Journal of Machine Learning Research 16 (2015) 3757-3811 Submitted 5/14; Revised 3/15; Published 12/15

Condition for Perfect Dimensionality Recovery by Variational Bayesian PCA*

Shinichi Nakajima NAKAJIMA@TU-BERLIN.DE
Berlin Big Data Center
Technische Universität Berlin
Berlin 10587 Germany

Ryota Tomioka TOMIOKA@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, IL 60637 USA

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
Department of Complexity Science and Engineering
The University of Tokyo
Tokyo 113-0033 Japan

S. Derin Babacan DBABACAN@GMAIL.COM
Google Inc.
Mountain View, CA 94043 USA

Editor: David Barber

Abstract

Having shown its good performance in many applications, variational Bayesian (VB) learning is known to be one of the best tractable approximations to Bayesian learning. However, its performance was not well understood theoretically. In this paper, we clarify the behavior of VB learning in probabilistic PCA (or fully-observed matrix factorization). More specifically, we establish a necessary and sufficient condition for perfect dimensionality (or rank) recovery in the large-scale limit when the matrix size goes to infinity. Our result theoretically guarantees the performance of VB-PCA. At the same time, it also reveals the conservative nature of VB learning—it offers a low false positive rate at the expense of low sensitivity. By contrasting with an alternative dimensionality selection method, we characterize VB learning in PCA. In our analysis, we obtain bounds of the noise variance estimator, and a new and simple analytic-form solution for the other parameters, which themselves are useful for implementation of VB-PCA.

Keywords: variational Bayesian learning, matrix factorization, principal component analysis, automatic relevance determination, perfect dimensionality recovery

1. Introduction

Variational Bayesian (VB) learning (Attias, 1999; Bishop, 2006) was proposed as a computationally efficient approximation to Bayesian learning. The key idea is to find the closest distribution to the Bayes posterior in a restricted function space, where the expectation—an often intractable operation in Bayesian learning—can be easily performed. VB learning

* This paper is an extended version of our earlier conference paper (Nakajima et al., 2012).

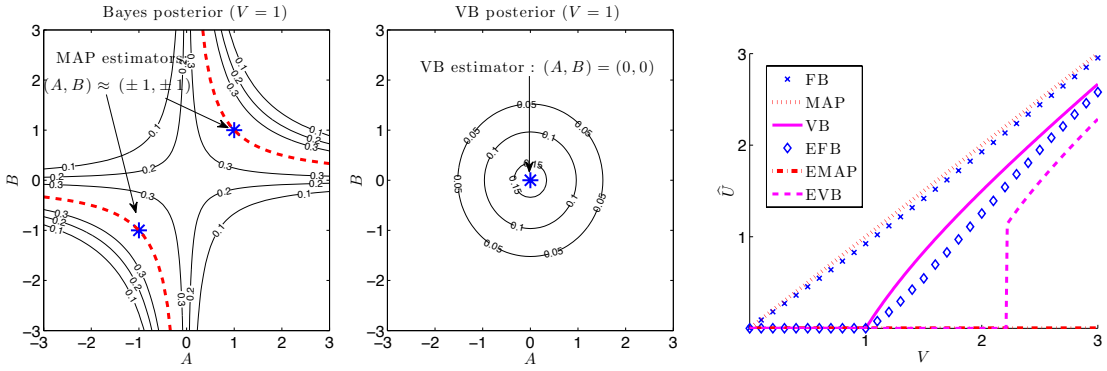


Figure 1: Dissimilarities between VB and rigorous Bayesian learning. (Left and Center) The Bayes posterior and the VB posterior of the 1×1 MF model $V = BA + \mathcal{E}$ with almost flat prior, when $V = 1$ is observed ($\mathcal{E}$ is Gaussian noise). VB approximates the Bayes posterior having two modes by an origin-centered Gaussian, which induces sparsity. (Right) Behavior of estimators of $U = BA$, given the observation V. The VB estimator (the magenta solid curve) is zero when $V \leq 1$, which indicates *exact* sparsity. On the other hand, FB (fully-Bayesian or rigorous Bayesian learning; blue crosses) shows no sign of sparsity. All graphs are quoted from Nakajima and Sugiyama (2011).

has been applied to many applications, and its good performance has been experimentally shown (Bishop, 1999a; Bishop and Tipping, 2000; Ghahramani and Beal, 2001; Jaakkola and Jordan, 2000; Blei et al., 2003; Sato et al., 2004; Lim and Teh, 2007; Seeger, 2009; Ilin and Raiko, 2010). Typically, the restriction is imposed as a factorized form of posterior, under which a tractable iterative algorithm is derived.

Although the VB algorithm is simple and efficient, it solves a non-convex optimization problem, which makes theoretical analysis difficult. An exceptional case is the matrix factorization (MF) model (Bishop, 1999a; Lim and Teh, 2007; Ilin and Raiko, 2010; Salakhutdinov and Mnih, 2008) with fully-observed matrices, in which the global VB solution has been analytically obtained (Nakajima et al., 2013b), and some properties have been theoretically revealed (Nakajima and Sugiyama, 2011). These works also posed thought-provoking relations between VB and rigorous Bayesian learning: The VB posterior is actually quite different from the true Bayes posterior (compare the left and the middle graphs in Figure 1), and VB induces sparsity in its solution but such sparsity is hardly observed in rigorous Bayesian learning (see the right graph in Fig. 1). Actually, Mackay (2001) has discussed the sparsity of VB as an artifact by showing *inappropriate* model pruning in mixture models. These facts might deprive the justification of VB based solely on the fact that it is one of the best tractable approximations to Bayesian learning.

The goal of this paper is to provide direct justification for VB learning. Focusing on the probabilistic PCA (Tipping and Bishop, 1999; Roweis and Ghahramani, 1999; Bishop, 1999a), an instance of fully-observed MF, we give a theoretical guarantee for the performance of VB learning. Our starting point is the global analytic solution derived by Nakajima

et al. (2013b). After describing our formulation in Section 2, we conduct the following three steps:

1. We derive a new and simple analytic-form of the global VB solution in Section 3.

   The analytic-form solution derived in Nakajima et al. (2013b) is expressed with a solution of a *quartic* equation, which obstructs further analysis. In this paper, we derive an alternative form, which consists of simple algebra.

2. We obtain a simple form of the objective function for noise variance estimation in Section 4.

   The previous analyses in Nakajima and Sugiyama (2011) and in Nakajima et al. (2013b) assumed that the noise variance is a given constant. In this paper, we assume that the noise variance is also estimated from observation, and derive an objective function, of which the minimizer gives the noise variance estimator. We also derive bounds of the rank estimator and the noise variance estimator.

3. We establish a necessary and sufficient condition for perfect dimensionality recovery in Section 5.

   Combining the results obtained in the former two steps with random matrix theory (Marčenko and Pastur, 1967; Wachter, 1978; Johnstone, 2001; Hoyle and Rattray, 2004; Baik and Silverstein, 2006), we establish a necessary and sufficient condition that VB-PCA perfectly recovers the true dimensionality in the *large-scale limit* when the matrix size goes to infinity.

To the best of our knowledge, this is the first theoretical result that guarantees the performance of VB learning. To give more insight into practical situations, we also derive a sufficient condition for perfect recovery, which approximately holds for moderate-sized matrices. It is worth noting that, although the objective function minimized for noise variance estimation is non-convex and possibly multimodal in general, only a local search algorithm is required for perfect recovery.

Section 6 is devoted to discussion on a few topics. First, we propose a simple implementation of VB-PCA, based on the new analytic-form solution and the bounds of the noise variance estimator, which are obtained in our analysis. After that, we consider the behavior of VB learning in more detail. Our result theoretically guarantees the performance of VB-PCA. At the same time, it also reveals the conservative nature of VB learning—it offers a low false positive rate at the expense of low sensitivity, due to which VB-PCA does not behave *optimally* in the large-scale limit. By contrasting with an alternative dimensionality selection method, called the *overlap* (OL) method (Hoyle, 2008), we characterize VB learning in PCA.

Section 7 concludes, and Appendix provides all technical details.

2. Formulation

In this section, we formulate variational Bayesian learning in the matrix factorization model.

2.1 Probabilistic Matrix Factorization

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

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

In the *matrix factorization* (MF) model, the target matrix is assumed to be low rank, and therefore can be factorized as

$$\boldsymbol{U} = \boldsymbol{B}\boldsymbol{A}^\top,$$

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

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

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

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

$$p(\boldsymbol{B}) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{C}_B^{-1}\boldsymbol{B}^\top\right)\right). \tag{3}$$

Here, we denote by $\|\cdot\|_{\mathrm{Fro}}$ the Frobenius norm, and by $\mathrm{tr}(\cdot)$ the trace of a matrix. Throughout the paper, we assume that

$$L \leq M. \tag{4}$$

If $L > M$, we may simply re-define the transpose $\boldsymbol{V}^\top$ as $\boldsymbol{V}$ so that $L \leq M$ holds. Therefore, the assumption (4) does not impose any restriction. We assume that the prior covariance matrices $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$ are diagonal and positive definite, i.e.,

$$\boldsymbol{C}_A = \mathbf{diag}(c_{a_1}^2, \ldots, c_{a_H}^2),$$
$$\boldsymbol{C}_B = \mathbf{diag}(c_{b_1}^2, \ldots, c_{b_H}^2),$$

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 $\boldsymbol{C}_A\boldsymbol{C}_B$ are arranged in non-increasing order, i.e., $c_{a_h}c_{b_h} \geq c_{a_{h'}}c_{b_{h'}}$ for any pair $h < h'$. We denote a column vector of a matrix by a bold lowercase letter, i.e.,

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

2.2 Variational Bayesian Approximation

The Bayes posterior is given by

$$p(\boldsymbol{A}, \boldsymbol{B}|\boldsymbol{V}) = \frac{p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B})p(\boldsymbol{A})p(\boldsymbol{B})}{p(\boldsymbol{V})}, \tag{5}$$

where $p(\boldsymbol{V}) = \langle p(\boldsymbol{V}|\boldsymbol{A},\boldsymbol{B})\rangle_{p(\boldsymbol{A})p(\boldsymbol{B})}$. Here, $\langle\cdot\rangle_p$ denotes the expectation over the distribution p. Since this expectation is intractable, we need to approximate the Bayes posterior.

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

$$F(r) = \left\langle \log \frac{r(\boldsymbol{A},\boldsymbol{B})}{p(\boldsymbol{V}|\boldsymbol{A},\boldsymbol{B})p(\boldsymbol{A})p(\boldsymbol{B})} \right\rangle_{r(\boldsymbol{A},\boldsymbol{B})} = \left\langle \log \frac{r(\boldsymbol{A},\boldsymbol{B})}{p(\boldsymbol{A},\boldsymbol{B}|\boldsymbol{V})} \right\rangle_{r(\boldsymbol{A},\boldsymbol{B})} - \log p(\boldsymbol{V}). \tag{6}$$

In the last equation, the first term is the Kullback-Leibler (KL) divergence from the trial distribution to the Bayes posterior (5), and the second term is constant. Therefore, minimizing the free energy amounts to finding a distribution closest to the Bayes posterior in the sense of the KL divergence. A general approach to Bayesian approximate inference is to find the minimizer of the free energy (6) with respect to r in some restricted function space.

In the VB approximation, the independence between the entangled parameter matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is assumed:

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

Under this constraint, an iterative algorithm for minimizing the free energy (6) was derived (Bishop, 1999a; Lim and Teh, 2007). Let $\widehat{r}$ be the obtained minimizer. We define the MF solution by the mean of the target matrix $\boldsymbol{U}$:

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

The MF model has hyperparameters $(\boldsymbol{C}_A, \boldsymbol{C}_B)$ in the priors (2) and (3). By manually choosing them, we can control regularization and sparsity of the solution (e.g., the PCA dimension in our setting). A popular way to set the hyperparameter in the Bayesian framework is again based on the minimization of the free energy (6):

$$(\widehat{\boldsymbol{C}}_A, \widehat{\boldsymbol{C}}_B) = \operatorname*{argmin}_{\boldsymbol{C}_A,\boldsymbol{C}_B} \left(\min_r F(r; \boldsymbol{C}_A, \boldsymbol{C}_B|\boldsymbol{V}) \right).$$

We refer to this method as an empirical VB (EVB) method. When the noise variance σ^2 is unknown, it can also be estimated as

$$\widehat{\sigma}^2 = \operatorname*{argmin}_{\sigma^2} \left(\min_{r,\boldsymbol{C}_A,\boldsymbol{C}_B} F(r; \boldsymbol{C}_A, \boldsymbol{C}_B, \sigma^2|\boldsymbol{V}) \right).$$

3. Simple Analytic-Form Solution

Recently, an analytic-form of the global VB, as well as EVB, solution for MF has been derived (Nakajima et al., 2013b), which enables us to reach the global solution easily. However, the form involves a solution of a *quartic* equation, which obstructs further analysis. In this section, we derive a simple alternative form of the global VB, as well as EVB, solution, which facilitates subsequent analysis.

3.1 VB Solution

Let

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

be the singular value decomposition (SVD) of $\boldsymbol{V}$, where γ_h (≥ 0) is the h-th largest singular value, and $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ are the associated right and left singular vectors. We denote by $\mathcal{N}_d(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, by $\boldsymbol{I}_d$ the d-dimensional identity matrix, and by $\mathbb{R}_{++}$ the set of the positive real numbers.

Under the independence assumption (7), it is easily shown that the VB posterior has the Gaussian form:

$$r(\boldsymbol{A}, \boldsymbol{B}) \propto \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}_A^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})^\top\right)}{2}\right) \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{B}-\widehat{\boldsymbol{B}})\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{B}-\widehat{\boldsymbol{B}})^\top\right)}{2}\right)$$

with the means $\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}$ and the covariances $\boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B$ minimizing the free energy (6), which is explicitly written as

$$\begin{aligned} 2F = LM\log(2\pi\sigma^2) + \frac{\left\|\boldsymbol{V} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^\top\right\|^2}{\sigma^2} + M\log\frac{|\boldsymbol{C}_A|}{|\boldsymbol{\Sigma}_A|} + L\log\frac{|\boldsymbol{C}_B|}{|\boldsymbol{\Sigma}_B|} - (L+M)H \\ + \mathrm{tr}\left(\boldsymbol{C}_A^{-1}\left(\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\boldsymbol{\Sigma}_A\right)\right) + \mathrm{tr}\left(\boldsymbol{C}_B^{-1}\left(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\boldsymbol{\Sigma}_B\right)\right) \\ + \frac{\mathrm{tr}\left(-\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + \left(\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\boldsymbol{\Sigma}_A\right)\left(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\boldsymbol{\Sigma}_{\boldsymbol{B}}\right)\right)}{\sigma^2}. \end{aligned} \tag{8}$$

Here $|\cdot|$ denotes the determinant of a matrix. The derivatives of the free energy (8) give the following stationary condition, which is used for constructing an iterative local search algorithm:

$$\widehat{\boldsymbol{A}} = \boldsymbol{V}^\top\widehat{\boldsymbol{B}}\frac{\boldsymbol{\Sigma}_A}{\sigma^2}, \qquad \boldsymbol{\Sigma}_A = \sigma^2\left(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\boldsymbol{\Sigma}_B + \sigma^2\boldsymbol{C}_A^{-1}\right)^{-1}, \tag{9}$$

$$\widehat{\boldsymbol{B}} = \boldsymbol{V}\widehat{\boldsymbol{A}}\frac{\boldsymbol{\Sigma}_B}{\sigma^2}, \qquad \boldsymbol{\Sigma}_B = \sigma^2\left(\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\boldsymbol{\Sigma}_A + \sigma^2\boldsymbol{C}_B^{-1}\right)^{-1}. \tag{10}$$

In our previous work, we proved that finding the solution with diagonal covariances is sufficient—any solution has an *equivalent transform* to the solution such that $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$ are diagonal (Theorem 1 in Nakajima et al. (2013b)). Under the focus on diagonal covariances, the stationary condition (9) and (10) implies that $\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}}$ are also diagonal, meaning that the column vectors of $\widehat{\boldsymbol{A}}$, as well as $\widehat{\boldsymbol{B}}$, are orthogonal to each other. Then, we find that the column vectors of $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{B}}$ only depend on the second term in Eq.(8), which coincides with the objective for (truncated) SVD. Consequently, the mean parameters are expressed as $\widehat{\boldsymbol{a}}_h = \widehat{a}_h\boldsymbol{\omega}_{a_h}$ and $\widehat{\boldsymbol{b}}_h = \widehat{b}_h\boldsymbol{\omega}_{b_h}$ (Lemma 8 in Nakajima and Sugiyama (2011)), and the following proposition thus holds:

Proposition 1 *(Nakajima et al., 2013b) The VB posterior can be written as*

$$r(\boldsymbol{A}, \boldsymbol{B}) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \widehat{a}_h \boldsymbol{\omega}_{a_h}, \sigma_{a_h}^2 \boldsymbol{I}_M) \mathcal{N}_L(\boldsymbol{b}_h; \widehat{b}_h \boldsymbol{\omega}_{b_h}, \sigma_{b_h}^2 \boldsymbol{I}_L), \tag{11}$$

where $\{\widehat{a}_h, \widehat{b}_h, \sigma_{a_h}^2, \sigma_{b_h}^2\}_{h=1}^H$ *are the solution of the following minimization problem:*

$$\begin{aligned} &\textit{Given} \quad \sigma^2 \in \mathbb{R}_{++}, \quad \{c_{a_h}^2, c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H, \\ &\min_{\{\widehat{a}_h, \widehat{b}_h, \sigma_{a_h}^2, \sigma_{b_h}^2\}_{h=1}^H} \quad 2F, \\ &\textit{s.t.} \quad \{\widehat{a}_h, \widehat{b}_h \in \mathbb{R}, \quad \sigma_{a_h}^2, \sigma_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H. \end{aligned} \tag{12}$$

Here, F is the free energy (6)*, which can be written as*

$$2F = LM \log(2\pi\sigma^2) + \frac{\sum_{h=1}^L \gamma_h^2}{\sigma^2} + \sum_{h=1}^H 2F_h, \tag{13}$$

$$\begin{aligned} \textit{where} \quad 2F_h = M \log \frac{c_{a_h}^2}{\sigma_{a_h}^2} + L \log \frac{c_{b_h}^2}{\sigma_{b_h}^2} + \frac{\widehat{a}_h^2 + M\sigma_{a_h}^2}{c_{a_h}^2} + \frac{\widehat{b}_h^2 + L\sigma_{b_h}^2}{c_{b_h}^2} \\ - (L+M) + \frac{-2\widehat{a}_h \widehat{b}_h \gamma_h + \left(\widehat{a}_h^2 + M\sigma_{a_h}^2\right)\left(\widehat{b}_h^2 + L\sigma_{b_h}^2\right)}{\sigma^2}. \end{aligned} \tag{14}$$

The minimization problem (12) has been analytically solved (Nakajima et al., 2013b), which provides an analytic-form of the global VB solution (see Proposition 18 in Appendix A). However, the form involves a solution of a *quartic* equation, with which further analysis is difficult. In this paper, finding a shortcut to an alternative *quadratic* equation, we obtain the following theorem, which provides a new and simple analytic-form of the global VB solution (the proof is given in Appendix A):

Theorem 2 *The VB solution can be written as truncated shrinkage SVD as follows:*

$$\widehat{\boldsymbol{U}}^{\mathrm{VB}} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top, \qquad \textit{where} \qquad \widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{VB}} & \textit{if } \gamma_h \geq \underline{\gamma}_h^{\mathrm{VB}}, \\ 0 & \textit{otherwise.} \end{cases} \tag{15}$$

Here, the truncation threshold and the shrinkage estimator are, respectively, given by

$$\underline{\gamma}_h^{\mathrm{VB}} = \sigma \sqrt{\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM}}, \tag{16}$$

$$\breve{\gamma}_h^{\mathrm{VB}} = \gamma_h \left(1 - \frac{\sigma^2}{2\gamma_h^2}\left(M + L + \sqrt{(M-L)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}\right)\right). \tag{17}$$

Our new form with the truncation threshold (16) and the shrinkage estimator (17) consisting of simple algebra facilitates further analysis.

The VB posterior is also written in a simple form (the proof is given in Appendix A):

Corollary 3 *The VB posterior is given by Eq.*(11) *with the following estimators: If* $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$,

$$\widehat{a}_h = \pm\sqrt{\breve{\gamma}_h^{\mathrm{VB}}\widehat{\delta}_h^{\mathrm{VB}}}, \qquad \widehat{b}_h = \pm\sqrt{\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\widehat{\delta}_h^{\mathrm{VB}}}}, \qquad \sigma_{a_h}^2 = \frac{\sigma^2\widehat{\delta}_h^{\mathrm{VB}}}{\gamma_h}, \qquad \sigma_{b_h}^2 = \frac{\sigma^2}{\gamma_h\widehat{\delta}_h^{\mathrm{VB}}}, \tag{18}$$

$$\textit{where} \qquad \widehat{\delta}_h^{\mathrm{VB}}\left(\equiv\frac{\widehat{a}_h}{\widehat{b}_h}\right) = \frac{c_{a_h}^2}{\sigma^2}\left(\gamma_h - \breve{\gamma}_h^{\mathrm{VB}} - \frac{L\sigma^2}{\gamma_h}\right), \tag{19}$$

and otherwise,

$$\widehat{a}_h = 0, \qquad \widehat{b}_h = 0, \qquad \sigma_{a_h}^2 = c_{a_h}^2\left(1 - \frac{L\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right), \qquad \sigma_{b_h}^2 = c_{b_h}^2\left(1 - \frac{M\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right), \tag{20}$$

$$\textit{where} \quad \widehat{\zeta}_h^{\mathrm{VB}}\left(\equiv\sigma_{a_h}^2\sigma_{b_h}^2\right) = \frac{\sigma^2}{2LM}\left(L+M+\frac{\sigma^2}{c_{a_h}^2c_{b_h}^2} - \sqrt{\left(L+M+\frac{\sigma^2}{c_{a_h}^2c_{b_h}^2}\right)^2 - 4LM}\right). \tag{21}$$

3.2 EVB Solution

The empirical VB (EVB) learning, where the hyperparameters $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$ are also estimated from observation, solves the following problem:

$$\begin{aligned} &\text{Given} \quad \sigma^2 \in \mathbb{R}_{++}, \\ &\min_{\{\widehat{a}_h,\widehat{b}_h,\sigma_{a_h}^2,\sigma_{b_h}^2,c_{a_h}^2,c_{b_h}^2\}_{h=1}^H} \quad 2F, \\ &\text{s.t.} \quad \{\widehat{a}_h,\widehat{b}_h \in \mathbb{R}, \quad \sigma_{a_h}^2,\sigma_{b_h}^2,c_{a_h}^2,c_{b_h}^2 \in \mathbb{R}_{++}\}_{h=1}^H. \end{aligned}$$

This problem has also been analytically solved (Nakajima et al., 2013b), which enables efficient computation of the global EVB solution (see Proposition 23 in Appendix B). However, the form requires to solve a quartic equation, and also to evaluate the free energy (14) to judge whether EVB discards each component. This again obstructs further analysis.

