Journal of Machine Learning Research 12 (2011) 2583-2648 Submitted 5/10; Revised 2/11; Published 9/11

Theoretical Analysis of Bayesian Matrix Factorization*

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

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

Editor: Inderjit Dhillon

Abstract

Recently, *variational Bayesian* (VB) techniques have been applied to probabilistic matrix factorization and shown to perform very well in experiments. In this paper, we theoretically elucidate properties of the VB matrix factorization (VBMF) method. Through finite-sample analysis of the VBMF estimator, we show that two types of shrinkage factors exist in the VBMF estimator: the *positive-part James-Stein (PJS)* shrinkage and the *trace-norm* shrinkage, both acting on each singular component separately for producing low-rank solutions. The trace-norm shrinkage is simply induced by non-flat prior information, similarly to the maximum a posteriori (MAP) approach. Thus, no trace-norm shrinkage remains when priors are non-informative. On the other hand, we show a counter-intuitive fact that the PJS shrinkage factor is kept activated even with flat priors. This is shown to be induced by the *non-identifiability* of the matrix factorization model, that is, the mapping between the target matrix and factorized matrices is not one-to-one. We call this *model-induced regularization*. We further extend our analysis to empirical Bayes scenarios where hyperparameters are also learned based on the VB free energy. Throughout the paper, we assume no missing entry in the observed matrix, and therefore collaborative filtering is out of scope.

Keywords: matrix factorization, variational Bayes, empirical Bayes, positive-part James-Stein shrinkage, non-identifiable model, model-induced regularization

1. Introduction

The goal of *matrix factorization* (MF) is to find a low-rank expression of a target matrix. MF can be used for learning linear relation between vectors such as *reduced rank regression* (Baldi and Hornik, 1995; Reinsel and Velu, 1998), *canonical correlation analysis* (Hotelling, 1936; Anderson, 1984), *partial least-squares* (Wold, 1966; Worsley et al., 1997; Rosipal and Krämer, 2006), and *multi-task learning* (Chapelle and Harchaoui, 2005; Yu et al., 2005). More recently, MF is applied to *collaborative filtering* for imputing missing entries of a target matrix, for example, in the context of *recommender systems* (Konstan et al., 1997; Funk, 2006) and *microarray data analysis* (Baldi and Brunak, 1998). For these reasons, MF has attracted considerable attention these days.

*. This paper is an extended version of our earlier conference paper (Nakajima and Sugiyama, 2010).

1.1 MF Methods

Srebro and Jaakkola (2003) proposed the *weighted low-rank approximation* method, which is based on the *expectation-maximization* (EM) algorithm: a matrix is fitted to the data without a rank constraint in the E-step and it is projected back to the set of low-rank matrices by *singular value decomposition* (SVD) in the M-step. Since the optimization problem of the weighted low-rank approximation method involves a low-rank constraint, it is non-convex and thus only a local optimal solution may be obtained. Furthermore, SVD of the target matrix needs to be carried out in each iteration, which may be computationally intractable for large-scale data.

Funk (2006) proposed the *regularized SVD* method that minimizes a goodness-of-fit term combined with the *Frobenius-norm* penalty under a low-rank constraint by gradient descent (see also Paterek, 2007). The regularized SVD method could be computationally more efficient than the weighted low-rank approximation method in the context of collaborative filtering since only observed entries are referred to in each gradient iteration.

Srebro et al. (2005) proposed to use the *trace-norm* penalty instead of the Frobenius-norm penalty, so that a low-rank solution can be obtained without having an explicit low-rank constraint. Thanks to the convexity of the *trace-norm*, a semi-definite programming formulation can be obtained when the *hinge-loss* (Schölkopf and Smola, 2002) is used. See also Rennie and Srebro (2005) for a computationally efficient variant using a gradient-based optimization method with smooth approximation.

Salakhutdinov and Mnih (2008) proposed a Bayesian *maximum a posteriori* (MAP) method based on the Gaussian noise model and Gaussian priors on the decomposed matrices. This method actually corresponds to minimizing the squared-loss with the trace-norm penalty (Srebro et al., 2005).

Recently, the *variational Bayesian* (VB) approach (Attias, 1999) has been applied to MF (Lim and Teh, 2007; Raiko et al., 2007), which we refer to as *VBMF*. The VBMF method was shown to perform very well in experiments. However, its good performance was not completely understood beyond its experimental success. The purpose of this paper is to provide new insight into Bayesian MF.

1.2 MF Models and Non-identifiability

The MF models can be regarded as re-parameterization of the target matrix using low-rank matrices. This kind of re-parameterization often significantly changes the statistical behavior of the estimator (Gelman, 2004). Indeed, MF models possess a special structure called *non-identifiability* (Watanabe, 2009), meaning that the mapping between the target matrix and the factorized matrices is not one-to-one .

Previous theoretical studies on non-identifiable models investigated the behavior of *multi-layer pereptrons*, *Gaussian mixture models*, and *hidden Markov models*. It was shown that when such non-identifiable models are trained using *full-Baysian* (FB) estimation, the regularization effect is significantly stronger than the MAP method (Watanabe, 2001; Yamazaki and Watanabe, 2003). Since a single point in the function space corresponds to a set of points in the (redundant) parameter space in non-identifiable models, simple distributions such as the Gaussian distribution in the function space produce highly complicated *multimodal* distributions in the parameter space. This causes the MAP and FB solutions to be significantly different. Thus the behavior of non-identifiable models is substantially different from that of identifiable models. For Gaussian mixture models and

reduced rank regression models, theoretical properties of VB have also been investigated (Watanabe and Watanabe, 2006; Nakajima and Watanabe, 2007).

1.3 Our Contribution

In this paper, following the line of Nakajima and Watanabe (2007) which investigated asymptotic behavior of VBMF estimators and the generalization error, we provide a more precise analysis of VB estimators. More specifically, we derive *non-asymptotic* bounds of the VBMF estimator. The obtained solution can be seen as a re-weighted singular value decomposition, and the weights include a factor induced by the *Bayesian* inference procedure, in the same way as *automatic relevance determination* (Neal, 1996; Wipf and Nagarajan, 2008).

We show that VBMF consists of two shrinkage factors, the *positive-part James-Stein* (PJS) shrinkage (James and Stein, 1961; Efron and Morris, 1973) and the *trace-norm* shrinkage (Srebro et al., 2005), operating on each singular component separately for producing low-rank solutions.

The trace-norm shrinkage is simply induced by non-flat prior information, as in the MAP approach (Salakhutdinov and Mnih, 2008). Thus, no trace-norm shrinkage remains when priors are non-informative. On the other hand, we show a counter-intuitive fact that the PJS shrinkage factor is still kept activated even with uniform priors. This allows the VBMF method to avoid overfitting (or in some cases, this may cause underfitting) even when non-informative priors are provided. We call this regularization effect *model-induced regularization* since it is caused by the structure of the model likelihood function.

We further extend the above analysis to *empirical VBMF* (EVBMF) scenarios, where hyperparameters in prior distributions are also learned based on the *VB free energy*. We derive bounds of the EVBMF estimator, and show that the effect of PJS shrinkage is at least doubled compared with the uniform prior cases.

Finally, we note that our analysis relies on the following three assumptions: First, we assume that the given matrix is *fully* observed, and no missing entry exists. This means that missing entry prediction is out of scope of our theory. Second, we require the noise to be independent Gaussian noise and the priors to be isotropic Gaussian. Third, we assume the column-wise independence on the VB posterior, which is different from the standard VB assumption that only the matrix-wise independence is required.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, we formulate the MF problem and review its Bayesian approaches including FB, MAP, VB methods, and their empirical variants. In Section 3, we analyze the behavior of MAPMF, VBMF, and their empirical variants, and elucidate the regularization mechanism. In Section 4, we illustrate the characteristic behavior of MF solutions through simple numerical experiments, highlighting the influence of non-identifiability of the MF models. Finally, we conclude in Section 5. A brief review of the James-Stein shrinkage estimator and all the technical details are provided in Appendix.

2. Bayesian Approaches to Matrix Factorization

In this section, we give a probabilistic formulation of the *matrix factorization* (MF) problem and review its Bayesian methods.

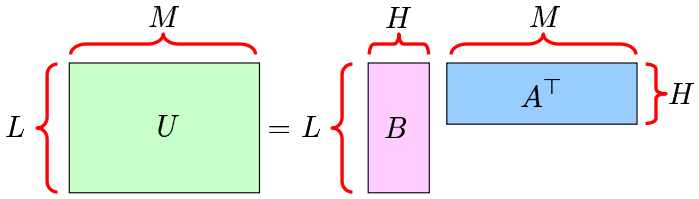

Figure 1: Matrix factorization model.

2.1 Formulation

The goal of the MF problem is to estimate a target matrix U ($\in \mathbb{R}^{L\times M}$) from its observation

$$V \in \mathbb{R}^{L\times M}.$$

Throughout the paper, we assume that

$$L \leq M.$$

If $L > M$, we may simply re-define the transpose $U^\top$ as U so that $L \leq M$ holds. Thus this does not impose any restriction.

A key assumption of MF is that U is a low-rank matrix. Let H ($\leq L$) be the rank of U. Then the matrix U can be decomposed into the product of $A \in \mathbb{R}^{M\times H}$ and $B \in \mathbb{R}^{L\times H}$ as follows (see Figure 1):

$$U = BA^\top.$$

With appropriate *pre-whitening* (Hyvärinen et al., 2001), *reduced rank regression* (Baldi and Hornik, 1995; Reinsel and Velu, 1998), *canonical correlation analysis* (Hotelling, 1936; Anderson, 1984), *partial least-squares* (Wold, 1966; Worsley et al., 1997; Rosipal and Krämer, 2006), and *multi-task learning* (Chapelle and Harchaoui, 2005; Yu et al., 2005) can be seen as special cases of the MF problem. *Collaborative filtering* (Konstan et al., 1997; Baldi and Brunak, 1998; Funk, 2006) and *image processing* (Lee and Seung, 1999) would be popular applications of MF. Note that, some of these applications such as *collaborative filtering* and *multi-task learning* with unshared input sets are out of scope of our theory, since they require missing entry prediction.

Assume that the observed matrix V is subject to the following additive-noise model:

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

where $\mathcal{E}$ ($\in \mathbb{R}^{L\times M}$) is a noise matrix. Each entry of $\mathcal{E}$ is assumed to independently follow the Gaussian distribution with mean zero and variance σ^2. Then, the likelihood $p(V|A,B)$ is given by

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

where $\|\cdot\|_{\mathrm{Fro}}$ denotes the *Frobenius norm* of a matrix.

2.2 Full-Bayesian Matrix Factorization (FBMF) and Its Empirical Variant (EFBMF)

We use the Gaussian priors on the parameters A and B:

$$\phi(U) = \phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B),$$

where

$$\phi_{\mathrm{A}}(A) \propto \exp\left(-\sum_{h=1}^{H} \frac{\|\boldsymbol{a}_h\|^2}{2c_{a_h}^2}\right) = \exp\left(-\frac{\mathrm{tr}(AC_A^{-1}A^\top)}{2}\right), \tag{2}$$

$$\phi_{\mathrm{B}}(B) \propto \exp\left(-\sum_{h=1}^{H} \frac{\|\boldsymbol{b}_h\|^2}{2c_{b_h}^2}\right) = \exp\left(-\frac{\mathrm{tr}(BC_B^{-1}B^\top)}{2}\right). \tag{3}$$

Here, $\boldsymbol{a}_h$ and $\boldsymbol{b}_h$ are the h-th column vectors of A and B, respectively, that is,

$$A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_H),$$
$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_H).$$

$c_{a_h}^2$ and $c_{b_h}^2$ are hyperparameters corresponding to the prior variances of those vectors. Without loss of generality, we assume that the product $c_{a_h}c_{b_h}$ is non-increasing with respect to h. We also denote them as covariance matrices:

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

where $\mathrm{diag}(\boldsymbol{c})$ denotes the diagonal matrix with its entries specified by vector $\boldsymbol{c}$. $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

With the Bayes theorem and the definition of marginal distributions, the *Bayes posterior* $p(A,B|V)$ can be written as

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

where $\langle\cdot\rangle_p$ denotes the expectation over p. The *full-Bayesian* (FB) solution is given by the *Bayes posterior mean*:

$$\widehat{U}^{\mathrm{FB}} = \langle BA^\top\rangle_{p(A,B|V)}. \tag{5}$$

We call this method *FBMF*.

The hyperparameters c_{a_h} and c_{b_h} may be determined so that the *Bayes free energy* $F(V)$ is minimized.

$$\begin{aligned} F(V) &= -\log p(V) \\ &= -\log\langle p(V|A,B)\rangle_{\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)}. \end{aligned} \tag{6}$$

We call this method the *empirical full-Bayesian MF* (EFBMF). The Bayes free energy is also referred to as the *marginal log-likelihood* (MacKay, 2003), the *evidence* (MacKay, 1992) or the *stochastic complexity* (Rissanen, 1986).

2.3 Maximum A Posteriori Matrix Factorization (MAPMF) and Its Empirical Variant (EMAPMF)

When computing the Bayes posterior (4), the expectation in the denominator of Equation (4) is often intractable due to high dimensionality of the parameters A and B. More importantly, computing the posterior mean (5) is also intractable. A simple approach to mitigating this problem is to use the *maximum a posteriori* (MAP) approximation, which we refer to as MAPMF. The MAP solution $\widehat{U}^{\mathrm{MAP}}$ is given by

$$\widehat{U}^{\mathrm{MAP}} = \widehat{B}^{\mathrm{MAP}}(\widehat{A}^{\mathrm{MAP}})^{\top},$$

where

$$(\widehat{A}^{\mathrm{MAP}}, \widehat{B}^{\mathrm{MAP}}) = \operatorname*{argmax}_{A,B} p(A,B|V).$$

In the MAP framework, one may determine the hyperparameters c_{a_h} and c_{b_h} so that the Bayes posterior $p(A,B|V)$ is maximized (equivalently, the negative log posterior is minimized). We call this method *empirical MAPMF* (EMAPMF). Note that EMAPMF does not work properly, as explained in Section 3.3.

2.4 Variational Bayesian Matrix Factorization (VBMF) and Its Empirical Variant (EVBMF)

Another approach to avoiding computational intractability of the FB method is to use the *variational Bayes* (VB) approximation (Attias, 1999; Bishop, 2006). Here, we review the VB-based MF method (Lim and Teh, 2007; Raiko et al., 2007).

