theorem 2 *The free energy of empirical MAP has a non-trivial local minimum such that*

$$a_h b_h = \breve{\gamma}_h^{\textit{local-MAP}} \textit{ if and only if } \gamma_h > \underline{\gamma}_h^{\textit{local-MAP}}, \quad (24)$$

where

$$\underline{\gamma}_h^{\textit{local-EMAP}} = \sigma \sqrt{2(M+L)}, \quad (25)$$

$$\breve{\gamma}_h^{\textit{local-EMAP}} = \frac{1}{2} \left(\gamma_h + \sqrt{\gamma_h^2 - 2\sigma^2(M+L)} \right). \quad (26)$$

(Sketch of proof) Similarly to the proof of Theorem 1, we substitute the MAP solution (Srebro et al., 2005; Nakajima and Sugiyama, 2011) given $c_{a_h}^2$ and $c_{b_h}^2$ into the MAP free energy or the negative log likelihood:

$$2F_h^{\text{MAP}} = M \log c_{a_h}^2 + L \log c_{b_h}^2 + \frac{a_h^2}{c_{a_h}^2} + \frac{b_h^2}{c_{b_h}^2} + \frac{-2a_h b_h \gamma_h + a_h^2 b_h^2}{\sigma^2} + \text{const.} \quad (27)$$

By taking the first derivative, we find that the stationary condition is written as a quadratic function of $c_{a_h} c_{b_h}$. Analyzing the stationary condition, we obtain the local minimum (24), which completes the proof. □

Figure 1 shows the free energy as a function of $c_{a_h} c_{b_h}$ (minimized with respect to the other parameters), i.e.,

$$\overline{F}_h(c_{a_h} c_{b_h}) = \min_{(a_h, b_h, \sigma_{a_h}^2, \sigma_{b_h}^2)} F_h. \quad (28)$$

This exhibits the behavior of local minima of empirical VB, empirical PB, and empirical MAP. We see that the free energy around the trivial solution $c_{a_h} c_{b_h} \to 0$

is different, but the behavior of a non-trivial local minimum is similar: it appears when the observed singular value γ_h exceeds a threshold.

The boundedness is essential when we stick to the global solution. The VB free energy is lower-bounded at the origin, which enables inference consistently based on the minimum free energy principle. However, as long as we rely on local search, the unboundedness at the origin is not essential in practice. Assume that a non-trivial local minimum exists, and we perform local search only once. Then, whether local search for empirical PB (empirical MAP) converges to the trivial global solution or the non-trivial local solution simply depends on the initialization. Note that this also applies to empirical VB, where local search is not guaranteed to converge to the global solution, because of the multi-modality of the free energy.

3.2 Behavior of Local Solutions

Let us define, for each singular component, a *local* empirical PB (*local* empirical MAP) estimator to be the estimator equal to the non-trivial local minimum if it exists, and the trivial global minimum otherwise. The analytic-form of the non-trivial local minimum is given by Theorem 1 (Theorem 2). We suppose that local search for empirical PB (empirical MAP) looks for this solution.

In the rest of this section, we investigate the local empirical PB and the local empirical MAP estimators with their analytic-forms. We will also experimentally investigate the behavior of local search algorithms in Section 5.

Let us consider the normalized singular values of the observation and the estimators:

$$\gamma'_h = \frac{\gamma_h}{\sqrt{K\sigma^2}}, \widehat{\gamma}'_h = \frac{\widehat{\gamma}_h}{\sqrt{K\sigma^2}}, \underline{\gamma}'_h = \frac{\underline{\gamma}_h}{\sqrt{K\sigma^2}}, \breve{\gamma}'_h = \frac{\breve{\gamma}_h}{\sqrt{K\sigma^2}}.$$

Then, the estimators can be written as functions of $\alpha = \underline{K}/\overline{K}$:

$$\underline{\gamma}'^{\text{EVB}}_h = \sigma\sqrt{1+\alpha+\sqrt{\alpha}\left(\underline{\kappa}+\frac{1}{\underline{\kappa}}\right)}, \tag{29}$$