By substituting the VB solution, given by Theorem 2 and Corollary 3, we can derive an explicit form of the free energy (13) as a function of $\{c_{a_h}^2, c_{b_h}^2\}_{h=1}^H$ and σ^2. Minimizing it with respect to $\{c_{a_h}^2, c_{b_h}^2\}_{h=1}^H$, we obtain the following theorem, which provides a new and simple analytic-form of the global EVB solution (the proof is given in Appendix B):

Theorem 4 *Let*

$$\alpha = \frac{L}{M} \qquad (0 < \alpha \leq 1), \tag{22}$$

and let $\underline{\tau} = \underline{\tau}(\alpha)$ *be the unique zero-cross point of the following decreasing function:*

$$\Xi(\tau;\alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right), \qquad \textit{where} \qquad \Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}. \tag{23}$$

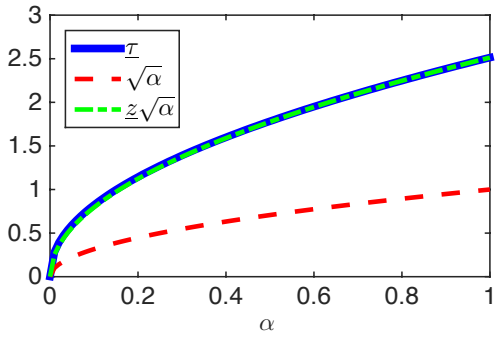


Figure 2: Values of $\underline{\tau}(\alpha)$, $\sqrt{\alpha}$, and $\underline{z}\sqrt{\alpha}$.

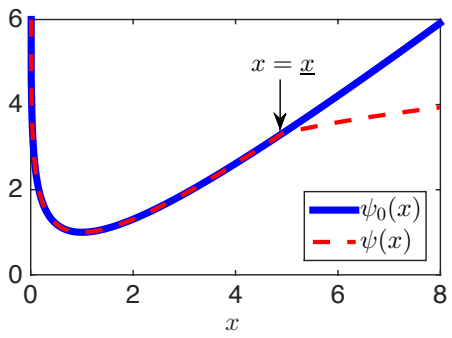


Figure 3: $\psi_0(x)$ and $\psi(x)$.

Then, the EVB solution can be written as truncated shrinkage SVD as follows:

$$\widehat{\boldsymbol{U}}^{\mathrm{EVB}} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{EVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}, \qquad \textit{where} \qquad \widehat{\gamma}_h^{\mathrm{EVB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{EVB}} & \textit{if } \gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}, \\ 0 & \textit{otherwise.} \end{cases} \tag{24}$$

Here, the truncation threshold and the shrinkage estimator are, respectively, given by

$$\underline{\gamma}^{\mathrm{EVB}} = \sigma \sqrt{M\,(1+\underline{\tau})\left(1+\frac{\alpha}{\underline{\tau}}\right)}, \tag{25}$$

$$\breve{\gamma}_h^{\mathrm{EVB}} = \frac{\gamma_h}{2}\left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2} + \sqrt{\left(1 - \frac{(M+L)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}\right). \tag{26}$$

The EVB threshold (25) involves $\underline{\tau}$, which needs to be numerically computed. However, we can easily prepare a table of the values for $0 < \alpha \leq 1$ beforehand, like the cumulative Gaussian probability used in statistical tests. Alternatively, $\underline{\tau} \approx \underline{z}\sqrt{\alpha}$ is a good approximation, where $\underline{z} \approx 2.5129$ is the unique zero-cross point of $\Phi(z)$, as seen in Figure 2. We can show that $\underline{\tau}$ lies in the following range (see Appendix B for its proof):

$$\sqrt{\alpha} < \underline{\tau} \leq \underline{z}. \tag{27}$$

We will see in Section 5 that $\underline{\tau}$ is an important quantity in describing the behavior of the EVB solution.

In the rest of this section, we summarize some intermediate results obtained in the proof of Theorem 4, which are useful in the subsequent analysis (see Appendix B for their proof):

Corollary 5 *The EVB shrinkage estimator* (26) *is a stationary point of the free energy* (14)*, which exists if and only if*

$$\gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}} \equiv (\sqrt{L} + \sqrt{M})\sigma, \tag{28}$$

and satisfies the following equation:

$$\left(\gamma_h \breve{\gamma}_h^{\mathrm{EVB}} + L\sigma^2\right)\left(1 + \frac{M\sigma^2}{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}\right) = \gamma_h^2. \tag{29}$$

It holds that

$$\gamma_h \breve{\gamma}_h^{\mathrm{EVB}} \geq \sqrt{LM}\sigma^2. \tag{30}$$

Corollary 6 *The minimum free energy achieved under EVB is given by Eq.*(13) *with*

$$2F_h = \begin{cases} M\log\left(\frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{M\sigma^2} + 1\right) + L\log\left(\frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{L\sigma^2} + 1\right) - \frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{\sigma^2} & \text{if } \gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

Corollary 5 together with Theorem 4 implies that, when

$$\underline{\gamma}^{\mathrm{local-EVB}} \leq \gamma_h < \underline{\gamma}^{\mathrm{EVB}},$$

a stationary point exists at Eq.(26), but it is not the global minimum. Actually, a local minimum (called a *null* stationary point in Appendix B) with $F_h = 0$ always exists, and the stationary point (26) (called a *positive* stationary point) is a *non-global* local minimum when $\underline{\gamma}^{\mathrm{local-EVB}} < \gamma_h < \underline{\gamma}^{\mathrm{EVB}}$ and the global minimum when $\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}$ (see Figure 8 and its caption for details). This phase transition induces the free energy thresholding observed in Corollary 6.

We define a *local*-EVB solution by

$$\widehat{\boldsymbol{U}}^{\mathrm{local-EVB}} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{local-EVB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}, \quad \text{where} \quad \widehat{\gamma}_h^{\mathrm{local-EVB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{EVB}} & \text{if } \gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}}, \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

and call $\underline{\gamma}^{\mathrm{local-EVB}}$ a local-EVB threshold. We will discuss an interesting relation between the *local*-EVB solution and an alternative dimensionality selection method (Hoyle, 2008) in Section 6.2.

Rescaling the quantities related to the squared singular value by $M\sigma^2$— to which the contribution from noise (each eigenvalue of $\boldsymbol{\mathcal{E}}^{\top}\boldsymbol{\mathcal{E}}$) scales linearly—simplifies expressions. Assume that the condition (28) holds, and define

$$x_h = \frac{\gamma_h^2}{M\sigma^2}, \tag{33}$$

$$\tau_h = \frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{M\sigma^2}, \tag{34}$$

which are used as a rescaled observation and a rescaled EVB estimator, respectively. Eqs.(29) and (26) specify the mutual relations between them:

$$x_h \equiv x(\tau_h; \alpha) = (1 + \tau_h)\left(1 + \frac{\alpha}{\tau_h}\right), \tag{35}$$

$$\tau_h \equiv \tau(x_h;\alpha) = \frac{1}{2}\left(x_h - (1+\alpha) + \sqrt{(x_h-(1+\alpha))^2 - 4\alpha}\right). \tag{36}$$

With these rescaled variables, the condition (28), as well as (30), for the existence of the *positive* local-EVB solution $\breve{\gamma}_h^{\mathrm{EVB}}$ is expressed as

$$x_h \geq \underline{x}^{\mathrm{local}} = \frac{(\underline{\gamma}^{\mathrm{local-EVB}})^2}{M\sigma^2} = x(\sqrt{\alpha};\alpha) = (1+\sqrt{\alpha})^2, \tag{37}$$

$$\tau_h \geq \underline{\tau}^{\mathrm{local}} = \sqrt{\alpha}. \tag{38}$$

The EVB threshold (25) is expressed as

$$\underline{x} = \frac{(\underline{\gamma}^{\mathrm{EVB}})^2}{M\sigma^2} = x(\underline{\tau};\alpha) = (1+\underline{\tau})\left(1+\frac{\alpha}{\underline{\tau}}\right), \tag{39}$$

and the free energy (31) is expressed as

$$F_h = M\tau_h \cdot \min\left(0, \Xi\left(\tau_h;\alpha\right)\right),$$

where $\Xi(\tau;\alpha)$ is defined by Eq.(23).

The rescaled expressions above give an intuition of Theorem 4: The EVB solution $\widehat{\gamma}_h^{\mathrm{EVB}}$ is positive, if and only if the positive local-EVB solution $\breve{\gamma}_h^{\mathrm{EVB}}$ exists (i.e., $x_h \geq \underline{x}^{\mathrm{local}}$), and the free energy $\Xi\left(\tau(x_h;\alpha);\alpha\right)$ at the local-EVB solution is non-positive (i.e., $\tau(x_h;\alpha) \geq \underline{\tau}$ or equivalently $x_h \geq \underline{x}$).

4. Objective Function for Noise Variance Estimation

In this section, we analyze EVB with noise variance estimation:

$$\min_{\{\widehat{a}_h,\widehat{b}_h,\sigma^2_{a_h},\sigma^2_{b_h},c^2_{a_h},c^2_{b_h}\}_{h=1}^H,\sigma^2} \quad 2F,$$
$$\text{s.t.} \quad \{\widehat{a}_h,\widehat{b}_h \in \mathbb{R}, \quad \sigma^2_{a_h},\sigma^2_{b_h},c^2_{a_h},c^2_{b_h} \in \mathbb{R}_{++}\}_{h=1}^H, \sigma^2 \in \mathbb{R}_{++}.$$

Again, by substituting the EVB solution, given by Theorem 4, with the help of Corollary 6, we can express the free energy (13) as a function of the noise variance σ^2. With the rescaled expressions (33)–(39), the free energy is written in a simple form (the proof is given in Appendix C):

Theorem 7 *The noise variance estimator $\widehat{\sigma}^{2\,\mathrm{EVB}}$ is the global minimizer of*

$$\Omega(\sigma^{-2})\left(\equiv \frac{2F(\sigma^{-2})}{LM} + \text{const.}\right) = \frac{1}{L}\left(\sum_{h=1}^{H}\psi\left(\frac{\gamma_h^2}{M\sigma^2}\right) + \sum_{h=H+1}^{L}\psi_0\left(\frac{\gamma_h^2}{M\sigma^2}\right)\right), \tag{40}$$

$$\textit{where} \qquad \psi\left(x\right) = \psi_0\left(x\right) + \theta\left(x > \underline{x}\right)\psi_1\left(x\right), \tag{41}$$

$$\psi_0\left(x\right) = x - \log x, \tag{42}$$

$$\psi_1\left(x\right) = \log\left(\tau(x;\alpha)+1\right) + \alpha\log\left(\frac{\tau(x;\alpha)}{\alpha}+1\right) - \tau(x;\alpha), \tag{43}$$

and $\theta(\cdot)$ denotes an indicator function such that $\theta(\mathit{condition}) = 1$ if the condition is true and $\theta(\mathit{condition}) = 0$ otherwise.

The functions $\psi_0(x)$ and $\psi(x)$ are depicted in Figure 3. We can confirm the convexity of $\psi_0(x)$ and the quasi-convexity of $\psi(x)$,[1] which are useful properties in our analysis.

Let $\widehat{H}^{\mathrm{EVB}}$ be the estimated rank by EVB, i.e., the rank of the EVB estimator $\widehat{U}^{\mathrm{EVB}}$, such that $\widehat{\gamma}_h^{\mathrm{EVB}} > 0$ for $h = 1, \ldots, \widehat{H}^{\mathrm{EVB}}$, and $\widehat{\gamma}_h^{\mathrm{EVB}} = 0$ for $h = \widehat{H}^{\mathrm{EVB}} + 1, \ldots, H$. By bounding the minimizer of the objective (40), we obtain the following theorem (the proof is given in Appendix D):

Theorem 8 *$\widehat{H}^{\mathrm{EVB}}$ is upper-bounded as*

$$\widehat{H}^{\mathrm{EVB}} \leq \overline{H} = \min\left(\left\lceil \frac{L}{1+\alpha} \right\rceil - 1, H\right),$$

and the noise variance estimator $\widehat{\sigma}^{2\,\mathrm{EVB}}$ is bounded as follows:

$$\max\left(\underline{\sigma}^2_{\overline{H}+1}, \frac{\sum_{h=\overline{H}+1}^{L} \gamma_h^2}{M\left(L-\overline{H}\right)}\right) \leq \widehat{\sigma}^{2\,\mathrm{EVB}} \leq \frac{1}{LM}\sum_{h=1}^{L}\gamma_h^2, \tag{44}$$

$$\textit{where} \quad \underline{\sigma}^2_h = \begin{cases} \infty & \textit{for } h = 0, \\ \frac{\gamma_h^2}{M\underline{x}} & \textit{for } h = 1, \ldots, L, \\ 0 & \textit{for } h = L+1. \end{cases} \tag{45}$$

Theorem 8 states that EVB discards the $(L - \lceil L/(1+\alpha)\rceil + 1)$ smallest components, regardless of the observed singular values $\{\gamma_h\}_{h=1}^L$. For example, half of the components are always discarded when the matrix is square (i.e., $\alpha = L/M = 1$). The smallest singular value γ_L is always discarded, and $\widehat{\sigma}^{2\,\mathrm{EVB}} \geq \gamma_L^2/M$ always holds.

Given the EVB estimators $\{\widehat{\gamma}_h^{\mathrm{EVB}}\}_{h=1}^H$ for the singular values, the noise variance estimator $\widehat{\sigma}^{2\,\mathrm{EVB}}$ is specified by the following corollary (the proof is also given in Appendix D):

Corollary 9 *The EVB estimator for the noise variance satisfies the following equality:*

$$\widehat{\sigma}^{2\,\mathrm{EVB}} = \frac{1}{LM}\left(\sum_{l=1}^{L}\gamma_l^2 - \sum_{h=1}^{H}\gamma_h\widehat{\gamma}_h^{\mathrm{EVB}}\right). \tag{46}$$

Theorem 8 and Corollary 9 are used for simple implementation of EVB-PCA in Section 6.1.

5. Performance Analysis

In this section, based on the results obtained in Section 3 and Section 4, we analyze the behavior of EVB with noise variance estimation. We also rely on random matrix theory (Marčenko and Pastur, 1967; Wachter, 1978; Johnstone, 2001; Hoyle and Rattray, 2004; Baik and Silverstein, 2006), which describes the distribution of the singular values of random matrices in the limit when the matrix size goes to infinity. We first introduce some results obtained in random matrix theory, and then apply them to our analysis.

[1] A function $\psi : \mathcal{D} \to \mathbb{R}$ is called quasi-convex if $\psi(\lambda x + (1-\lambda)y) \leq \max(\psi(x), \psi(y)), \forall x, y \in \mathcal{D}, \forall \lambda \in [0,1]$. In other words, $\psi(x)$ is quasi-convex if $-\psi(x)$ is unimodal.

5.1 Random Matrix Theory

Random matrix theory originates from nuclear physics (Wigner, 1957; Mehta, 2000), where the eigenvalue distribution of (infinitely large) symmetric random matrices was investigated to analyze the spectra of heavy atoms. In statistical applications, Wishart matrices play an important role, of which the eigenvalue distribution (or equivalently, the singular value distribution of random data matrices) was derived (Marčenko and Pastur, 1967; Wachter, 1978). Under appropriate scaling, those distributions typically have a finite support, which enables us to clean noisy data and bound quantities related to randomness. Results from random matrix theory have been used in many research fields, including financial risk analysis, where the observed covariance matrix is cleaned for stable prediction (Bouchaud and Potters, 2003), information theory, where the capacity of noisy communication channel was evaluated (Tulino and Verdu, 2004), and signal processing, where the restricted isometry property of random projection was proved for guaranteeing the performance of compressed sensing (Candès and Tao, 2006; Recht et al., 2010). Development of random matrix theory is still actively on going, and new important results are being reported (Bai and Silverstein, 2010).

To analyze the performance of EVB-PCA, we assume that the observed matrix $\boldsymbol{V}$ is generated from the *spiked covariance* model (Johnstone, 2001):

$$\boldsymbol{V} = \boldsymbol{U}^* + \boldsymbol{\mathcal{E}},$$

where $\boldsymbol{U}^* \in \mathbb{R}^{L \times M}$ is a *true* signal matrix with rank H^* and singular values $\{\gamma_h^*\}_{h=1}^{H^*}$, and $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{L \times M}$ is a random matrix such that each element is independently drawn from a distribution with mean zero and variance σ^{*2} (not necessarily Gaussian). As the observed singular values $\{\gamma_h\}_{h=1}^{L}$ of $\boldsymbol{V}$, the true singular values $\{\gamma_h^*\}_{h=1}^{H^*}$ are also assumed to be arranged in the non-increasing order.

We define rescaled versions of the observed and the true singular values:

$$\begin{aligned} y_h &= \frac{\gamma_h^2}{M\sigma^{*2}} && \text{for} && h = 1, \ldots, L, \\ \nu_h^* &= \frac{\gamma_h^{*2}}{M\sigma^{*2}} && \text{for} && h = 1, \ldots, H^*. \end{aligned}$$

In other words, $\{y_h\}_{h=1}^{L}$ are the eigenvalues of $\boldsymbol{V}\boldsymbol{V}^\top/(M\sigma^{*2})$, and $\{\nu_h^*\}_{h=1}^{H^*}$ are the eigenvalues of $\boldsymbol{U}^*\boldsymbol{U}^{*\top}/(M\sigma^{*2})$. Note the difference between x_h, defined by Eq.(33), and y_h: x_h is the squared observed singular value rescaled with the model noise variance σ^2 to be estimated, while y_h is the one rescaled with the true noise variance σ^{*2}.

Define the empirical distribution of the observed eigenvalues $\{y_h\}_{h=1}^{L}$ by

$$p(y) = \frac{1}{L}\sum_{h=1}^{L} \delta(y - y_h),$$

where $\delta(y)$ denotes the Dirac delta function. When $H^* = 0$, the observed matrix $\boldsymbol{V} = \boldsymbol{\mathcal{E}}$ consists only of noise, and its singular value distribution in the large-scale limit is specified by the following proposition:

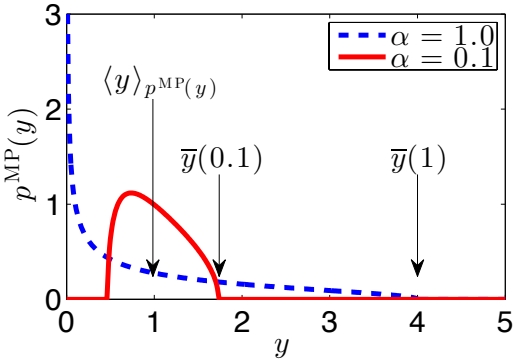


Figure 4: Marčenko-Pastur distribution.

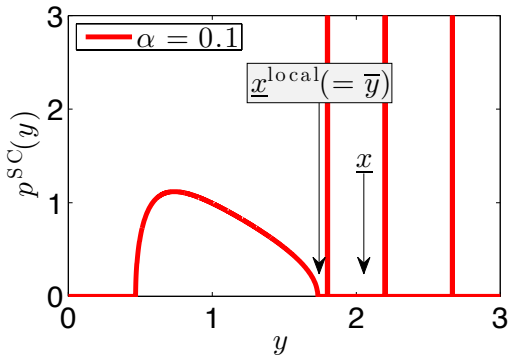


Figure 5: Spiked covariance distribution when $\{\nu_h^*\}_{h=1}^{H^{**}} = \{1.5, 1.0, 0.5\}$.

Proposition 10 *(Marčenko and Pastur, 1967; Wachter, 1978) In the large-scale limit when L and M go to infinity with its ratio $\alpha = L/M$ fixed, the empirical distribution of the eigenvalue y of $\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top/(M\sigma^{*2})$ converges almost surely to*

$$p(y) \to p^{\mathrm{MP}}(y) \equiv \frac{\sqrt{(y-\underline{y})(\overline{y}-y)}}{2\pi\alpha y}\theta(\underline{y} < y < \overline{y}), \tag{47}$$

$$\textit{where} \qquad \overline{y} = (1+\sqrt{\alpha})^2, \qquad \underline{y} = (1-\sqrt{\alpha})^2, \tag{48}$$

and $\theta(\cdot)$ is the indicator function, defined in Theorem 7.

Figure 4 shows Eq.(47), which we call the Marčenko-Pastur (MP) distribution, for $\alpha = 0.1, 1$. The mean $\langle y \rangle_{p^{\mathrm{MP}}(y)} = 1$ (which is constant for any $0 < \alpha \leq 1$) and the upper-limits $\overline{y} = \overline{y}(\alpha)$ of the support for $\alpha = 0.1, 1$ are indicated by arrows. Proposition 10 states that the probability mass is concentrated in the range between $\underline{y} \leq y \leq \overline{y}$. Note that the MP distribution appears for a *single* sample matrix; different from standard "large-sample" theories, Proposition 10 does not require to average over many sample matrices (this property is called *self-averaging*). This single-sample property of the MP distribution is highly useful in our analysis because we are working with a single observation matrix in the PCA scenario.