Let $r(A,B|V)$ be a *trial* distribution for A and B, and we define the following functional F_{VB} called the *VB free energy* with respect to $r(A,B|V)$:

$$F_{\mathrm{VB}}(r|V) = \left\langle \log \frac{r(A,B|V)}{p(V,A,B)} \right\rangle_{r(A,B|V)}. \tag{7}$$

Using $p(V,A,B) = p(A,B|V)p(V)$, we can decompose Equation (7) into two terms:

$$F_{\mathrm{VB}}(r|V) = \left\langle \log \frac{r(A,B|V)}{p(A,B|V)} \right\rangle_{r(A,B|V)} + F(V), \tag{8}$$

where $F(V)$ is the Bayes free energy defined by Equation (6). The first term in Equation (8) is the *Kullback-Leibler divergence* (Kullback and Leibler, 1951) from $r(A,B|V)$ to the Bayes posterior $p(A,B|V)$. This is non-negative and vanishes if and only if the two distributions agree with each other. Therefore, the VB free energy $F_{\mathrm{VB}}(r|V)$ is lower-bounded by the Bayes free energy $F(V)$:

$$F_{\mathrm{VB}}(r|V) \geq F(V),$$

where the equality is satisfied if and only if $r(A,B|V)$ agrees with $p(A,B|V)$.

The VB approach minimizes the VB free energy $F_{\mathrm{VB}}(r|V)$ with respect to the trial distribution $r(A,B|V)$, by restricting the search space of $r(A,B|V)$ so that the minimization is computationally tractable. Typically, dissolution of probabilistic dependency between entangled parameters (A and B in the case of MF) makes the calculation feasible:

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

Then, the VB free energy (7) is written as

$$F_{\mathrm{VB}}(r|V) = \left\langle \log \frac{r_{\mathrm{A}}(A|V) r_{\mathrm{B}}(B|V)}{p(V|A,B)\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)} \right\rangle_{r_{\mathrm{A}}(A|V) r_{\mathrm{B}}(B|V)}. \tag{10}$$

The resulting distribution is called the *VB posterior*. The VB solution $\widehat{U}^{\mathrm{VB}}$ is given by the *VB posterior mean*:

$$\widehat{U}^{\mathrm{VB}} = \langle BA^{\top} \rangle_{r(A,B|V)}. \tag{11}$$

We call this method *VBMF*.

Applying the variational method to the VB free energy shows that the VB posterior satisfies the following conditions:

$$r_{\mathrm{A}}(A|V) \propto \phi_{\mathrm{A}}(A) \exp\left(\langle \log p(V|A,B) \rangle_{r_{\mathrm{B}}(B|V)} \right), \tag{12}$$

$$r_{\mathrm{B}}(B|V) \propto \phi_{\mathrm{B}}(B) \exp\left(\langle \log p(V|A,B) \rangle_{r_{\mathrm{A}}(A|V)} \right). \tag{13}$$

Recall that we are using the Gaussian priors (2) and (3). Also, Equation (1) implies that the log-likelihood $\log p(V|A,B)$ is a quadratic function of A when B is fixed, and vice versa. Then the conditions (12) and (13) imply that the VB posteriors $r_{\mathrm{A}}(A|V)$ and $r_{\mathrm{B}}(B|V)$ are also Gaussian. This enables one to derive a computationally efficient algorithm called the *iterated conditional modes* (Besag, 1986; Bishop, 2006), where the mean and the covariance of the parameters A and B are iteratively updated using Equations (12) and (13) (Lim and Teh, 2007; Raiko et al., 2007). This amounts to alternating between minimizing the free energy (10) with respect to $r_{\mathrm{A}}(A|V)$ and $r_{\mathrm{B}}(B|V)$.

As in Raiko et al. (2007), we assume in our theoretical analysis that the trial distribution $r(A,B|V)$ can be further factorized as

$$r(A,B|V) = \prod_{h=1}^{H} r_{\mathrm{a}_h}(\boldsymbol{a}_h|V) r_{\mathrm{b}_h}(\boldsymbol{b}_h|V). \tag{14}$$

Then the update rules (12) and (13) are simplified as

$$r_{\mathrm{a}_h}(\boldsymbol{a}_h|V) \propto \phi_{\mathrm{a}_h}(\boldsymbol{a}_h) \exp\left(\langle \log p(V|A,B) \rangle_{r_{\backslash \mathrm{a}_h}(A \backslash \boldsymbol{a}_h, B|V)} \right), \tag{15}$$

$$r_{\mathrm{b}_h}(\boldsymbol{b}_h|V) \propto \phi_{\mathrm{b}_h}(\boldsymbol{b}_h) \exp\left(\langle \log p(V|A,B) \rangle_{r_{\backslash \mathrm{b}_h}(A, B \backslash \boldsymbol{b}_h|V)} \right), \tag{16}$$

where $r_{\backslash \mathrm{a}_h}$ and $r_{\backslash \mathrm{b}_h}$ denote the VB posterior of the parameters A and B except $\boldsymbol{a}_h$ and $\boldsymbol{b}_h$, respectively.

The VB free energy also allows us to determine the hyperparameters $c_{a_h}^2$ and $c_{b_h}^2$ in a computationally tractable way. That is, instead of the Bayes free energy $F(V)$, the VB free energy $F_{\mathrm{VB}}(r|V)$ is minimized with respect to $c_{a_h}^2$ and $c_{b_h}^2$. We call this method *empirical VBMF* (EVBMF).

3. Analysis of Bayesian MF Methods

In this section, we theoretically analyze the behavior of MAPMF, VBMF, EMAPMF, and EVBMF solutions, and elucidate their regularization mechanism.

3.1 MAPMF

The MAP estimator $(\widehat{A}^{\mathrm{MAP}}, \widehat{B}^{\mathrm{MAP}})$ is the maximizer of the Bayes posterior. In our model (1), (2), and (3), the negative log of the Bayes posterior is expressed as

$$-\log p(A,B|V) = \frac{LM\log\sigma^2}{2} + \frac{1}{2}\sum_{h=1}^{H}\left(M\log c_{a_h}^2 + L\log c_{b_h}^2 + \frac{\|\boldsymbol{a}_h\|^2}{c_{a_h}^2} + \frac{\|\boldsymbol{b}_h\|^2}{c_{b_h}^2}\right) + \frac{1}{2\sigma^2}\left\|V - \sum_{h=1}^{H}\boldsymbol{b}_h\boldsymbol{a}_h^\top\right\|_{\mathrm{Fro}}^2 + \mathrm{Const.} \tag{17}$$

Differentiating Equation (17) with respect to A and B and setting the derivatives to zero, we have the following conditions:

$$\boldsymbol{a}_h = \left(\|\boldsymbol{b}_h\|^2 + \frac{\sigma^2}{c_{a_h}^2}\right)^{-1}\left(V - \sum_{h'\neq h}\boldsymbol{b}_{h'}\boldsymbol{a}_{h'}^\top\right)^\top\boldsymbol{b}_h, \tag{18}$$

$$\boldsymbol{b}_h = \left(\|\boldsymbol{a}_h\|^2 + \frac{\sigma^2}{c_{b_h}^2}\right)^{-1}\left(V - \sum_{h'\neq h}\boldsymbol{b}_{h'}\boldsymbol{a}_{h'}^\top\right)\boldsymbol{a}_h. \tag{19}$$

One may search a local solution (i.e., a local minimum of the negative log posterior (17)) by iterating Equations (18) and (19). However, as shown below, the optimal solution can be obtained analytically in the current setup.

When the hyperparameters are homogeneous, that is, $\{c_{a_h}c_{b_h} = c; {}^\forall h = 1,\ldots,H\}$, a closed-form expression of the MAP estimator can be immediately obtained by combining the results given in Srebro et al. (2005) and Cai et al. (2010). The following theorem is its slight extension that covers heterogeneous cases (its proof is given in Appendix B):

Theorem 1 *Let γ_h (≥ 0) be the h-th largest singular value of V. Let $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ be the associated right and left singular vectors:*

$$V = \sum_{h=1}^{L}\gamma_h\boldsymbol{\omega}_{b_h}\boldsymbol{\omega}_{a_h}^\top. \tag{20}$$

The MAP estimator $\widehat{U}^{\mathrm{MAP}}$ is given by

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

where

$$\widehat{\gamma}_h^{\mathrm{MAP}} = \max\left\{0, \gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right\}. \tag{21}$$

The theorem implies that the MAP solution cuts off the singular values less than $\sigma^2/(c_{a_h}c_{b_h})$; otherwise it reduces the singular values by $\sigma^2/(c_{a_h}c_{b_h})$ (see Figure 2). This shrinkage effect allows the MAPMF method to avoid overfitting.

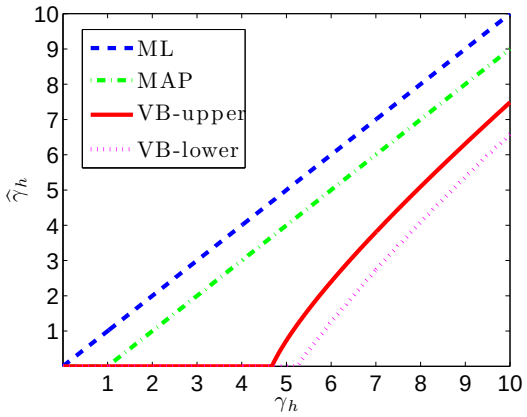

Figure 2: Shrinkage of the ML estimator (22), the MAP estimator (21), and the VB estimator (28) when $\sigma^2 = 0.1$, $c_{a_h} c_{b_h} = 0.1$, $L = 100$, and $M = 200$.

Similarly to Theorem 1, we can show that the *maximum likelihood* (ML) estimator is given by

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

where

$$\widehat{\gamma}_h^{\mathrm{ML}} = \gamma_h \text{ for all } h. \tag{22}$$

Thus the ML solution is reduced to V when $H = L$ (see Figure 2):

$$\widehat{U}^{\mathrm{ML}} = \sum_{h=1}^{L} \widehat{\gamma}_h^{\mathrm{ML}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top} = V.$$

A parametric model is said to be *identifiable* if the mapping between parameters and functions is one-to-one; otherwise the model is said to be *non-identifiable* (Watanabe, 2001). Since the decomposition $U = BA^{\top}$ is redundant, the MF model is non-identifiable (Nakajima and Watanabe, 2007). For identifiable models, the MAP estimator with the uniform prior is reduced to the ML estimator (Bishop, 2006). On the other hand, in the MF model, a single point in the space of U corresponds to a set of points in the joint space of A and B. For this reason, the uniform priors on A and B do not produce the uniform prior on U. Nevertheless, Equations (21) and (22) imply that MAP is reduced to ML when the priors on A and B are uniform (i.e., $c_{a_h}, c_{b_h} \to \infty$).

More precisely, Equations (21) and (22) show that the product $c_{a_h} c_{b_h} \to \infty$ is sufficient for MAP to be reduced to ML, which is weaker than both $c_{a_h}, c_{b_h} \to \infty$. This implies that both priors on A and B do not have to be uniform; only the condition that one of the priors is uniform is sufficient for MAP to be reduced to ML in the MF model. This phenomenon is distinctively different from the case of identifiable models.

If the prior is uniform and the likelihood is Gaussian, then the posterior is also Gaussian. Thus the mean and mode of the posterior agree with each other due to the symmetry of the Gaussian

density. For identifiable models, this fact implies that the FB and MAP solutions agree with each other. However, the FB and MAP solutions are generally different in non-identifiable models since the symmetry of the Gaussian density in the space of U is no longer kept in the joint space of A and B. In Section 4.1, we will further investigate these distinctive features of the MF model using illustrative examples.

3.2 VBMF

Substituting Equations (1), (2), and (3) into Equations (15) and (16), we find that the VB posteriors can be expressed as follows:

$$r_{\mathrm{A}}(A|V) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \boldsymbol{\mu}_{a_h}, \Sigma_{a_h}),$$
$$r_{\mathrm{B}}(B|V) = \prod_{h=1}^{H} \mathcal{N}_L(\boldsymbol{b}_h; \boldsymbol{\mu}_{b_h}, \Sigma_{b_h}),$$

where $\mathcal{N}_d(\cdot; \boldsymbol{\mu}, \Sigma)$ denotes the d-dimensional Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ. $\boldsymbol{\mu}_{a_h}$, $\boldsymbol{\mu}_{b_h}$, Σ_{a_h}, and Σ_{b_h} satisfy

$$\boldsymbol{\mu}_{a_h} = \frac{1}{\sigma^2}\Sigma_{a_h}\left(V - \sum_{h' \neq h} \boldsymbol{\mu}_{b_{h'}}\boldsymbol{\mu}_{a_{h'}}^{\top}\right)^{\top}\boldsymbol{\mu}_{b_h}, \tag{23}$$

$$\boldsymbol{\mu}_{b_h} = \frac{1}{\sigma^2}\Sigma_{b_h}\left(V - \sum_{h' \neq h} \boldsymbol{\mu}_{b_{h'}}\boldsymbol{\mu}_{a_{h'}}^{\top}\right)\boldsymbol{\mu}_{a_h}, \tag{24}$$

$$\Sigma_{a_h} = \left(\frac{1}{\sigma^2}\left(\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})\right) + c_{a_h}^{-2}\right)^{-1} I_M, \tag{25}$$

$$\Sigma_{b_h} = \left(\frac{1}{\sigma^2}\left(\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h})\right) + c_{b_h}^{-2}\right)^{-1} I_L. \tag{26}$$

I_d denotes the d-dimensional identity matrix. One may search a local solution (i.e., a local minimum of the free energy (10)) by iterating Equations (23)–(26).

It is straightforward to see that the VB solution $\widehat{U}^{\mathrm{VB}}$ (see Equation (11)) can be expressed as

$$\widehat{U}^{\mathrm{VB}} = \sum_{h=1}^{H} \boldsymbol{\mu}_{b_h}\boldsymbol{\mu}_{a_h}^{\top}. \tag{27}$$

Then we have the following theorem (its proof is given in Appendix C):[1]

Theorem 2 *$\widehat{U}^{\mathrm{VB}}$ is expressed as*

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

1. This theorem could be regarded as a more precise version of Theorem 1 given in Nakajima and Watanabe (2007).

where $\boldsymbol{\omega}_{a_h}$ and $\boldsymbol{\omega}_{b_h}$ are the right and the left singular vectors of V (see Equation (20)*). When $\gamma_h > \sqrt{M\sigma^2}$, $\widehat{\gamma}_h^{\mathrm{VB}} (= \|\boldsymbol{\mu}_{a_h}\| \|\boldsymbol{\mu}_{b_h}\|)$ is bounded as*

$$\max\left\{0, \left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)\gamma_h - \frac{\sigma^2\sqrt{M/L}}{c_{a_h}c_{b_h}}\right\} \leq \widehat{\gamma}_h^{\mathrm{VB}} < \left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)\gamma_h. \tag{28}$$

Otherwise, $\widehat{\gamma}_h^{\mathrm{VB}} = 0$.

