On Bayesian PCA: Automatic Dimensionality Selection and Analytic Solution

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

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

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

Abstract

In probabilistic PCA, the *fully* Bayesian estimation is computationally intractable. To cope with this problem, two types of approximation schemes were introduced: the *partially Bayesian* PCA (PB-PCA) where only the latent variables are integrated out, and the variational Bayesian PCA (VB-PCA) where the loading vectors are also integrated out. The VB-PCA was proposed as an improved variant of PB-PCA for enabling automatic dimensionality selection (ADS). In this paper, we investigate whether VB-PCA is really the best choice from the viewpoints of computational efficiency and ADS. We first show that ADS is not the unique feature of VB-PCA—PB-PCA is also actually equipped with ADS. We further show that PB-PCA is more advantageous in computational efficiency than VB-PCA because the global solution of PB-PCA can be computed analytically. However, we also show the negative fact that PB-PCA results in a trivial solution in the empirical Bayesian framework. We next consider a simplified variant of VB-PCA, where the latent variables and loading vectors are assumed to be mutually independent (while the ordinary VB-PCA only requires matrix-wise independence). We show that this simplified VB-PCA is the most advantageous in practice because its empirical Bayes solution experimentally works as well as the original VB-PCA, and its global optimal solution can be computed efficiently in a closed form.

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

1. Introduction

Principal component analysis (PCA) is a well-established technique for unsupervised dimensionality reduction (Hotelling, 1933). A decade ago, PCA was given its probabilistic interpretation as a latent variable model called *probabilistic PCA* (PPCA) (Tipping & Bishop, 1999; Roweis & Ghahramani, 1999). In the first formulation of PPCA, only the latent variables are integrated out and the loading vectors are estimated by maximizing the marginal likelihood (Tipping & Bishop, 1999). Since the loading vectors are point-estimated, we refer to this approach as *partially Bayesian* PCA (PB-PCA) in the following.

Subsequently, a variant of PPCA was proposed (Bishop, 1999b), which has the following two features:

(A) The *fully* Bayesian treatment: both the latent variables and the loading vectors are integrated out. We refer to this as *fully Bayesian* PCA (FB-PCA).

(B) The *empirical* Bayesian procedure: the prior variances of the loading vectors can also be estimated from observation.

A notable advantage of FB-PCA is that it offers *automatic dimensionality selection* (ADS). However, since exact FB-PCA is computationally intractable, the Laplace approximation (Bishop, 1999b; Hoyle, 2008), Markov chain Monte Carlo (Bishop, 1999b), and the variational approximation (Bishop, 1999a) were used for approximate inference in practice. Among them, the variational approximation (which we refer to as VB-PCA) seems to be the most popular choice.

The purpose of this paper is to revisit PPCA, and investigate whether VB-PCA is really the best choice from the viewpoints of computational efficiency and ADS, within a

unified framework of free energy minimization under different constraints on the posterior distribution.

First, in defense of PB-PCA, we theoretically show the following two positive facts:

- Neither (A) nor (B) is essential for PPCA to induce ADS. PB-PCA is actually equipped with ADS. Thus, the original unique advantage of VB-PCA is lost.
- The global solution of PB-PCA can be computed very efficiently in a closed form (which is an extension of the result given in Tipping & Bishop (1999)). On the other hand, VB-PCA requires a number of iterations to find a local optimal solution (Bishop, 1999a).

These facts encourage the use of PB-PCA. However, we next show that the following negative fact:

- PB-PCA has a critical drawback when hyperparameters are learned from data in the empirical Bayesian framework. More precisely, the empirical Bayes method in PB-PCA results in a trivial solution (i.e., zero), and thus cannot be used in practice. On the other hand, the empirical VB-PCA works well.

Based on this fact, we conclude that PB-PCA cannot be a practical alternative to VB-PCA unfortunately.

We next consider a simplified variant of VB-PCA, where the latent variables and loading vectors are mutually independent (cf. the latent variables and loading matrix are matrix-wise independent in the original VB-PCA). We refer to this method as *simple-VB-PCA*. For simple-VB-PCA, we show the following two positive facts:

- The global optimal solution of simple-VB-PCA can be computed analytically (which can be immediately obtained from the result given in Nakajima et al. (2010)). On the other hand, the ordinary VB-PCA requires a number of iterations to find a local optimal solution (Bishop, 1999a).
- The mutual independence assumption of the latent variables and loading vectors is not restrictive in practice. More specifically, we experimentally show that the performance of simple-VB-PCA is comparable to that of the original VB-PCA.

Based on these observations, we conclude that simple-VB-PCA is the most attractive as a Bayesian PCA method.

This paper is organized as follows. In Section 2, we formulate PPCA, and introduce approximation methods to the fully Bayesian inference. In Section 3, we derive the analytic-form solution for PB-PCA, and discuss its behavior in comparison with VB-PCA. In Section 4, we introduce a simplified variant of VB-PCA, and compare its behavior with the original VB-PCA. We further discuss various issues of PPCA in Section 5, and conclude in Section 6.