When $H^* > 0$, the true signal matrix $\boldsymbol{U}^*$ affects the singular value distribution of $\boldsymbol{V}$. However, if $H^* \ll L$, the distribution can be approximated by a mixture of spikes (delta functions) and the MP distribution $p^{\mathrm{MP}}(y)$. Let H^{**} $(\leq H^*)$ be the number of singular values of $\boldsymbol{U}^*$ greater than $\gamma_h^* > \alpha^{1/4}\sqrt{M}\sigma^*$, i.e.,

$$\nu_{H^{**}}^* > \sqrt{\alpha} \qquad \text{and} \qquad \nu_{H^{**}+1}^* \leq \sqrt{\alpha}.$$

Then, the following proposition holds:

Proposition 11 *(Baik and Silverstein, 2006) In the large-scale limit when L and M go to infinity with finite α and H^*, it almost surely holds that*

$$y_h = y_h^{\mathrm{Sig}} \equiv (1+\nu_h^*)\left(1+\frac{\alpha}{\nu_h^*}\right) \quad \text{for} \quad h = 1, \ldots, H^{**}, \tag{49}$$

$$y_{H^{**}+1} = \overline{y}, \quad \text{and} \quad y_L = \underline{y}.$$

Furthermore, Hoyle and Rattray (2004) argued that, when L and M are large (but finite) and $H^* \ll L$, the empirical distribution of the eigenvalue y of $\boldsymbol{V}\boldsymbol{V}^\top/(M\sigma^{*2})$ is accurately approximated by

$$p(y) \approx p^{\mathrm{SC}}(y) \equiv \frac{1}{L}\sum_{h=1}^{H^{**}} \delta\left(y - y_h^{\mathrm{Sig}}\right) + \frac{L-H^{**}}{L} p^{\mathrm{MP}}(y). \tag{50}$$

Figure 5 shows Eq.(50), which we call the spiked covariance (SC) distribution, for $\alpha = 0.1$, $H^{**} = 3$, and $\{\nu_h^*\}_{h=1}^{H^{**}} = \{1.5, 1.0, 0.5\}$. The SC distribution is irrespective of $\{\nu_h^*\}_{h=H^{**}+1}^{H^*}$, which satisfy $0 < \nu_h^* \leq \sqrt{\alpha}$ by definition.

Proposition 11 states that, in the large-scale limit, the large signal components such that $\nu_h^* > \sqrt{\alpha}$ appear outside the support of the MP distribution as spikes, while the other small signals are indistinguishable from the MP distribution (note that $y_h^{\mathrm{Sig}} > \overline{y}$ for $\nu_h^* > \sqrt{\alpha}$). This implies that any PCA method fails to recover the true dimensionality, unless

$$\nu_{H^*}^* > \sqrt{\alpha}. \tag{51}$$

The condition (51) requires that U^* has no small positive singular value such that $0 < \nu_h^* \leq \sqrt{\alpha}$, and therefore $H^{**} = H^*$.

The approximation (50) allows us to investigate more practical situations when the matrix size is finite. Based on this approximation, Hoyle (2008) analyzed the performance of the *overlap* method, an alternative dimensionality selection method which will be introduced and discussed in Section 6.2. In Section 5.2, we provide two theorems: One is based on Proposition 11, and guarantees the perfect dimensionality recovery of EVB in the large-scale limit, and the other one relies on the approximation (50), and provides a more realistic condition for perfect recovery.

5.2 Perfect Dimensionality Recovery Condition

Now, we are almost ready for clarifying the behavior of EVB-PCA. We assume that the model rank is set to be large enough, i.e., $H^* \leq H \leq L$, and all model parameters including the noise variance are estimated from observation. The last proposition on which our analysis relies is related to the property, called the *strong unimodality*,[2] of the log-concave distributions:

Proposition 12 *(Ibragimov, 1956; Dharmadhikari and Joag-dev, 1988) The convolution*

$$g(s) = \langle f(s+t) \rangle_{p(t)} = \int f(s+t)p(t)dt$$

is quasi-convex, if $p(t)$ is a log-concave distribution, and $f(t)$ is a quasi-convex function.

[2] A distribution $p(t)$ is called strongly unimodal if the convolution of $p(t)$ with any unimodal function is unimodal.

In the large-scale limit, the summation over $h = 1, \ldots, L$ in the objective $\Omega(\sigma^{-2})$, given by Eq.(40), for noise variance estimation can be replaced with an expectation over the MP distribution $p^{\mathrm{MP}}(y)$. By scaling variables, the objective can be written as a convolution with a scaled version of the MP distribution, which turns out to be log-concave. Accordingly, we can use Proposition 12 to show that $\Omega(\sigma^{-2})$ is quasi-convex, and therefore, the noise variance estimation by EVB is accurate. Combining this result with Proposition 11, we obtain the following theorem (the proof is given in Appendix E):

Theorem 13 *In the large-scale limit when L and M go to infinity with finite α and H^*, EVB almost surely recovers the true rank, i.e., $\widehat{H}^{\mathrm{EVB}} = H^*$, if and only if*

$$\nu^*_{H^*} \geq \underline{\tau}, \tag{52}$$

where $\underline{\tau}$ is defined in Theorem 4.

Furthermore, the following corollary completely describes the behavior of EVB in the large-scale limit (the proof is also given in Appendix E):

Corollary 14 *In the large-scale limit, the objective $\Omega(\sigma^{-2})$, defined by Eq.(40), for the noise variance estimation converges to a quasi-convex function, and it almost surely holds that*

$$\widehat{\tau}^{\mathrm{EVB}}_h \left(\equiv \frac{\gamma_h \widehat{\gamma}^{\mathrm{EVB}}_h}{M \widehat{\sigma}^{2\,\mathrm{EVB}}} \right) = \begin{cases} \nu^*_h & \text{if } \nu^*_h \geq \underline{\tau}, \\ 0 & \text{otherwise}, \end{cases} \tag{53}$$
$$\widehat{\sigma}^{2\,\mathrm{EVB}} = \sigma^{*2}.$$

One may get intuition of Eqs.(52) and (53) from comparing Eqs.(39) and (35) with Eq.(49): The estimator τ_h has the same relation to the observation x_h as the true signal ν^*_h, and hence is an unbiased estimator of the signal. However, Theorem 13 does not even approximately hold in practical situations with moderate-sized matrices (see the numerical simulation below). The following theorem, which relies on the approximation (50), provides a more practical condition for perfect recovery (the proof is given in Appendix F):

Theorem 15 *Let*

$$\xi = \frac{H^*}{L}$$

be the relevant rank (dimensionality) ratio, and assume that

$$p(y) = p^{\mathrm{SC}}(y). \tag{54}$$

Then, EVB recovers the true rank, i.e., $\widehat{H}^{\mathrm{EVB}} = H^*$, if the following two inequalities hold:

$$\xi < \frac{1}{\underline{x}}, \tag{55}$$

$$\nu^*_{H^*} > \frac{\left(\frac{\underline{x}-1}{1-\underline{x}\xi} - \alpha \right) + \sqrt{\left(\frac{\underline{x}-1}{1-\underline{x}\xi} - \alpha \right)^2 - 4\alpha}}{2}, \tag{56}$$

where $\underline{x}$ is defined by Eq.(39).

Note that, in the large-scale limit, ξ converges to zero, and the sufficient condition, (55) and (56), in Theorem 15 is equivalent to the necessary and sufficient condition (52) in Theorem 13.

Theorem 15 only requires that the SC distribution (50) well approximates the observed singular value distribution. Accordingly, it well describes the dependency of the behavior of EVB on ξ, as shown in the numerical simulation below. Theorem 15 states that, if the true rank H^* is small enough compared with L and the smallest signal $\nu^*_{H^*}$ is large enough, EVB perfectly recovers the true dimensionality.

The following corollary also supports EVB (the proof is also given in Appendix F):

Corollary 16 *Under the assumption* (54) *and the conditions* (55) *and* (56)*, the objective* $\Omega(\sigma^{-2})$ *for the noise variance estimation has no local minimum (no stationary point if* $\xi > 0$*) that results in a wrong estimated rank* $\widehat{H}^{\mathrm{EVB}} \neq H^*$*.*

This corollary states that, although the objective function (40) is non-convex and possibly multimodal in general, any local minimum leads to the correct estimated rank. Therefore, perfect recovery does not require global search, but only local search, for noise variance estimation, if L and M are sufficiently large so that we can assume Eq.(54).

Figure 6 shows numerical simulation results for $M = 200$ and $L = 20, 100, 200$. $\boldsymbol{\mathcal{E}}$ was drawn from the independent Gaussian distribution with variance $\sigma^{*2} = 1$, and *true* signal singular values $\{\gamma^*_h\}_{h=1}^{H^*}$ were drawn from the uniform distribution on $[z\sqrt{M}\sigma^*, 10\sqrt{M}\sigma^*]$ for different z, which is indicated by the horizontal axis. The vertical axis indicates the success rate of dimensionality recovery, i.e., $\widehat{H}^{\mathrm{EVB}} = H^*$, over 100 trials. If the condition (55) on ξ is violated, the corresponding curve is depicted with markers. Otherwise, the condition (56) on $\nu^*_{H^*}(= \gamma^{*2}_{H^*}/(M\sigma^{*2}))$ is indicated by a vertical bar with the same color and line style for each ξ. In other words, Theorem 15 states that the success rate should be equal to one if z $(> \gamma^*_{H^*}/(\sqrt{M}\sigma^*))$ is larger than the value indicated by the vertical bar. The solid cyan bar, which lies at the left-most in each graph, indicates the condition (52) given by Theorem 13.

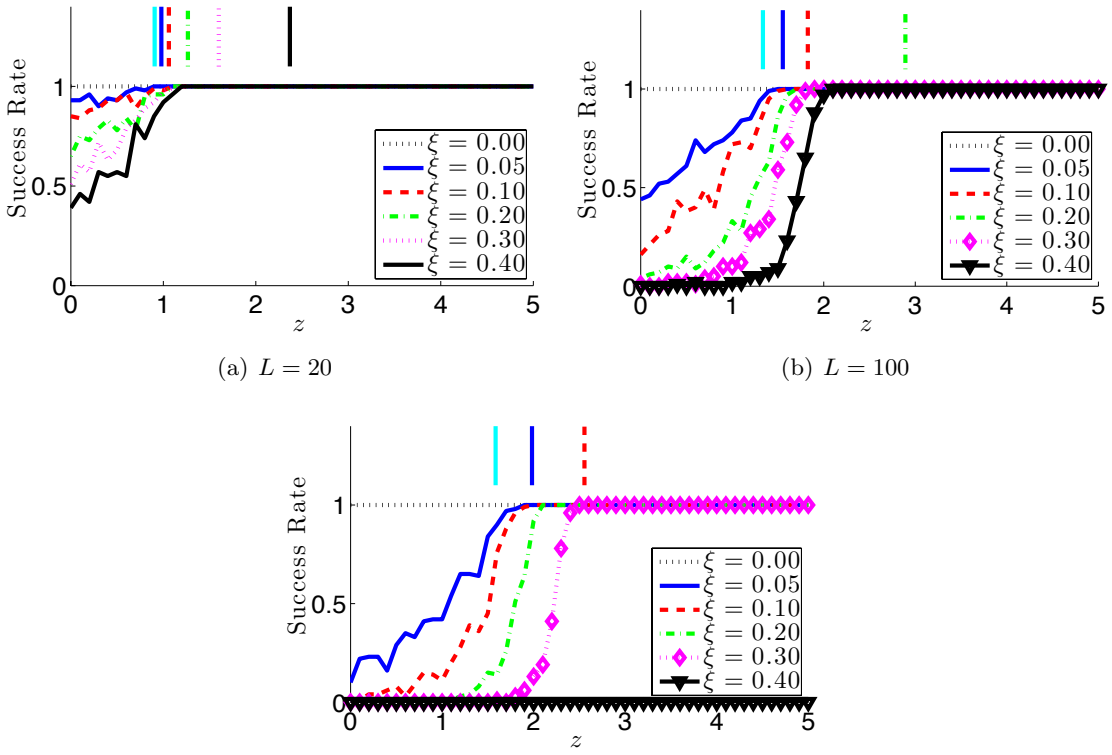
We see that Theorem 15 with the condition (56) approximately holds for these moderate-sized matrices, while Theorem 13 with the condition (52), which does not depend on the relevant rank ratio ξ, immediately breaks for positive ξ.

6. Discussion

In this section, we first propose a few implementations of EVB-PCA. After that, by contrasting with an alternative dimensionality selection method, we characterize the behavior of EVB-PCA, and discuss the optimality in the large-scale limit.

6.1 Implementation

The analytic-form solution derived in Nakajima et al. (2013b) involves a solution of a *quartic* equation. To implement EVB-PCA based on that form, we needed to use a highly complicated analytic-form solution, derived by, e.g., Ferrari's method, or rely on a numerical quartic solver. Our new analytic-form solution can greatly simplify the implementation. Note that, since our theory of performance guarantee assumes that the observed matrix has no missing entry, its applicability is mostly limited to the standard use of PCA—dimensionality



(a) $L = 20$

(b) $L = 100$

(c) $L = 200$

Figure 6: Success rate of dimensionality recovery in numerical simulation for $M = 200$. The horizontal axis indicates the lower limit of the support of the simulated true signal distribution, i.e., $z \approx \sqrt{\nu^*_{H^*}}$. The recovery condition (56) for finite-sized matrices is indicated by a vertical bar with the same color and line style for each ξ. The recovery condition (52), which does not depend on ξ, for infinite-sized matrices is also indicated by a solid cyan bar.

reduction for preprocessing (Bishop, 2006). However, our simple implementation introduced below can be applied to more general cases where the global VB solver is used as a subroutine, e.g., in non-conjugate matrix factorization with missing entries (Seeger and Bouchard, 2012), and in *sparse additive matrix factorization* (Nakajima et al., 2013a), an extension of robust PCA.

A table of $\underline{\tau}$ defined in Theorem 4 should be prepared beforehand (or use a simple approximation $\underline{\tau} \approx \underline{z}\sqrt{\alpha} \approx 2.5129\sqrt{\alpha}$). Given an observed matrix $\boldsymbol{V}$, we perform SVD and obtain the singular values $\{\gamma_h\}_{h=1}^{L}$. After that, in our new implementation, we first directly estimate the noise variance based on Theorem 7, using any 1-D local search algorithm with the search range restricted by Theorem 8. Thus, we obtain the noise variance estimator $\widehat{\sigma}^{2\,\mathrm{EVB}}$. Discarding all the components such that $\underline{\sigma}^2_h < \widehat{\sigma}^{2\,\mathrm{EVB}}$, where $\underline{\sigma}^2_h$ is defined by

Algorithm 1 Global EVB-PCA algorithm.

1: Transpose $\boldsymbol{V} \to \boldsymbol{V}^\top$ if $L > M$.
2: Refer to the table of $\underline{\tau}(\alpha)$ at $\alpha = L/M$ (or use a simple approximation $\underline{\tau} \approx 2.5129\sqrt{\alpha}$).
3: Set H ($\leq L$) to a sufficiently large value, and compute the SVD of $\boldsymbol{V} = \sum_{h=1}^{H} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top$.
4: Locally search the minimizer $\widehat{\sigma}^{2\,\mathrm{EVB}}$ of Eq.(40), which lies in the range (44).
5: Discard the components such that $\underline{\sigma}_h^2 < \widehat{\sigma}^{2\,\mathrm{EVB}}$, where $\underline{\sigma}_h^2$ is defined by Eq.(45).

Eq.(45), gives a dimensionality reduction result. Algorithm 1 describes a pseudo code.[3] If necessary, Theorem 4 gives the EVB estimator $\widehat{\boldsymbol{U}}^{\mathrm{EVB}}$ for $\sigma^2 = \widehat{\sigma}^{2\,\mathrm{EVB}}$. The EVB posterior is also easily computed by using Corollary 3. In this way, we can easily perform EVB-PCA equipped with the guaranteed automatic dimensionality selection functionality at little expense—computation time of Algorithm 1 is dominated by SVD, which the plain PCA also requires to perform.

Another implementation, which we refer to as EVB(Ite), is to iterate Eqs.(24) and (46) in turn. Although it is not guaranteed, EVB(Ite) tends to converge to the global solution if we initialize the noise variance $\widehat{\sigma}^{2\,\mathrm{EVB}}$ sufficiently small (see Section 6.2).

Finally, we introduce an iterative algorithm for the local-EVB solution, defined by Eq.(32). This solution can be obtained by iterating Eq.(32) and

$$\widehat{\sigma}^{2\,\mathrm{local-EVB}} = \frac{1}{LM}\left(\sum_{l=1}^{L}\gamma_l^2 - \sum_{h=1}^{H}\gamma_h \widehat{\gamma}_h^{\mathrm{local-EVB}}\right) \tag{57}$$

in turn. If we initialize the noise variance $\widehat{\sigma}^{2\,\mathrm{local-EVB}}$ sufficiently small, this algorithm can be trapped at the *positive* stationary point for each h even if it is not the global minimum, and tends to converge to the local-EVB solution.

6.2 Comparison with Laplace Approximation

Here, we compare EVB with the *overlap* method (Hoyle, 2008), an alternative dimensionality selection method based on the Laplace approximation (LA). Consider the PCA application, where D denotes the dimensionality of the observation space, and N denotes the number of samples, i.e., in our MF notation to keep $L \leq M$,

$$\begin{aligned} L = D, M = N \qquad & \text{if} \qquad D \leq N, \\ L = N, M = D \qquad & \text{if} \qquad D > N. \end{aligned}$$

Just after Tipping and Bishop (1999) proposed the probabilistic PCA, Bishop (1999b) proposed to select the PCA dimension by maximizing the marginal likelihood:[4]

$$p(\boldsymbol{V}) = \langle p(\boldsymbol{V}|\boldsymbol{A},\boldsymbol{B})\rangle_{p(\boldsymbol{A})p(\boldsymbol{B})}\,. \tag{58}$$

[3] The MATLAB® code will be available at `http://sites.google.com/site/shinnkj23/`.

[4] Tipping and Bishop (1999) adopted *partially Bayesian* (PB) learning, where $\boldsymbol{A}$ is marginalized out and $\boldsymbol{B}$ is point-estimated. Although PB has some similarities to VB (Nakajima et al., 2011; Nakajima and Sugiyama, 2014), it does not offer automatic dimensionality selection when all hyperparameters $(\boldsymbol{C}_A, \boldsymbol{C}_B, \sigma^2)$ are unknown.

Since the marginal likelihood (58) is computationally intractable, he approximated it by LA, and suggested Gibbs sampling and VB learning as alternatives. The VB variant, of which the model is almost the same as ours (1)–(3), was proposed by himself (Bishop, 1999a). A standard local search algorithm, where the means and the covariances of $\boldsymbol{A}$ and $\boldsymbol{B}$ are iteratively updated, was used for inference.

The LA-based approach was polished in Minka (2001), by introducing a conjugate prior on $\boldsymbol{B}$ to $p(\boldsymbol{V}|\boldsymbol{B}) = \langle p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B})\rangle_{p(\boldsymbol{A})}$, and ignoring the non-leading terms that do not grow fast as the number N of samples goes to infinity. Hoyle (2008) pointed out that Minka's method is inaccurate when $D \gg N$, and proposed the overlap (OL) method, a further polished variant of the LA-based approach. A notable difference of OL from most of the LA-based methods is that OL applies LA to a more accurate estimator than the MAP estimator, while the other methods apply LA simply to the MAP estimator. Thanks to the use of an accurate estimator, OL behaves *optimally* in the large-scale limit when D and N go to infinity, while Minka's method does not. We will clarify the meaning of optimality, and discuss it in more detail in Section 6.3.

OL minimizes an approximation to the negative log of the marginal likelihood (58), which depends on estimators of $\lambda_h = b_h^2 + \sigma^2$ and σ^2 computed by an iterative algorithm, over the hypothetical model rank $H = 1, \ldots, L$ (see Appendix H for details). Figure 7 shows numerical simulation results that compare EVB and OL: Figure 7(a) shows the success rate for the no signal case $\xi = 0$ ($H^* = 0$), while Figures 7(b)–7(f) show the success rate for $\xi = 0.05$ and $D = 20, 100, 200, 400, 1000$, respectively.

We also show the performance of EVB(Ite) and local-EVB. As mentioned in Section 6.1, EVB(Ite) gives almost the same results as EVB. Local-EVB behaves similarly to OL except the case when D/N is small (Figure 7(b)). The reason of this similarity will be elucidated in Section 6.3. For OL, EVB(Ite), and local-EVB, we initialized the noise variance estimator to $10^{-4} \cdot \sum_{h=1}^{L} \gamma_h^2/(LM)$.

Comparing EVB with OL, we observe the conservative nature of EVB: It exhibits almost zero false positive rate at the expense of low sensitivity. Because of the low sensitivity, EVB actually does not behave optimally in the large-scale limit, which is discussed in Section 6.3.