The upper and lower bounds given in Equation (28) are illustrated in Figure 2. Theorem 2 states that, in the limit of $c_{a_h}c_{b_h} \to \infty$, the lower bound agrees with the upper bound and we have

$$\lim_{c_{a_h}c_{b_h}\to\infty} \widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \max\left\{0, \left(1 - \dfrac{M\sigma^2}{\gamma_h^2}\right)\gamma_h\right\} & \text{if } \gamma_h > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

This is the same form as the *positive-part James-Stein (PJS) shrinkage estimator* (James and Stein, 1961; Efron and Morris, 1973) (see Appendix A for the details of the PJS estimator). The factor $M\sigma^2$ is the expected contribution of the noise to γ_h^2—when the target matrix is $U = 0$, the expectation of γ_h^2 over all h is given by $M\sigma^2$. When $\gamma_h^2 < M\sigma^2$, Equation (29) implies that $\widehat{\gamma}_h^{\mathrm{VB}} = 0$. Thus, the PJS estimator cuts off the singular components dominated by noise. As γ_h^2 increases, the PJS shrinkage factor $M\sigma^2/\gamma_h^2$ tends to 0, and thus the estimated singular value $\widehat{\gamma}_h^{\mathrm{VB}}$ becomes close to the original singular value γ_h.

Let us compare the behavior of the VB solution (29) with that of the MAP solution (21) when $c_{a_h}c_{b_h} \to \infty$. In this case, the MAP solution merely results in the ML solution where no regularization is incorporated. In contrast, VB offers PJS-type regularization even when $c_{a_h}c_{b_h} \to \infty$. Thus VB can still mitigate overfitting (or it can possibly cause underfitting). This fact is in good agreement with the experimental results reported in Raiko et al. (2007), where no overfitting was observed when $c_{a_h}^2 = 1$ and $c_{b_h}^2$ is set to large values. This counter-intuitive fact stems again from the non-identifiability of the MF model—the Gaussian noise $\mathcal{E}$ imposed in the space of U possesses a very complex surface in the joint space of A and B, in particular, *multimodal* structure. This causes the MAP solution to be distinctively different from the VB solution. We call this regularization effect *model-induced regularization*. In Section 4.2, we investigate the effect of model-induced regularization in more detail using illustrative examples.

The following theorem more precisely specifies under which condition the VB estimator is strictly positive or zero (its proof is also included in Appendix C):

Theorem 3 *It holds that*

$$\widehat{\gamma}_h^{\mathrm{VB}} = 0 \text{ if } \gamma_h \leq \widetilde{\gamma}_h^{\mathrm{VB}},$$
$$\widehat{\gamma}_h^{\mathrm{VB}} > 0 \text{ if } \gamma_h > \widetilde{\gamma}_h^{\mathrm{VB}},$$

where

$$\widetilde{\gamma}_h^{\mathrm{VB}} = \sqrt{\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2c_{b_h}^2} + \sqrt{\left(\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2c_{b_h}^2}\right)^2 - LM\sigma^4}}. \tag{30}$$

$\widetilde{\gamma}_h^{\mathrm{VB}}$ is monotone decreasing with respect to $c_{a_h}c_{b_h}$, and is lower-bounded as

$$\widetilde{\gamma}_h^{\mathrm{VB}} > \lim_{c_{a_h}c_{b_h}\to\infty} \widetilde{\gamma}_h^{\mathrm{VB}} = \sqrt{M\sigma^2}.$$

As shown in Equation (21), $\widehat{\gamma}_h^{\mathrm{MAP}}$ satisfies

$$\widehat{\gamma}_h^{\mathrm{MAP}} = 0 \text{ if } \gamma_h \le \widetilde{\gamma}_h^{\mathrm{MAP}},$$
$$\widehat{\gamma}_h^{\mathrm{MAP}} > 0 \text{ if } \gamma_h > \widetilde{\gamma}_h^{\mathrm{MAP}},$$

where

$$\widetilde{\gamma}_h^{\mathrm{MAP}} = \frac{\sigma^2}{c_{a_h}c_{b_h}}.$$

Since

$$\widetilde{\gamma}_h^{\mathrm{VB}} > \sqrt{\frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2}} = \widetilde{\gamma}_h^{\mathrm{MAP}},$$

VB has a stronger shrinkage effect than MAP in terms of the vanishing condition of singular values.

We can derive another upper bound of $\widehat{\gamma}_h^{\mathrm{VB}}$, which depends on hyperparameters c_{a_h} and c_{b_h} (its proof is also included in Appendix C):

Theorem 4 *When $\gamma_h > \sqrt{M\sigma^2}$, $\widehat{\gamma}_h^{\mathrm{VB}}$ is upper-bounded as*

$$\widehat{\gamma}_h^{\mathrm{VB}} \le \sqrt{\left(1-\frac{L\sigma^2}{\gamma_h^2}\right)\left(1-\frac{M\sigma^2}{\gamma_h^2}\right)}\cdot\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}}. \tag{31}$$

When $L = M$ and $\gamma_h > \sqrt{M\sigma^2}$, the lower bound in Equation (28) and the upper bound in Equation (31) agree with each other. Thus, we have an analytic-form expression of $\widehat{\gamma}_h^{\mathrm{VB}}$ as follows:

$$\widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \max\left\{0, \left(1-\frac{M\sigma^2}{\gamma_h^2}\right)\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right\} & \text{if } \gamma_h > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Then, the complete VB posterior can also be obtained analytically (its proof is given in Appendix D):

Corollary 1 *When $L = M$, the VB posteriors are given by*

$$r_{\mathrm{A}}(A|V) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{a}_h; \boldsymbol{\mu}_{a_h}, \Sigma_{a_h}),$$
$$r_{\mathrm{B}}(B|V) = \prod_{h=1}^{H} \mathcal{N}_M(\boldsymbol{b}_h; \boldsymbol{\mu}_{b_h}, \Sigma_{b_h}),$$

where, for $\widehat{\gamma}_h^{\mathrm{VB}}$ *given by Equation* (32)*,*

$$\boldsymbol{\mu}_{a_h} = \pm\sqrt{\frac{c_{a_h}}{c_{b_h}}\widehat{\gamma}_h^{\mathrm{VB}}}\cdot\boldsymbol{\omega}_{a_h}, \tag{33}$$

$$\boldsymbol{\mu}_{b_h} = \pm\sqrt{\frac{c_{b_h}}{c_{a_h}}\widehat{\gamma}_h^{\mathrm{VB}}}\cdot\boldsymbol{\omega}_{b_h}, \tag{34}$$

$$\Sigma_{a_h} = \frac{c_{a_h}}{2c_{b_h}M}\left(\sqrt{\left(\widehat{\gamma}_h^{\mathrm{VB}}+\frac{\sigma^2}{c_{a_h}c_{b_h}}\right)^2+4\sigma^2M}-\left(\widehat{\gamma}_h^{\mathrm{VB}}+\frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\right)I_M, \tag{35}$$

$$\Sigma_{b_h} = \frac{c_{b_h}}{2c_{a_h}M}\left(\sqrt{\left(\widehat{\gamma}_h^{\mathrm{VB}}+\frac{\sigma^2}{c_{a_h}c_{b_h}}\right)^2+4\sigma^2M}-\left(\widehat{\gamma}_h^{\mathrm{VB}}+\frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\right)I_M. \tag{36}$$

3.3 EMAPMF

In the EMAPMF framework, the hyperparameters c_{a_h} and c_{b_h} are determined so that the Bayes posterior $p(A,B|V)$ is maximized (equivalently, the negative log posterior is minimized).

Differentiating the negative log posterior (17) with respect to $c_{a_h}^2$ and $c_{b_h}^2$ and setting the derivatives to zero lead to the following optimality conditions.

$$c_{a_h}^2 = \frac{\|\boldsymbol{a}_h\|^2}{M}, \tag{37}$$

$$c_{b_h}^2 = \frac{\|\boldsymbol{b}_h\|^2}{L}. \tag{38}$$

Alternating Equations (18), (19), (37), and (38), one may learn the parameters A,B and the hyperparameters c_{a_h},c_{b_h} at the same time.

However, as pointed out in Raiko et al. (2007), EMAPMF does not work properly since its objective (17) is unbounded from below at $\boldsymbol{a}_h,\boldsymbol{b}_h=\mathbf{0}$ and $c_{a_h},c_{b_h}\to 0$. Thus we end up in merely finding the trivial solution $(\boldsymbol{a}_h,\boldsymbol{b}_h=\mathbf{0})$ unless the iterative algorithm is stuck at some local optimum.

3.4 EVBMF

For the trial distribution (14), the VB free energy (10) can be written as follows:

$$\begin{aligned}F_{\mathrm{VB}}(r|V,\{c_{a_h}^2,c_{b_h}^2\}) = {}& \frac{LM}{2}\log\sigma^2+\sum_{h=1}^{H}\left(\frac{M}{2}\log c_{a_h}^2-\frac{1}{2}\log|\Sigma_{a_h}|+\frac{\|\boldsymbol{\mu}_{a_h}\|^2+\mathrm{tr}(\Sigma_{a_h})}{2c_{a_h}^2}\right.\\ &\left.+\frac{L}{2}\log c_{b_h}^2-\frac{1}{2}\log|\Sigma_{b_h}|+\frac{\|\boldsymbol{\mu}_{b_h}\|^2+\mathrm{tr}(\Sigma_{b_h})}{2c_{b_h}^2}\right)\\ &+\frac{1}{2\sigma^2}\left\|V-\sum_{h=1}^{H}\boldsymbol{\mu}_{b_h}\boldsymbol{\mu}_{a_h}^\top\right\|_{\mathrm{Fro}}^2\\ &+\frac{1}{2\sigma^2}\sum_{h=1}^{H}\left(\|\boldsymbol{\mu}_{a_h}\|^2\mathrm{tr}(\Sigma_{b_h})+\mathrm{tr}(\Sigma_{a_h})\|\boldsymbol{\mu}_{b_h}\|^2+\mathrm{tr}(\Sigma_{a_h})\mathrm{tr}(\Sigma_{b_h})\right),\end{aligned} \tag{39}$$

where $|\cdot|$ denotes the determinant of a matrix. Differentiating Equation (39) with respect to $c_{a_h}^2$ and $c_{b_h}^2$ and setting the derivatives to zero, we obtain the following optimality conditions:

$$c_{a_h}^2 = \frac{\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h})}{M}, \tag{40}$$

$$c_{b_h}^2 = \frac{\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})}{L}. \tag{41}$$

Here, we observe the invariance of Equation (39) with respect to the transform

$$\left\{(\boldsymbol{\mu}_{a_h}, \boldsymbol{\mu}_{b_h}, \Sigma_{a_h}, \Sigma_{b_h}, c_{a_h}^2, c_{b_h}^2)\right\} \to \left\{(s_h^{1/2}\boldsymbol{\mu}_{a_h}, s_h^{-1/2}\boldsymbol{\mu}_{b_h}, s_h\Sigma_{a_h}, s_h^{-1}\Sigma_{b_h}, s_h c_{a_h}^2, s_h^{-1}c_{b_h}^2)\right\} \tag{42}$$

for any $\{s_h \in \mathbb{R}; s_h > 0, h = 1, \ldots, H\}$. This redundancy can be eliminated by fixing the ratio between the hyperparameters to some constant—we choose 1 without loss of generality:

$$\frac{c_{a_h}}{c_{b_h}} = 1. \tag{43}$$

Then, Equations (40) and (41) yield

$$c_{a_h}^2 = \sqrt{\frac{(\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h}))\,(\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h}))}{LM}}, \tag{44}$$

$$c_{b_h}^2 = \sqrt{\frac{(\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h}))\,(\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h}))}{LM}}. \tag{45}$$

One may learn the parameters A, B and the hyperparameters c_{a_h}, c_{b_h} by applying Equations (44) and (45) after every iteration of Equations (23)–(26) (this gives a local minimum of Equation (39) at convergence).

For the EVB solution $\widehat{U}^{\mathrm{EVB}}$, we have the following theorem (its proof is provided in Appendix E):

Theorem 5 *The EVB estimator is given by the following form:*

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

$\widehat{\gamma}_h^{\mathrm{EVB}} = 0$ *if* $\gamma_h < \underline{\gamma}_h^{\mathrm{EVB}}$, *where*

$$\underline{\gamma}_h^{\mathrm{EVB}} = \left(\sqrt{L} + \sqrt{M}\right)\sigma.$$

If $\gamma_h \geq \underline{\gamma}_h^{\mathrm{EVB}}$, $\widehat{\gamma}_h^{\mathrm{EVB}}$ *is upper-bounded as*

$$\widehat{\gamma}_h^{\mathrm{EVB}} < \left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)\gamma_h. \tag{46}$$

If $\gamma_h \geq \overline{\gamma}_h^{\mathrm{EVB}}$, *where*

$$\overline{\gamma}_h^{\mathrm{EVB}} = \sqrt{7M} \cdot \sigma > \underline{\gamma}_h^{\mathrm{EVB}},$$

$\widehat{\gamma}_h^{\mathrm{EVB}}$ *is lower-bounded as*

$$\widehat{\gamma}_h^{\mathrm{EVB}} > \max\left\{0, \left(1 - \frac{2M\sigma^2}{\gamma_h^2 - \sqrt{\gamma_h^2(L+M+\sqrt{LM})\sigma^2}}\right)\gamma_h\right\}. \tag{47}$$

Theorem 5 implies that

$$\widehat{\gamma}_h^{\mathrm{EVB}} = 0 \text{ if } \gamma_h < \underline{\gamma}_h^{\mathrm{EVB}},$$
$$\widehat{\gamma}_h^{\mathrm{EVB}} > 0 \text{ if } \gamma_h \geq \underline{\gamma}_h^{\mathrm{EVB}}.$$

When

$$\underline{\gamma}_h^{\mathrm{EVB}} \leq \gamma_h < \bar{\gamma}_h^{\mathrm{EVB}},$$

our theoretical analysis is not precise enough to conclude whether $\widehat{\gamma}_h^{\mathrm{EVB}}$ is zero or not. As explained in Section 3.3, EMAP always results in the trivial solution (i.e., $\widehat{\gamma}_h^{\mathrm{EMAP}} = 0$). In contrast, Theorem 5 states that EVB gives a non-trivial solution (i.e., $\widehat{\gamma}_h^{\mathrm{EVB}} > 0$) when $\gamma_h \geq \bar{\gamma}_h^{\mathrm{EVB}}$. Since $\lim_{c_{a_h}c_{b_h}\to\infty}\widetilde{\gamma}_h^{\mathrm{VB}} = \sqrt{M\sigma^2} < \underline{\gamma}_h^{\mathrm{EVB}}$ (see Theorem 3), EVB has stronger shrinkage effect than VB with flat priors in terms of the vanishing condition of singular values.