2. Formulation

In this section, we formulate PPCA, and review approximation methods to the fully Bayesian inference. Then, we describe the empirical Bayesian procedure.

2.1. Probabilistic PCA

PPCA assumes that the observation $\boldsymbol{y} \in \mathbb{R}^L$ is driven by a latent vector $\widetilde{\boldsymbol{a}} \in \mathbb{R}^H$ as follows:

$$\boldsymbol{y} = B\widetilde{\boldsymbol{a}} + \boldsymbol{\varepsilon},$$

where $B \in \mathbb{R}^{L\times H}$ specifies the linear relationship between $\widetilde{\boldsymbol{a}}$ and $\boldsymbol{y}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^L$ is a Gaussian noise subject to $\mathcal{N}_L(\boldsymbol{0}, \sigma^2 I_L)$. Here, we denote by $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and by I_d the d-dimensional identity matrix. We assume that the latent vector $\widetilde{\boldsymbol{a}}$ is subject to $\mathcal{N}_H(\boldsymbol{0}, I_H)$.

Suppose that we are given M observed samples $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$, which are generated from the latent vectors $\{\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M\}$. Define the following matrices:

$$Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M) \in \mathbb{R}^{L\times M}, \quad A^\top = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M) \in \mathbb{R}^{H\times M}.$$

Then the PPCA model is written as follows[1]:

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

$$\phi_{\mathrm{A}}(A) \propto \exp\left(-\frac{1}{2}\|A\|_{\mathrm{Fro}}^2\right), \quad (2)$$

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

where $\|\cdot\|_{\mathrm{Fro}}$ and $\mathrm{tr}(\cdot)$ denote the Frobenius norm and the trace of a matrix, respectively, and $C_B \in \mathbb{R}^{H\times H}$ is a diagonal matrix with positive entries. The column vectors of $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_H)$ correspond to the loading vectors, and the diagonal elements of $C_B = \mathrm{diag}(c_{b_1}^2, \ldots, c_{b_H}^2)$ correspond to the prior variances of the loading vectors. Without loss of generality, we assume that $\{c_{b_h}\}$ are arranged in non-increasing order.

Note that the above PPCA model can be seen as Bayesian matrix factorization (Salakhutdinov & Mnih, 2008), if $U = BA^\top$ is regarded as a low rank matrix approximating Y.

[1] Note that, for controlling the regularization effect and introducing an *empirical Bayesian* variant, we added the prior (3) on B to the original PPCA formulation given in Tipping & Bishop (1999). When the diagonal elements of C_B tend to infinity, the model (1)–(3) is reduced to the original formulation.

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

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

2.2. Approximate Bayesian Inference

The Bayes posterior is given by

$$p(A, B|Y) = \frac{p(Y|A, B)\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)}{Z(Y)}, \tag{4}$$

where $Z(Y) = \langle p(Y|A, B)\rangle_{\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)}$. Here, $\langle\cdot\rangle_p$ denotes the expectation over the distribution p. Since this expectation is intractable, it needs to be approximated. Here we review methods of approximate Bayesian inference.

2.2.1. Free Energy Minimization

First, we describe an approximation framework.

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

$$F(r|Y) = -\left\langle \log \frac{r(A, B)}{p(Y|A, B)\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)} \right\rangle_{r(A,B)} \tag{5}$$
$$= -\left\langle \log \frac{r(A, B)}{p(A, B|Y)} \right\rangle_{r(A,B)} - \log Z(Y).$$

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 Bayesian approximate inference is to find the minimizer of the free energy (5) with respect to r in some restricted function space.

Let $\widehat{r}$ be such a minimizer. In the context of PCA, the solution is the subspace spanned by the estimated loading vectors:

$$\mathcal{S} = \mathrm{span}\left(\langle B\rangle_{\widehat{r}(A,B)}\right), \tag{6}$$

where $\mathrm{span}(\cdot)$ denotes the subspace spanned by the column vectors of a matrix. Let

$$Y = \sum_{h=1}^{H} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top \tag{7}$$

be the singular value decomposition of Y. In many methods, the PCA solution can be written in the following form:

$$\mathcal{S} = \mathrm{span}\left(\sum_{h=1}^{H} \theta(\gamma_h > \underline{\gamma}_h)\boldsymbol{\omega}_{b_h}\boldsymbol{\omega}_{a_h}^\top\right), \tag{8}$$

where $\theta(\cdot)$ is the function taking one if the event is true and zero otherwise, and $\underline{\gamma}_h$ is a threshold.

Below, we review two types of approximation methods.