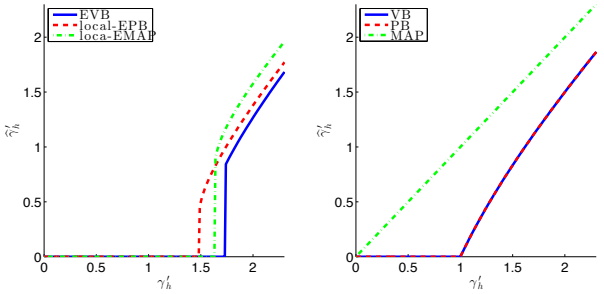
$$\breve{\gamma}'^{\text{EVB}}_h = \frac{\gamma'_h}{2}\left(1-\frac{(1+\alpha)\sigma^2}{\gamma'^2_h}+\sqrt{\left(1-\frac{(1+\alpha)\sigma^2}{\gamma'^2_h}\right)^2-\frac{4\alpha\sigma^4}{\gamma'^4_h}}\right), \tag{30}$$

$$\underline{\gamma}'^{\text{local-EPB}}_h = \sigma\sqrt{1+\alpha+\sqrt{2\alpha+\alpha^2}}, \tag{31}$$

$$\breve{\gamma}'^{\text{local-EPB}}_h = \frac{\gamma'_h}{2}\left(1+\frac{-\sigma^2+\sqrt{\gamma'^4_h-2(1+\alpha)\sigma^2\gamma'^2_h+\sigma^4}}{\gamma'^2_h}\right), \tag{32}$$

$$\underline{\gamma}'^{\text{local-EMAP}}_h = \sigma\sqrt{2(1+\alpha)}, \tag{33}$$

$$\breve{\gamma}'^{\text{local-EMAP}}_h = \frac{1}{2}\left(\gamma'_h+\sqrt{\gamma'^2_h-2\sigma^2(1+\alpha)}\right). \tag{34}$$



(a) Empirical Bayes. (b) Non-empirical Bayes with $c_{a_h} = c_{b_h} = 100$.

Figure 2: Behavior of empirical and non-empirical Bayesian estimators for $\alpha = \underline{K}/\overline{K} = 1/3$.

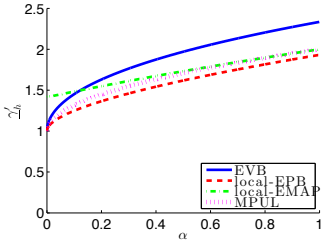


Figure 3: Truncation thresholds.

Note that $\underline{\kappa}$ is also a function of α.

The left graph in Figure 2 compares the behavior of the empirical VB, the local empirical PB, and the local empirical MAP estimators. We see the similarity between those three empirical Bayesian estimators. This is in contrast to the non-empirical Bayesian estimators, where the hyperparameters C_A, C_B are given (see the right graph in Figure 2). There, VB and PB behave similarly, but MAP behaves like the maximum likelihood estimator (Nakajima et al., 2011).

Figure 3 compares the truncation thresholds (29), (31), and (33) of the estimators. We find that the thresholds behave similarly. However, an essential difference of local empirical PB from empirical VB and local empirical MAP is found: It holds that, for any α,

$$\underline{\gamma}'^{\text{local-EPB}} < \overline{\gamma}'^{\text{MPUL}} \leq \underline{\gamma}'^{\text{EVB}}, \underline{\gamma}'^{\text{local-EMAP}}, \tag{35}$$
$$\text{where} \quad \overline{\gamma}'^{\text{MPUL}} = (1+\sqrt{\alpha})$$

is the Marčenko-Pastur upper limit (MPUL) (Marčenko and Pastur, 1967; Nakajima et al., 2012), which is also shown in Figure 3. MPUL is the largest singular value of an $L \times M$ random matrix when the matrix consists of zero-mean independent noise, and its size L and M goes to infinity with fixed α. Inequalities (35) imply that, with high probability, empirical VB and local empirical MAP discard the singular components dominated by noise, while local empirical PB overestimates the rank when the ratio $\xi = H^*/\underline{K}$ between the (unknown) true rank and the possible largest rank is small, i.e., when most of the

singular components consist of noise.