It is also note worthy that the upper bound in Equation (46) is the same as that in Theorem 2. Thus, even when the hyperparameters c_{a_h} and c_{b_h} are learned from data by EVB, the same upper bound as the fixed-hyperparameter case in VB holds.

Another upper bound of $\widehat{\gamma}_h^{\mathrm{EVB}}$ is given as follows (its proof is also included in Appendix E):

Theorem 6 *When* $\gamma_h \geq \underline{\gamma}_h^{\mathrm{EVB}}$ $(= (\sqrt{L}+\sqrt{M})\sigma)$, $\widehat{\gamma}_h^{\mathrm{EVB}}$ *is upper-bounded as*

$$\widehat{\gamma}_h^{\mathrm{EVB}} < \sqrt{\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)}\gamma_h - \frac{\sqrt{LM}\sigma^2}{\gamma_h}. \tag{48}$$

Note that the right-hand side of (48) is strictly positive under $\gamma_h \geq \underline{\gamma}_h^{\mathrm{EVB}}$.

When $L = M$, the upper bound in Equation (48) is sharper than that in Equation (46), resulting in

$$\widehat{\gamma}_h^{\mathrm{EVB}} < \left(1 - \frac{2M\sigma^2}{\gamma_h^2}\right)\gamma_h. \tag{49}$$

The PJS shrinkage factor of the upper bound (49) is $2M\sigma^2/\gamma_h^2$. On the other hand, as shown in Equation (29), the PJS shrinkage factor of the plain VB with uniform priors on A and B (i.e., $c_a, c_b \to \infty$) is $M\sigma^2/\gamma_h^2$, which is *less than a half* of EVB. Thus, EVB provides substantially stronger regularization effect than the plain VB with uniform priors. Furthermore, from Equation (32), we can confirm that the upper bound (49) is equivalent to the VB solution when $c_{a_h}c_{b_h} = \gamma_h/M$.

When $L = M$, the complete EVB posterior is obtained analytically by using the following corollary (the proof is given in Appendix F):

Corollary 2 *For* $\gamma_h \geq 2\sqrt{M}\sigma$, *we define*

$$\varphi(\gamma_h) = \log\left(\frac{\gamma_h^2}{M\sigma^2}(1-\rho_-)\right) - \frac{\gamma_h^2}{M\sigma^2}(1-\rho_-) + \left(1 + \frac{\gamma_h^2}{2M\sigma^2}\rho_+^2\right), \tag{50}$$

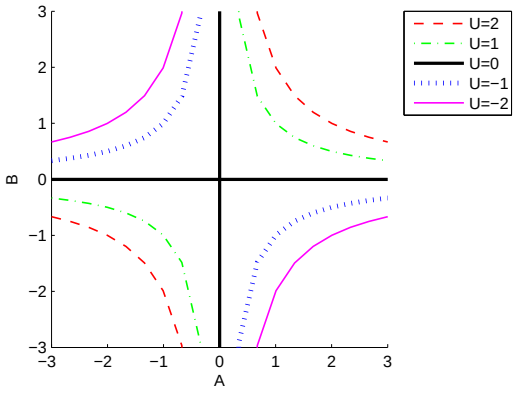

Figure 3: Equivalence class. Any A and B such that their product is unchanged give the same U.

where

$$\rho_{\pm} = \sqrt{\frac{1}{2}\left(1 - \frac{2M\sigma^2}{\gamma_h^2} \pm \sqrt{1 - \frac{4M\sigma^2}{\gamma_h^2}}\right)}.$$

Suppose $L = M$. If $\gamma_h \geq 2\sqrt{M}\sigma$ and $\varphi(\gamma_h) \leq 0$, then the EVB estimator of $c_{a_h}c_{b_h}$ is given by

$$\widehat{c}_{a_h}^{EVB}\widehat{c}_{b_h}^{EVB} = \frac{\gamma_h}{M}\rho_{+}. \tag{51}$$

Otherwise, $\widehat{c}_{a_h}^{EVB}\widehat{c}_{b_h}^{EVB} \to 0$. The EVB posterior is obtained by Corollary 1 with

$$(c_{a_h}^2, c_{b_h}^2) = \left(\widehat{c}_{a_h}^{EVB}\widehat{c}_{b_h}^{EVB}, \widehat{c}_{a_h}^{EVB}\widehat{c}_{b_h}^{EVB}\right).$$

Furthermore, when $\gamma_h \geq \sqrt{7M}\sigma$, it holds that

$$\varphi(\gamma_h) < 0. \tag{52}$$

Given γ_h, Equation (50) and then Equation (51) are computed analytically. By substituting Equations (51) and (43) into Equations (33)–(36), the complete EVB posterior is obtained. In Section 4.3, properties of EVBMF along with the behavior of the function (50) are further investigated through numerical examples.

4. Illustration of Influence of Non-identifiability

In order to understand the regularization mechanism of the Bayesian MF methods more intuitively, we illustrate the influence of non-identifiability when $L = M = H = 1$ (i.e., U, V, A, and B are merely scalars). In this case, any A and B such that their product is unchanged form an *equivalence class* and give the same U (see Figure 3). When $U = 0$, the equivalence class has a 'cross-shape' profile on the A- and B-axes; otherwise, it forms a pair of hyperbolic curves.

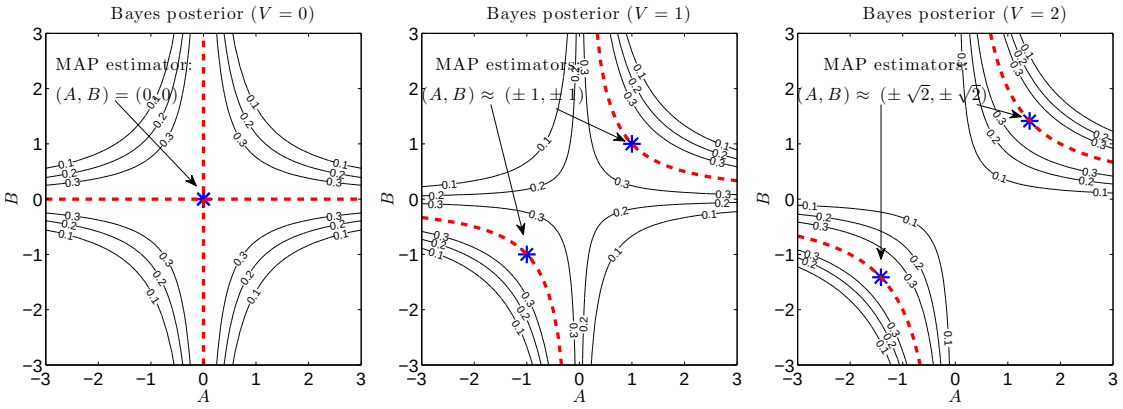

Figure 4: Bayes posteriors with $c_a = c_b = 100$ (i.e., almost flat priors). The asterisks are the MAP solutions, and the dashed lines indicate the ML solutions (the modes of the contour when $c_a = c_b = c \to \infty$).

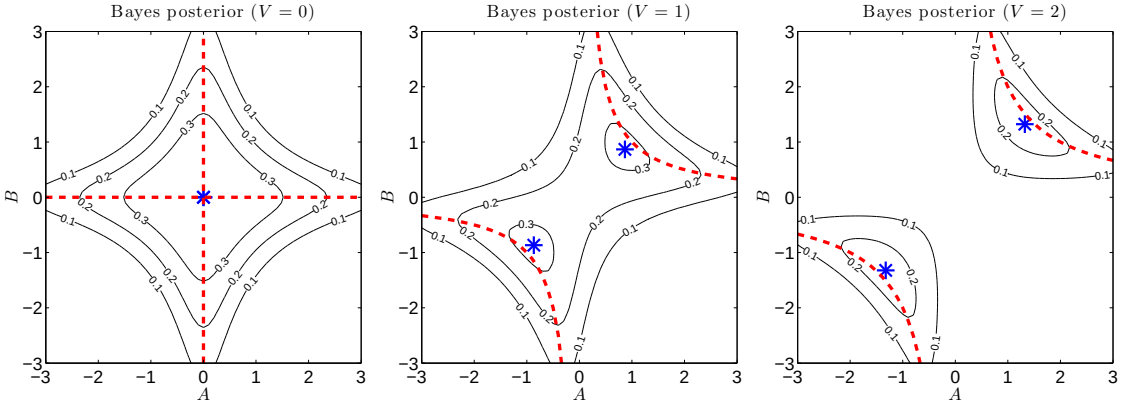

Figure 5: Bayes posteriors with $c_a = c_b = 2$. The dashed lines indicating the ML solutions are identical to those in Figure 4.

4.1 MAPMF

First, we illustrate the behavior of the MAP estimator.

When $L = M = H = 1$, Equation (17) yields that the Bayes posterior $p(A, B|V)$ is given as

$$p(A, B|V) \propto \exp\left(-\frac{1}{2\sigma^2}(V - BA)^2 - \frac{A^2}{2c_a^2} - \frac{B^2}{2c_b^2}\right). \tag{53}$$

Figure 4 shows the contour of the above Bayes posterior when $V = 0, 1, 2$ are observed, where the noise variance is $\sigma^2 = 1$ and the hyperparameters are $c_a = c_b = 100$ (i.e., almost flat priors). When $V = 0$, the surface of the Bayes posterior has a cross-shape profile and its maximum is at the origin. When $V > 0$, the surface is divided into the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$), and the two 'modes' get farther as V increases.

For finite c_a and c_b, Theorem 1 and Equation (66) (in Appendix B) imply that the MAP solution can be expressed as

$$\widehat{A}^{\mathrm{MAP}} = \pm\sqrt{\frac{c_a}{c_b}\max\left\{0, |V| - \frac{\sigma^2}{c_a c_b}\right\}},$$
$$\widehat{B}^{\mathrm{MAP}} = \pm\mathrm{sign}(V)\sqrt{\frac{c_b}{c_a}\max\left\{0, |V| - \frac{\sigma^2}{c_a c_b}\right\}},$$

where $\mathrm{sign}(\cdot)$ denotes the sign of a scalar. In Figure 4, the asterisks indicate the MAP estimators, and the dashed lines indicate the ML estimators (the modes of the contour of Equation (53) when $c_a = c_b = c \to \infty$). When $V = 0$, the Bayes posterior takes the maximum value on the A- and B-axes, which results in $\widehat{U}^{\mathrm{MAP}} = 0$. When $V = 1$, the profile of the Bayes posterior is hyperbolic and the maximum value is achieved on the hyperbolic curves in the positive orthant (i.e., $A, B > 0$) and the negative orthant (i.e., $A, B < 0$); in either case, $\widehat{U}^{\mathrm{MAP}} \approx 1$ (and $\widehat{U}^{\mathrm{MAP}} \to 1$ as $c_a, c_b \to \infty$). When $V = 2$, a similar multimodal structure is observed and the solution is $\widehat{U}^{\mathrm{MAP}} \approx 2$ (and $\widehat{U}^{\mathrm{MAP}} \to 2$ as $c_a, c_b \to \infty$). From these plots, we can visually confirm that the MAP solution with almost flat priors ($c_a = c_b = 100$) approximately agrees with the ML solution: $\widehat{U}^{\mathrm{MAP}} \approx \widehat{U}^{\mathrm{ML}} = V$ (and $\widehat{U}^{\mathrm{MAP}} \to \widehat{U}^{\mathrm{ML}}$ as $c_a, c_b \to \infty$).

Furthermore, these graphs illustrate the reason why the product $c_a c_b \to \infty$ is sufficient for MAP to agree with ML in the MF setup (see Section 3.1). Suppose c_a is kept small, say $c_a = 1$, in Figure 4. Then the Gaussian 'decay' remains along the horizontal axis in the profile of the Bayes posterior. However, the MAP solution $\widehat{U}^{\mathrm{MAP}}$ does not change since the mode of the Bayes posterior is kept lying on the dashed line (equivalence class). Thus, MAP agrees with ML if either c_a or c_b tends to infinity.

Figure 5 shows the contour of the Bayes posterior when $c_a = c_b = 2$. The MAP estimators are shifted from the ML estimators (dashed lines) toward the origin, and they are more clearly contoured as peaks.

4.2 VBMF

Here, we illustrate the behavior of the VB estimator, where the Bayes posterior is approximated by a spherical Gaussian.