2.2.2. Partially Bayesian PCA (PB-PCA)

In the first formulation of PPCA by Tipping & Bishop (1999), only the latent matrix A is integrated out, and the loading matrix B is point-estimated. This amounts to restricting the posterior to the following form:

$$r^{\text{PB-A}}(A, B) = r_{\mathrm{A}}^{\text{PB-A}}(A)\delta(B; B^*), \tag{9}$$

where $\delta(B; B^*)$ denotes a (*pseudo*-)Dirac delta function of B located at $B = B^*$.[2]

We refer to this method as PB-A-PCA since A is integrated out. An analytic-form solution for PB-A-PCA when $\{c_{b_h} \to \infty\}$ is given as follows (this proposition is obtained by combing Eqs.(6) and (7) in Tipping & Bishop (1999).):

Proposition 1 *(Tipping & Bishop, 1999) The PB-A-PCA solution when $\{c_{b_h} \to \infty\}$ is given by Eq.(8) with the following threshold:*

$$\underline{\gamma}_h^{\text{PB-A}} = \sigma\sqrt{M}. \tag{10}$$

In the above formulation, A is integrated out and B is point-estimated. On the other hand, we can naturally think of the opposite variant, i.e., B is integrated out and A is point-estimated (which we refer to as PB-B-PCA):

$$r^{\text{PB-B}}(A, B) = \delta(A; A^*)r_{\mathrm{B}}^{\text{PB-B}}(B). \tag{11}$$

In this paper, we slightly extend the formulation by Tipping & Bishop (1999) and allow the posterior to be chosen from (9) or (11) that gives a smaller free energy. We call this adaptive method *partially* Bayesian PCA (PB-PCA).

In Section 3, we derive an analytic-form solution of PB-PCA for arbitrary $\{c_{b_h}\}$.

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

2.2.3. VARIATIONAL BAYESIAN PCA (VB-PCA)

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

$$r^{\mathrm{VB}}(A,B) = r_{\mathrm{A}}^{\mathrm{VB}}(A) r_{\mathrm{B}}^{\mathrm{VB}}(B). \tag{12}$$

With this constraint, an iterative algorithm for minimizing the free energy (5) was derived (Bishop, 1999a; Lim & Teh, 2007).

The VB posterior can be written as

$$r^{\mathrm{VB}}(A,B) = \prod_{m=1}^{M} \mathcal{N}_H(\widetilde{\boldsymbol{a}}_m; \boldsymbol{\mu}_{\widetilde{a}_m}, \Sigma_{\widetilde{A}}) \prod_{l=1}^{L} \mathcal{N}_H(\widetilde{\boldsymbol{b}}_l; \boldsymbol{\mu}_{\widetilde{b}_l}, \Sigma_{\widetilde{B}}),$$

where the means and the covariances necessarily satisfy

$$\boldsymbol{\mu}_{\widetilde{a}_m} = \frac{\Sigma_{\widetilde{A}}}{\sigma^2} \left(\boldsymbol{\mu}_{\widetilde{b}_1}, \ldots, \boldsymbol{\mu}_{\widetilde{b}_L} \right) \boldsymbol{y}_m, \tag{13}$$

$$\boldsymbol{\mu}_{\widetilde{b}_l} = \frac{\Sigma_{\widetilde{B}}}{\sigma^2} \left(\boldsymbol{\mu}_{\widetilde{a}_1}, \ldots, \boldsymbol{\mu}_{\widetilde{a}_M} \right) \widetilde{\boldsymbol{y}}_l, \tag{14}$$

$$\Sigma_{\widetilde{A}} = \sigma^2 \left(\sum_{l=1}^{L} \left(\boldsymbol{\mu}_{\widetilde{b}_l} \boldsymbol{\mu}_{\widetilde{b}_l}^{\top} + \Sigma_{\widetilde{B}} \right) + \sigma^2 I_H \right)^{-1}, \tag{15}$$

$$\Sigma_{\widetilde{B}} = \sigma^2 \left(\sum_{l=1}^{M} \left(\boldsymbol{\mu}_{\widetilde{a}_m} \boldsymbol{\mu}_{\widetilde{a}_m}^{\top} + \Sigma_{\widetilde{A}} \right) + \sigma^2 C_B^{-1} \right)^{-1}. \tag{16}$$

We may obtain a local minimizer by iterating (13)–(16). After convergence, the VB-PCA solution is given as

$$\mathcal{S}^{\mathrm{VB}} = \mathrm{span}\left((\boldsymbol{\mu}_{\widetilde{b}_1}, \ldots, \boldsymbol{\mu}_{\widetilde{b}_L})^{\top} \right).$$

2.3. Empirical Bayesian Procedure

PPCA has a hyperparameter C_B in the prior (3), which controls the sparsity of the result (i.e., the PCA dimensions). A popular way to set the hyperparameter in the Bayesian framework is again based on the minimization of the free energy (5).

$$\widehat{C}_B = \operatorname*{argmin}_{C_B} \left(\min_{r} F(r; C_B | Y) \right).$$

We refer to this method as an empirical Bayes method.

3. Behavior of PB-PCA

Among the PPCA methods reviewed in the previous section, VB-PCA seems to be a popular choice. The purpose of this paper is to revisit PPCA, and investigate whether VB-PCA is really the best choice from the viewpoints of computational efficiency and automatic dimensionality selection (ADS).