4 PRACTICAL APPLICATIONS

In fully-observed MF, the analytic solution of VB (Proposition 1) is available, and therefore, one takes no advantage by substituting PB or MAP for VB. However, in other models, e.g., collaborative filtering (or partially-observed MF) and tensor factorization, no analytic-form solution has been derived, and all the state-of-the-art algorithms are local search. In such cases, approximating VB by PB or MAP reduces the memory consumption and calculation time, because inverse calculations for estimating posterior covariances can be skipped. In this section, we show that, also in collaborative filtering and tensor factorization, the global solutions of empirical PB and empirical MAP are trivial and useless. In Section 5, we will experimentally show that local search algorithms however tend to find useful local minima. We also mention the possibility of noise variance estimation.

4.1 Collaborative Filtering

In the collaborative filtering (CF) setting, the observed matrix V has missing entries, and the likelihood (1) is replaced with

$$p(V|A,B) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|\mathcal{P}_{\Lambda}(V) - \mathcal{P}_{\Lambda}(BA^\top)\|_{\mathrm{Fro}}^2\right), \quad (36)$$

where Λ denotes the set of observed indices, and

$$(\mathcal{P}_{\Lambda}(V))_{l,m} = \begin{cases} V_{l,m} & \text{if } (l,m) \in \Lambda, \\ 0 & \text{otherwise.} \end{cases}$$

An iterative local search algorithm for empirical VB has been derived (Lim and Teh, 2007), similarly to the fully-observed case.

We can show that the global solutions for empirical PB and empirical MAP are useless also in this case:

Lemma 1 *All elements of the global empirical PB and the global empirical MAP solutions in CF are zero, regardless of observation.*

Nevertheless, we experimentally show in Section 5 that local search for empirical PB and empirical MAP behaves like local search for empirical VB.

4.2 Tensor Factorization

By extending MF to N-mode tensor data, the Bayesian Tucker factorization (TF) model was proposed (Chu and Ghahramani, 2009; Mørup and Hansen, 2009):

$$p(\mathcal{Y}|\mathcal{G},\{A^{(n)}\}) \propto \exp\left(-\frac{\|\mathcal{Y} - \mathcal{G}\times_1 A^{(1)} \cdots \times_N A^{(N)}\|^2}{2\sigma^2}\right), \quad (37)$$

$$p(\mathcal{G}) \propto \exp\left(-\frac{\mathrm{vec}(\mathcal{G})^\top \left(C_{G^{(N)}} \otimes \cdots \otimes C_{G^{(1)}}\right)^{-1} \mathrm{vec}(\mathcal{G})}{2}\right), \quad (38)$$

$$p(\{A^{(n)}\}) \propto \exp\left(-\frac{\sum_{n=1}^N \mathrm{tr}(A^{(n)} C_{A^{(n)}}^{-1} A^{(n)\top})}{2}\right), \quad (39)$$

where $\mathcal{Y} \in \mathbb{R}^{\prod_{n=1}^N M^{(n)}}$, $\mathcal{G} \in \mathbb{R}^{\prod_{n=1}^N H^{(n)}}$, and $\{A^{(n)} \in \mathbb{R}^{M^{(n)} \times H^{(n)}}\}$ are an observed tensor, a core tensor, and factor matrices, respectively. Here, $\times_n$, $\otimes$, and $\mathrm{vec}(\cdot)$ denote the n-mode tensor product, the Kronecker product, and the vectorization operator, respectively (Kolda and Bader, 2009). $\{C_{G^{(n)}}\}$ and $\{C_{A^{(n)}}\}$ are the prior covariances restricted to be diagonal. We fix the ratio between the prior covariances to $C_{G^{(n)}} C_{A^{(n)}}^{-1} = I_{H^{(n)}}$, where I_d denotes the d-dimensional identity matrix.

Chu and Ghahramani (2009) applied *non-empirical* PB learning, where the core tensor is integrated out, and the factor matrices are point-estimated. Their model corresponds to Eqs.(37)–(39) with $C_{G^{(n)}} = C_{A^{(n)}} = I_{H^{(n)}}$. Mørup and Hansen (2009) applied empirical MAP to this model, and experimentally showed its model selection ability through the ARD property.