In the current one-dimensional setup, Corollary 1 implies that the VB posteriors $r_{\mathrm{A}}(A|V)$ and $r_{\mathrm{B}}(B|V)$ can be expressed as

$$r_{\mathrm{A}}(A|V) = \mathcal{N}(A; \pm\sqrt{\widehat{\gamma}^{\mathrm{VB}} c_a/c_b}, \zeta c_a/c_b),$$
$$r_{\mathrm{B}}(B|V) = \mathcal{N}(B; \pm\mathrm{sign}(V)\sqrt{\widehat{\gamma}^{\mathrm{VB}} c_b/c_a}, \zeta c_b/c_a),$$

where $\mathcal{N}(\cdot; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2, and

$$\zeta = \sqrt{\left(\frac{\widehat{\gamma}^{\mathrm{VB}}}{2} + \frac{\sigma^2}{2c_a c_b}\right)^2 + \sigma^2} - \left(\frac{\widehat{\gamma}^{\mathrm{VB}}}{2} + \frac{\sigma^2}{2c_a c_b}\right),$$
$$\widehat{\gamma}^{\mathrm{VB}} = \begin{cases} \max\left\{0, \left(1 - \frac{\sigma^2}{V^2}\right)|V| - \frac{\sigma^2}{c_a c_b}\right\} & \text{if } V \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

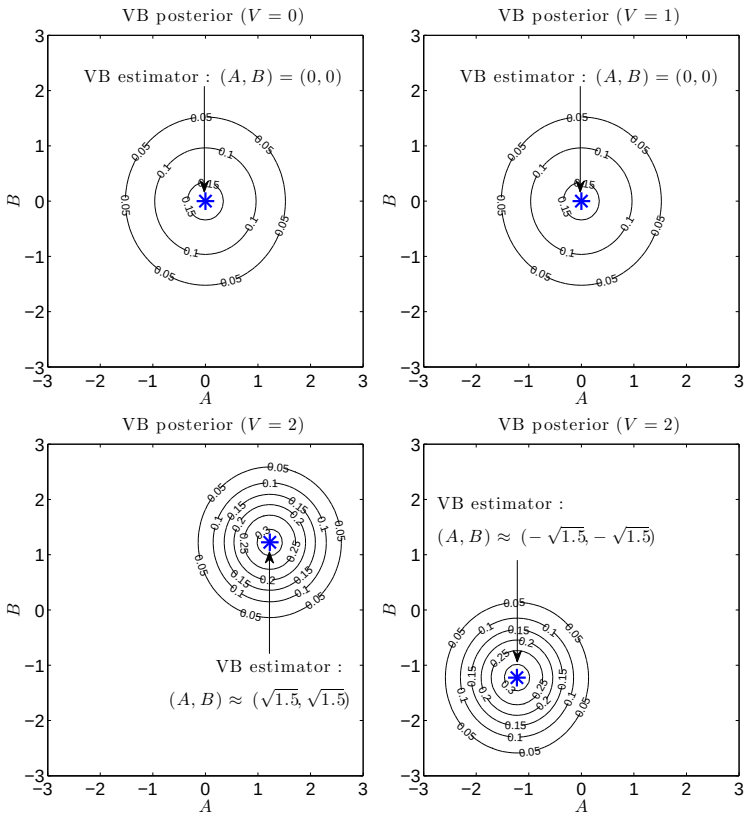

Figure 6: VB posteriors and VB solutions when $L = M = 1$ (i.e., the matrices V, U, A, and B are scalars). When $V = 2$, VB gives either one of the two solutions shown in the bottom row.

Figure 6 shows the contour of the VB posterior $r(A,B|V) = r_{\mathrm{A}}(A|V)r_{\mathrm{B}}(B|V)$ when $V = 0,1,2$ are observed, where the noise variance is $\sigma^2 = 1$ and the hyperparameters are $c_a = c_b = 100$ (i.e., almost flat priors). When $V = 0$, the cross-shaped contour of the Bayes posterior (see Figure 4) is approximated by a spherical Gaussian function located at the origin. Thus, the VB estimator is $\widehat{U}^{\mathrm{VB}} = 0$, which is equivalent to the MAP solution. When $V = 1$, two hyperbolic 'modes' of the Bayes posterior are approximated again by a spherical Gaussian function located at the origin. Thus, the VB estimator is still $\widehat{U}^{\mathrm{VB}} = 0$, which is different from the MAP solution.

$V = \widetilde{\gamma}_h^{\mathrm{VB}} \approx \sqrt{M\sigma^2} = 1$ ($\widetilde{\gamma}_h^{\mathrm{VB}} \to \sqrt{M\sigma^2}$ as $c_a, c_b \to \infty$) is actually a transition point of the behavior of the VB estimator. When V is not larger than the threshold $\sqrt{M\sigma^2}$, the VB method tries to approximate the two 'modes' of the Bayes posterior by the origin-centered Gaussian function. When V goes beyond the threshold $\sqrt{M\sigma^2}$, the 'distance' between two hyperbolic modes of the Bayes posterior becomes so large that the VB method chooses to approximate one of the two modes in the positive and negative orthants. As such, the symmetry is broken spontaneously and the VB solution is detached from the origin. Note that, as discussed in Section 3, $M\sigma^2$ amounts to the expected contribution of noise $\mathcal{E}$ to the squared singular value γ^2 ($= V^2$ in the current setup).

The bottom row of Figure 6 shows the contour of two possible VB posteriors when $V = 2$. Note that, in either case, the VB solution is the same: $\widehat{U}^{\mathrm{VB}} \approx 3/2$. The VB solution is closer to the origin

than the MAP solution $\widehat{U}^{\mathrm{MAP}} = 2$, and the difference between the VB and MAP solutions tends to shrink as V increases.

4.3 EVBMF

Next, we illustrate the behavior of the EVB estimator.

In the current one-dimensional setup, the free energy (39) is expressed as

$$F_{\mathrm{VB}}(r|V,c_a^2,c_b^2) = \log\frac{c_a^2c_b^2}{\Sigma_a\Sigma_b} + \frac{\mu_a^2+\Sigma_a}{2c_a^2} + \frac{\mu_b^2+\Sigma_b}{2c_b^2} - \frac{1}{\sigma^2}V\mu_a\mu_b + \frac{1}{2\sigma^2}\left(\mu_a^2+\Sigma_a\right)\left(\mu_b^2+\Sigma_b\right) + \mathrm{Const.}$$

According to Corollary 2, if $|V| \geq 2\sigma$ and $\varphi(|V|) \leq 0$, the EVB estimator of the hyperparameters is given by

$$(\widehat{c}_a^{\mathrm{EVB}})^2 = (\widehat{c}_b^{\mathrm{EVB}})^2 = |V|\rho_+, \tag{54}$$

where

$$\varphi(|V|) = \log\left(\frac{|V|^2}{\sigma^2}(1-\rho_-)\right) - \frac{|V|^2}{\sigma^2}(1-\rho_-) + \left(1+\frac{|V|^2}{2\sigma^2}\rho_+^2\right),$$
$$\rho_\pm = \sqrt{\frac{1}{2}\left(1-\frac{\sigma^2}{|V|^2} \pm \sqrt{1-\frac{4\sigma^2}{|V|^2}}\right)}.$$

Based on a simple numerical evaluation (Figure 7) of $\varphi(|V|)$, we can confirm that Equation (54) holds if $|V| \geq \widetilde{\gamma}^{\mathrm{EVB}}$, where

$$\widetilde{\gamma}^{\mathrm{EVB}} \approx 2.22.$$

Otherwise $\widehat{c}_{a_h}^{\mathrm{EVB}}, \widehat{c}_{b_h}^{\mathrm{EVB}} \to 0$. Note that $\widetilde{\gamma}^{\mathrm{EVB}}$ is theoretically bounded as

$$\left(2 = 2\sigma^2 =\right) \underline{\gamma}^{\mathrm{EVB}} \leq \widetilde{\gamma}^{\mathrm{EVB}} \leq \bar{\gamma}^{\mathrm{EVB}} \left(= \sqrt{7}\sigma^2 \approx 2.64\right),$$

as shown in Equation (52).

Using Corollary 1 with Equation (54), we can plot the EVB posterior. When

$$|V| < \widetilde{\gamma}^{\mathrm{EVB}} \approx 2.22,$$

the infimum of the free energy with respect to $(\mu_a, \mu_b, \Sigma_a, \Sigma_b, c_a^2, c_b^2)$ is attained by $c_a^2 = c_b^2 = \varepsilon$, $\mu_a = \mu_b = 0$, and

$$\Sigma_a = \Sigma_b = \frac{\sigma^2}{2\varepsilon}\left(\sqrt{1+\frac{4n\varepsilon^2}{\sigma^2}} - 1\right),$$

where $\varepsilon \to 0$ (i.e., $c_a^2 = c_b^2 \to 0$, $\mu_a = \mu_b = 0$, and $\Sigma_a = \Sigma_b \to 0$). Therefore, the Gaussian width of the EVB posterior approaches zero (i.e., *Dirac's delta function* located at the origin). The left graph of Figure 8 illustrates the contour of the EVB posterior $r(A,B|V) = r_{\mathrm{A}}(A|V)r_{\mathrm{B}}(B|V)$ when $V = 2$

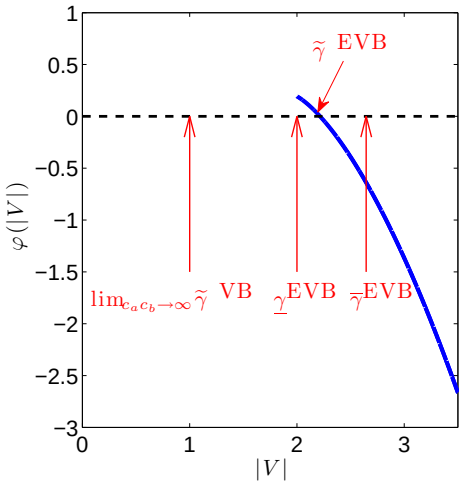

Figure 7: Numerical evaluation of $\varphi(|V|)$ when $L = M = 1$ and $\sigma^2 = 1$ (the blue solid curve). The blue solid curve crosses the black dashed line ($\varphi(|V|) = 0$) at $|V| = \widetilde{\gamma}^{\mathrm{EVB}} \approx 2.22$.

is observed, where the noise variance is $\sigma^2 = 1$. Since $\widehat{U}^{\mathrm{MAP}} \approx 2$ and $\widehat{U}^{\mathrm{VB}} \approx 1.5$ under almost flat priors (see Figure 4 and Figure 6), $\widehat{U}^{\mathrm{EVB}} = 0$ is more strongly regularized than VB and MAP.

On the other hand, when

$$|V| \geq \widetilde{\gamma}^{\mathrm{EVB}} \approx 2.22,$$

the EVB posteriors $r_{\mathrm{A}}(A|V)$ and $r_{\mathrm{B}}(B|V)$ can be expressed as

$$r_{\mathrm{A}}(A|V) = \mathcal{N}(A; \pm\sqrt{\widehat{\gamma}^{\mathrm{EVB}}}, \zeta),$$
$$r_{\mathrm{B}}(B|V) = \mathcal{N}(B; \pm\mathrm{sign}(V)\sqrt{\widehat{\gamma}^{\mathrm{EVB}}}, \zeta),$$

where

$$\zeta = \sqrt{\left(\frac{\widehat{\gamma}^{\mathrm{EVB}}}{2} + \frac{|V|\rho_-}{2}\right)^2 + \sigma^2} - \left(\frac{\widehat{\gamma}^{\mathrm{EVB}}}{2} + \frac{|V|\rho_-}{2}\right),$$
$$\rho_- = \sqrt{\frac{1}{2}\left(1 - \frac{2\sigma^2}{\gamma_h^2} - \sqrt{1 - \frac{4\sigma^2}{\gamma_h^2}}\right)},$$
$$\widehat{\gamma}^{\mathrm{EVB}} = \left(1 - \frac{\sigma^2}{V^2} - \rho_-\right)|V|.$$

When $V = 3$ is observed, we have $\widehat{U}^{\mathrm{EVB}} \approx 2.28$ ($c_a^2 = c_b^2 \approx 2.62$, $\mu_a = \mu_b \approx \sqrt{2.28}$, and $\Sigma_a = \Sigma_b \approx 0.33$). The possible posteriors are plotted in the middle and the right graphs of Figure 8. Since $\widehat{U}^{\mathrm{MAP}} \approx 3$ and $\widehat{U}^{\mathrm{VB}} = 3/8 \approx 2.67$ under almost flat priors, EVB has stronger regularization effect than VB and MAP.

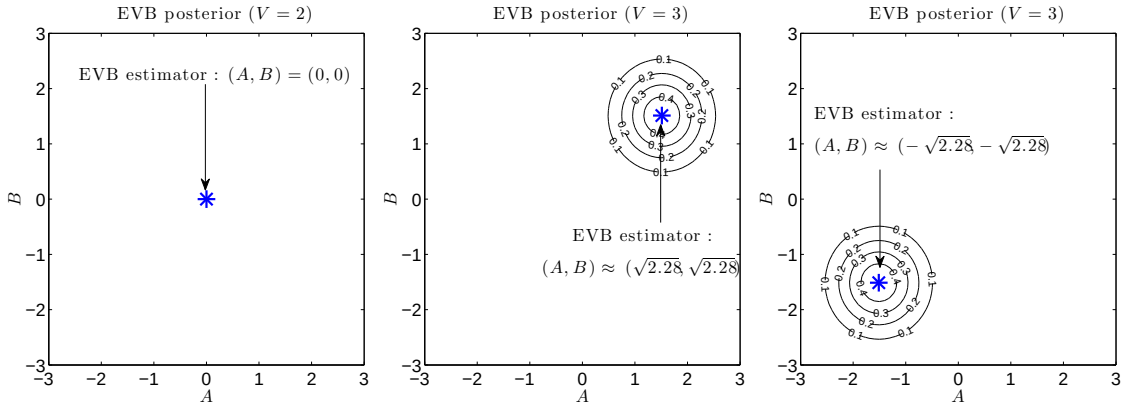

Figure 8: EVB posteriors and EVB solutions when $L = M = 1$. Left: When $V = 2$, the EVB posterior is reduced to Dirac's delta function located at the origin. Right: When $V = 3$, the solution is detached from the origin and given by $(A,B) \approx (\sqrt{2.28}, \sqrt{2.28})$ or $(A,B) \approx (-\sqrt{2.28}, -\sqrt{2.28})$, which both yields the same solution $\widehat{U}^{\mathrm{EVB}} \approx 2.28$.

4.4 FBMF

Here, we illustrate the behavior of the FB estimator.

When $L = M = H = 1$, the FB solution (5) is expressed as

$$\widehat{U}^{\mathrm{FB}} = \langle AB \rangle_{p(V|A,B)\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)}. \tag{55}$$

If $V = 0, 1, 2, 3$ are observed, the FB solutions with almost flat priors are $0, 0.92, 1.93, 2.95$, respectively, which were numerically computed.[2] Since the corresponding MAP solutions (with the almost flat priors) are $0, 1, 2, 3$, FB and MAP were shown to produce different solutions.