In this section, we investigate properties of PB-PCA.

3.1. Analytic-Form Solution for PB-PCA

Here, we derive an analytic-form solution for PB-PCA and discuss its properties.

We first consider PB-A-PCA. Substituting Eq.(9) into Eq.(5), we have

$$\begin{aligned} F(r_{\mathrm{A}}^{\text{PB-A}}, B^* | Y) &= -\left\langle \log \frac{r_{\mathrm{A}}^{\text{PB-A}}(A)}{p(Y|A,B^*)\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B^*)} \right\rangle_{r_{\mathrm{A}}^{\text{PB-A}}(A)} + \chi_B \\ &= -\left\langle \log \frac{r_{\mathrm{A}}^{\text{PB-A}}(A)}{p(A|Y,B^*)} \right\rangle_{r_{\mathrm{A}}^{\text{PB-A}}(A)} - \log Z(Y|B^*)\phi_{\mathrm{B}}(B^*) + \chi_B, \end{aligned} \tag{17}$$

where

$$p(A|Y,B) = \frac{p(Y|A,B)\phi_{\mathrm{A}}(A)}{Z(Y|B)}, \tag{18}$$

$$Z(Y|B) = \langle p(Y|A,B) \rangle_{\phi_{\mathrm{A}}(A)}, \tag{19}$$

$$\chi_B = -\langle \log \delta(B; B^*) \rangle_{\delta(B;B^*)}. \tag{20}$$

Note that Eq.(17) is a functional of $r_{\mathrm{A}}^{\text{PB-A}}$ and B^*, and χ_B is a constant with respect to them. Since only the first term depends on $r_{\mathrm{A}}^{\text{PB-A}}$ on which we impose no restriction, Eq.(17) is minimized when

$$\widehat{r}_{\mathrm{A}}^{\text{PB-A}}(A) = p(A|Y,B^*) \tag{21}$$

for any B^*. With Eq.(21), the first term in Eq.(17) vanishes, and thus an estimator for B^* is given by

$$\widehat{B}^{\text{PB-A}} = \operatorname*{argmin}_{B^*} F^{\text{PB-A}}(B^*|Y),$$

where

$$F^{\text{PB-A}}(B|Y) = -\log Z(Y|B)\phi_{\mathrm{B}}(B) + \chi_B. \tag{22}$$

We call Eq.(22) the PB-A free energy.

Minimizing Eq.(22), we obtain the following solution.

Theorem 1 *The PB-A-PCA solution is given by Eq.(8) with the following threshold:*

$$\underline{\gamma}_h^{\text{PB-A}} = \sigma \sqrt{M + \sigma^2 / c_{b_h}^2}. \tag{23}$$

When $c_{b_h} \to \infty$, Eq.(23) converges to Eq.(10). Therefore, Theorem 1 is an extension of Proposition 1 for arbitrary $\{c_{b_h}\}$.

The PB-B-PCA solution with the constraint (11) is similarly obtained by minimizing the PB-B free energy:

$$F^{\text{PB-B}}(A|Y) = -\log Z(Y|A)\phi_{\mathrm{A}}(A) + \chi_A. \tag{24}$$

The threshold for PB-B-PCA is similarly given by

$$\underline{\gamma}_h^{\text{PB-B}} = \sigma \sqrt{L + \sigma^2 / c_{b_h}^2}. \tag{25}$$

By comparing Eqs.(22) and (24), we obtain the following lemma:

Lemma 1 *In PB-PCA, the constraint* (9) *is chosen when* $M > L$*, and the constraint* (11) *is chosen when* $M < L$*.*

Based on this, we can express the PB-PCA subspace analytically as follows (we can also derive the PB-PCA posterior analytically, but we omit the details due to lack of space):

Theorem 2 *The PB-PCA solution is given by Eq.*(8) *with the following threshold:*

$$\underline{\gamma}_h^{PB} = \sigma\sqrt{\max(L,M) + \sigma^2/c_{b_h}^2}. \tag{26}$$

Thanks to Theorem 2, the solution of PB-PCA can be computed analytically in a very efficient way. On the other hand, VB-PCA requires a number of iterations to find a local optimal solution (see Section 2.2.3). Thus, PB-PCA is computationally more attractive than VB-PCA.

Theorem 2 also shows that PB-PCA has a thresholding effect, i.e., the components with singular values smaller than $\underline{\gamma}_h^{\text{PB}}$ are eliminated. Moreover, this thresholding effect remains even when the prior is flat ($c_{b_h} \to \infty$).[3] Actually, such a thresholding effect is also observed in the existing work (see Proposition 1) for PB-A-PCA with $c_{b_h} \to \infty$. However, it was regarded as an artifact in the original paper by Tipping & Bishop (1999).

Bishop (1999a) proposed VB-PCA (see Section 2.2.3) for the purpose of enabling ADS. However, our analysis above shows that PB-PCA also has the thresholding effect. We further show in Section 5.2 that PB-PCA behaves similarly to a simplified variant of VB-PCA.