6.3 Optimality in Large-scale Limit

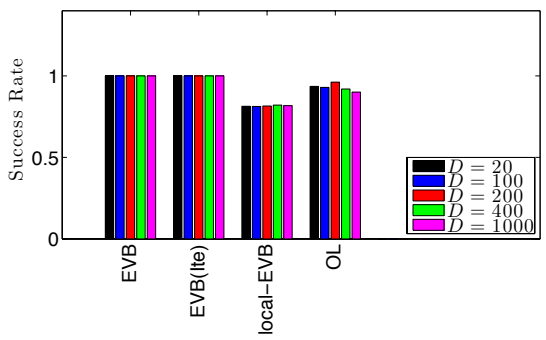
Consider the large-scale limit, i.e., $L, M \to \infty, \alpha = L/M$, and assume that the model rank H is set to be large enough but finite so that $H \geq H^*$ and $H/L \to 0$. Then, OL is equivalent to counting the number of components such that $\widehat{\lambda}_h^{\mathrm{OL-limit}} > \widehat{\sigma}^{2\,\mathrm{OL-limit}}$, i.e.,

$$\widehat{H}^{\mathrm{OL-limit}} = \sum_{h=1}^{L} \theta\left(\widehat{\lambda}_h^{\mathrm{OL-limit}} > \widehat{\sigma}^{2\,\mathrm{OL-limit}}\right), \tag{59}$$

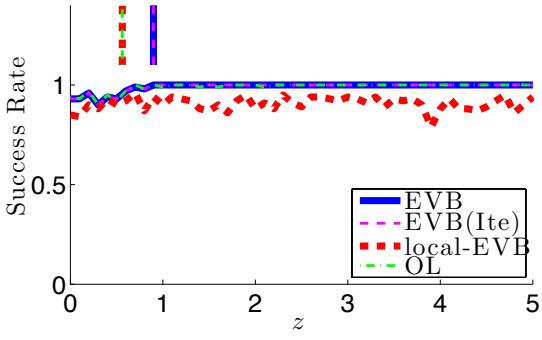
after the following updates converge:

$$\widehat{\lambda}_h^{\mathrm{OL-limit}} = \begin{cases} \breve{\lambda}_h^{\mathrm{OL-limit}} & \text{if } \gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}}, \\ \widehat{\sigma}^{2\,\mathrm{OL-limit}} & \text{otherwise,} \end{cases} \quad \text{for} \quad h = 1, \ldots, H, \tag{60}$$

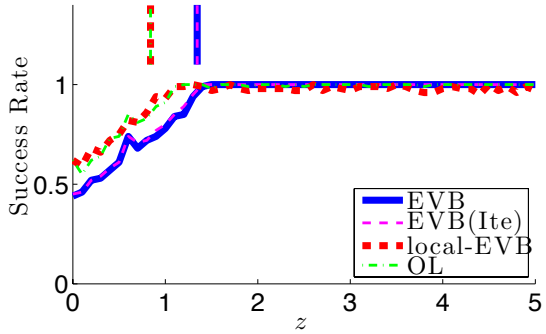
$$\widehat{\sigma}^{2\,\mathrm{OL-limit}} = \frac{1}{(M-H)}\left(\sum_{l=1}^{L}\frac{\gamma_l^2}{L} - \sum_{h=1}^{H}\widehat{\lambda}_h^{\mathrm{OL-limit}}\right), \tag{61}$$



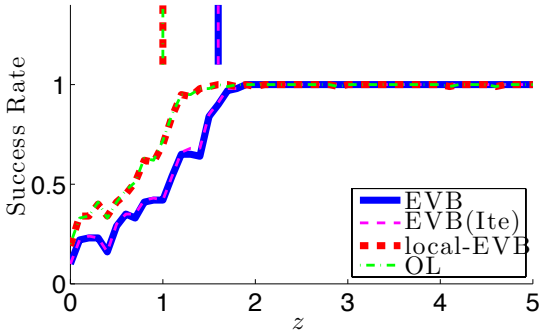
(a) $\xi = 0$



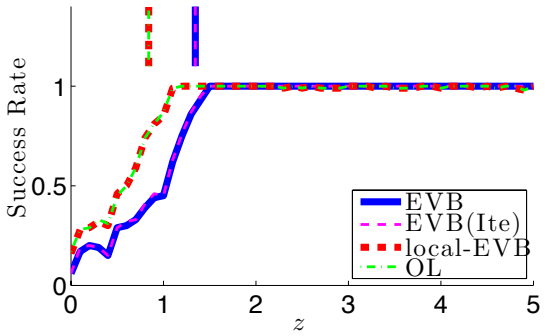
(b) $\xi = 0.05, D = 20$



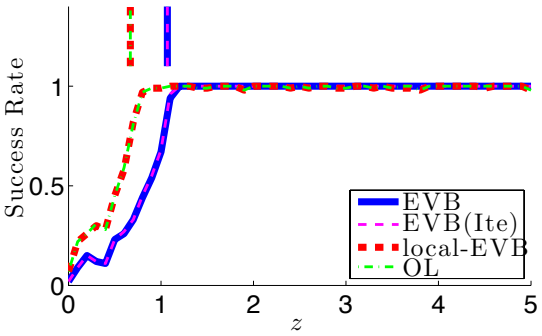
(c) $\xi = 0.05, D = 100$



(d) $\xi = 0.05, D = 200$



(e) $\xi = 0.05, D = 400$



(f) $\xi = 0.05, D = 1000$

Figure 7: Success rate of dimensionality recovery by EVB, EVB(Ite), local-EVB, and OL for $N = 200$. Vertical bars indicate the recovery conditions, Eq.(52) for EVB and EVB(Ite), and Eq.(63) for OL and local-EVB, in the large-scale limit.

where $\breve{\lambda}_h^{\mathrm{OL-limit}} = \dfrac{\gamma_h^2}{2L}\left(1 - \dfrac{(M-L)\widehat{\sigma}^{2\,\mathrm{OL-limit}}}{\gamma_h^2}\right.$

$$+\sqrt{\left(1-\frac{(M-L)\widehat{\sigma}^{2\ \mathrm{OL-limit}}}{\gamma_h^2}\right)^2-\frac{4L\widehat{\sigma}^{2\ \mathrm{OL-limit}}}{\gamma_h^2}}\Bigg). \quad (62)$$

OL evaluates its objective, which approximates the negative log of the marginal likelihood (58), after the updates (60) and (61) converge for each hypothetical H, and adopts the minimizer $\widehat{H}^{\mathrm{OL-limit}}$ as the rank estimator. However, Hoyle (2008) proved that, in the large-scale limit, the objective decreases as H increases, as long as Eq.(62) is a real number (or equivalently $\gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}}$ holds) for all $h = 1, \ldots, H$ at the convergence. Accordingly, Eq.(59) suffices.

Interestingly, the threshold in Eq.(60) coincides with the local-EVB threshold (28). Moreover, the updates (60) and (61) for OL are equivalent to the updates (32) and (57) for local-EVB with the following correspondence:

$$\widehat{\lambda}_h^{\mathrm{OL-limit}} = \frac{\gamma_h \widehat{\gamma}_h^{\mathrm{local-EVB}}}{L} + \widehat{\sigma}^{2\ \mathrm{local-EVB}},$$
$$\widehat{\sigma}^{2\ \mathrm{OL-limit}} = \widehat{\sigma}^{2\ \mathrm{local-EVB}}.$$

Thus, the dimensionality selection by OL and local-EVB are equivalent in the large-scale limit, i.e., $\widehat{H}^{\mathrm{OL-limit}} = \widehat{H}^{\mathrm{local-EVB}}$.

The optimality of OL in the large-scale limit was shown:

Proposition 17 *(Hoyle, 2008) In the large-scale limit when L and M go to infinity with finite α, H^*, and H ($\geq H^*$)[5], OL almost surely recovers the true rank, i.e., $\widehat{H}^{\mathrm{OL-limit}} = H^*$, if and only if*

$$\nu_{H^*}^* > \sqrt{\alpha}. \quad (63)$$

It almost surely holds that

$$\frac{\widehat{\lambda}_h^{\mathrm{OL-limit}}}{\widehat{\sigma}^{2\ \mathrm{OL-limit}}} - 1 = \nu_h^*,$$
$$\widehat{\sigma}^{2\ \mathrm{OL-limit}} = \sigma^{*2}.$$

Note that the condition (63) coincides with the condition (51)—random matrix theory states that any signal component violating this condition is indistinguishable from the noise distribution, and therefore, any PCA method fails to recover the correct dimensionality if such a signal component exists. In this sense, OL, as well as local-EVB, is *optimal* in the large-scale limit.

On the other hand, Theorem 13 implies that (global) EVB is not optimal in the large-scale limit, and more conservative (see the difference between $\underline{\tau}$ and $\sqrt{\alpha}$ in Figure 2). In Figure 7, the conditions for perfect dimensionality recovery in the large-scale limit are indicated by vertical bars:

$$z = \sqrt{\underline{\tau}} \ \text{ for EVB and EVB(Ite)}, \qquad \text{and} \qquad z = \sqrt{\underline{\tau}^{\mathrm{local}}} = \alpha^{1/4} \ \text{ for OL and local-EVB.}$$

[5] Unlike our analysis in Section 5, Hoyle (2008) assumes that $H/L \to 0$, which trivially guarantees that the noise variance is accurately estimated.

All methods accurately estimate the noise variance in the large-scale limit, i.e.,

$$\widehat{\sigma}^{2\,\mathrm{EVB}} = \widehat{\sigma}^{2\,\mathrm{OL-limit}} = \widehat{\sigma}^{2\,\mathrm{local-EVB}} = \sigma^{*2}.$$

Taking this into account, we indicate the recovery conditions in Figure 5 by arrows at

$$y = \underline{x} \ \text{ for EVB and EVB(Ite)}, \qquad \text{and} \qquad y = \underline{x}^{\mathrm{local}} (= \overline{y}) \ \text{ for OL and local-EVB},$$

respectively. Figure 5 implies that, in this particular case, EVB discards the third spike coming from the third true signal $\nu_3^* = 0.5$, while OL and local-EVB successfully capture it as a signal.

When the matrix size is finite, the conservative nature of EVB is not always bad, since it offers almost zero false positive rate, which makes Theorem 15 approximately hold for finite cases, as seen in Figure 6 and Figure 7. However, the fact that not (global) EVB but local-EVB is optimal in the large-scale limit should be a consequence of inaccurate approximation of VB learning under the independence assumption. We will further investigate the difference between VB and Bayesian learning in our future work.

7. Conclusion

In this paper, we analyzed the variational Bayesian (VB) learning in probabilistic PCA. More specifically, we considered empirical VB (EVB) learning with noise variance estimation, i.e., all model parameters are estimated from observed data. We established a necessary and sufficient condition for perfect dimensionality recovery by EVB-PCA, which theoretically guarantees its performance. At the same time, our result also revealed the conservative nature of EVB-PCA—it offers a low false positive rate at the expense of low sensitivity, due to which EVB-PCA does not behave optimally in the large-scale limit.

By contrasting with an alternative dimensionality selection method, called the overlap (OL) method, we characterized the behavior of EVB. We also pointed out the equivalence between OL and local-EVB, a slight modification of EVB, in the large scale limit.

In our analysis, we derived bounds of the noise variance estimator, and a new and simple analytic-form solution for the other parameters, with which we proposed a new simple implementation of EVB-PCA.

Acknowledgments

The authors thank anonymous reviewers for helpful comments. Shinichi Nakajima thanks the support from Nikon Corporation, the support from Grant-in-Aid for Scientific Research on Innovative Areas: Prediction and Decision Making, 23120004, and the support from the German Research Foundation (GRK 1589/1) by the Federal Ministry of Education and Research (BMBF) under the Berlin Big Data Center project (FKZ 01IS14013A). Masashi Sugiyama was supported by the CREST program.

Appendix A. Proof of Theorem 2 and Corollary 3

The global VB solution is known:

Proposition 18 *(Nakajima et al., 2013b) The VB solution can be written as truncated shrinkage SVD as follows:*

$$\widehat{\boldsymbol{U}}^{\mathrm{VB}} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}, \qquad \textit{where} \qquad \widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{VB}} & \textit{if } \gamma_h \geq \underline{\gamma}_h^{\mathrm{VB}}, \\ 0 & \textit{otherwise.} \end{cases}$$

Here, the truncation threshold is given by

$$\underline{\gamma}_h^{\mathrm{VB}} = \sigma \sqrt{\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM}},$$

and the shrinkage estimator $\breve{\gamma}_h^{\mathrm{VB}}$ *is the second largest real solution of a* quartic *equation.*[6]

With Proposition 18, it is sufficient to obtain the new analytic-form (17) of the shrinkage estimator for proving Theorem 2. However, we give a proof, starting not from Proposition 18 but from Proposition 1. Thanks to the new analytic-form of the shrinkage estimator, our new proof is much more intuitive than the proof given in Nakajima and Sugiyama (2011) and in Nakajima et al. (2013b), for example, in choosing the global solution from two stationary points: the free energy is directly compared in the new proof, while it was shown that one of the stationary points is a saddle point by evaluating the Hessian in Nakajima and Sugiyama (2011).

Proposition 1 states that the VB estimator can be obtained by minimizing the free energy (14) for each singular component separately. Clearly, Eq.(14) is differentiable, and diverges to $F_h \to \infty$ as any variable approaches to any point on the domain boundary. Therefore, any minimizer is stationary point.

The stationary condition of Eq.(14) is given by

$$\widehat{a}_h = \frac{1}{\sigma^2} \gamma_h \widehat{b}_h \sigma_{a_h}^2, \tag{64}$$

$$\sigma_{a_h}^2 = \sigma^2 \left(\widehat{b}_h^2 + L\sigma_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2}\right)^{-1}, \tag{65}$$

$$\widehat{b}_h = \frac{1}{\sigma^2} \gamma_h \widehat{a}_h \sigma_{b_h}^2, \tag{66}$$

$$\sigma_{b_h}^2 = \sigma^2 \left(\widehat{a}_h^2 + M\sigma_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2}\right)^{-1}. \tag{67}$$

By using Eqs.(65) and (67), the free energy (14) can be written as

$$F_h = M \log \frac{c_{a_h}^2}{\sigma_{a_h}^2} + L \log \frac{c_{b_h}^2}{\sigma_{b_h}^2} + \frac{\sigma^2}{\sigma_{a_h}^2 \sigma_{b_h}^2} - \frac{2\widehat{a}_h \widehat{b}_h \gamma_h}{\sigma^2} - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right). \tag{68}$$

The stationary condition, Eqs.(64)–(67), implies two possibilities of stationary points.

[6] The quartic equation is omitted, since it is complicated and no longer important.

A.1 Null Stationary Point

If $\widehat{a}_h = 0$ or $\widehat{b}_h = 0$, Eqs.(64) and (66) require that $\widehat{a}_h = 0$ *and* $\widehat{b}_h = 0$. In this case, Eqs.(65) and (67) lead to

$$\sigma_{a_h}^2 = c_{a_h}^2 \left(1 - \frac{L\sigma_{a_h}^2\sigma_{b_h}^2}{\sigma^2}\right), \tag{69}$$

$$\sigma_{b_h}^2 = c_{b_h}^2 \left(1 - \frac{M\sigma_{a_h}^2\sigma_{b_h}^2}{\sigma^2}\right). \tag{70}$$

Multiplying Eqs.(69) and (70), we have

$$\left(1 - \frac{L\sigma_{a_h}^2\sigma_{b_h}^2}{\sigma^2}\right)\left(1 - \frac{M\sigma_{a_h}^2\sigma_{b_h}^2}{\sigma^2}\right) = \frac{\sigma_{a_h}^2\sigma_{b_h}^2}{c_{a_h}^2 c_{b_h}^2}, \tag{71}$$

and therefore

$$\frac{LM}{\sigma^2}\sigma_{a_h}^4\sigma_{b_h}^4 - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right)\sigma_{a_h}^2\sigma_{b_h}^2 + \sigma^2 = 0. \tag{72}$$

Solving the quadratic equation (72) with respect to $\sigma_{a_h}^2\sigma_{b_h}^2$, and checking the signs of $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$, we have the following lemma (the proof is given in Appendix G.1):

Lemma 19 *For any $\gamma_h \geq 0$ and $c_{a_h}^2, c_{b_h}^2, \sigma^2 \in \mathbb{R}_{++}$, the null stationary point given by Eq.*(20) *exists with the following free energy:*

$$F_h^{\mathrm{VB-Null}} = -M\log\left(1 - \frac{L}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right) - L\log\left(1 - \frac{M}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right) - \frac{LM}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}, \tag{73}$$

$$\textit{where} \quad \widehat{\zeta}_h^{\mathrm{VB}}\left(\equiv \sigma_{a_h}^2\sigma_{b_h}^2\right) = \frac{\sigma^2}{2LM}\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2} - \sqrt{\left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right)^2 - 4LM}\right). \tag{21}$$

A.2 Positive Stationary Point

Assume that $\widehat{a}_h, \widehat{b}_h \neq 0$. In this case, Eqs.(64) and (66) imply that $\widehat{a}_h$ and $\widehat{b}_h$ have the same sign. Define

$$\widehat{\gamma}_h = \widehat{a}_h\widehat{b}_h > 0,$$
$$\widehat{\delta}_h = \frac{\widehat{a}_h}{\widehat{b}_h} > 0.$$

From Eqs.(64) and (66), we have

$$\sigma_{a_h}^2 = \frac{\sigma^2}{\gamma_h}\widehat{\delta}_h, \tag{74}$$

$$\sigma_{b_h}^2 = \frac{\sigma^2}{\gamma_h}\widehat{\delta}_h^{-1}. \tag{75}$$

Substituting Eqs.(74) and (75) into Eqs.(65) and (67) gives

$$\widehat{\delta}_h = \frac{c_{a_h}^2}{\sigma^2}\left(\gamma_h - \widehat{\gamma}_h - \frac{L\sigma^2}{\gamma_h}\right), \tag{76}$$

$$\widehat{\delta}_h^{-1} = \frac{c_{b_h}^2}{\sigma^2}\left(\gamma_h - \widehat{\gamma}_h - \frac{M\sigma^2}{\gamma_h}\right). \tag{77}$$

Multiplying Eqs.(76) and (77), we have

$$\left(\gamma_h - \widehat{\gamma}_h - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \widehat{\gamma}_h - \frac{M\sigma^2}{\gamma_h}\right) = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}, \tag{78}$$

and therefore

$$\widehat{\gamma}_h^2 - \left(2\gamma_h - \frac{(L+M)\sigma^2}{\gamma_h}\right)\widehat{\gamma}_h + \left(\gamma_h - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \frac{M\sigma^2}{\gamma_h}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = 0. \tag{79}$$

By solving the quadratic equation (79) with respect to $\widehat{\gamma}_h$, and checking the signs of $\widehat{\gamma}_h, \widehat{\delta}_h, \sigma_{a_h}^2$ and $\sigma_{b_h}^2$, we have the following lemma (the proof is given in Appendix G.2):

Lemma 20 *If and only if* $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$, *where*

$$\underline{\gamma}_h^{\mathrm{VB}} = \sigma\sqrt{\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)}{2} + \frac{\sigma^2}{2c_{a_h}^2 c_{b_h}^2}\right)^2 - LM}}, \tag{16}$$

the positive stationary point given by Eq.(18) *exists with the following free energy:*

$$F_h^{\mathrm{VB-Posi}} = -M\log\left(1 - \left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} + \frac{L\sigma^2}{\gamma_h^2}\right)\right) - L\log\left(1 - \left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} + \frac{M\sigma^2}{\gamma_h^2}\right)\right) - \frac{\gamma_h^2}{\sigma^2}\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} + \frac{L\sigma^2}{\gamma_h^2}\right)\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} + \frac{M\sigma^2}{\gamma_h^2}\right), \tag{80}$$

where $\breve{\gamma}_h^{\mathrm{VB}} = \gamma_h\left(1 - \frac{\sigma^2}{2\gamma_h^2}\left(M + L + \sqrt{(M-L)^2 + \frac{4\gamma_h^2}{c_{a_h}^2 c_{b_h}^2}}\right)\right)$. (17)

A.3 Useful Relations

Here, we summarize some useful relations between variables, which are used in the subsequent sections. $\widehat{\zeta}_h^{\mathrm{VB}}, \breve{\gamma}_h^{\mathrm{VB}}$, and $\underline{\gamma}_h^{\mathrm{VB}}$, derived from Eqs.(71), (78), and the constant part of Eq.(79), respectively, satisfy the following:

$$\left(1 - \frac{L\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right)\left(1 - \frac{M\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right) - \frac{\widehat{\zeta}_h^{\mathrm{VB}}}{c_{a_h}^2 c_{b_h}^2} = 0, \tag{81}$$

$$\left(\gamma_h - \breve{\gamma}_h^{\mathrm{VB}} - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \breve{\gamma}_h^{\mathrm{VB}} - \frac{M\sigma^2}{\gamma_h}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = 0, \tag{82}$$

$$\left(\underline{\gamma}_h^{\mathrm{VB}} - \frac{L\sigma^2}{\underline{\gamma}_h^{\mathrm{VB}}}\right)\left(\underline{\gamma}_h^{\mathrm{VB}} - \frac{M\sigma^2}{\underline{\gamma}_h^{\mathrm{VB}}}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} = 0. \tag{83}$$

From Eqs.(21) and (16), we find that

$$\underline{\gamma}_h^{\mathrm{VB}} = \sqrt{\left((L+M)\sigma^2 + \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}\right) - LM\widehat{\zeta}_h^{\mathrm{VB}}}, \tag{84}$$

which is useful when comparing the free energies of the null and the positive stationary points.

A.4 Free Energy Comparison

Lemma 19 and Lemma 20 imply that, when $\gamma_h \leq \underline{\gamma}_h^{\mathrm{VB}}$, the null stationary point is only the stationary point, and therefore the global solution. When $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$, both of the null and the positive stationary points exist, and therefore, identifying the global solution requires to compare the free energies, given by Eqs.(73) and (80), at them.