The theory by Jeffreys (1946) explains the origin of *model-induced regularization* in FB. Let us consider the *non-factorizing* model

$$p(V|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}\|V - U\|_{\mathrm{Fro}}^2\right), \tag{56}$$

where U itself is the parameter to be estimated. The Jeffreys (non-informative) prior for this model is uniform

$$\phi_{\mathrm{U}}^{\mathrm{Jef}}(U) \propto 1. \tag{57}$$

On the other hand, the Jeffreys prior for the MF model (1) is given by

$$\phi_{\mathrm{A,B}}^{\mathrm{Jef}}(A,B) \propto \sqrt{A^2 + B^2}, \tag{58}$$

which is illustrated in Figure 9 (see Appendix I for the derivation of Equations (57) and (58)). Note that $\phi_{\mathrm{U}}^{\mathrm{Jef}}(U)$ and $\phi_{\mathrm{A,B}}^{\mathrm{Jef}}(A,B)$ are both *improper*.

2. More precisely, we numerically calculated the FB solution (55) by sampling A and B from the almost flat prior distributions $\phi_{\mathrm{A}}(A)\phi_{\mathrm{B}}(B)$ with $c_a = c_b = 100$ and taking the sample average of $AB \cdot p(V|A,B)$.

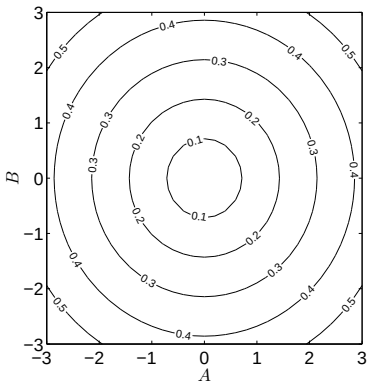

Figure 9: The Jeffreys non-informative prior of the MF model in the joint space of A and B: $\phi_{\mathrm{Jef}}(A,B) \propto \sqrt{A^2+B^2}$. The scaling of the density value in the graph is arbitrary due to impropriety.

Jeffreys (1946) states that the both combinations, the *non-factorizing* model (56) with its Jeffreys prior (57) and the MF model (1) with its Jeffreys prior (58), give the equivalent FB solution. We can easily show that the former combination, Equations (56) and (57), gives an unregularized solution. Thus, the FB solution in the MF model (1) with its Jeffreys prior (58) is also unregularized. Since the flat prior on (A,B) has more probability mass around the origin than the Jeffreys prior (58) (see Figure 9), it favors smaller $|U|$ and regularizes the FB solution.

4.5 EMAPMF

As explained in Section 3.3, EMAPMF always results in the trivial solution, $A,B=0$ and $c_{a_h}, c_{b_h} \to 0$.

4.6 EFBMF

The EFBMF solution is written as follows:

$$\widehat{U}^{\mathrm{EFB}} = \langle AB \rangle_{p(V|A,B)\phi_{\mathrm{A}}(A;\widehat{c}_a)\phi_{\mathrm{B}}(B;\widehat{c}_b)},$$

where

$$(\widehat{c}_a, \widehat{c}_b) = \operatorname*{argmin}_{(c_a,c_b)} F(V;c_a,c_b).$$

Here $F(V;c_a,c_b)$ is the Bayes free energy (6).

When $V=0,1,2,3$ are observed, the EFB solutions are $0, 0.00, 1.25, 2.58$ ($\widehat{c}_a=\widehat{c}_b \approx 0, 0.0, 1.4$, 2.1), respectively, which were numerically computed.[3] Since $F(V;c_a,c_b) \to \infty$ when $c_a c_b \to \infty$, the

3. The model (1) and the priors (2) and (3) are invariant under the following parameter transformation

$$(\boldsymbol{a}_h, \boldsymbol{b}_h, c_{a_h}, c_{b_h}) \to (s_h^{1/2}\boldsymbol{a}_h, s_h^{-1/2}\boldsymbol{b}_h, s_h^{1/2}c_{a_h}, s_h^{-1/2}c_{b_h})$$

for any $\{s_h \in \mathbb{R}; s_h > 0, h=1,\ldots,H\}$. Here, we fixed the ratio to $c_a/c_b=1$. For $c_a c_b = 10^{-2.00}, 10^{-1.99}, \ldots, 10^{1.00}$, we numerically computed the free energy (6), and chose the minimizer $\widehat{c}_a\widehat{c}_b$, with which the FB solution is computed.

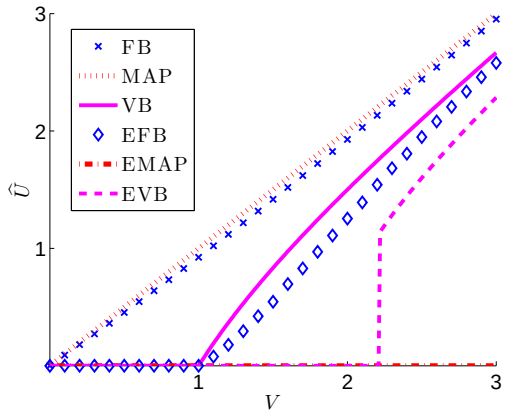

Figure 10: Numerical results of the FBMF solution $\widehat{U}^{\mathrm{FB}}$, the MAPMF solution $\widehat{U}^{\mathrm{MAP}}$, the VBMF solution $\widehat{U}^{\mathrm{VB}}$, the EFBMF solution $\widehat{U}^{\mathrm{EFB}}$, the EMAPMF solution $\widehat{U}^{\mathrm{EMAP}}$, and the EVBMF solution $\widehat{U}^{\mathrm{EVB}}$ when the noise variance is $\sigma^2 = 1$. For MAPMF, VBMF, and FBMF, the hyperparameters are set to $c_a = c_b = 100$ (i.e., almost flat priors).

minimizer of $F(V; c_a, c_b)$ with respect to $\widehat{c}_a$ and $\widehat{c}_b$ are always finite. This implies that EFBMF is more strongly regularized than FBMF with almost flat priors ($c_a c_b \to \infty$).

4.7 Summary

Finally, we summarize the numerical results of all Bayes estimators in Figure 10, including the FBMF solution $\widehat{U}^{\mathrm{FB}}$, the MAPMF solution $\widehat{U}^{\mathrm{MAP}}$, the VBMF solution $\widehat{U}^{\mathrm{VB}}$, the EFBMF solution $\widehat{U}^{\mathrm{EFB}}$, the EMAPMF solution $\widehat{U}^{\mathrm{EMAP}}$, and the EVBMF solution $\widehat{U}^{\mathrm{EVB}}$ when the noise variance is $\sigma^2 = 1$. For MAPMF, VBMF, and FBMF, the hyperparameters are set to $c_a = c_b = 100$ (i.e., almost flat priors). Overall, the solutions satisfy

$$\widehat{U}^{\mathrm{EMAP}} \leq \widehat{U}^{\mathrm{EVB}} \leq \widehat{U}^{\mathrm{EFB}} \leq \widehat{U}^{\mathrm{VB}} \leq \widehat{U}^{\mathrm{FB}} \leq \widehat{U}^{\mathrm{MAP}},$$

which shows the strength of regularization effect of each method.

5. Conclusion

In this paper, we theoretically analyzed the behavior of Bayesian matrix factorization methods. More specifically, in Section 3, we derived *non-asymptotic* bounds of the *maximum a posteriori matrix factorization* (MAPMF) estimator and the *variational Bayesian matrix factorization* (VBMF) estimator. Then we showed that MAPMF consists of the *trace-norm* shrinkage alone, while VBMF consists of the *positive-part James-Stein* (PJS) shrinkage and the trace-norm shrinkage.

An interesting finding was that, while the trace-norm shrinkage does not take effect when the priors are flat, the PJS shrinkage remains activated even with flat priors. The fact that the PJS shrinkage remains activated even with flat priors is induced by the non-identifiability of the MF models, where parameters form equivalent classes. Thus, flat priors in the space of factorized matrices are no longer flat in the space of the target (composite) matrix. Furthermore, simple distributions such

as the Gaussian distribution in the space of the target matrix produce highly complicated *multimodal* distributions in the space of factorized matrices.

We further extended the above analysis to *empirical VBMF* scenarios where hyperparameters included in priors are optimized based on the VB free energy. We showed that the 'strength' of the PJS shrinkage is more than doubled compared with the flat prior cases. We also illustrated the behavior of Bayesian matrix factorization methods using one-dimensional examples in Section 4.

Our theoretical analysis relies on the assumption that a fully observed matrix is provided as a training sample. Thus, our results are not directly applicable to the collaborative filtering scenarios where an observed matrix with missing entries is given. Our important future work is to extend the current analysis so that the behavior of the collaborative filtering algorithms can also be explained. The correspondence between MAPMF and the trace-norm regularization still holds even if missing entries exist. Likewise, we hope to find a relation between VBMF and a regularization term acting on a matrix, which results in the PJS shrinkage if a fully observed matrix is given.

Our analysis also relies on the column-wise independence constraint (14), which was also used in Raiko et al. (2007), on the VB posterior. In principle, the weaker matrix-wise constraint (9) which was used in Lim and Teh (2007) allows non-zero covariances between column vectors, and can achieve a better approximation to the true Bayes posterior. How this affects the performance and when the difference is substantial are to be investigated.

As explained in Appendix A, the PJS estimator dominates (i.e., uniformly better than) the maximum likelihood (ML) estimator in vector estimation. This means that, when $L = 1$, VBMF with (almost) flat priors dominates MLMF. Another interesting future direction is to investigate whether this nice property is inherited to matrix estimation. For matrix estimation ($L > 1$), a variety of estimators which shrink singular values have been proposed (Stein, 1975; Ledoit and Wolf, 2004; Daniels and Kass, 2001), and were shown to possess nice properties under different criteria. Discussing the superiority of such shrinkage estimators including VBMF is interesting future work.

Our investigation revealed a gap between the *fully-Bayesian* (FB) estimator and the VB estimator (see Section 4.7). Figure 10 showed that the VB estimator tends to be strongly regularized. This could cause underfitting and degrade the performance. On the other hand, it is also possible that, in some cases, this stronger regularization could work favorably to suppress overfitting, if we take into account the fact that practitioners do not always choose their prior distributions based on explicit prior information (it is often the case that conjugate priors are chosen only for computational convenience). Further theoretical analysis and empirical investigation are needed to clarify when the stronger regularization of the VB estimator is harmful or helpful.

Tensor factorization is a high-dimensional extension of matrix factorization, which gathers considerable attention recently as a novel data analysis tool (Cichocki et al., 2009). Among various methods, Bayesian methods of tensor factorization have been shown to be promising (Tao et al., 2008; Yu et al., 2008; Hayashi et al., 2009; Chu and Ghahramani, 2009). In our future work, we will elucidate the behavior of tensor factorization methods based on a similar line of discussion to the current work.

Acknowledgments

We would like to thank anonymous reviewers for helpful comments and suggestions for future work. Masashi Sugiyama thanks the support from the FIRST program.

Appendix A. James-Stein Shrinkage Estimator

Here, we briefly introduce the *James-Stein* (JS) shrinkage estimator and its variants (James and Stein, 1961; Efron and Morris, 1973).

Let us consider the problem of estimating the mean $\boldsymbol{\mu}$ ($\in \mathbb{R}^d$) of the d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$ from its independent and identically distributed samples

$$\mathcal{X}^n = \{\boldsymbol{x}_i \in \mathbb{R}^d \mid i = 1, \ldots, n\}.$$

We measure the generalization error (or the risk) of an estimator $\widehat{\boldsymbol{\mu}}$ by the expected squared error:

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2,$$

where $\mathbb{E}$ denotes the expectation over the samples $\mathcal{X}^n$.

An estimator $\widehat{\boldsymbol{\mu}}$ is said to *dominate* another estimator $\widehat{\boldsymbol{\mu}}'$ if

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \mathbb{E}\|\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}\|^2 \text{ for all } \boldsymbol{\mu},$$

and

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 < \mathbb{E}\|\widehat{\boldsymbol{\mu}}' - \boldsymbol{\mu}\|^2 \text{ for some } \boldsymbol{\mu}.$$

An estimator is said to be *admissible* if no estimator dominates it.

Stein (1956) proved the inadmissibility of the maximum likelihood (ML) estimator (or equivalently the least-squares estimator),

$$\widehat{\boldsymbol{\mu}}^{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$$

when $d \geq 3$. This discovery was surprising because the ML estimator had been believed to be a *good* estimator. James and Stein (1961) subsequently proposed the JS shrinkage estimator $\widehat{\boldsymbol{\mu}}^{\mathrm{JS}}$, which was proved to dominate the ML estimator:

$$\widehat{\boldsymbol{\mu}}^{\mathrm{JS}} = \left(1 - \frac{\chi\sigma^2}{n\|\widehat{\boldsymbol{\mu}}^{\mathrm{ML}}\|^2}\right)\widehat{\boldsymbol{\mu}}^{\mathrm{ML}}, \tag{59}$$

where $\chi = d - 2$. Efron and Morris (1973) showed that the JS shrinkage estimator can be derived as an empirical Bayes estimator. In the current paper, we refer to all estimators of the form (59) with arbitrary $\chi > 0$ as the JS shrinkage estimators.

The *positive-part James-Stein* (PJS) shrinkage estimator, which was shown to dominate the JS estimator, is given as follows (Baranchik, 1964):

$$\widehat{\boldsymbol{\mu}}^{\mathrm{PJS}} = \max\left\{0, \left(1 - \frac{\chi\sigma^2}{n\|\widehat{\boldsymbol{\mu}}^{\mathrm{ML}}\|^2}\right)\widehat{\boldsymbol{\mu}}^{\mathrm{ML}}\right\}.$$

Note that the PJS estimator itself is also inadmissible, following the fact that admissible estimators are necessarily smooth (Lehmann, 1983). Indeed, there exist several estimators that dominate the PJS estimator (Strawderman, 1971; Guo and Pal, 1992; Shao and Strawderman, 1994). However, their improvement is rather minor, and they are not as simple as the PJS estimator. Moreover, none of these estimators is admissible.