3.2. Empirical PB-PCA

A critical drawback of PB-PCA appears when the hyperparameter C_B is estimated from data. Specifically, the empirical Bayesian procedure for PB-PCA always results in the trivial solution, i.e., $c_{b_h} \to 0$. This is because the first term of the free energy (22), as well as (24), is not lower-bounded (i.e., it tends to minus infinity). More generally, in order to obtain a meaningful solution when a hyperparameter is estimated through the empirical Bayes procedure, the corresponding parameter must be integrated out.

Consequently, despite the existence of the ADS effect, PB-PCA cannot be a practical alternative to VB-PCA unfortunately.

[3] This applies only to the case when the noise variance σ^2 is strictly positive. When σ^2 is unkown, the free energy minimization fails to estimate it in PB-PCA with $c_{b_h} \to \infty$, because $\sigma^2 \to 0$ is the global minimizer (see Appendix A.2 in Tipping & Bishop (1999)). We would say that it is essential for inducing ADS to correctly estimate σ^2 when it is unknown. When c_{b_h} is finite, σ^2 is estimated to be positive, although we experimentally found that it tends to be underestimated.

4. Behavior of Simplified VB-PCA (Simple-VB-PCA)

In this section, we introduce a simplified variant of VB-PCA, and investigate its properties.

4.1. Analytic-Form Solution for simple-VB-PCA

In the context of collaborative filtering, Raiko et al. (2007) proposed a simple VB iterative algorithm under the following stronger decomposability constraint:

$$r^{\text{simple-VB}}(A,B) = \prod_{h=1}^{H} r_{a_h}^{\text{simple-VB}}(\boldsymbol{a}_h) r_{b_h}^{\text{simple-VB}}(\boldsymbol{b}_h). \tag{27}$$

That is, all column-vectors of A and B are assumed to be mutually independent. Under this stronger constraint, Nakajima et al. (2010) derived the VB global solution in a closed-form.

From their result, the following theorem can be immediately obtained.

Theorem 3 *The simple-VB-PCA solution is given by Eq.*(8) *with the following threshold:*

$$\underline{\gamma}_h^{\textit{simple-VB}} = \sigma\sqrt{\kappa_h + \sqrt{\kappa_h^2 - LM}}, \tag{28}$$
$$\kappa_h = (L+M)/2 + \sigma^2/(2c_{b_h}^2).$$

Theorem 3 shows that the solution of simple-VB-PCA can be computed analytically in a very efficient way. On the other hand, the ordinary VB-PCA (with matrix-wise independence) requires a number of iterations to find a local optimal solution (see Section 2.2.3). Thus, simple-VB-PCA is computationally more attractive than the ordinary VB-PCA.

4.2. Simple Empirical VB-PCA (Simple-EVB-PCA)

As opposed to PB-PCA, the *empirical* Bayesian variant of simple-VB-PCA is well-defined, and its global solution can be obtained in a closed form as follows:

Theorem 4 *The simple-EVB-PCA solution is given by Eq.*(8) *with the following threshold:*

$$\underline{\gamma}_h^{\textit{simple-EVB}} = \begin{cases} \sigma(\sqrt{L}+\sqrt{M}) & \textit{if } \Delta_h < 0, \\ \infty & \textit{otherwise}, \end{cases} \tag{29}$$

where

$$\Delta_h = M\log\left(\frac{\gamma_h}{M\sigma^2}\breve{\gamma}_h^{\textit{simple-VB}} + 1\right) + L\log\left(\frac{\gamma_h}{L\sigma^2}\breve{\gamma}_h^{\textit{simple-VB}} + 1\right)$$
$$+ \frac{1}{\sigma^2}\left(LM\breve{c}_{b_h}^2 - 2\gamma_h\breve{\gamma}_h^{\textit{simple-VB}}\right),$$
$$\breve{c}_{b_h}^2 = \frac{\tau_h + \sqrt{\tau_h^2 - 4LM\sigma^4}}{2LM}, \quad \tau_h = \gamma_h^2 - (L+M)\sigma^2.$$

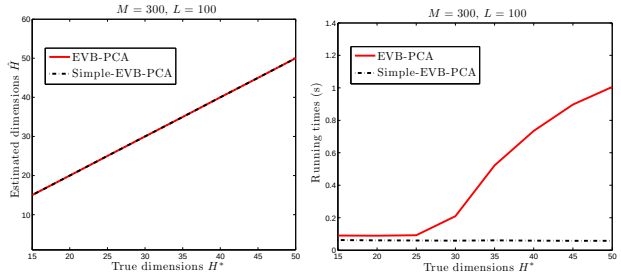

Figure 1. EVB-PCA vs. simple-EVB-PCA. Performance of the PCA dimensionality selection is highly comparable to each other (left), while simple-EVB-PCA is computationally more efficient than EVB-PCA (right).