Given the observed singular value $\gamma_h \geq 0$, we view the free energy as a function of $c_{a_h}^2 c_{b_h}^2$. We also view the threshold $\underline{\gamma}_h^{\mathrm{VB}}$ as a function of $c_{a_h}^2 c_{b_h}^2$. We find from Eq.(16) that $\underline{\gamma}_h^{\mathrm{VB}}$ is decreasing and lower-bounded by $\underline{\gamma}_h^{\mathrm{VB}} > \sqrt{M}\sigma$. Therefore, when $\gamma_h \leq \sqrt{M}\sigma$, $\underline{\gamma}_h^{\mathrm{VB}}$ never gets smaller than γ_h for any $c_{a_h}^2 c_{b_h}^2 > 0$. When $\gamma_h > \sqrt{M}\sigma$ on the other hand, there is a threshold $\underline{c}_{a_h}^2 \underline{c}_{b_h}^2$ such that $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$ if $c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h}^2 \underline{c}_{b_h}^2$. Eq.(83) implies that the threshold is given by

$$\underline{c}_{a_h}^2 \underline{c}_{b_h}^2 = \frac{\sigma^4}{\gamma_h^2\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)}.$$

We have the following lemma (the proof is given in Appendix G.3):

Lemma 21 *For any $\gamma_h \geq 0$ and $c_{a_h}^2 c_{b_h}^2 > 0$, the derivative of the free energy* (73) *at the null stationary point with respect to $c_{a_h}^2 c_{b_h}^2$ is given by*

$$\frac{\partial F_h^{\mathrm{VB-Null}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{LM\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2 c_{a_h}^2 c_{b_h}^2}. \tag{85}$$

For $\gamma_h > M/\sigma^2$ and $c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h}^2 \underline{c}_{b_h}^2$, the derivative of the free energy (80) *at the positive stationary point with respect to $c_{a_h}^2 c_{b_h}^2$ is given by*

$$\frac{\partial F_h^{\mathrm{VB-Posi}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{\gamma_h^2}{\sigma^2 c_{a_h}^2 c_{b_h}^2}\left(\frac{(\breve{\gamma}_h^{\mathrm{VB}})^2}{\gamma_h^2} - \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} + \frac{LM\sigma^4}{\gamma_h^4}\right). \tag{86}$$

The derivative of the difference is negative, i.e.,

$$\frac{\partial(F_h^{\mathrm{Posi}} - F_h^{\mathrm{Null}})}{\partial c_{a_h}^2 c_{b_h}^2} = -\frac{1}{\sigma^2 c_{a_h}^2 c_{b_h}^2}\left(\gamma_h\left(\gamma_h - \breve{\gamma}_h^{\mathrm{VB}}\right) - (\underline{\gamma}_h^{\mathrm{VB}})^2\right) < 0. \tag{87}$$

It is easy to show that the null stationary point (20) and the positive stationary point (18) coincide with each other at $c_{a_h}^2 c_{b_h}^2 \to \underline{c}_{a_h}^2 \underline{c}_{b_h}^2 + 0$. Therefore,

$$\lim_{c_{a_h}^2 c_{b_h}^2 \to \underline{c}_{a_h}^2 \underline{c}_{b_h}^2 + 0} \left(F_h^{\mathrm{VB-Posi}} - F_h^{\mathrm{VB-Null}} \right) = 0. \tag{88}$$

Eqs.(87) and (88) together imply that

$$F_h^{\mathrm{VB-Posi}} - F_h^{\mathrm{VB-Null}} < 0 \qquad \text{for} \qquad c_{a_h}^2 c_{b_h}^2 > \underline{c}_{a_h}^2 \underline{c}_{b_h}^2,$$

which results in the following lemma:

Lemma 22 *The positive stationary point is the global solution (the global minimizer of the free energy* (14) *for fixed c_{a_h} and c_{b_h}) whenever it exists.*

Figure 8 illustrates the behavior of the free energies.

Combining Lemma 19, Lemma 20, and Lemma 22 completes the proof of of Theorem 2 and Corollary 3. ■

Appendix B. Proof of Theorem 4, Corollary 5, and Corollary 6

The EVB solution was also previously obtained:

Proposition 23 *(Nakajima et al., 2013b) The EVB solution is given by*

$$\widehat{\gamma}_h^{\mathrm{EVB}} = \begin{cases} \breve{\gamma}_h^{\mathrm{VB}} & \text{if } \gamma_h > (\sqrt{L} + \sqrt{M})\sigma \text{ and } F_h \leq 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $\breve{\gamma}_h^{\mathrm{VB}}$ is the VB solution for $c_{a_h}^2 c_{b_h}^2 = \widehat{c}_{a_h}^2 \widehat{c}_{b_h}^2$, and

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

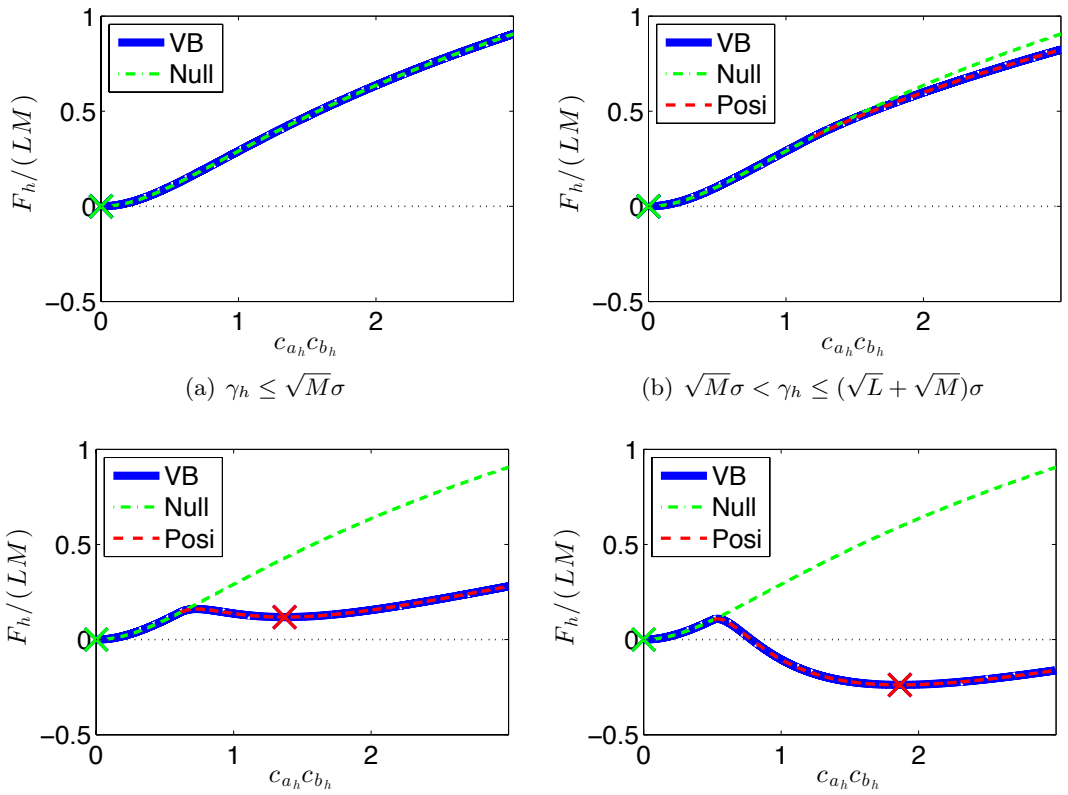
$$F_h = M \log \left(\frac{\gamma_h}{M\sigma^2} \breve{\gamma}_h^{\mathrm{VB}} + 1 \right) + L \log \left(\frac{\gamma_h}{L\sigma^2} \breve{\gamma}_h^{\mathrm{VB}} + 1 \right) + \frac{-2\gamma_h \breve{\gamma}_h^{\mathrm{VB}} + LM \widehat{c}_{a_h}^2 \widehat{c}_{b_h}^2}{\sigma^2}.$$

However, Proposition 23 requires to solve a quartic equation for obtaining $\breve{\gamma}_h^{\mathrm{VB}}$, and moreover, to evaluate the free energy F_h at the obtained $\breve{\gamma}_h^{\mathrm{VB}}$. This obstructs further analysis.

In this appendix, we prove Theorem 4, which provides explicit-forms, (25) and (26), of the EVB threshold $\underline{\gamma}^{\mathrm{EVB}}$ and the EVB shrinkage estimator $\breve{\gamma}_h^{\mathrm{EVB}}$. Without relying on Proposition 23, we can easily obtain Eq.(26) in an intuitive way, by using some of the results obtained in Appendix A. After that, by expressing the free energy F_h with rescaled observation and estimator, we derive Eq.(25).

B.1 EVB Shrinkage Estimator

Eqs.(73) and (80) imply that the free energy does not depend on the ratio c_{a_h}/c_{b_h} between the hyperparameters. Accordingly, we fix the ratio to $c_{a_h}/c_{b_h} = 1$. Lemma 21 allows us to minimize the free energy with respect to $c_{a_h} c_{b_h}$ in a straight-forward way.



(a) $\gamma_h \leq \sqrt{M}\sigma$ (b) $\sqrt{M}\sigma < \gamma_h \leq (\sqrt{L}+\sqrt{M})\sigma$

(c) $(\sqrt{L}+\sqrt{M})\sigma < \gamma_h < \underline{\gamma}^{\mathrm{EVB}}$ (d) $\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}$

Figure 8: Behavior of the free energies (73) and (80) at the null and the positive stationary points as functions of $c_{a_h}c_{b_h}$, when $L = M = H = 1$ and $\sigma^2 = 1$. The blue curve shows the VB free energy $F_h = \min(F_h^{\mathrm{VB-Null}}, F_h^{\mathrm{VB-Posi}})$ at the global solution, given $c_{a_h}c_{b_h}$. If $\gamma_h \leq \sqrt{M}\sigma$, only the null stationary point exists for any $c_{a_h}c_{b_h} > 0$. Otherwise, the positive stationary point exists for $c_{a_h}c_{b_h} > \underline{c}_{a_h}\underline{c}_{b_h}$, and it is the global minimum whenever it exists. In EVB where $c_{a_h}c_{b_h}$ is also optimized, $c_{a_h}c_{b_h} \to 0$ (indicated by a green cross) is the unique local minimum if $\gamma_h \leq (\sqrt{L}+\sqrt{M})\sigma$. Otherwise, a positive local minimum also exists (indicated by a red cross), and it is the global minimum if and only if $\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}$.

We see the free energies (73) and (80) at the null and the positive stationary points as function of $c_{a_h}c_{b_h}$ (see Figure 8). We find from Eq.(85) that

$$\frac{\partial F_h^{\mathrm{VB-Null}}}{\partial c_{a_h}^2 c_{b_h}^2} > 0,$$

which implies that the free energy (73) at the null stationary point is increasing. Using Lemma 19, we thus have the following lemma:

Lemma 24 *For any given $\gamma_h \geq 0$ and $\sigma^2 > 0$, the null EVB local solution given by*

$$\widehat{a}_h = 0, \qquad \widehat{b}_h = 0, \qquad \sigma^2_{a_h} = \sqrt{\widehat{\zeta}^{\mathrm{EVB}}}, \qquad \sigma^2_{b_h} = \sqrt{\widehat{\zeta}^{\mathrm{EVB}}}, \qquad c_{a_h} c_{b_h} = \sqrt{\widehat{\zeta}^{\mathrm{EVB}}},$$
$$\textit{where} \qquad \widehat{\zeta}^{\mathrm{EVB}} \to +0,$$

exists with the free energy that converges to

$$F_h^{\mathrm{EVB-Null}} \to +0. \tag{89}$$

When $\gamma_h \geq (\sqrt{L} + \sqrt{M})\sigma$, the derivative (86) of the free energy (80) at the positive stationary point can be further factorized as

$$\frac{\partial F_h^{\mathrm{VB-Posi}}}{\partial c_{a_h}^2 c_{b_h}^2} = \frac{\gamma_h}{\sigma^2 c_{a_h}^2 c_{b_h}^2} \left(\breve{\gamma}_h^{\mathrm{VB}} - \acute{\gamma}_h\right) \left(\breve{\gamma}_h^{\mathrm{VB}} - \breve{\gamma}_h^{\mathrm{EVB}}\right), \tag{90}$$

$$\text{where} \qquad \acute{\gamma}_h = \frac{\gamma_h}{2} \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2} - \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}\right), \tag{91}$$

$$\breve{\gamma}_h^{\mathrm{EVB}} = \frac{\gamma_h}{2} \left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2} + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}\right). \tag{26}$$

The VB shrinkage estimator (17) is an increasing function of $c_{a_h} c_{b_h}$ ranging over

$$0 < \breve{\gamma}_h^{\mathrm{VB}} < \gamma_h - \frac{M\sigma^2}{\gamma_h},$$

and both of Eqs.(91) and (26) are in this range, i.e.,

$$0 < \acute{\gamma}_h \leq \breve{\gamma}_h^{\mathrm{EVB}} < \gamma_h - \frac{M\sigma^2}{\gamma_h}.$$

Therefore Eq.(90) leads to the following lemma:

Lemma 25 *If $\gamma_h \leq (\sqrt{L} + \sqrt{M})\sigma$, the free energy $F_h^{\mathrm{VB-Posi}}$ at the positive stationary point is monotonically increasing. Otherwise,*

$$F_h^{\mathrm{VB-Posi}} \textit{ is } \begin{cases} \textit{increasing} & \textit{for} \quad \breve{\gamma}_h^{\mathrm{VB}} < \acute{\gamma}_h, \\ \textit{decreasing} & \textit{for} \quad \acute{\gamma}_h < \breve{\gamma}_h^{\mathrm{VB}} < \breve{\gamma}_h^{\mathrm{EVB}}, \\ \textit{increasing} & \textit{for} \quad \breve{\gamma}_h^{\mathrm{VB}} > \breve{\gamma}_h^{\mathrm{EVB}}, \end{cases}$$

and therefore, minimized at $\breve{\gamma}_h^{\mathrm{VB}} = \breve{\gamma}_h^{\mathrm{EVB}}$.

We can see this behavior of the free energy in Figure 8.

The derivative (86) is zero when $\breve{\gamma}_h^{\mathrm{VB}} = \breve{\gamma}_h^{\mathrm{EVB}}$, which leads to

$$\left(\breve{\gamma}_h^{\mathrm{EVB}} + \frac{L\sigma^2}{\gamma_h}\right) \left(\breve{\gamma}_h^{\mathrm{EVB}} + \frac{M\sigma^2}{\gamma_h}\right) = \gamma_h \breve{\gamma}_h^{\mathrm{EVB}}. \tag{92}$$

Using Eq.(92), we obtain the following lemma (the proof is given in Appendix G.4):

Lemma 26 *If and only if*

$$\gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}} \equiv (\sqrt{L}+\sqrt{M})\sigma, \tag{28}$$

the positive EVB local solution given by

$$\widehat{a}_h = \pm\sqrt{\breve{\gamma}_h^{\mathrm{EVB}}\widehat{\delta}_h^{\mathrm{EVB}}}, \qquad \widehat{b}_h = \pm\sqrt{\frac{\breve{\gamma}_h^{\mathrm{EVB}}}{\widehat{\delta}_h^{\mathrm{EVB}}}}, \qquad \sigma_{a_h}^2 = \frac{\sigma^2\widehat{\delta}_h^{\mathrm{EVB}}}{\gamma_h}, \qquad \sigma_{b_h}^2 = \frac{\sigma^2}{\gamma_h\widehat{\delta}_h^{\mathrm{EVB}}}, \tag{93}$$

$$c_{a_h}c_{b_h} = \sqrt{\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{LM}}, \qquad \textit{where} \qquad \widehat{\delta}_h^{\mathrm{EVB}} = \sqrt{\frac{M\breve{\gamma}_h^{\mathrm{EVB}}}{L\gamma_h}}\left(1+\frac{L\sigma^2}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}\right), \tag{94}$$

$$\breve{\gamma}_h^{\mathrm{EVB}} = \frac{\gamma_h}{2}\left(1-\frac{(M+L)\sigma^2}{\gamma_h^2}+\sqrt{\left(1-\frac{(M+L)\sigma^2}{\gamma_h^2}\right)^2-\frac{4LM\sigma^4}{\gamma_h^4}}\right), \tag{26}$$

exists with the following free energy:

$$F_h^{\mathrm{EVB-Posi}} = M\log\left(\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{M\sigma^2}+1\right)+L\log\left(\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{L\sigma^2}+1\right)-\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\sigma^2}. \tag{95}$$

In Figure 8, the positive EVB local solution at $c_{a_h}c_{b_h} = \sqrt{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}/(LM)}$ is indicated by a red cross if it exists.

B.2 EVB Threshold

Lemma 24 and Lemma 26 state that, if $\gamma_h \leq \underline{\gamma}^{\mathrm{local-EVB}}$, only the null EVB local solution exists, and therefore it is the global EVB solution. Below, assuming that $\gamma_h \geq \underline{\gamma}^{\mathrm{local-EVB}}$, we compare the free energy (89) at the null EVB local solution and the free energy (95) at the positive EVB local solution. Since $F_h^{\mathrm{EVB-Null}} \to +0$, we simply clarify when $F_h^{\mathrm{EVB-Posi}} \leq 0$. Eq.(92) gives

$$\left(\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}+L\sigma^2\right)\left(1+\frac{M\sigma^2}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}\right) = \gamma_h^2. \tag{29}$$

By using Eqs.(26) and (28), we have

$$\begin{aligned}\gamma_h\breve{\gamma}_h^{\mathrm{EVB}} = \frac{1}{2}\Bigg(&\gamma_h^2-\left(\underline{\gamma}^{\mathrm{local-EVB}}\right)^2+2\sqrt{LM}\sigma^2 \\ &+\sqrt{\left(\gamma_h^2-\left(\underline{\gamma}^{\mathrm{local-EVB}}\right)^2\right)\left(\gamma_h^2-\left(\underline{\gamma}^{\mathrm{local-EVB}}\right)^2+4\sqrt{LM}\sigma^2\right)}\Bigg) \\ \geq\ &\sqrt{LM}\sigma^2.\end{aligned} \tag{30}$$

Let

$$\alpha = \frac{L}{M} \qquad (0<\alpha\leq 1), \tag{22}$$

$$x_h = \frac{\gamma_h^2}{M\sigma^2}, \tag{33}$$

$$\tau_h = \frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{M\sigma^2}. \tag{34}$$

Eqs.(29) and (26) imply the following mutual relations between x_h and τ_h:

$$x_h \equiv x(\tau_h;\alpha) = (1+\tau_h)\left(1+\frac{\alpha}{\tau_h}\right), \tag{35}$$

$$\tau_h \equiv \tau(x_h;\alpha) = \frac{1}{2}\left(x_h - (1+\alpha) + \sqrt{(x_h-(1+\alpha))^2 - 4\alpha}\right). \tag{36}$$

Eqs.(28) and (30) lead to

$$x_h \geq \underline{x}^{\mathrm{local}} = \frac{(\underline{\gamma}^{\mathrm{local-EVB}})^2}{M\sigma^2} = x(\sqrt{\alpha};\alpha) = (1+\sqrt{\alpha})^2, \tag{37}$$

$$\tau_h \geq \underline{\tau}^{\mathrm{local}} = \sqrt{\alpha}. \tag{38}$$

Then, using

$$\Xi(\tau;\alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right), \qquad \text{where} \qquad \Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}, \tag{23}$$

we can rewrite Eq.(95) as

$$\begin{aligned} F_h^{\mathrm{EVB-Posi}} &= M\log(\tau_h+1) + L\log\left(\frac{\tau_h}{\alpha}+1\right) - M\tau_h \\ &= M\tau_h \Xi(\tau;\alpha). \end{aligned} \tag{96}$$

The following holds for $\Phi(z)$ (the proof is given in Appendix G.5):

Lemma 27 *$\Phi(z)$ is decreasing for $z > 0$.*

Figure 9 shows $\Phi(z)$. Since $\Phi(z)$ is decreasing, $\Xi(\tau;\alpha)$ is also decreasing with respect to τ. It holds that, for any $0 < \alpha \leq 1$,

$$\lim_{\tau\to 0} \Xi(\tau;\alpha) = 1,$$
$$\lim_{\tau\to\infty} \Xi(\tau;\alpha) = -1.$$

Therefore, $\Xi(\tau;\alpha)$ has a unique zero-cross point $\underline{\tau}$, such that

$$\Xi(\tau;\alpha) \leq 0 \qquad \text{if and only if} \qquad \tau \geq \underline{\tau}. \tag{97}$$

We can prove the following lemma (the proof is given in Appendix G.6):

Lemma 28 *The unique zero-cross point $\underline{\tau}$ of $\Xi(\tau;\alpha)$ lies in the following range:*

$$\sqrt{\alpha} < \underline{\tau} \leq \underline{z}, \tag{27}$$

where $\underline{z} \approx 2.5129$ is the unique zero-cross point of $\Phi(z)$.

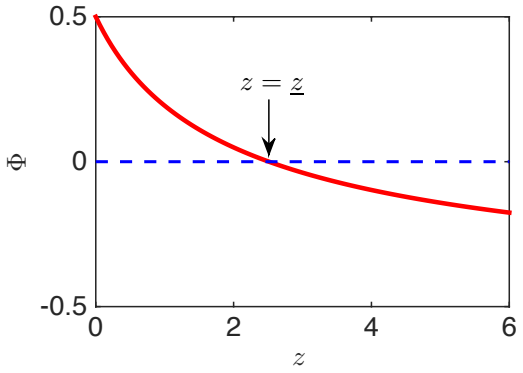


Figure 9: $\Phi(z) = \frac{\log(z+1)}{z} - \frac{1}{2}$. $\underline{z} \approx 2.5129$ is the unique zero cross point, i.e., $\Phi(\underline{z}) = 0$.

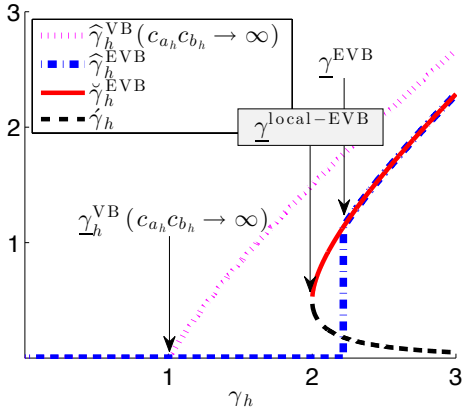


Figure 10: Estimators and thresholds for $L = M = H = 1$ and $\sigma^2 = 1$.