However, we can show that the global solutions of empirical PB and empirical MAP are useless also in TF:

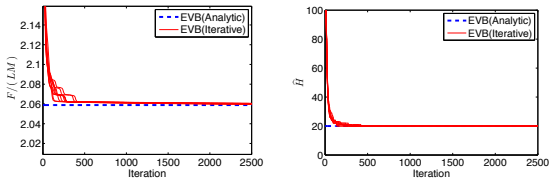
Lemma 2 *All elements of the global empirical PB and the global empirical MAP solutions in TF are zero, regardless of observation.*

Nevertheless, our experiments in Section 5, as well as the experiments in Mørup and Hansen (2009), show that empirical PB and empirical MAP with local search are useful alternatives to empirical VB.

4.3 Noise Variance Estimation

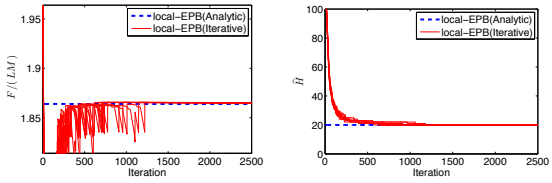
Generally, the noise variance, σ^2 in Eq.(1), can be estimated based on free energy minimization. Nakajima et al. (2012) obtained a sufficient condition for the perfect rank recovery by empirical VB with the noise variance estimated. However, we experimentally found that, when the noise variance is estimated, local search for empirical PB tends to underestimate the noise variance. Even worse, local search for empirical MAP tends to result in $\widehat{\sigma}^2 \to 0$.

This unreliability was pointed out by Mørup and Hansen (2009) in empirical MAP for TF. They suggested to set the noise variance under the assumption that the signal to noise ratio (SNR) is 0 db, i.e., the signal and the noise have equal energy. In this paper, we follow their approach in the experiments with real datasets, where the noise variance is unknown, and leave further investigation as future work.



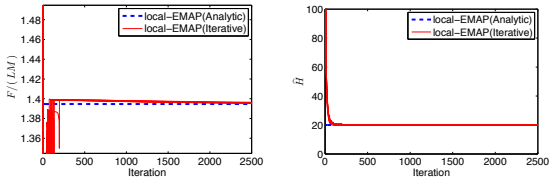
(a) Free energy (b) Estimated rank

Figure 4: Empirical VB result on *Artificial1* ($L = 100, M = 300, H^* = 20$).



(a) Free energy (b) Estimated rank

Figure 5: *Local* empirical PB result on *Artificial1*.



(a) Free energy (b) Estimated rank

Figure 6: *Local* empirical MAP result on *Artificial1*.

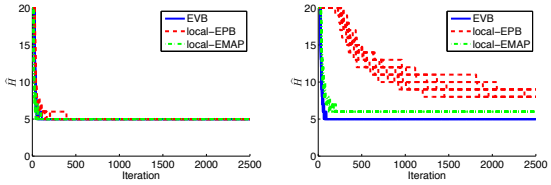
5 EXPERIMENT

In this section, we experimentally show that empirical PB and empirical MAP with local search can be alternatives to empirical VB. We start from fully-observed MF, where we can see how often local search finds the analytic *local* empirical solution. After that, we conduct experiments in CF and TF.

For local search, efficient algorithms can be implemented based on the gradient descent algorithm (Chu and Ghahramani, 2009; Mørup and Hansen, 2009). However, it requires pruning, and we found that the pruning strategy can significantly affect the result, especially in the estimated rank. Accordingly, we used the *iterated conditional modes* (ICM) algorithm (Besag, 1986; Bishop, 2006) without pruning, which we found is more stable, for all empirical Bayesian methods. In all experiments, the entries of the mean parameters, e.g., $\widehat{A}, \widehat{B}$, and $\widehat{\mathcal{G}}$, were initialized with independent draws from $\mathcal{N}_1(0,1)$, while the covariance parameters were initialized to the identity, e.g., $\Sigma_A = \Sigma_B = C_A = C_B = I_H$.