$\breve{\gamma}_h^{\textit{simple-VB}} = \langle \|\boldsymbol{a}_h\| \|\boldsymbol{b}_h\| \rangle_{\widehat{r}^{\textit{simple-VB}}(A,B)}$ *is the simple-VB solution for* $c_{b_h} = \breve{c}_{b_h}$.

Below, we experimentally show that the above analytic-form solution of simple-VB-PCA works as well as the ordinary VB-PCA.

4.3. Experimental Comparison

In this section, we experimentally compare the performance of EVB-PCA and simple-EVB-PCA on artificial and real-world datasets. Through all the experiments, H is set to be the full rank ($H = \min(L, M)$), and the noise variance σ^2 is estimated based on the free energy minimization, similarly to the hyperparameter estimation through the empirical Bayes method (see Section 2.3).

In EVB-PCA, hyperparameters are updated in each iteration as

$$c_{b_h}^2 = \|\boldsymbol{\mu}_{b_h}\|^2 / L + \left(\Sigma_{\widetilde{B}}\right)_{hh}, \tag{30}$$

which is obtained as a stationary point of the free energy with respect to $\{c_{b_h}^2\}$ (Bishop, 1999a). The noise variance σ^2 is estimated by using the following update rule in each iteration

$$\sigma^2 = \frac{\|Y - \sum_{h=1}^{H} \boldsymbol{\mu}_{b_h} \boldsymbol{\mu}_{a_h}^\top\|_{\mathrm{Fro}}^2}{LM} + \frac{\sum_{l=1}^{L} \boldsymbol{\mu}_{\widetilde{b}_l}^\top \Sigma_{\widetilde{A}} \boldsymbol{\mu}_{\widetilde{b}_l}}{L} + \frac{\sum_{m=1}^{M} \boldsymbol{\mu}_{\widetilde{a}_m}^\top \Sigma_{\widetilde{B}} \boldsymbol{\mu}_{\widetilde{a}_m}}{M} + \mathrm{tr}\left(\Sigma_{\widetilde{A}} \Sigma_{\widetilde{B}}\right), \tag{31}$$

which is obtained again as a stationary point of the free energy with respect to σ^2 (Bishop, 1999a).

In simple-EVB-PCA, we use Theorem 4 combined with a naive 1-dimensional search strategy over the free energy for the estimation of σ^2 (Nakajima et al., 2010).

The strong independence assumption of simple-VB-PCA given in Eq.(27) naturally leads to the question, whether simple-VB-PCA can provide accurate estimation when the data has a correlated structure. To investigate this, we created an artificial dataset having variances 5 and covariances 1 in the first H^* directions, and variances 0.1 and no covariance in the remaining $L - H^*$ directions.

Table 1. Estimated effective data dimensions of real datasets. $\widehat{H}^{\mathrm{EVB}}$ and $\widehat{H}^{\text{simple-EVB}}$ denote the PCA dimensions estimated by EVB-PCA and simple-EVB-PCA, respectively.

Data set	M	L	$\widehat{H}^{\mathrm{EVB}}$	$\widehat{H}^{\text{simple-EVB}}$
Chart	600	60	11	10
Glass	214	9	7	7
Wine	178	13	7	8
Optical Digits	5620	64	56	56
Satellite	6435	36	32	31
Segmentation	2310	19	12	13
Letter	20000	16	15	15

Figure 1 shows the estimated PCA dimensions with varying true dimension H^*. It can be observed that EVB-PCA and simple-EVB-PCA have very similar dimension estimation performance, while simple-EVB-PCA is computationally more efficient than EVB-PCA[4]. We also performed similar experiments with various settings (e.g., different levels of correlation, different number of samples M, and different dimensionality L), and empirically observed similar trends to Figure 1.[5]

We we further tested both methods on seven datasets taken from the UCI repository (Asuncion & Newman, 2007). The specification of the dataset as well as the number of PCA dimensions retained by EVB-PCA and simple-EVB-PCA is shown in Table 1. As with the synthetic data, the behavior of the two methods is very similar to each other.

In summary, we observed that the stronger independence assumption (27) does not significantly change the performance. Given that simple-EVB-PCA and EVB-PCA perform similarly, we conclude that simple-EVB-PCA is more attractive than EVB-PCA because it has an analytic-form solution that can be computed very efficiently.

5. Discussion

In this section, we further discuss various issues in several PPCA methods, and give insights.

5.1. Maximum A Posteriori PCA (MAP-PCA)

For comparison purposes, we define the maximum a posteriori PCA (MAP-PCA), where both A and B are point-

[4] Note that slightly different hyperpriors are used in the original EVB-PCA papers (Bishop, 1999a;b). We also tested them and confirmed that they also behave similarly to the simple-EVB-PCA.

[5] We also observed that the iterative algorithm (Eqs.(2) and (3) in Nakajima et al. (2010)) for simple-EVB-PCA gives similar dimension estimation performance.