Since Eq.(35) is increasing with respect to τ_h ($> \sqrt{\alpha}$), the thresholding condition $\tau \geq \underline{\tau}$ in Eq.(97) can be expressed in terms of x:

$$\Xi(\tau(x);\alpha) \leq 0 \qquad \text{if and only if} \qquad x \geq \underline{x},$$
$$\text{where} \qquad \underline{x} \equiv x(\underline{\tau};\alpha) = (1+\underline{\tau})\left(1+\frac{\alpha}{\underline{\tau}}\right). \tag{39}$$

Using Eqs.(33) and (96), we have

$$F_h^{\mathrm{EVB-Posi}} \leq 0 \qquad \text{if and only if} \qquad \gamma_h \geq \underline{\gamma}^{\mathrm{EVB}},$$
$$\text{where} \qquad \underline{\gamma}^{\mathrm{EVB}} = \sigma\sqrt{M(1+\underline{\tau})\left(1+\frac{\alpha}{\underline{\tau}}\right)}. \tag{25}$$

Thus, we have the following lemma:

Lemma 29 *The positive EVB local solution is the global EVB solution if and only if* $\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}$.

Combining Lemma 24, Lemma 26, and Lemma 29 completes the proof of Theorem 4 and Corollary 6. All formulas in Corollary 5 have already been derived. ■

Figure 10 shows estimators and thresholds for $L = M = H = 1$ and $\sigma^2 = 1$. The curves indicate the VB solution $\widehat{\gamma}_h^{\mathrm{VB}}$, given by Eq.(15), the EVB solution $\widehat{\gamma}_h^{\mathrm{EVB}}$, given by Eq.(24), the EVB positive local minimizer $\breve{\gamma}_h^{\mathrm{EVB}}$, given by Eq.(26), and the EVB positive local maximizer $\acute{\gamma}_h$, given by Eq.(91), respectively. The arrows indicate the VB threshold $\underline{\gamma}_h^{\mathrm{VB}}$, given by Eq.(16), the local-EVB threshold $\underline{\gamma}^{\mathrm{local-EVB}}$, given by Eq.(28), and the EVB threshold $\underline{\gamma}^{\mathrm{EVB}}$, given by Eq.(25), respectively.

Appendix C. Proof of Theorem 7

By using Lemma 24 and Lemma 26, the free energy (13) can be written as a function of σ^2:

$$2F = LM\log(2\pi\sigma^2) + \frac{\sum_{h=1}^{L}\gamma_h^2}{\sigma^2} + \sum_{h=1}^{H}\theta\left(\gamma_h > \underline{\gamma}^{\mathrm{EVB}}\right)F_h^{\mathrm{EVB-Posi}}, \tag{98}$$

$$\text{where} \quad F_h^{\mathrm{EVB-Posi}} = M\log\left(\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{M\sigma^2}+1\right) + L\log\left(\frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{L\sigma^2}+1\right) - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\sigma^2}. \tag{95}$$

By using Eqs.(34) and (36), Eq.(95) can be written as

$$\begin{aligned} F_h^{\mathrm{EVB-Posi}} &= M\log\left(\tau_h+1\right) + L\log\left(\frac{\tau_h}{\alpha}+1\right) - M\tau_h \\ &= M\psi_1(x_h). \end{aligned} \tag{99}$$

Therefore, Eq.(98) is written as

$$\begin{aligned} 2F &= M\left\{\sum_{h=1}^{L}\log\left(\frac{2\pi\gamma_h^2}{M}\right) + \sum_{h=1}^{L}\left(\log\left(\frac{M\sigma^2}{\gamma_h^2}\right)+\frac{\gamma_h^2}{M\sigma^2}\right) + \sum_{h=1}^{H}\theta\left(\gamma_h > \underline{\gamma}^{\mathrm{EVB}}\right)\frac{F_h^{\mathrm{EVB-Posi}}}{M}\right\} \\ &= M\left\{\sum_{h=1}^{L}\log\left(\frac{2\pi\gamma_h^2}{M}\right) + \sum_{h=1}^{L}\psi_0(x_h) + \sum_{h=1}^{H}\theta\left(x_h > \underline{x}\right)\psi_1(x_h)\right\}. \end{aligned}$$

Note that the first term in the curly braces is constant with respect to σ^2. By defining

$$\Omega = \frac{2F}{LM} - \frac{1}{L}\sum_{h=1}^{L}\log\left(\frac{2\pi\gamma_h^2}{M}\right),$$

we obtain Eq.(40), which completes the proof of Theorem 7. ■

Appendix D. Proof of Theorem 8 and Corollary 9

First, we investigate properties of the following functions, which are depicted in Fig. 3:

$$\psi(x) = \psi_0(x) + \theta\left(x > \underline{x}\right)\psi_1(x), \tag{41}$$

$$\psi_0(x) = x - \log x, \tag{42}$$

$$\text{where} \quad \psi_1(x) = \log\left(\tau(x;\alpha)+1\right) + \alpha\log\left(\frac{\tau(x;\alpha)}{\alpha}+1\right) - \tau(x;\alpha). \tag{43}$$

They have nice properties (the proof is given in Appendix G.7):

Lemma 30 *The following hold for $x > 0$: $\psi_0(x)$ is differentiable and strictly convex; $\psi(x)$ is continuous and strictly quasi-convex; $\psi(x)$ is differentiable except $x = \underline{x}$, at which $\psi(x)$ has a discontinuously decreasing derivative, i.e., $\lim_{x\to\underline{x}-0}\partial\psi/\partial x > \lim_{x\to\underline{x}+0}\partial\psi/\partial x$; Both of $\psi_0(x)$ and $\psi(x)$ are minimized at $x = 1$. For $x > \underline{x}$, $\psi_1(x)$ is negative and decreasing.*

Lemma 30 implies that our objective

$$\Omega(\sigma^{-2}) = \frac{1}{L}\left(\sum_{h=1}^{H}\psi\left(\frac{\gamma_h^2}{M\sigma^2}\right) + \sum_{h=H+1}^{L}\psi_0\left(\frac{\gamma_h^2}{M\sigma^2}\right)\right) \tag{40}$$

is a sum of quasi-convex functions with respect to σ^{-2}. Therefore, its minimizer can be bounded by the smallest and the largest ones of the minimizers of each quasi-convex function (the proof is given in Appendix G.8):

Lemma 31 *$\Omega(\sigma^{-2})$ has at least one global minimizer, and any of its local minimizers is bounded as*

$$\frac{M}{\gamma_1^2} \leq \widehat{\sigma}^{-2} \leq \frac{M}{\gamma_L^2}.$$

$\Omega(\sigma^{-2})$ has at most H non-differentiable points, which come from the non-differentiable point $x = \underline{x}$ of $\psi(x)$. The values

$$\underline{\sigma}_h^{-2} = \begin{cases} 0 & \text{for } h = 0, \\ \frac{M\underline{x}}{\gamma_h^2} & \text{for } h = 1, \ldots, L, \\ \infty & \text{for } h = L+1, \end{cases} \tag{100}$$

defined in Eq.(45), for $h = 1, \ldots, H$ actually correspond to those points.

Lemma 30 states that, at $x = \underline{x}$, $\psi(x)$ has a discontinuously decreasing derivative and neither $\psi_0(x)$ nor $\psi(x)$ has discontinuously increasing derivative at any point. Therefore, none of those non-differentiable points can be local minimum. Consequently, we have the following lemma:

Lemma 32 *$\Omega(\sigma^{-2})$ has no local minimizer at $\sigma^{-2} = \underline{\sigma}_h^{-2}$ for $h = 1, \ldots, H$, and therefore, any of its local minimizer is stationary point.*

Then, Theorem 4 leads to the following lemma:

Lemma 33 *The estimated rank is $\widehat{H} = h$, if and only if the inverse noise variance estimator lies in the range*

$$\widehat{\sigma}^{-2} \in \mathbb{B}_h \equiv \left\{\sigma^{-2}; \underline{\sigma}_h^{-2} < \sigma^{-2} < \underline{\sigma}_{h+1}^{-2}\right\}.$$

Figure 11 shows quasi-convex functions $\{\psi(\gamma_h^2\sigma^{-2}/M)\}_{h=1}^{H}$ and their sum $\Omega(\sigma^{-2})$ in two example cases for $H = L$. In the left case, the inverse noise variance estimator $\widehat{\sigma}^{-2}$ is smaller than the inverse threshold $\underline{\sigma}_1^{-2}$ for the largest singular value, and therefore, no EVB estimator $\widehat{\gamma}_h$ is positive, i.e., $\widehat{H} = 0$. In the right case, it holds that $\underline{\sigma}_1^{-2} < \widehat{\sigma}^{-2} < \underline{\sigma}_2^{-2}$, and therefore, $\widehat{\gamma}_1$ is positive and the others are zero, i.e., $\widehat{H} = 1$.

We have the following lemma (the proof is given in Appendix G.9):

Lemma 34 *The derivative of $\Omega(\sigma^{-2})$ is given by*

$$\Theta \equiv \frac{\partial\Omega}{\partial\sigma^{-2}} = -\sigma^2 + \frac{\sum_{h=1}^{\widehat{H}}\gamma_h\left(\gamma_h - \breve{\gamma}_h^{\mathrm{EVB}}\right) + \sum_{h=\widehat{H}+1}^{L}\gamma_h^2}{LM}, \tag{101}$$

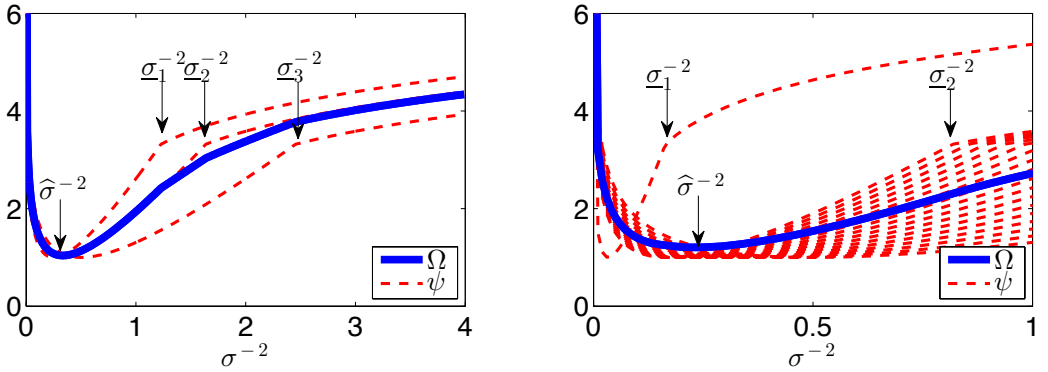


Figure 11: $\{\psi(\gamma_h^2\sigma^{-2}/M)\}_{h=1}^H$ and $\Omega(\sigma^{-2})$ in two example cases for $H = L$. (Left) The case when $\gamma_h^2/M = 4, 3, 2$ for $h = 1, 2, 3$. (Right) The case when $\gamma_1^2/M = 30$, $\gamma_h^2/M = 6.0, 5.75, 5.5, \ldots, 2.0$ for $h = 2, \ldots, 18$.

where $\widehat{H}$ *is a function of* σ^{-2} *defined by*

$$\widehat{H} = \widehat{H}(\sigma^{-2}) = h \qquad \textit{if} \qquad \sigma^{-2} \in \mathbb{B}_h. \tag{102}$$

Note that Eq.(101) involves the shrinkage estimator $\breve{\gamma}_h^{\mathrm{EVB}}$, which is a function of σ^{-2} (see Eq.(26)). For each hypothetical $\widehat{H}$, the solutions of the equation

$$\Theta = 0 \tag{103}$$

lying in $\sigma^{-2} \in \mathbb{B}_{\widehat{H}}$ are stationary points, and hence candidates for the global minimum. If we can solve Eq.(103) for all $\widehat{H} = 1, \ldots, H$, we can obtain the global solution by evaluating the objective (40) at each obtained stationary points. However, solving Eq.(103) is difficult unless $\widehat{H}$ is small (it is easy to derive a closed-form solution for $\widehat{H} = 0, 1$). Based on Lemma 34, we will obtain tighter bounds than Lemma 31.

Since

$$\gamma_h - \breve{\gamma}_h^{\mathrm{EVB}} > 0,$$

Eq.(101) is upper-bounded by

$$\Theta \leq -\sigma^2 + \sum_{h=1}^{L} \frac{\gamma_h^2}{LM},$$

which leads to the upper-bound given in Eq.(44). Actually, if

$$\left(\sum_{h=1}^{L} \frac{\gamma_h^2}{LM}\right)^{-1} \in \mathbb{B}_0,$$

then

$$\widehat{H} = 0,$$

$$\widehat{\sigma}^2 = \sum_{h=1}^{L} \frac{\gamma_h^2}{LM},$$

is a local minimum.

The following lemma is easily obtained from Eq.(26) by using $z_1 < \sqrt{z_1^2 - z_2^2} < z_1 - z_2$ for $z_1 > z_2 > 0$:

Lemma 35 *For $\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}$, the EVB shrinkage estimator* (26) *can be bounded as follows:*

$$\gamma_h - \frac{(\sqrt{M}+\sqrt{L})^2\sigma^2}{\gamma_h} < \breve{\gamma}_h^{\mathrm{EVB}} < \gamma_h - \frac{(M+L)\sigma^2}{\gamma_h}.$$

This lemma is important for our analysis, because it allows us to bound the most complicated part of Eq.(101) by terms independent of γ_h, i.e.,

$$(M+L)\sigma^2 < \gamma_h \left(\gamma_h - \breve{\gamma}_h^{\mathrm{EVB}}\right) < (\sqrt{M}+\sqrt{L})^2\sigma^2. \tag{104}$$

Using Eq.(104), we obtain the following lemma (the proof is given in Appendix G.10):

Lemma 36 *Any local minimizer exists in $\sigma^{-2} \in \mathbb{B}_{\widehat{H}}$ such that*

$$\widehat{H} < \frac{L}{1+\alpha},$$

and the following holds for any local minimizer lying in $\sigma^{-2} \in \mathbb{B}_{\widehat{H}}$:

$$\widehat{\sigma}^2 \geq \frac{\sum_{h=\widehat{H}+1}^{L} \gamma_h^2}{LM - \widehat{H}(M+L)}.$$

It holds that

$$\frac{\sum_{h=\widehat{H}+1}^{L} \gamma_h^2}{LM - \widehat{H}(M+L)} \geq \frac{\sum_{h=\widehat{H}+1}^{L} \gamma_h^2}{M(L-\widehat{H})}, \tag{105}$$

of which the right-hand side is decreasing with respect to $\widehat{H}$. Combining Lemma 31, Lemma 32, Lemma 33, Lemma 36, and Eq.(105) completes the proof of Theorem 8. Corollary 9 is easily obtained from Lemma 32 and Lemma 34.

■

Appendix E. Proof of Theorem 13 and Corollary 14

In the large-scale limit, we can substitute the expectation $\langle f(y) \rangle_{p(y)}$ for the summation $L^{-1}\sum_{h=1}^{L} f(y_h)$. We can also substitute the MP distribution $p^{\mathrm{MP}}(y)$ for $p(y)$ in the expectation, since the contribution from the H^* signal components converges to zero. Accordingly, our objective (40) converges to

$$\Omega(\sigma^{-2}) \to \Omega^{\mathrm{LSL}}(\sigma^{-2}) \equiv \int_{\kappa}^{\overline{y}} \psi\left(\sigma^{*2}\sigma^{-2}y\right) p^{\mathrm{MP}}(y)dy + \int_{\underline{y}}^{\kappa} \psi_0\left(\sigma^{*2}\sigma^{-2}y\right) p^{\mathrm{MP}}(y)dy$$

$$= \Omega^{\mathrm{LSL-Full}}(\sigma^{-2}) - \int_{\max(\underline{x}\sigma^2/\sigma^{*2}, \underline{y})}^{\kappa} \psi_1\left(\sigma^{*2}\sigma^{-2}y\right) p^{\mathrm{MP}}(y)dy, \quad (106)$$

$$\text{where} \qquad \Omega^{\mathrm{LSL-Full}}(\sigma^{-2}) \equiv \int_{\underline{y}}^{\overline{y}} \psi\left(\sigma^{*2}\sigma^{-2}y\right) p^{\mathrm{MP}}(y)dy, \quad (107)$$

and κ is a constant satisfying

$$\frac{H}{L} = \int_{\kappa}^{\overline{y}} p^{\mathrm{MP}}(y)dy \qquad (\underline{y} \le \kappa \le \overline{y}).$$

Note that $\underline{x}$, $\underline{y}$, and $\overline{y}$ are defined by Eqs.(39) and (48), and it holds that

$$\underline{x} > \overline{y}. \quad (108)$$

We first investigate Eq.(107), which corresponds to the objective for the full-rank $H = L$ model. Let

$$s = \log(\sigma^{-2}),$$
$$t = \log y \qquad \left(dt = \tfrac{1}{y}dy\right).$$

Then, Eq.(107) is written as a convolution:

$$\widetilde{\Omega}^{\mathrm{LSL-Full}}(s) \equiv \Omega^{\mathrm{LSL-Full}}(e^s) = \int \psi\left(\sigma^{*2}e^{s+t}\right) e^t p^{\mathrm{MP}}(e^t)dt$$
$$= \int \widetilde{\psi}(s+t)p^{\mathrm{LSMP}}(t)dt,$$

where

$$\widetilde{\psi}(s) = \psi(\sigma^{*2}e^s),$$
$$p^{\mathrm{LSMP}}(t) = e^t p^{\mathrm{MP}}(e^t)$$
$$= \frac{\sqrt{(e^t - \underline{y})(\overline{y} - e^t)}}{2\pi\alpha}\theta(\underline{y} < e^t < \overline{y}). \quad (109)$$

Since Lemma 30 states that $\psi(x)$ is quasi-convex, its composition $\widetilde{\psi}(s)$ with the non-decreasing function $\sigma^{*2}e^s$ is also quasi-convex.

The following holds for $p^{\mathrm{LSMP}}(t)$, which we call a log-scaled MP (LSMP) distribution (the proof is given in Appendix G.11):

Lemma 37 *The LSMP distribution* (109) *is log-concave.*

Lemma 37 and Proposition 12 imply that $\widetilde{\Omega}^{\mathrm{LSL-Full}}(s)$ is quasi-convex, and therefore, its composition $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$ with the non-decreasing function $\log(\sigma^{-2})$ is quasi-convex. The minimizer of $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$ can be found by evaluating the derivative Θ, given by Eq.(101), in the large-scale limit:

$$\Theta^{\mathrm{Full}} \to \Theta^{\mathrm{LSL-Full}} = -\sigma^2 + \sigma^{*2}\int_{\underline{y}}^{\overline{y}} y \cdot p^{\mathrm{MP}}(y)dy - \int_{\underline{x}\sigma^2/\sigma^{*2}}^{\overline{y}} \tau(\sigma^{*2}\sigma^{-2}y; \alpha)p^{\mathrm{MP}}(y)dy. \quad (110)$$

Here, we used Eqs.(34) and (36). In the range

$$0 < \sigma^{-2} < \frac{\underline{x}\sigma^{*-2}}{\overline{y}} \qquad \left(i.e., \quad \frac{\underline{x}\sigma^2}{\sigma^{*2}} > \overline{y}\right), \tag{111}$$

the third term in Eq.(110) is zero. Therefore, Eq.(110) is increasing with respect to σ^{-2}, and zero when

$$\sigma^2 = \sigma^{*2} \int_{\underline{y}}^{\overline{y}} y \cdot p^{\mathrm{MP}}(y) dy = \sigma^{*2}.$$

Accordingly, $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$ is strictly convex in the range (111). Eq.(108) implies that the point $\sigma^{-2} = \sigma^{*-2}$ is contained in the region (111), and therefore, it is a local minimum of $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$. Combined with the quasi-convexity of $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$, we have the following lemma:

Lemma 38 *The objective $\Omega^{\mathrm{LSL-Full}}(\sigma^{-2})$ for the full rank model $H = L$ in the large-scale limit is quasi-convex with its minimizer at $\sigma^{-2} = \sigma^{*-2}$. It is strictly convex in the range* (111).

For any κ ($\underline{y} < \kappa < \overline{y}$), the second term in Eq.(106) is zero in the range (111), which includes its minimizer at $\sigma^{-2} = \sigma^{*-2}$. Since Lemma 30 states that $\psi_1(x)$ is decreasing for $x > \underline{x}$, the second term in Eq.(106) is non-decreasing in the region where

$$\left(\sigma^{*-2} <\right) \frac{\underline{x}\sigma^{*-2}}{\overline{y}} \leq \sigma^{-2} < \infty.$$

Therefore, the quasi-convexity of $\Omega^{\mathrm{LSL-Full}}$ is inherited to Ω^{LSL}:

Lemma 39 *The objective $\Omega^{\mathrm{LSL}}(\sigma^{-2})$ for noise variance estimation in the large-scale limit is quasi-convex with its minimizer at $\sigma^{-2} = \sigma^{*-2}$. $\Omega^{\mathrm{LSL}}(\sigma^{-2})$ is strictly convex in the range* (111).