5.1 Fully-observed MF

We first conducted an experiment with an artificial (*Artificial1*) dataset, following Nakajima et al. (2013).



(a) *ArtificialCF*. (b) *MovieLens*.

Figure 7: Estimated rank in collaborative filtering with 99% missing ratio.

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 $B^* A^{*\top}$. Figures 4–6 show the free energy and the estimated rank over iterations for the *Artificial1* dataset with the data matrix size $L = 100$ and $M = 300$, and the true rank $H^* = 20$. The noise variance was assumed to be known, $\sigma^2 = 1$. We performed iterative local search 10 times, starting from different initial points, and each trial is plotted by a solid line in the figures. We see that iterative local search tends to successfully find the analytic *local* empirical solution, although it is not the global solution.

We also conducted experiments on another artificial dataset and benchmark datasets. The results are summarized in Table 1. *Artificial2* was created in the same way as *Artificial1* with $L = 400, M = 500$, and $H^* = 5$. The benchmark datasets were collected from the UCI repository (Asuncion and Newman, 2007). We set the noise variance under the assumption that the SNR is 0 db, following Mørup and Hansen (2009).

In the table, the estimated ranks by the analytic-form solution and the ones by iterative local search are shown. The percentages for iterative local search indicate the frequencies over 10 trials. We can observe the following: First, iterative local search tends to consistently estimate the same rank as the analytic solution over the trials.[3] Second, the estimated rank tends to be consistent over empirical VB, *local* empirical PB, and *local* empirical MAP, and the correct rank is found on the artificial datasets. Exceptions are *Artificial2*, where *local* empirical PB overestimates the rank, and *Optical Digits* and *Satellite*, where *local* empirical MAP estimates a smaller rank than the others. These phenomena can be explained by the theoretical implications in Section 3.2: In *Artificial2*, the ratio $\xi = H^*/\underline{K} = 5/400$ between the true rank and the pos-

[3]The results of empirical VB are different from the ones reported in Nakajima et al. (2013), where the noise variance is also estimated from observation. Generally, the 0 db SNR assumption (Mørup and Hansen, 2009) tends to result in a smaller rank than empirical VB with noise variance estimation (Nakajima et al., 2013) in our experiments.

Table 1: Estimated rank in fully-observed MF experiments.

Data set	M	L	H^*	$\widehat{H}^{\mathrm{EVB}}$		$\widehat{H}^{\text{local-EPB}}$		$\widehat{H}^{\text{local-EMAP}}$	
				Analytic	Iterative	Analytic	Iterative	Analytic	Iterative
Artificial1	200	100	20	2	2(100%)	2	2 (100%)	2	2 (100%)
Artificial2	500	400	5	5	5 (100%)	8	8 (90%) 9 (10%)	5	5 (100%)
Chart	600	60	–	2	2 (100%)	2	2 (100%)	2	2 (100%)
Glass	214	9	–	1	1 (100%)	1	1 (100%)	1	1 (100%)
Optical Digits	5620	64	–	10	10 (100%)	10	10 (100%)	6	6 (100%)
Satellite	6435	36	–	2	2 (100%)	2	2 (100%)	1	1 (100%)

Table 2: Estimated rank (effective size of core tensor) in TF experiments.

Data set	M	H^*	$\widehat{H}^{\mathrm{EVB}}$	$\widehat{H}^{\text{local-EPB}}$	$\widehat{H}^{\text{local-EMAP}}$	$\widehat{H}^{\text{ARD-Tucker}}$
ArtificialTF	(30, 40, 50)	(3, 4, 5)	(3, 4, 5): 100%	(3, 4, 5): 100%	(3, 4, 5): 90% (3, 7, 5): 10%	(3, 4, 4): 80% (3, 4, 5): 20%
FIA	(12, 100, 89)	(3, 6, 4)	(3, 5, 3): 100%	(3, 5, 3): 100%	(3, 5, 2): 50% (4, 5, 2): 20% (5, 4, 2): 10% (4, 4, 2): 10% (8, 5, 2): 10%	(3, 2, 2): 40% (2, 1, 2): 20% (2, 3, 2): 10% (2, 2, 2): 10% (1, 1, 1): 10% (10, 4, 3): 10%

sible largest rank is small, which means that most of the singular components consist of noise. In this case, *local* empirical PB with its lower truncation threshold than MPUL fails to discard noise components (see Figure 3). In *Optical Digits* and *Satellite*, $\alpha(= 64/5620$ for *Optical Digits* and $= 36/6435$ for *Satellite*) is extremely small, and therefore *local* empirical MAP with its higher truncation threshold tends to discard more components than the others, as Figure 3 implies.