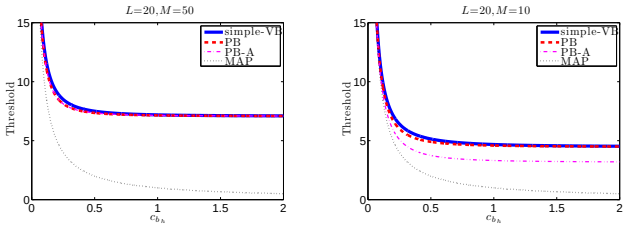

Figure 2. Behavior of the thresholds, $\gamma_h^{\text{simple-VB}}$, γ_h^{PB}, $\gamma_h^{\text{PB-A}}$, and γ_h^{MAP}. The noise variance is set to $\sigma^2 = 1$.

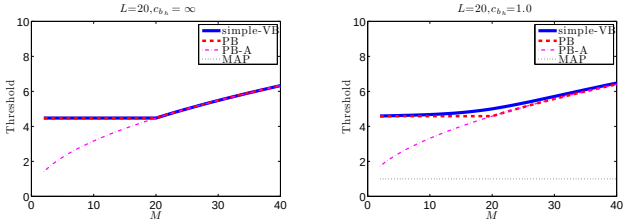

Figure 3. Behavior of the thresholds when the number M of samples is changed.

estimated. This corresponds to the posterior restricted to the product of delta functions:

$$r^{\text{MAP}}(A, B) = \delta(A; A^*)\delta(B; B^*). \tag{32}$$

Its solution is given as follows.

Proposition 2 *(Srebro et al., 2005) The MAP-PCA solution is given by Eq.*(8) *with the following threshold:*

$$\gamma_h^{MAP} = \sigma^2 / c_{b_h}. \tag{33}$$

When $c_{b_h} \to \infty$, MAP-PCA is reduced to the classical PCA, i.e., no automatic dimensionality selection (ADS).

5.2. Comparison between PB-A-PCA, PB-PCA, and simple-VB-PCA

The threshold (26) becomes simpler when the prior is flat:

$$\lim_{c_{b_h} \to \infty} \gamma_h^{\text{PB}} = \sigma \sqrt{\max(L, M)}. \tag{34}$$

Interestingly, the simple-VB-PCA solution agrees with the PB-PCA solution under the flat prior:

$$\lim_{c_{b_h} \to \infty} \gamma_h^{\text{simple-VB}} = \sigma \sqrt{\max(L, M)}. \tag{35}$$

Below, we numerically investigate the behavior of the PPCA thresholds in more general settings.

Figure 2 compares the thresholds of simple-VB-PCA (28), PB-PCA (26), PB-A-PCA (23), and MAP-PCA (33) in two situations. In the left graph (when the dimensionality of the original space is $L = 20$ and the number of samples is $M = 50$), PB-PCA and PB-A-PCA coincide, and simple-VB-PCA behaves similarly to them. On the other hand, MAP-PCA behaves differently. This is because of the effect of *model-induced regularization* (MIR) (Nakajima et al., 2010), which indicates the fact that, when the model is *non-identifiable*, even if the prior is *almost* flat ($c_{b_h} \to \infty$), Bayesian methods have a regularization effect as long as (a part of) parameters are integrated out. PB-PCA and VB-PCA belong to this class, but MAP-PCA does not.

In the right graph (when $L = 20$ and $M = 10$), PB-PCA and simple-VB-PCA behave almost identically. Although PB-A-PCA also shows its MIR effect (i.e., the threshold do not converge to zero even in the limit $c_{b_h} \to \infty$), a significant difference from simple-VB-PCA and PB-PCA is observed.

Figure 3 shows how the number M of samples affects the thresholds. In the limit $c_{b_h} \to \infty$ (left graph), Eqs.(34) and (35) together state that the solutions of PB-PCA and simple-VB-PCA agrees with each other. PB-A-PCA also gives the identical solution to them when $M \geq L$. However, its behavior changes at $M = L$; the threshold of PB-A-PCA smoothly goes down as M decreases, while those of PB-PCA and simple-VB-PCA make a sudden turn and becomes constant. The right graph in Figure 3 shows the case with a non-flat prior ($c_{b_h} = 1$), which also shows a similar phenomenon.

A question is which is more desirable, simple-VB-PCA/PB-PCA (which are with a sudden turn in the threshold curve), or PB-A-PCA (which is with smooth behavior). When $M < L$, PB-PCA and VB-PCA more strongly regularize the solutions than PB-A-PCA. We argue that the behavior of PB-PCA and VB-PCA is more reasonable because of the following reason. Let us consider the case where no driving latent variable exists, i.e., the true dimension is $H^* = 0$. In this case, we merely observe pure noise, and the average of the squared singular values of Y over all the components is given by

$$\frac{\langle \text{tr}(YY^\top) \rangle_{\mathcal{N}(0,\sigma^2)}}{\min(L, M)} = \sigma^2 \max(L, M).$$

Comparing this with the thresholds (34) and (35) of PB-PCA and simple-VB-PCA, we find that they eliminate the components with singular values equal to the average noise contribution. The sudden turn in the threshold curve actually follows the behavior of the noise contribution, which would be reasonable. In this sense, PB-A-PCA is suboptimal since it strongly overfits the noise when $L \gg M$.