Thus, we have proved that EVB accurately estimates the noise variance in the large-scale limit:

$$\widehat{\sigma}^{2\,\mathrm{EVB}} = \sigma^{*2}.$$

Assume that

$$\nu^*_{H^*} > \sqrt{\alpha}. \tag{51}$$

Then, Proposition 11 guarantees that, in the large-scale limit, it holds that

$$\frac{\gamma^2_{H^*}}{M\sigma^{*2}} \equiv y_{H^*} = (1 + \nu^*_{H^*}) \left(1 + \frac{\alpha}{\nu^*_{H^*}}\right), \tag{112}$$

$$\frac{\gamma^2_{H^*+1}}{M\sigma^{*2}} \equiv y_{H^*+1} = \overline{y} = (1 + \sqrt{\alpha})^2. \tag{113}$$

The EVB threshold is given by

$$\frac{(\underline{\gamma}^{\mathrm{EVB}})^2}{M\widehat{\sigma}^{2\,\mathrm{EVB}}} \equiv \underline{x} = (1+\underline{\tau})\left(1+\frac{\alpha}{\underline{\tau}}\right). \tag{39}$$

Since Lemma 39 states that $\widehat{\sigma}^{2\,\mathrm{EVB}} = \sigma^{*2}$, comparing Eqs.(112) and (113) with Eq.(39) results in the following lemma:

Lemma 40 *It almost surely holds that*

$$\begin{aligned} \gamma_{H^*} \geq \underline{\gamma}^{\mathrm{EVB}} \qquad & \textit{if and only if} \qquad \nu^*_{H^*} \geq \underline{\tau}, \\ \gamma_{H^*+1} < \underline{\gamma}^{\mathrm{EVB}} \qquad & \textit{for any} \qquad \{\nu^*_h\}. \end{aligned}$$

This completes the proof of Theorem 13. Comparing Eqs.(35) and (49) under Lemma 39 and Lemma 40 proves Corollary 14. ■

Appendix F. Proof of Theorem 15 and Corollary 16

We regroup the terms in Eq.(40) as follows:

$$\Omega(\sigma^{-2}) = \Omega_1(\sigma^{-2}) + \Omega_0(\sigma^{-2}), \tag{114}$$

where

$$\Omega_1(\sigma^{-2}) = \frac{1}{H^*}\sum_{h=1}^{H^*} \psi\left(\frac{\gamma_h^2}{M}\sigma^{-2}\right), \tag{115}$$

$$\Omega_0(\sigma^{-2}) = \frac{1}{L-H^*}\left(\sum_{h=H^*+1}^{H} \psi\left(\frac{\gamma_h^2}{M}\sigma^{-2}\right) + \sum_{h=H+1}^{L} \psi_0\left(\frac{\gamma_h^2}{M}\sigma^{-2}\right)\right). \tag{116}$$

Below, assuming that

$$p(y) = p^{\mathrm{SC}}(y), \tag{54}$$

and

$$y_{H^*} > \overline{y}, \tag{117}$$

we derive a sufficient condition for any local minimizer to lie only in $\sigma^{-2} \in \mathbb{B}_{H^*}$, with which Lemma 33 proves the theorem.

Under the assumption (54) and the condition (117), $\Omega_0(\sigma^{-2})$, defined by Eq.(116), is equivalent to the objective $\Omega^{LSL}(\sigma^{-2})$ in the large-scale limit. Using Lemma 39, and noting that

$$\underline{\sigma}^{-2}_{H^*+1} = \frac{M\underline{x}}{\gamma_{H^*+1}}^2 = \frac{\underline{x}\sigma^{*-2}}{\overline{y}} > \sigma^{*-2}, \tag{118}$$

we have the following lemma:

Lemma 41 *$\Omega_0(\sigma^{-2})$ is quasi-convex with its minimizer at*

$$\sigma^{-2} = \sigma^{*-2}.$$

$\Omega_0(\sigma^{-2})$ is strictly convex in the range

$$0 < \sigma^{-2} < \underline{\sigma}_{H^*+1}^{-2}.$$

Using Lemma 41 and the strict quasi-convexity of $\psi(x)$, we can deduce the following lemma (the proof is given in Appendix G.12):

Lemma 42 *$\Omega(\sigma^{-2})$ is non-decreasing (increasing if $\xi > 0$) in the range $\underline{\sigma}_{H^*+1}^2 < \sigma^{-2} < \infty$.*

Using the bounds given by Eq.(104) and Lemma 41, we also obtain the following lemma (the proof is given in Appendix G.13):

Lemma 43 *$\Omega(\sigma^{-2})$ is increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$. It is decreasing at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0$ if the following hold:*

$$\xi < \frac{1}{(1+\sqrt{\alpha})^2}, \tag{119}$$

$$y_{H^*} > \frac{\underline{x}(1-\xi)}{1-\xi(1+\sqrt{\alpha})^2}. \tag{120}$$

Finally, we obtain the following lemma (the proof is given in Appendix G.14):

Lemma 44 *$\Omega(\sigma^{-2})$ is decreasing in the range $0 < \sigma^{-2} < \underline{\sigma}_{H^*}^2$ if the following hold:*

$$\xi < \frac{1}{\underline{x}}, \tag{121}$$

$$y_{H^*} > \frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}. \tag{122}$$

Lemma 42, Lemma 43, and Lemma 44 together state that, if all the conditions (117), (119)–(122) hold, at least one local minimum exists in the correct range $\sigma^{-2} \in \mathbb{B}_{H^*}$, and no local minimum (no stationary point if $\xi > 0$) exists outside the correct range. Therefore, we can estimate the correct rank $\widehat{H}^{\mathrm{EVB}} = H^*$ by using a local search algorithm for noise variance estimation. Choosing the tightest conditions, we have the following lemma:

Lemma 45 *$\Omega(\sigma^{-2})$ has a global minimum in $\sigma^{-2} \in \mathbb{B}_{H^*}$, and no local minimum (no stationary point if $\xi > 0$) outside $\mathbb{B}_{H^*}$, if the following hold:*

$$\xi < \frac{1}{\underline{x}},$$

$$y_{H^*} = \frac{\gamma_{H^*}^2}{M\sigma^{*2}} > \frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}. \tag{123}$$

Using Eq.(49), Eq.(123) can be written with the *true* signal amplitude as follows:

$$(1+\nu^*_{H^*})\left(1+\frac{\alpha}{\nu^*_{H^*}}\right)-\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}>0.$$

The left-hand side can be factorized as follows:

$$\frac{1}{\nu^*_{H^*}}\left(\nu^*_{H^*}-\frac{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)+\sqrt{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)^2-4\alpha}}{2}\right)$$
$$\cdot\left(\nu^*_{H^*}-\frac{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)-\sqrt{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)^2-4\alpha}}{2}\right)>0. \tag{124}$$

When Eq.(51) holds, the last factor in the left-hand side in Eq.(124) is positive. Therefore, we have the following condition:

$$\begin{aligned}\nu^*_{H^*}&>\frac{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)+\sqrt{\left(\frac{\underline{x}(1-\xi)}{1-\underline{x}\xi}-(1+\alpha)\right)^2-4\alpha}}{2}\\&=\frac{\left(\frac{\underline{x}-1}{1-\underline{x}\xi}-\alpha\right)+\sqrt{\left(\frac{\underline{x}-1}{1-\underline{x}\xi}-\alpha\right)^2-4\alpha}}{2}.\end{aligned} \tag{125}$$

Lemma 45 with the condition (123) replaced with the condition (125) leads to Theorem 15 and Corollary 16. ■

Appendix G. Proof of Lemmas

Here, we give proofs of the lemmas used in Appendices.

G.1 Proof of Lemma 19

Eq.(72) has two positive real solutions:

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

The larger solution (with the plus sign) is decreasing with respect to $c^2_{a_h}c^2_{b_h}$, and lower-bounded as $\sigma^2_{a_h}\sigma^2_{b_h}>\sigma^2/L$. The smaller solution (with the minus sign) is increasing with respect to $c^2_{a_h}c^2_{b_h}$, and upper-bounded as $\sigma^2_{a_h}\sigma^2_{b_h}<\sigma^2/M$.

For $\sigma^2_{a_h}$ and $\sigma^2_{b_h}$ to be positive, Eqs.(69) and (70) require that

$$\sigma^2_{a_h}\sigma^2_{b_h}<\frac{\sigma^2}{M},$$

which is violated by the larger solution, while satisfied by the smaller solution. With the smaller solution (21), Eqs.(69) and (70) give the stationary point given by (20).

Using Eq.(72), we can easily derive Eq.(73) from Eq.(68), which completes the proof of Lemma 19. ■

G.2 Proof of Lemma 20

Since $\widehat{\delta} > 0$, Eqs.(76) and (77) require that

$$\widehat{\gamma}_h < \gamma_h - \frac{M\sigma^2}{\gamma_h}, \tag{126}$$

and therefore, the positive stationary point exists only when

$$\gamma_h > \sqrt{M}\sigma. \tag{127}$$

Below, we assume that Eq.(127) holds.

Eq.(79) has two solutions:

$$\widehat{\gamma}_h = \frac{1}{2}\left(2\gamma_h - \frac{(L+M)\sigma^2}{\gamma_h} \pm \sqrt{\left(\frac{(M-L)\sigma^2}{\gamma_h}\right)^2 + \frac{4\sigma^4}{c_{a_h}^2 c_{b_h}^2}}\right).$$

The larger solution with the plus sign is positive, decreasing with respect to $c_{a_h}^2 c_{b_h}^2$, and lower-bounded as $\widehat{\gamma}_h > \gamma_h - L\sigma^2/\gamma_h$, which violates the condition (126).

The smaller solution, Eq.(17), with the minus sign is positive if the intercept of the left-hand side in Eq.(79) is positive, i.e.,

$$\left(\gamma_h - \frac{L\sigma^2}{\gamma_h}\right)\left(\gamma_h - \frac{M\sigma^2}{\gamma_h}\right) - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} > 0. \tag{128}$$

From the condition (128), we obtain the threshold (16) for the existence of the positive stationary point. Note that $\underline{\gamma}_h^{\mathrm{VB}} > \sqrt{M}\sigma$, and therefore, Eq.(127) holds whenever $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$.

Assume that $\gamma_h > \underline{\gamma}_h^{\mathrm{VB}}$. Then, with the solution (17), $\widehat{\delta}_h$, given by Eq.(76), and $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$, given by Eqs.(74) and (75), are all positive. Thus, we obtain the positive stationary point (18).

Substituting Eqs.(74) and (75), and then Eqs.(76) and (77), into the free energy (68), we have

$$F_h^{\mathrm{VB-Posi}} = -M\log\left(1 - \frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} - \frac{L\sigma^2}{\gamma_h^2}\right) - L\log\left(1 - \frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h} - \frac{M\sigma^2}{\gamma_h^2}\right) + \frac{-2\gamma_h\breve{\gamma}_h^{\mathrm{VB}}}{\sigma^2} + \frac{\gamma_h^2}{\sigma^2} - \left(L + M + \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right). \tag{129}$$

Using Eq.(78), we can eliminate the direct dependency on $c_{a_h}^2 c_{b_h}^2$, and express the free energy (129) as a function of $\breve{\gamma}_h^{\mathrm{VB}}$. This results in Eq.(80), and completes the proof of Lemma 20. ■

G.3 Proof of Lemma 21

By differentiating Eqs.(73), (21), (80), and (17), we have

$$\frac{\partial F_h^{\mathrm{VB-Null}}}{\partial \widehat{\zeta}_h^{\mathrm{VB}}} = \frac{LM}{\sigma^2\left(1-\frac{L}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right)} + \frac{LM}{\sigma^2\left(1-\frac{M}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right)} - \frac{LM}{\sigma^2}$$
$$= \frac{LMc_{a_h}^2c_{b_h}^2\left(1+\frac{\sqrt{LM}}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right)\left(1-\frac{\sqrt{LM}}{\sigma^2}\widehat{\zeta}_h^{\mathrm{VB}}\right)}{\sigma^2\widehat{\zeta}_h^{\mathrm{VB}}}, \tag{130}$$

$$\frac{\partial \widehat{\zeta}_h^{\mathrm{VB}}}{\partial c_{a_h}^2c_{b_h}^2} = \frac{\sigma^2}{2LM}\left(-\frac{\sigma^2}{c_{a_h}^4c_{b_h}^4} + \frac{2\sigma^2\left(L+M+\frac{\sigma^2}{c_{a_h}^2c_{b_h}^2}\right)}{2c_{a_h}^4c_{b_h}^4\sqrt{\left(L+M+\frac{\sigma^2}{c_{a_h}^2c_{b_h}^2}\right)^2-4LM}}\right)$$
$$= \frac{1}{c_{a_h}^4c_{b_h}^4}\left(\frac{(\widehat{\zeta}_h^{\mathrm{VB}})^2}{\left(1-\frac{\sqrt{LM}\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right)\left(1+\frac{\sqrt{LM}\widehat{\zeta}_h^{\mathrm{VB}}}{\sigma^2}\right)}\right), \tag{131}$$

$$\frac{\partial F_h^{\mathrm{VB-Posi}}}{\partial \breve{\gamma}_h^{\mathrm{VB}}} = \frac{M}{\gamma_h\left(1-\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{L\sigma^2}{\gamma_h^2}\right)\right)} + \frac{L}{\gamma_h\left(1-\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{M\sigma^2}{\gamma_h^2}\right)\right)} - \frac{\gamma_h}{\sigma^2}\left(\frac{2\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{(L+M)\sigma^2}{\gamma_h^2}\right)$$
$$= \frac{2c_{a_h}^2c_{b_h}^2\gamma_h^3\left(1-\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{(L+M)\sigma^2}{2\gamma_h^2}\right)\right)\left(\frac{(\breve{\gamma}_h^{\mathrm{VB}})^2}{\gamma_h^2}-\left(1-\frac{(L+M)\sigma^2}{\gamma_h^2}\right)\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{LM\sigma^4}{\gamma_h^4}\right)}{\sigma^6}, \tag{132}$$

$$\frac{\partial \widehat{\gamma}_h}{\partial c_{a_h}^2c_{b_h}^2} = \frac{4\gamma_h^2\sigma^2}{4\gamma_hc_{a_h}^4c_{b_h}^4\sqrt{(M-L)^2+\frac{4\gamma_h^2}{c_{a_h}^2c_{b_h}^2}}}$$
$$= \frac{\sigma^4}{2\gamma_hc_{a_h}^4c_{b_h}^4\left(1-\left(\frac{\breve{\gamma}_h^{\mathrm{VB}}}{\gamma_h}+\frac{(M+L)\sigma^2}{2\gamma_h^2}\right)\right)}. \tag{133}$$

Here, we used Eqs.(21) and (81) to obtain Eqs.(130) and (131), and Eqs.(17) and (82) to obtain Eqs.(132) and (133), respectively. Eq.(85) is obtained by multiplying Eqs.(130) and (131), while Eq.(86) is obtained by multiplying Eqs.(132) and (133).

Taking the difference between the derivatives (85) and (86), and then using Eqs.(82) and (84), we have

$$\frac{\partial(F_h^{\mathrm{Posi}}-F_h^{\mathrm{Null}})}{\partial c_{a_h}^2c_{b_h}^2} = \frac{\partial F_h^{\mathrm{Posi}}}{\partial c_{a_h}^2c_{b_h}^2} - \frac{\partial F_h^{\mathrm{Null}}}{\partial c_{a_h}^2c_{b_h}^2}$$
$$= -\frac{1}{\sigma^2c_{a_h}^2c_{b_h}^2}\left(\gamma_h\left(\gamma_h-\widehat{\gamma}_h\right)-(\underline{\gamma}_h^{\mathrm{VB}})^2\right). \tag{134}$$

The following can be obtained from Eqs.(82) and (83), respectively:

$$\left(\gamma_h(\gamma_h-\breve{\gamma}_h^{\mathrm{VB}})-\frac{(L+M)\sigma^2}{2}\right)^2 = \frac{(L+M)^2\sigma^4}{4}-LM\sigma^4+\frac{\sigma^4}{c_{a_h}^2c_{b_h}^2}\gamma_h^2, \tag{135}$$

$$\left((\underline{\gamma}_h^{\mathrm{VB}})^2 - \frac{(L+M)\sigma^2}{2}\right)^2 = \frac{(L+M)^2\sigma^4}{4} - LM\sigma^4 + \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}(\underline{\gamma}_h^{\mathrm{VB}})^2. \tag{136}$$

Eqs.(135) and (136) imply that

$$\gamma_h(\gamma_h - \breve{\gamma}_h^{\mathrm{VB}}) > (\underline{\gamma}_h^{\mathrm{VB}})^2 \qquad \text{when} \qquad \gamma_h > \underline{\gamma}_h^{\mathrm{VB}}.$$

Therefore, Eq.(134) is negative, which completes the proof of Lemma 21. ■

G.4 Proof of Lemma 26

Lemma 25 immediately leads to the EVB shrinkage estimator (26). We can find the value of $c_{a_h}c_{b_h}$ at the positive EVB local solution by combining the condition (82) for the VB estimator and the condition (92) for the EVB estimator:

$$\left(\gamma_h - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\breve{\gamma}_h^{\mathrm{EVB}} + \frac{M\sigma^2}{\gamma_h}}\right)\left(\gamma_h - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\breve{\gamma}_h^{\mathrm{EVB}} + \frac{L\sigma^2}{\gamma_h}}\right) = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}$$
$$\frac{LM\sigma^4}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}} = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2},$$

which gives the former equation in Eq.(94). Similarly, using Eqs.(19) and (92), we have

$$\widehat{\delta}_h = \frac{c_{a_h}^2}{\sigma^2}\left(\gamma_h - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\breve{\gamma}_h^{\mathrm{EVB}} + \frac{M\sigma^2}{\gamma_h}}\right)$$
$$= \frac{c_{a_h}^2 M}{\gamma_h}\left(1 + \frac{L\sigma^2}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}\right).$$

Using the assumption that $c_{a_h} = c_{b_h}$ and therefore $c_{a_h}^2 = c_{a_h}c_{b_h}$, we obtain the latter equation in Eq.(94). The equations in Eq.(93) are simply obtained from Lemma 20.

Finally, applying Eq.(92) to the free energy (80), we have

$$F_h^{\mathrm{EVB-Posi}} = -M\log\left(1 - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}} + M\sigma^2}\right) - L\log\left(1 - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}} + L\sigma^2}\right) - \frac{\gamma_h\breve{\gamma}_h^{\mathrm{EVB}}}{\sigma^2},$$

which leads to Eq.(95). This completes the proof of Lemma 26. ■

G.5 Proof of Lemma 27

The derivative is

$$\frac{\partial\Phi}{\partial z} = \frac{1 - \frac{1}{z+1} - \log(z+1)}{z^2},$$

which is negative for $z > 0$ because

$$\frac{1}{z+1} + \log(z+1) > 1.$$

This completes the proof of Lemma 27. ■

G.6 Poof of Lemma 28

Since $\Phi(z)$ is decreasing, $\Xi(\tau;\alpha)$ is upper-bounded by

$$\Xi(\tau;\alpha) = \Phi(\tau) + \Phi\left(\frac{\tau}{\alpha}\right) \leq 2\Phi(\tau) = \Xi(\tau;1).$$

Therefore, the unique zero-cross point $\underline{\tau}$ of $\Xi(\tau;\alpha)$ is no greater than the unique zero-cross point $\underline{z}$ of $\Phi(z)$:

$$\underline{\tau} \leq \underline{z}.$$

For obtaining the lower-bound $\underline{\tau} > \sqrt{\alpha}$, it suffices to show that $\Xi(\sqrt{\alpha};\alpha) > 0$. Below, we prove that the following function is decreasing and positive for $0 < \alpha \leq 1$:

$$g(\alpha) \equiv \frac{\Xi(\sqrt{\alpha};\alpha)}{\sqrt{\alpha}}.$$

From the definition (23) of $\Xi(\tau;\alpha)$, we have

$$g(\alpha) = \left(1 + \frac{1}{\alpha}\right)\log(\sqrt{\alpha}+1) - \log\sqrt{\alpha} - \frac{1}{\sqrt{\alpha}}.$$

The derivative is given by

$$\begin{aligned}\frac{\partial g}{\partial\sqrt{\alpha}} &= \frac{\left(1+\frac{1}{\alpha}\right)}{\sqrt{\alpha}+1} - \frac{2}{\alpha^{3/2}}\log(\sqrt{\alpha}+1) - \frac{1}{\sqrt{\alpha}} + \frac{1}{\alpha} \\ &= -\frac{2}{\alpha^{3/2}}\left(\log(\sqrt{\alpha}+1) + \frac{1}{\sqrt{\alpha}+1} - 1\right) \\ &< 0,\end{aligned}$$

which implies that $g(\alpha)$ is decreasing. Since

$$g(1) = 2\log 2 - 1 \approx 0.3863 > 0,$$

$g(\alpha)$ is positive for $0 < \alpha \leq 1$, which completes the proof of Lemma 28. ■

G.7 Proof of Lemma 30

Since

$$\begin{aligned}\frac{\partial\psi_0}{\partial x} &= 1 - \frac{1}{x}, \\ \frac{\partial^2\psi_0}{\partial x^2} &= \frac{1}{x^2} > 0,\end{aligned} \tag{137}$$

$\psi_0(x)$ is differentiable and strictly convex for $x > 0$ with its minimizer at $x = 1$. $\psi_1(x)$ is continuous for $x \geq \underline{x}$, and Eq.(99) implies that $\psi_1(x_h) \propto F_h^{\mathrm{EVB-Posi}}$. Accordingly, $\psi_1(x) \leq 0$ for $x \geq \underline{x}$, where the equality holds when $x = \underline{x}$. This equality implies that $\psi(x)$ is continuous. Since $\underline{x} > 1$, $\psi(x)$ shares the same minimizer with $\psi_0(x)$ at $x = 1$ (see Figure 3).

Hereafter, we investigate $\psi_1(x)$ and $\psi(x)$ for $x \geq \underline{x}$. By differentiating Eqs.(43) and (36), respectively, we have

$$\frac{\partial \psi_1}{\partial \tau} = -\left(\frac{\frac{\tau^2}{\alpha} - 1}{(\tau+1)\left(\frac{\tau}{\alpha}+1\right)} \right) < 0, \tag{138}$$

$$\frac{\partial \tau}{\partial x} = \frac{1}{2}\left(1 + \frac{x-(1+\alpha)}{\sqrt{(x-(1+\alpha))^2 - 4\alpha}} \right) > 0. \tag{139}$$

Substituting

$$x = x(\tau;\alpha) = (1+\tau)\left(1+\frac{\alpha}{\tau}\right) = 1+\alpha+\tau+\alpha\tau^{-1} \tag{35}$$

into Eq.(139), we have

$$\frac{\partial \tau}{\partial x} = \frac{\tau^2}{\alpha\left(\frac{\tau^2}{\alpha}-1\right)}. \tag{140}$$

Multiplying Eqs.(138) and (140) gives

$$\frac{\partial \psi_1}{\partial x} = \frac{\partial \psi_1}{\partial \tau}\frac{\partial \tau}{\partial x} = -\left(\frac{\tau^2}{\alpha(\tau+1)\left(\frac{\tau}{\alpha}+1\right)} \right) = -\frac{\tau}{x} < 0, \tag{141}$$

which implies that $\psi_1(x)$ is decreasing for $x > \underline{x}$.