5.2 Collaborative Filtering

To investigate the behavior of the estimators in CF, we conducted experiments with an artificial (*ArtificialCF*) and the *MovieLens* datasets.[4] The *ArtificialCF* dataset was created in the same way as the fully-observed case for $L = 2000, M = 5000$, and $H^* = 5$, and masked 99% of the entries as missing values. For the *MovieLens* dataset (with $L = 943, M = 1682$), we randomly selected observed values so that 99% of the entries are missing.

Figure 7 shows the estimated rank over iterations. We see that all three methods tend to estimate the same rank, but empirical PB tends to give a larger rank, as in the fully-observed case.

5.3 Tensor Factorization

Finally, we conducted experiments on TF. We created an artificial (*ArtificialTF*) dataset, following Mørup and Hansen (2009): We drew a 3-mode random tensor of the size of $(M^{(1)}, M^{(2)}, M^{(3)}) = (30, 40, 50)$ with the signal components with $(H^{(1)*}, H^{(2)*}, H^{(3)*}) = (3, 4, 5)$. The noise is added so that the SNR is 0 db. As a benchmark, we used the *Flow Injection Analysis* (*FIA*) dataset.[5] Table 2 shows the estimated rank with frequencies over 10 trials. Here, we also show the result by ARD Tucker with the ridge prior (Mørup and Hansen, 2009), of which the objective function is exactly the same as our empirical MAP. A slight difference comes from the iterative algorithm (ICM vs. gradient descent) and the initialization strategy.

We generally observe that the three empirical Bayesian methods provide reasonable results, although *local* empirical MAP is less stable than the others.

6 CONCLUSION

In this paper, we analyzed partially Bayesian (PB) learning and MAP estimation under the empirical Bayesian scenario, where the hyperparameters are also estimated from observation. Our theoretical analysis in fully-observed matrix factorization (MF) revealed a notable fact: Although the global solutions of empirical PB and empirical MAP are trivial and useless, their local solutions behave similarly to the global solution of variational Bayesian (VB) learning.

We also conducted experiments in collaborative filtering (or partially observed MF) and tensor factorization, and showed that empirical PB and empirical MAP solved by *local search* can be good alternatives to empirical VB.

Acknowledgments

The authors thank reviewers for helpful comments. SN and MS thank the support from MEXT Kakenhi 23120004 and the CREST program, respectively.

[4] `http://www.grouplens.org/`

[5] `http://www.models.kvl.dk/datasets`

References

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 *Proc. of UAI*, pages 21–30, 1999.

D. Barber and S. Chiappa. Unified inference for variational Bayesian linear Gaussian state-space models. In *Advances in NIPS*, volume 19, 2006.

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 International Conference on Artificial Neural Networks*, volume 1, pages 514–509, 1999.

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

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.

M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo Methods for Bayesian Computation*. Springer, 2001.

W. Chu and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2009.

Z. Ghahramani and M. J. Beal. Graphical models and variational methods. In *Advanced Mean Field Methods*, pages 161–177. MIT Press, 2001.

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.

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

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

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

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

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proc. of UAI*, pages 362–369, 2001.

M. Mørup and L. R. Hansen. Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23:352–363, 2009.

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 S. D. Babacan. On Bayesian PCA: Automatic dimensionality selection and analytic solution. In *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pages 497–504, Bellevue, WA, USA, Jun. 28–Jul.2 2011.

S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Perfect dimensionality recovery by variational Bayesian PCA. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 980–988, 2012.

S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *Journal of Machine Learning Research*, 14:1–37, 2013.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.

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.

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

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

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