Appendix B. Proof of Theorem 1

The MAP estimator is defined as the minimizer of the negative log (17) of the Bayes posterior. Let us double Equation (17) and neglect some constant terms which are irrelevant to its minimization with respect to $\{\boldsymbol{a}_h, \boldsymbol{b}_h\}_{h=1}^H$:

$$\mathcal{L}^{\mathrm{MAP}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h\}_{h=1}^H) = \sum_{h=1}^H \left(\frac{\|\boldsymbol{a}_h\|^2}{c_{a_h}^2} + \frac{\|\boldsymbol{b}_h\|^2}{c_{b_h}^2} \right) + \frac{1}{\sigma^2} \left\| V - \sum_{h=1}^H \boldsymbol{b}_h \boldsymbol{a}_h^\top \right\|_{\mathrm{Fro}}^2. \tag{60}$$

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

Lemma 7 *For arbitrary matrices $A \in \mathbb{R}^{M\times H}$ and $B \in \mathbb{R}^{L\times H}$, let*

$$BA^\top = \Omega_L \Gamma \Omega_R^\top$$

be the singular value decomposition of the product $BA^\top$, where $\Gamma = diag(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_H)$ ($\{\widehat{\gamma}_h\}$ are in non-increasing order). Remember that $\{c_{a_h} c_{b_h}\}$, where $C_A = diag(c_{a_1}^2, \ldots, c_{a_H}^2)$ and $C_B = diag(c_{b_1}^2, \ldots, c_{b_H}^2)$ are positive-definite, are also arranged in non-increasing order. Then, it holds that

$$\mathrm{tr}(AC_A^{-1}A^\top) + \mathrm{tr}(BC_B^{-1}B^\top) \geq \sum_{h=1}^H \frac{2\widehat{\gamma}_h}{c_{a_h} c_{b_h}}. \tag{61}$$

Using Lemma 7, we obtain the following lemma (its proof is given in Appendix G.2):

Lemma 8 *The MAP solution $\widehat{U}^{\mathrm{MAP}}$ is written in the following form:*

$$\widehat{U}^{\mathrm{MAP}} = \widehat{B}\widehat{A}^\top = \sum_{h=1}^H \widehat{\gamma}_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top. \tag{62}$$

There exists at least one minimizer that can be written as

$$\boldsymbol{a}_h = a_h \boldsymbol{\omega}_{a_h}, \tag{63}$$
$$\boldsymbol{b}_h = b_h \boldsymbol{\omega}_{b_h}, \tag{64}$$

where $\{a_h, b_h\}$ are scalars such that

$$\widehat{\gamma}_h = a_h b_h \geq 0.$$

Lemma 8 implies that the minimization of Equation (60) amounts to a re-weighted singular value decomposition.

We can also prove the following lemma (its proof is given in Appendix G.3):

Lemma 9 *Let $\{\mathcal{H}_k; k = 1, \ldots, K(\leq H)\}$ be the partition of $\{1, \ldots, H\}$ such that $c_{a_h} c_{b_h} = c_{a_{h'}} c_{b_{h'}}$ if and only if h and h' belong to the same group (i.e., $^\exists k$ such that $h, h' \in \mathcal{H}_k$). Suppose that $(\widehat{A}, \widehat{B})$ is a MAP solution. Then,*

$$\widehat{A}' = \widehat{A}\Theta^\top,$$
$$\widehat{B}' = \widehat{B}\Theta^{-1},$$

is also a MAP solution, for any Θ *defined by*

$$\begin{aligned}\Theta &= C_A^{1/2}\Xi C_A^{-1/2}\\ &= C_B^{-1/2}\Xi C_B^{1/2}.\end{aligned}$$

Here, Ξ *is a block diagonal matrix such that the blocks are organized based on the partition* $\{\mathcal{H}_k\}$*, and each block consists of an arbitrary orthogonal matrix.*

Lemma 9 states that non-orthogonal solutions (i.e., $\{\boldsymbol{a}_h\}$, as well as $\{\boldsymbol{b}_h\}$, are not orthogonal with each other) can exist. However, Lemma 8 guarantees that any non-orthogonal solution has its *equivalent* orthogonal solution, which is written in the form of Equations (63) and (64). Here, by *equivalent* solution, we denote a solution resulting in the identical $\widehat{U}^{\mathrm{MAP}}$ in Equation (62). Since we are interested in finding $\widehat{U}^{\mathrm{MAP}}$, we regard the orthogonal solution as the representative of the *equivalent* solutions, and focus on it.

The expression (63) and (64) allows us to decompose the minimization of Equation (60) into the minimization of the following H separate objective functions: for $h = 1, \ldots, H$,

$$\mathcal{L}_h^{\mathrm{MAP}}(a_h, b_h) = \left(\frac{a_h^2}{c_{a_h}^2} + \frac{b_h^2}{c_{b_h}^2}\right) + \frac{1}{\sigma^2}\left(\gamma_h - a_h b_h\right)^2.$$

This can be written as

$$\mathcal{L}_h^{\mathrm{MAP}}(a_h, b_h) = \frac{b_h^2}{c_{a_h}^2}\left(\frac{a_h}{b_h} - \frac{c_{a_h}}{c_{b_h}}\right)^2 + \frac{1}{\sigma^2}\left(a_h b_h - \left(\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\right)^2 + \left(\frac{2\gamma_h}{c_{a_h}c_{b_h}} - \frac{\sigma^2}{c_{a_h}^2 c_{b_h}^2}\right). \quad (65)$$

The third term is constant with respect to a_h and b_h. The first nonnegative term vanishes by setting the ratio a_h/b_h to

$$\frac{a_h}{b_h} = \frac{c_{a_h}}{c_{b_h}} \quad (\text{or } b_h = 0). \quad (66)$$

Minimizing the second term in Equation (65), which is quadratic with respect to the product $a_h b_h$ (≥ 0), we can easily obtain Equation (21), which completes the proof. ■

Appendix C. Proof of Theorem 2, Theorem 3, and Theorem 4

We denote by $\mathbb{R}_+^d$ the set of the d-dimensional vectors with non-negative elements, by $\mathbb{R}_{++}^d$ the set of the d-dimensional vectors with positive elements, by $\mathbb{S}_+^d$ the set of $d \times d$ positive semi-definite symmetric matrices, and by $\mathbb{S}_{++}^d$ the set of $d \times d$ positive definite symmetric matrices. The VB free energy to be minimized can be expressed as Equation (39). Neglecting constant terms, we define

the objective function as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h,\boldsymbol{b}_h,\Sigma_{a_h},\Sigma_{b_h}\}) &= 2F_{\mathrm{VB}}(r|V,\{c_{a_h}^2,c_{b_h}^2\})+\mathrm{Const.}\\ &= \sum_{h=1}^{H}\left(-\log|\Sigma_{a_h}|+\frac{\|\boldsymbol{\mu}_{a_h}\|^2+\mathrm{tr}(\Sigma_{a_h})}{c_{a_h}^2}-\log|\Sigma_{b_h}|+\frac{\|\boldsymbol{\mu}_{b_h}\|^2+\mathrm{tr}(\Sigma_{b_h})}{c_{b_h}^2}\right)\\ &\quad+\frac{1}{\sigma^2}\left\|V-\sum_{h=1}^{H}\boldsymbol{\mu}_{b_h}\boldsymbol{\mu}_{a_h}^{\top}\right\|_{\mathrm{Fro}}^2\\ &\quad+\frac{1}{\sigma^2}\sum_{h=1}^{H}\left(\|\boldsymbol{\mu}_{a_h}\|^2\mathrm{tr}(\Sigma_{b_h})+\mathrm{tr}(\Sigma_{a_h})\|\boldsymbol{\mu}_{b_h}\|^2+\mathrm{tr}(\Sigma_{a_h})\mathrm{tr}(\Sigma_{b_h})\right). \qquad (67)\end{aligned}$$

We solve the following problem:

$$\begin{aligned}&\text{Given } (c_{a_h}^2,c_{b_h}^2)\in\mathbb{R}_{++}^2(^{\forall}h=1,\dots,H),\sigma^2\in\mathbb{R}_{++},\\ &\quad\min\quad \mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{\mu}_{a_h},\boldsymbol{\mu}_{b_h},\Sigma_{a_h},\Sigma_{b_h};h=1,\dots,H\}) \qquad (68)\\ &\quad\text{s.t. } \boldsymbol{\mu}_{a_h}\in\mathbb{R}^M,\boldsymbol{\mu}_{b_h}\in\mathbb{R}^L,\Sigma_{a_h}\in\mathbb{S}_{++}^M,\Sigma_{b_h}\in\mathbb{S}_{++}^L(^{\forall}h=1,\dots,H). \qquad (69)\end{aligned}$$

First, we have the following lemma (its proof is given in Appendix G.4):

Lemma 10 *At least one minimizer always exists, and any minimizer is a stationary point.*

Given fixed $\{(\Sigma_{a_h},\Sigma_{b_h})\}$, the objective function (67) is of the same form as Equation (60) if we replace $\{(c_{a_h}^2,c_{b_h}^2)\}$ in Equation (60) with $\{(c_{a_h}'^2,c_{b_h}'^2)\}$ defined by

$$c_{a_h}'^2=\left(\frac{1}{c_{a_h}^2}+\frac{\mathrm{tr}(\Sigma_{b_h})}{\sigma^2}\right)^{-1}, \qquad (70)$$

$$c_{b_h}'^2=\left(\frac{1}{c_{b_h}^2}+\frac{\mathrm{tr}(\Sigma_{a_h})}{\sigma^2}\right)^{-1}. \qquad (71)$$

Therefore, Lemma 8 implies that the minimizers of $\boldsymbol{\mu}_{a_h}$ and $\boldsymbol{\mu}_{b_h}$ are parallel (or zero) to the singular vectors of V associated with the H largest singular values.[4] On the other hand, Lemma 10 guarantees that Equations (23)–(26), which together form a necessary and sufficient condition to be a stationary point, hold at any minimizer. Equations (25) and (26) suggest that Σ_{a_h} and Σ_{b_h} are proportional to I_M and I_L, respectively. Accordingly, any minimizer can be written as $\boldsymbol{\mu}_{a_h}=\mu_{a_h}\boldsymbol{\omega}_{a_h}$, $\boldsymbol{\mu}_{b_h}=\mu_{b_h}\boldsymbol{\omega}_{b_h}$, $\Sigma_{a_h}=\sigma_{a_h}^2I_M$, and $\Sigma_{b_h}=\sigma_{b_h}^2I_L$, where μ_{a_h}, μ_{b_h}, $\sigma_{a_h}^2$, and $\sigma_{b_h}^2$ are scalars. This allows us to decompose the problem (68) into H separate problems: for $h=1,\dots,H$,

$$\begin{aligned}&\text{Given } (c_{a_h}^2,c_{b_h}^2)\in\mathbb{R}_{++}^2,\sigma^2\in\mathbb{R}_{++},\\ &\quad\min\quad \mathcal{L}_h^{\mathrm{VB}}(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2)\\ &\quad\text{s.t. } (\mu_{a_h},\mu_{b_h})\in\mathbb{R}^2,(\sigma_{a_h}^2,\sigma_{b_h}^2)\in\mathbb{R}_{++}^2, \qquad (72)\end{aligned}$$

4. As in Appendix B, we regard the orthogonal solution of the form (63) and (64) as the representative of the *equivalent* solutions, and focus on it. See Lemma 9 and its subsequent paragraph.

where

$$\mathcal{L}_h^{\mathrm{VB}}(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2) = -M\log\sigma_{a_h}^2 + \frac{\mu_{a_h}^2+M\sigma_{a_h}^2}{c_{a_h}^2} - L\log\sigma_{b_h}^2 + \frac{\mu_{b_h}^2+L\sigma_{b_h}^2}{c_{b_h}^2} - \frac{2}{\sigma^2}\gamma_h\mu_{a_h}\mu_{b_h} + \frac{1}{\sigma^2}\left(\mu_{a_h}^2+M\sigma_{a_h}^2\right)\left(\mu_{b_h}^2+L\sigma_{b_h}^2\right). \tag{73}$$

Moreover, the necessary and sufficient condition (23)–(26) is reduced to

$$\mu_{a_h} = \frac{1}{\sigma^2}\sigma_{a_h}^2\gamma_h\mu_{b_h}, \tag{74}$$

$$\mu_{b_h} = \frac{1}{\sigma^2}\sigma_{b_h}^2\gamma_h\mu_{a_h}, \tag{75}$$

$$\sigma_{a_h}^2 = \sigma^2\left(\mu_{b_h}^2+L\sigma_{b_h}^2+\frac{\sigma^2}{c_{a_h}^2}\right)^{-1}, \tag{76}$$

$$\sigma_{b_h}^2 = \sigma^2\left(\mu_{a_h}^2+M\sigma_{a_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right)^{-1}. \tag{77}$$

We use the following definition:

$$\widehat{\gamma}_h = \mu_{a_h}\mu_{b_h}, \tag{78}$$

Note that Equations (27) and (78) imply that the VB solution $\widehat{U}^{\mathrm{VB}}$ can be expressed as

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

Equations (74) and (75) imply that μ_{a_h} and μ_{b_h} have the same sign (or both are zero), since $\gamma_h \geq 0$ by definition. Therefore, Equation (78) yields

$$\widehat{\gamma}_h \geq 0.$$

In the following, we investigate two types of stationary points. We say that $(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2) = (\mathring{\mu}_{a_h},\mathring{\mu}_{b_h},\mathring{\sigma}_{a_h}^2,\mathring{\sigma}_{b_h}^2)$ is a *null* stationary point if it is a stationary point resulting in the null output ($\widehat{\gamma}_h = \mathring{\mu}_{a_h}\mathring{\mu}_{b_h} = 0$). On the other hand, we say that $(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2) = (\breve{\mu}_{a_h},\breve{\mu}_{b_h},\breve{\sigma}_{a_h}^2,\breve{\sigma}_{b_h}^2)$ is a *positive* stationary point if it is a stationary point resulting in a positive output ($\widehat{\gamma}_h = \breve{\mu}_{a_h}\breve{\mu}_{b_h} > 0$).