Let us focus on the thresholding point of $\psi(x)$ at $x = \underline{x}$. Eq.(141) does not converge to zero for $x \to \underline{x} + 0$ but stay negative. On the other hand, $\psi_0(x)$ is differentiable at $x = \underline{x}$. Consequently, $\psi(x)$ has a discontinuously decreasing derivative, i.e., $\lim_{x\to\underline{x}-0} \partial\psi/\partial x > \lim_{x\to\underline{x}+0} \partial\psi/\partial x$, at $x = \underline{x}$.

Finally, we prove the strict quasi-convexity of $\psi(x)$. Taking the sum of Eqs.(137) and (141) gives

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi_0}{\partial x} + \frac{\partial \psi_1}{\partial x} = 1 - \frac{1+\tau}{x} = 1 - \frac{1+\tau}{1+\tau+\alpha+\alpha\tau^{-1}} > 0.$$

This means that $\psi(x)$ is increasing for $x > \underline{x}$. Since $\psi_0(x)$ is strictly convex and increasing at $x = \underline{x}$, and $\psi(x)$ is continuous, $\psi(x)$ is strictly quasi-convex. This completes the proof of Lemma 30. ■

G.8 Proof of Lemma 31

The strict convexity of $\psi_0(x)$ and the strict quasi-convexity of $\psi(x)$ also hold for $\psi_0(\gamma_h^2\sigma^{-2}/M)$ and $\psi(\gamma_h^2\sigma^{-2}/M)$ as functions of σ^{-2} (for $\gamma_h > 0$). Because of the different scale factor γ_h^2/M for each $h = 1, \ldots, L$, each of $\psi_0(\gamma_h^2\sigma^{-2}/M)$ and $\psi(\gamma_h^2\sigma^{-2}/M)$ has a minimizer at a different position:

$$\sigma^{-2} = \frac{M}{\gamma_h^2}.$$

The strict quasi-convexity of ψ_0 and ψ guarantees that $\Omega(\sigma^{-2})$ is decreasing for

$$0 < \sigma^{-2} < \frac{M}{\gamma_1^2},$$

and increasing for

$$\frac{M}{\gamma_L^2} < \sigma^{-2} < \infty.$$

This proves Lemma 31. ■

G.9 Proof of Lemma 34

The derivative of Eq.(40) with respect to σ^{-2} is given by

$$\frac{\partial \Omega}{\partial \sigma^{-2}} = \frac{1}{L}\left(\sum_{h=1}^{H} \frac{\gamma_h^2}{M}\frac{\partial \psi}{\partial x} + \sum_{h=H+1}^{L} \frac{\gamma_h^2}{M}\frac{\partial \psi_0}{\partial x}\right). \tag{142}$$

By using Eqs.(137) and (141), Eq.(142) can be written as

$$\begin{aligned}
\frac{\partial \Omega}{\partial \sigma^{-2}} &= \frac{1}{L}\left(\sum_{h=1}^{L} \frac{\gamma_h^2}{M}\frac{\partial \psi_0}{\partial x} + \sum_{h=1}^{H} \theta\left(x_h \geq \underline{x}\right)\frac{\gamma_h^2}{M}\frac{\partial \psi_1}{\partial x}\right) \\
&= \frac{1}{L}\left(\sum_{h=1}^{L} \frac{\gamma_h^2}{M}\left(1 - \frac{1}{x_h}\right) - \sum_{h=1}^{H} \theta\left(x_h \geq \underline{x}\right)\frac{\gamma_h^2 \tau_h}{M x_h}\right) \\
&= \frac{\sum_{h=1}^{L}\gamma_h^2}{LM} - \sigma^2 - \frac{1}{L}\sum_{h=1}^{H} \theta\left(\tau_h \geq \underline{\tau}\right)\sigma^2 \tau_h.
\end{aligned} \tag{143}$$

Here, we also used the definition (33) of x_h. Using Eq.(34), Eq.(143) can be written as

$$\begin{aligned}
\frac{\partial \Omega}{\partial \sigma^{-2}} &= \frac{\sum_{h=1}^{L}\gamma_h^2}{LM} - \sigma^2 - \sum_{h=1}^{H} \theta\left(\gamma_h \geq \underline{\gamma}^{\mathrm{EVB}}\right)\frac{\gamma_h \breve{\gamma}_h^{\mathrm{EVB}}}{LM} \\
&= -\sigma^2 + \frac{\sum_{h=1}^{H}\gamma_h\left(\gamma_h - \widehat{\gamma}_h^{\mathrm{EVB}}\right) + \sum_{h=H+1}^{L}\gamma_h^2}{LM}.
\end{aligned}$$

Here, we also used the definition (24) of $\widehat{\gamma}_h^{\mathrm{EVB}}$. Using the definition (102) and Lemma 33, we can replace $\widehat{\gamma}_h^{\mathrm{EVB}}$ and H with $\breve{\gamma}_h^{\mathrm{EVB}}$ and $\widehat{H}$, respectively, which completes the proof of Lemma 34. ■

G.10 Proof of Lemma 36

By substituting the lower-bound in Eq.(104) into Eq.(101), we obtain

$$\Theta \geq -\sigma^2 + \frac{\widehat{H}(M+L)\sigma^2 + \sum_{h=\widehat{H}+1}^{L}\gamma_h^2}{LM}.$$

This implies that $\Theta > 0$ unless the following hold:

$$\widehat{H} < \frac{LM}{M+L} = \frac{L}{1+\alpha},$$
$$\sigma^2 \geq \frac{\sum_{h=\widehat{H}+1}^{L} \gamma_h^2}{LM - \widehat{H}(M+L)}.$$

Therefore, no local minimum exists if either of these conditions is violated. This completes the proof of Lemma 36. ■

G.11 Proof of Lemma 37

Focusing on the support

$$\log \underline{y} < t < \log \overline{y}$$

of the LSMP distribution (109), we define

$$\begin{aligned} f(t) \equiv 2 \log p^{\mathrm{LSMP}}(t) &= 2 \log \frac{\sqrt{(e^t - \underline{y})(\overline{y} - e^t)}}{2\pi\alpha} \\ &= \log(-e^{2t} + (\underline{y} + \overline{y})e^t - \underline{y}\overline{y}) + \text{const.}. \end{aligned}$$

Let

$$u(t) \equiv (e^t - \underline{y})(\overline{y} - e^t) = -e^{2t} + (\underline{y} + \overline{y})e^t - \underline{y}\overline{y} > 0,$$

and let

$$\begin{aligned} v(t) \equiv \frac{\partial u}{\partial t} &= -2e^{2t} + (\underline{y} + \overline{y})e^t = u - e^{2t} + \underline{y}\overline{y}, \\ w(t) \equiv \frac{\partial^2 u}{\partial t^2} &= -4e^{2t} + (\underline{y} + \overline{y})e^t = v - 2e^{2t}, \end{aligned}$$

be the first and the second derivatives of u.

Therefore, the first and the second derivatives of $f(t)$ are given by

$$\begin{aligned} \frac{\partial f}{\partial t} &= \frac{v}{u}, \\ \frac{\partial^2 f}{\partial t^2} &= \frac{uw - v^2}{u^2} \\ &= -\frac{e^t\left((\underline{y} + \overline{y})e^{2t} - 4\underline{y}\overline{y}e^t + (\underline{y} + \overline{y})\underline{y}\overline{y}\right)}{u^2} \\ &= -\frac{e^t(\underline{y} + \overline{y})}{u^2}\left(\left(e^t - \frac{2\underline{y}\overline{y}}{(\underline{y} + \overline{y})}\right)^2 + \frac{\underline{y}\overline{y}\left(\overline{y} - \underline{y}\right)^2}{(\underline{y} + \overline{y})^2}\right) \\ &\leq 0. \end{aligned}$$

This proves the log-concavity of the LSMP distribution $p^{\mathrm{LSMP}}(t)$, and completes the proof of Lemma 37. ■

G.12 Proof of Lemma 42

Lemma 41 states that $\Omega_0(\sigma^{-2})$, defined by Eq.(116), is quasi-convex with its minimizer at

$$\sigma^{-2} = \left(\frac{\sum_{h=H^*+1}^{L} \gamma_h^2}{(L-H^*)M} \right)^{-1} = \sigma^{*-2}.$$

Since $\Omega_1(\sigma^{-2})$, defined by Eq.(115), is a sum of strictly quasi-convex functions with their minimizers at $\sigma^{-2} = M/\gamma_h^2 < \sigma^{*-2}$ for $h = 1, \ldots, H^*$, our objective $\Omega(\sigma^{-2})$, given by Eq.(114), is non-decreasing (increasing if $H^* > 0$) for

$$\sigma^{-2} \geq \sigma^{*-2}.$$

Since Eq.(118) implies that $\underline{\sigma}_{H^*+1}^{-2} > \sigma^{*-2}$, $\Omega(\sigma^{-2})$ is non-decreasing (increasing if $\xi > 0$) for $\sigma^{-2} > \underline{\sigma}_{H^*+1}^{-2}$, which completes the proof of Lemma 42. ■

G.13 Proof of Lemma 43

Lemma 41 states that $\Omega_0(\sigma^{-2})$ is strictly convex in the range $0 < \sigma^{-2} < \underline{\sigma}_{H^*+1}^2$, and minimized at $\sigma^{-2} = \sigma^{*-2}$. Since Eq.(118) implies that $\sigma^{*-2} < \underline{\sigma}_{H^*+1}^2$, $\Omega_0(\sigma^{-2})$ is increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$. Since $\Omega_1(\sigma^{-2})$ is a sum of strictly quasi-convex functions with their minimizers at $\sigma^{-2} = M/\gamma_h^2 < \sigma^{*-2}$ for $h = 1, \ldots, H^*$, $\Omega(\sigma^{-2})$ is also increasing at $\sigma^{-2} = \underline{\sigma}_{H^*+1}^2 - 0$.

Let us investigate the sign of the derivative Θ of $\Omega(\sigma^{-2})$ at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0 \in \mathbb{B}_{H^*}$. Substituting the upper-bound in Eq.(104) into Eq.(101), we have

$$\begin{aligned} \Theta &< -\sigma^2 + \frac{H^*(\sqrt{M}+\sqrt{L})^2\sigma^2 + \sum_{h=H^*+1}^{L} \gamma_h^2}{LM} \\ &= -\sigma^2 + \frac{H^*(\sqrt{M}+\sqrt{L})^2\sigma^2 + (L-H^*)M\sigma^{*2}}{LM}. \end{aligned} \tag{144}$$

The right-hand side of Eq.(144) is negative if the following hold:

$$\xi = \frac{H^*}{L} < \frac{M}{(\sqrt{M}+\sqrt{L})^2} = \frac{1}{(1+\sqrt{\alpha})^2}, \tag{145}$$

$$\sigma^2 > \frac{(L-H^*)M\sigma^{*2}}{LM - H^*(\sqrt{M}+\sqrt{L})^2} = \frac{(1-\xi)\sigma^{*2}}{1-\xi(1+\sqrt{\alpha})^2}. \tag{146}$$

Assume that the first condition (145) holds. Then, the second condition (146) holds at $\sigma^{-2} = \underline{\sigma}_{H^*}^2 + 0$, if

$$\underline{\sigma}_{H^*}^{-2} < \frac{1-\xi(1+\sqrt{\alpha})^2}{(1-\xi)}\sigma^{*-2},$$

or equivalently,

$$y_{H^*} = \frac{\gamma_{H^*}^2}{M\sigma^{*2}} = \underline{x} \cdot \frac{\underline{\sigma}_{H^*}^2}{\sigma^{*2}} > \frac{\underline{x}(1-\xi)}{1-\xi(1+\sqrt{\alpha})^2},$$

which completes the proof of Lemma 43. ■

G.14 Proof of Lemma 44

In the range $0 < \sigma^{-2} < \underline{\sigma}^2_{H^*}$, the estimated rank (102) is bounded as

$$0 \leq \widehat{H} \leq H^* - 1.$$

Substituting the upper-bound in Eq.(104) into Eq.(101), we have

$$\begin{aligned}\Theta &< -\sigma^2 + \frac{\widehat{H}(\sqrt{M}+\sqrt{L})^2\sigma^2 + \sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2 + \sum_{h=H^*+1}^{L}\gamma_h^2}{LM} \\ &= -\sigma^2 + \frac{\widehat{H}(\sqrt{M}+\sqrt{L})^2\sigma^2 + \sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2 + (L-H^*)M\sigma^{*2}}{LM}. \end{aligned} \tag{147}$$

The right-hand side of Eq.(147) is negative, if the following hold:

$$\frac{\widehat{H}}{L} < \frac{M}{(\sqrt{M}+\sqrt{L})^2} = \frac{1}{(1+\sqrt{\alpha})^2}, \tag{148}$$

$$\sigma^2 > \frac{\sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2 + (L-H^*)M\sigma^{*2}}{LM - \widehat{H}(\sqrt{M}+\sqrt{L})^2}. \tag{149}$$

Assume that

$$\xi = \frac{H^*}{L} < \frac{1}{(1+\sqrt{\alpha})^2}.$$

Then, both of the conditions (148) and (149) hold anywhere in $0 < \sigma^{-2} < \underline{\sigma}^2_{H^*}$, if the following holds

$$\underline{\sigma}^{-2}_{\widehat{H}+1} < \frac{LM - \widehat{H}(\sqrt{M}+\sqrt{L})^2}{\sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2 + (L-H^*)M\sigma^{*2}} \qquad \text{for} \qquad \widehat{H} = 0, \ldots, H^*-1. \tag{150}$$

Since the sum $\sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2$ in the right-hand side of Eq.(150) is upper-bounded as

$$\sum_{h=\widehat{H}+1}^{H^*}\gamma_h^2 \leq (H^* - \widehat{H})\gamma^2_{\widehat{H}+1},$$

Eq.(150) holds if

$$\begin{aligned}\underline{\sigma}^{-2}_{\widehat{H}+1} &< \frac{LM - \widehat{H}(\sqrt{M}+\sqrt{L})^2}{(H^*-\widehat{H})\gamma^2_{\widehat{H}+1} + (L-H^*)M\sigma^{*2}} \\ &= \frac{1 - \frac{\widehat{H}}{L}(1+\sqrt{\alpha})^2}{(\xi - \frac{\widehat{H}}{L})\frac{\gamma^2_{\widehat{H}+1}}{M} + (1-\xi)\sigma^{*2}} \qquad \text{for} \qquad \widehat{H} = 0, \ldots, H^*-1. \end{aligned} \tag{151}$$

Using Eq.(100), the condition (151) is rewritten as

$$\frac{\gamma^2_{\widehat{H}+1}}{M\underline{x}} > \frac{(\xi - \frac{\widehat{H}}{L})\frac{\gamma^2_{\widehat{H}+1}}{M} + (1-\xi)\sigma^{*2}}{1 - \frac{\widehat{H}}{L}(1+\sqrt{\alpha})^2}$$

$$\left(1-\frac{\widehat{H}}{L}(1+\sqrt{\alpha})^2\right)\frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}} > (\xi\underline{x}-\frac{\widehat{H}}{L}\underline{x})\frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}}+(1-\xi)\underline{x},$$

or equivalently

$$y_{\widehat{H}+1}=\frac{\gamma_{\widehat{H}+1}^2}{M\sigma^{*2}} > \frac{(1-\xi)\,\underline{x}}{\left(1-\xi\underline{x}+\frac{\widehat{H}}{L}\,(\underline{x}-(1+\sqrt{\alpha})^2)\right)} \qquad \text{for} \qquad \widehat{H}=0,\ldots,H^*-1. \quad (152)$$

Note that $\underline{x} > \overline{y} = (1+\sqrt{\alpha})^2$. Further bounding both sides, we have the following sufficient condition for Eq.(152) to hold:

$$y_{H^*} > \frac{(1-\xi)\underline{x}}{\max\,(0,1-\xi\underline{x})}.$$

Thus, we obtain the conditions (121) and (122) for Θ to be negative anywhere in $0<\sigma^{-2}<\underline{\sigma}_{H^*}^2$, which completes the proof of Lemma 44. ■

Appendix H. Detailed Description of Overlap Method

The overlap (OL) method (Hoyle, 2008) minimizes the following approximation to the negative log of the marginal likelihood (58) over the hypothetical model rank $H=1,\ldots,L$:[7]

$$\begin{aligned}
2F^{\mathrm{OL}} &\approx -2\log p(\boldsymbol{V}) \\
&= (LM-H(L-H-2))\log(2\pi)+L\log\pi-2\sum_{h=1}^{H}\log\left(\frac{\Gamma\left((M-h+1)/2\right)}{\Gamma\left((M-L-h+1)/2\right)}\right) \\
&\quad + H(M-L)\left(1-\log\left(M-L\right)\right)+\sum_{h=1}^{H}\sum_{l=H+1}^{L}\log\left(\gamma_h^2-\gamma_l^2\right)+(M-L)\sum_{h=1}^{H}\log\gamma_h^2 \\
&\quad + (M-H)\sum_{h=1}^{H}\log\left(\frac{1}{\widehat{\sigma}^{2\ \mathrm{OL}}}-\frac{1}{\widehat{\lambda}_h^{\mathrm{OL}}}\right)-\sum_{h=1}^{H}\left(\frac{1}{\widehat{\sigma}^{2\ \mathrm{OL}}}-\frac{1}{\widehat{\lambda}_h^{\mathrm{OL}}}\right)\gamma_h^2 \\
&\quad + (L+2)\left(\sum_{h=1}^{H}\log\widehat{\lambda}_h^{\mathrm{OL}}+(M-H)\log\widehat{\sigma}^{2\ \mathrm{OL}}\right)+\sum_{l=1}^{L}\frac{\gamma_l^2}{\widehat{\sigma}^{2\ \mathrm{OL}}},
\end{aligned}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\{\widehat{\lambda}_h^{\mathrm{OL}}\}$ and $\widehat{\sigma}^{2\ \mathrm{OL}}$ are estimators for $\lambda_h = b_h^2+\sigma^2$ and σ^2, computed by iterating the following equations until convergence:

$$\begin{aligned}
\widehat{\lambda}_h^{\mathrm{OL}} = \frac{\gamma_h^2}{2(L+2)}\Bigg(1-\frac{(M-H-(L+2))\widehat{\sigma}^{2\ \mathrm{OL}}}{\gamma_h^2} \\
+\sqrt{\left(1-\frac{(M-H-(L+2))\widehat{\sigma}^{2\ \mathrm{OL}}}{\gamma_h^2}\right)^2-\frac{4(L+2)\widehat{\sigma}^{2\ \mathrm{OL}}}{\gamma_h^2}}\Bigg),
\end{aligned} \quad (153)$$

[7] Our description is slightly different from Hoyle (2008), because our model (1) does not have the mean parameter shared over the samples.

$$\widehat{\sigma}^{2\,\mathrm{OL}} = \frac{1}{(M-H)}\left(\sum_{l=1}^{L}\frac{\gamma_l^2}{L} - \sum_{h=1}^{H}\widehat{\lambda}_h^{\mathrm{OL}}\right). \tag{154}$$

When iterating Eqs.(153) and (154), $\widehat{\lambda}_h^{\mathrm{OL}}$ can be a complex number. In such a case, the hypothetical H is rejected. Otherwise, F^{OL} is evaluated after convergence, and $\widehat{H}^{\mathrm{OL}}$ that minimizes F^{OL} is chosen.

For the null hypothesis, the negative log likelihood is given by

$$2F^{\mathrm{OL}} = -2\log P(\boldsymbol{V}) = LM\left(\log\left(\frac{2\pi}{LM}\sum_{l=1}^{L}\gamma_l^2\right)+1\right) \qquad \text{for} \qquad H=0.$$

References

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

Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

C. M. Bishop. Variational principal components. In *Proceedings of International Conference on Artificial Neural Networks*, volume 1, pages 514–509, 1999a.

C. M. Bishop. Bayesian principal components. In *Advances in Neural Information Processing Systems*, volume 11, pages 382–388, 1999b.

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

C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *Proceedings of the Sixteenth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 46–53, 2000.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

J. P. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing—From Statistical Physics to Risk Management, Second Edition*. Cambridge University Press, 2003.

E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions Information Theory*, 52(12):5406–5425, 2006.

S. Dharmadhikari and K. Joag-dev. *Unimodality, Convexity, and Applications*. Academic Press, 1988.

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

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

D. C. Hoyle and M. Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(026124), 2004.

I. A. Ibragimov. On the composition of unimodal distributions. *Theory of Probability and Its Applications*, 1(2):255–260, 1956.

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.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327, 2001.

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

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

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.

M. L. Mehta. *Random Matrices, Third Edition*. Academic Press, 2000.

T. P. Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems*, volume 13, pages 598–604. MIT Press, 2001.

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

S. Nakajima and M. Sugiyama. Analysis of empirical MAP and empirical partially Bayes: Can they be alternatives to variational Bayes? In *Proceedings of International Conference on Artificial Intelligence and Statistics*, volume 33, pages 20–28, 2014.

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, and S. D. Babacan. Variational Bayesian sparse additive matrix factorization. *Machine Learning*, 92:319–1347, 2013a.

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, 2013b.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Reveiw*, 52(3):471–501, 2010.

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.

M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian estimation for MEG inverse problem. *Neuro Image*, 23:806–826, 2004.

M. Seeger. Sparse linear models: Variational approximate inference and Bayesian experimental design. In *Journal of Physics: Conference Series*, volume 197, 2009.

M. Seeger and G. Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, La Palma, Spain, 2012.

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

A. M. Tulino and S. Verdu. *Random Matrix Theory and Wireless Communications*. Now Publishers, 2004.

K. W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability*, 6:1–18, 1978.

E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1957.