5.3. Changing Behavior of Simple-VB-PCA

In the left graph of Figure 3, the behavior of PB-PCA and simple-VB-PCA suddenly changes at $L = M$. Here, we theoretically investigate the reason for this phenomenon.

Table 2. Properties of approximate PPCA methods. Upper methods have weaker constraints on the posterior distribution. '◯' means the method positively possesses the corresponding property, '△' means weakly, and '×' means not.

Method	Automatic dimensionality selection	Empirical Bayes	Analytic-form solution
VB	◯	◯	×
simple-VB	◯	◯	◯
PB	◯	×	◯
PB-A (PB-B)	△	×	◯
MAP	×	×	◯

As Lemma 1 states, the PB-PCA posterior drastically changes depending on $L > M$ or $L < M$. This leads to the sudden turn in the thresholding curve. We can show that a similar effect also occurs in simple-VB-PCA, which is explained below. Let $\sigma_{a_h}^{\text{simple-VB}}$ and $\sigma_{b_h}^{\text{simple-VB}}$ be the standard deviations of the simple-VB posteriors of $\|\boldsymbol{a}_h\|$ and $\|\boldsymbol{b}_h\|$, respectively. Then the following lemma holds:

Lemma 2 *When $c_{b_h} \to \infty$ and $\gamma_h = \underline{\gamma}_h^{\text{simple-VB}}$, it holds that*

$$\frac{\sigma_{a_h}^{\text{simple-VB}}}{\sigma_{b_h}^{\text{simple-VB}}} = \begin{cases} \frac{(M-L)}{\sigma\sqrt{M}} + o(1) = O(1) & \text{if } M > L, \\ \frac{1}{c_{b_h}} + o(c_{b_h}^{-1}) = O(c_{b_h}^{-1}) & \text{if } M = L, \\ \frac{\sigma\sqrt{L}}{(L-M)c_{b_h}^2} + o(c_{b_h}^{-2}) = O(c_{b_h}^{-2}) & \text{if } M < L. \end{cases}$$

This lemma implies that with the (almost) flat prior, the shape of the simple-VB posterior suddenly changes. More specifically, the shape is spherical when $M > L$, while it is strongly elliptical in the $\boldsymbol{b}_h$-space when $M \leq L$.

6. Conclusion

In this paper, we revisited approximate Bayesian methods in PPCA, which are summarized in Table 2.

Although VB-PCA is the most general approximate PPCA method, it is computationally less efficient because it requires a number of iterations to find a local optimal solution. On the other hand, we showed by providing an analytic-form solution that PB-PCA is computationally more efficient (Section 3.1). The analytic-form solution also showed that, despite the fact that VB-PCA was proposed to induce automatic dimensionality selection (ADS), PB-PCA is already equipped with ADS. We also numerically showed that PB-PCA behaves similarly to simple-VB-PCA (Section 5.2). However, as shown in Section 3.2, PB-PCA has a critical disadvantage that its empirical Bayesian variant is practically useless.

On the other hand, simple-VB-PCA (whose 'complexity' is between PB-PCA and VB-PCA) still has an analytic-form solution (Section 4.1), and its empirical Bayesian version, simple-EVB-PCA, also has an analytic-form solution (Section 4.2). Experimentally, simple-EVB-PCA was shown to perform similarly to the computationally more demanding counterpart, EVB-PCA (Section 4.3). Based on these findings, we concluded that simple-VB-PCA is the most attractive PPCA method.

Our future work is to extend the current discussion to more complex models such as a mixture of PPCAs and missing value estimation. Also, we will further investigate the role of column-wise independence in simple-VB-PCA.

Acknowledgments

The authors thank anonymous reviewers for helpful comments to improve the paper. They also thank Ryota Tomioka of The University of Tokyo for discussion. MS was supported by the FIRST program. DB was supported by a Beckman Postdoctoral Fellowship.

References

Asuncion, A. and Newman, D.J. UCI machine learning repository, 2007.

Bishop, C. M. Variational principal components. *ICANN1999*, 1999a.

Bishop, C. M. Bayesian principal components. *NIPS1998*, 1999b.

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

Hoyle, D. C. Autotmaric PCA dimension selection for high dimensional data and small sample sizes. *Journal of Machine Learning Research*, 9:2733–2759, 2008.

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

Nakajima, S., Sugiyama, M., and Tomioka, R. Global analytic solution for variational Bayesian matrix factorization. *NIPS2010*.

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

Raiko, T., Ilin, A., and Karhunen, J. Principal component analysis for large scale problems with lots of missing values. *ECML2007*.

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

Salakhutdinov, R. and Mnih, A. Probabilistic matrix factorization. *NIPS2007*.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum margin matrix factorization. *NIPS2004*.

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