Let

$$\widehat{\eta}_h = \sqrt{\left(\mu_{a_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right)\left(\mu_{b_h}^2+\frac{\sigma^2}{c_{a_h}^2}\right)}. \tag{79}$$

The explicit form of the *null* stationary point is derived as follows (its proof is given in Appendix G.5):

Lemma 11 *The unique* null *stationary point always exists, and it is given by*

$$\mathring{\mu}_{a_h} = 0, \tag{80}$$

$$\mathring{\mu}_{b_h} = 0, \tag{81}$$

$$\mathring{\sigma}^2_{a_h} = \frac{c_{a_h}}{2Mc_{b_h}} \left\{ -\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} - c_{a_h}c_{b_h}(M-L) \right) + \sqrt{\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} - c_{a_h}c_{b_h}(M-L) \right)^2 + 4M\sigma^2} \right\}, \tag{82}$$

$$\mathring{\sigma}^2_{b_h} = \frac{c_{b_h}}{2Lc_{a_h}} \left\{ -\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} + c_{a_h}c_{b_h}(M-L) \right) + \sqrt{\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} + c_{a_h}c_{b_h}(M-L) \right)^2 + 4L\sigma^2} \right\}. \tag{83}$$

Next, we investigate the *positive* stationary points, assuming that $\mu_{a_h} \neq 0, \mu_{b_h} \neq 0$. Equations (74) and (75) suggest that no *positive* stationary point exists when $\gamma_h = 0$. Below, we focus on the case when $\gamma_h > 0$. Let

$$\widehat{\delta}_h = \frac{\mu_{a_h}}{\mu_{b_h}}. \tag{84}$$

We can transform the necessary and sufficient condition (74)–(77) as follows (its proof is given in Appendix G.6):

Lemma 12 *No* positive *stationary point exists if*

$$\gamma_h^2 \leq \sigma^2 M.$$

When

$$\gamma_h^2 > \sigma^2 M, \tag{85}$$

at least one positive *stationary point exists if and only if the following five equations*

$$\widehat{\eta}_h = \sqrt{\left(\widehat{\gamma}_h \widehat{\delta}_h + \frac{\sigma^2}{c_{b_h}^2} \right) \left(\widehat{\gamma}_h \widehat{\delta}_h^{-1} + \frac{\sigma^2}{c_{a_h}^2} \right)}, \tag{86}$$

$$\widehat{\eta}_h^2 = \left(1 - \frac{\sigma^2 L}{\gamma_h^2} \right) \left(1 - \frac{\sigma^2 M}{\gamma_h^2} \right) \gamma_h^2, \tag{87}$$

$$\sigma^2 \left(\frac{M\widehat{\delta}_h}{c_{a_h}^2} - \frac{L}{c_{b_h}^2 \widehat{\delta}_h} \right) = (M-L)(\gamma_h - \widehat{\gamma}_h), \tag{88}$$

$$\sigma^2_{a_h} = \frac{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 - \sigma^2(M-L))^2 + 4M\sigma^2\widehat{\eta}_h^2}}{2M(\widehat{\gamma}_h \widehat{\delta}_h^{-1} + \sigma^2 c_{a_h}^{-2})}, \tag{89}$$

$$\sigma^2_{b_h} = \frac{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(M-L))^2 + 4L\sigma^2\widehat{\eta}_h^2}}{2L(\widehat{\gamma}_h \widehat{\delta}_h + \sigma^2 c_{b_h}^{-2})} \tag{90}$$

have a solution with respect to $(\widehat{\gamma}_h, \widehat{\delta}_h, \sigma_{a_h}^2, \sigma_{b_h}^2, \widehat{\eta}_h)$ *such that*

$$(\widehat{\gamma}_h, \widehat{\delta}_h, \sigma_{a_h}^2, \sigma_{b_h}^2, \widehat{\eta}_h) \in \mathbb{R}_{++}^5. \tag{91}$$

When a solution exists, the corresponding pair of positive *stationary points*

$$(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2) = (\pm\sqrt{\widehat{\gamma}_h \widehat{\delta}_h}, \pm\sqrt{\widehat{\gamma}_h \widehat{\delta}_h^{-1}}, \sigma_{a_h}^2, \sigma_{b_h}^2) \tag{92}$$

exist.

Then we obtain a simpler necessary and sufficient condition for existence of *positive* stationary points (its proof is given in Appendix G.7):

Lemma 13 *At least one* positive *stationary point exists if and only if Equation* (85) *holds and*

$$\widehat{\gamma}_h^2 + q_1(\widehat{\gamma}_h) \cdot \widehat{\gamma}_h + q_0 = 0 \tag{93}$$

has any positive real solution with respect to $\widehat{\gamma}_h$*, where*

$$q_1(\widehat{\gamma}_h) = \frac{-(M-L)^2(\gamma_h - \widehat{\gamma}_h) + (L+M)\sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h)^2 + \frac{4\sigma^4 LM}{c_{a_h}^2 c_{b_h}^2}}}{2LM}, \tag{94}$$

$$q_0 = \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} - \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2. \tag{95}$$

Any positive solution $\widehat{\gamma}_h$ *satisfies*

$$0 < \widehat{\gamma}_h < \gamma_h. \tag{96}$$

Equation (96) guarantees that

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

Recall that a quadratic equation

$$\widehat{\gamma}^2 + q_1\widehat{\gamma} + q_0 = 0 \text{ for } q_1 > 0 \tag{97}$$

has only one positive solution when $q_0 < 0$ (otherwise no positive solution exists) (see Figure 11). The condition for the negativity of Equation (95) leads to the following lemma:

Lemma 14 *At least one* positive *stationary point exists if and only if*

$$\gamma_h^2 > \sigma^2 M \qquad \textit{and} \qquad \sqrt{\left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}} > 0. \tag{98}$$

The following lemma also holds (its proof is given in Appendix G.8):

Lemma 15 *Equation* (98) *holds if and only if*

$$\gamma_h > \widetilde{\gamma}_h^{\mathrm{VB}},$$

where $\widetilde{\gamma}_h^{VB}$ *is defined by Equation* (30)*.*

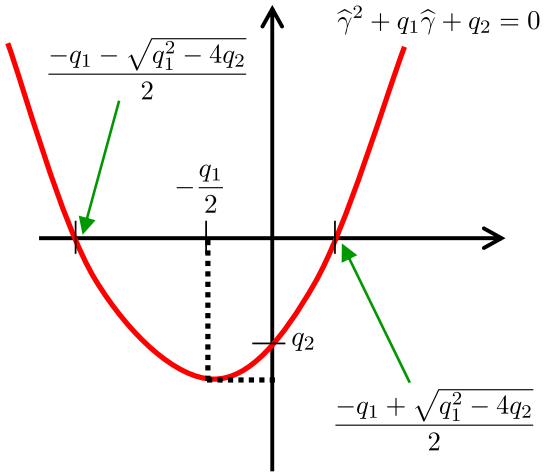

Figure 11: Quadratic function $f(\widehat{\gamma}) = \widehat{\gamma}^2 + q_1\widehat{\gamma} + q_0$, where $q_1 > 0$ and $q_0 < 0$.

Combining Lemma 10 and Lemma 14 together, we conclude that the *null* stationary point (which always exists) is the minimizer when Equation (98) does not hold. On the other hand, when a *positive* stationary point exists, we have to clarify which stationary point is the minimum. The following lemma holds (its proof is given in Appendix G.9).

Lemma 16 *The* null *stationary point is a saddle point when any* positive *stationary point exists.*

Combining Lemma 10, Lemma 14, and Lemma 16 together, we obtain the following lemma:

Lemma 17 *When Equation* (98) *holds, the minimizers consist of* positive *stationary points. Otherwise, the minimizer is the* null *stationary point.*

Combining Lemma 15 and Lemma 17 completes the proof of Theorem 3.

Finally, we derive bounds of the *positive* stationary points (its proof is given in Appendix G.10):

Lemma 18 *Equations* (28) *and* (31) *hold for any* positive *stationary point.*

Combining Lemma 17 and Lemma 18 completes the proof of Theorem 2 and Theorem 4. ■

Appendix D. Proof of Corollary 1

From Equations (78) and (84), we have $\mu_{a_h}^2 = \widehat{\gamma}_h\widehat{\delta}_h$ and $\mu_{b_h}^2 = \widehat{\gamma}_h/\widehat{\delta}_h$. When $L = M$, $\widehat{\gamma}_h$ is expressed analytically by Equation (32) and $\widehat{\delta}_h = c_a/c_b$ follows from Equation (88). From these, we have Equations (33) and (34).

When $L = M$, Equations (137) and (138) are reduced to

$$\sigma_{a_h}^2 = \frac{\widehat{\eta}_h \sqrt{\widehat{\eta}_h^2 + 4\sigma^2 M} - \widehat{\eta}_h^2}{2M\left(\mu_{b_h}^2 + \sigma^2 / c_{a_h}^2\right)}, \tag{99}$$

$$\sigma_{b_h}^2 = \frac{\widehat{\eta}_h \sqrt{\widehat{\eta}_h^2 + 4\sigma^2 M} - \widehat{\eta}_h^2}{2M\left(\mu_{a_h}^2 + \sigma^2 / c_{b_h}^2\right)}. \tag{100}$$

Substituting Equation (79) into Equations (99) and (100) and using Equations (33) and (34) give Equations (35) and (36). Because of the symmetry of the objective function (73), the two *positive* stationary points (33)–(36) give the same objective value, which completes the proof. ■

Note that *equivalent* nonorthogonal (with respect to $\{\boldsymbol{\mu}_{a_h}\}$, as well as $\{\boldsymbol{\mu}_{b_h}\}$) solutions may exist in principle. We neglect such solutions, because they almost surely do not exist; Equations (70), (71), (35), and (36) together imply that any pair $\{(h,h'); h \neq h'\}$ such that $\max(\widehat{\gamma}_h^{\mathrm{VB}}, \widehat{\gamma}_{h'}^{\mathrm{VB}}) > 0$ and $c'_{a_h} c'_{b_h} = c'_{a_{h'}} c'_{b_{h'}}$ can exist only when $c_{a_h} c_{b_h} = c_{a_{h'}} c_{b_{h'}}$ and $\gamma_h = \gamma_{h'}$ (i.e., two singular values of a random matrix coincide with each other).

Appendix E. Proof of Theorem 5 and Theorem 6

The EVB estimator is the minimizer of the VB free energy (39). Neglecting constant terms, we define the objective function as follows:

$$\begin{aligned}
\mathcal{L}^{\mathrm{EVB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \Sigma_{a_h}, \Sigma_{b_h}, c_{a_h}^2, c_{b_h}^2\}) &= 2F_{\mathrm{VB}}(r|V, \{c_{a_h}^2, c_{b_h}^2\}) + \mathrm{Const.} \\
&= \sum_{h=1}^{H} \left(\log \frac{c_{a_h}^{2M}}{|\Sigma_{a_h}|} + \frac{\|\boldsymbol{\mu}_{a_h}\|^2 + \mathrm{tr}(\Sigma_{a_h})}{c_{a_h}^2} + \log \frac{c_{b_h}^2}{|\Sigma_{b_h}|} + \frac{\|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})}{c_{b_h}^2} \right) \\
&\quad + \frac{1}{\sigma^2} \left\| V - \sum_{h=1}^{H} \boldsymbol{\mu}_{b_h} \boldsymbol{\mu}_{a_h}^\top \right\|_{\mathrm{Fro}}^2 \\
&\quad + \frac{1}{\sigma^2} \sum_{h=1}^{H} \left(\|\boldsymbol{\mu}_{a_h}\|^2 \mathrm{tr}(\Sigma_{b_h}) + \mathrm{tr}(\Sigma_{a_h}) \|\boldsymbol{\mu}_{b_h}\|^2 + \mathrm{tr}(\Sigma_{a_h}) \mathrm{tr}(\Sigma_{b_h}) \right).
\end{aligned}$$

We solve the following problem:

$$\begin{aligned}
&\text{Given } \sigma^2 \in \mathbb{R}_{++}, \\
&\quad \min \quad \mathcal{L}^{\mathrm{EVB}}(\{\boldsymbol{\mu}_{a_h}, \boldsymbol{\mu}_{b_h}, \Sigma_{a_h}, \Sigma_{b_h}, c_{a_h}^2, c_{b_h}^2; h = 1, \ldots, H\}) && (101) \\
&\quad \text{s.t. } \boldsymbol{\mu}_{a_h} \in \mathbb{R}^M, \boldsymbol{\mu}_{b_h} \in \mathbb{R}^L, \Sigma_{a_h} \in \mathbb{S}_{++}^M, \Sigma_{b_h} \in \mathbb{S}_{++}^L, (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}_{++}^2 (^\forall h = 1, \ldots, H). && (102)
\end{aligned}$$

Define a partial minimization problem of (101) with fixed $\{c_{a_h}^2, c_{b_h}^2\}$:

$$\begin{aligned}
\tilde{\mathcal{L}}^{\mathrm{EVB}}(\{c_{a_h}^2, c_{b_h}^2\}) = \min_{(\boldsymbol{\mu}_{a_h}, \boldsymbol{\mu}_{b_h}, \Sigma_{a_h}, \Sigma_{b_h})} & \mathcal{L}_h^{\mathrm{EVB}}(\{\boldsymbol{\mu}_{a_h}, \boldsymbol{\mu}_{b_h}, \Sigma_{a_h}, \Sigma_{b_h}\}; \{c_{a_h}^2, c_{b_h}^2\}) && (103) \\
\text{s.t. } \boldsymbol{\mu}_{a_h} \in \mathbb{R}^M, \boldsymbol{\mu}_{b_h} \in \mathbb{R}^L, \Sigma_{a_h} \in \mathbb{S}_{++}^M, \Sigma_{b_h} \in \mathbb{S}_{++}^L & (^\forall h = 1, \ldots, H).
\end{aligned}$$

This is identical to the VB estimation problem (68), and therefore, we can use the results proved in Appendix C. According to Lemma 10, at least one solution of the problem (103) exists. Therefore, the following problem is equivalent to the original problem (101):

$$\min_{\{c_{a_h}^2, c_{b_h}^2\}} \tilde{\mathcal{L}}^{\mathrm{EVB}}(\{c_{a_h}^2, c_{b_h}^2\}) \tag{104}$$
$$\text{s.t. } (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}_{++}^2 (^\forall h = 1, \ldots, H).$$

We have proved in Appendix C that any solution of the problem (103) can be written as $\boldsymbol{\mu}_{a_h} = \mu_{a_h}\boldsymbol{\omega}_{a_h}$, $\boldsymbol{\mu}_{b_h} = \mu_{b_h}\boldsymbol{\omega}_{b_h}$, $\Sigma_{a_h} = \sigma_{a_h}^2 I_M$, and $\Sigma_{b_h} = \sigma_{b_h}^2 I_L$, where μ_{a_h}, μ_{b_h}, $\sigma_{a_h}^2$, and $\sigma_{b_h}^2$ are scalars. This allows us to decompose the problem (101) into H separate problems: for $h = 1, \ldots, H$,

$$\begin{aligned} &\text{Given } \sigma^2 \in \mathbb{R}_{++}, \\ &\quad \min \quad \mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \\ &\quad \text{s.t. } (\mu_{a_h}, \mu_{b_h}) \in \mathbb{R}^2, (\sigma_{a_h}^2, \sigma_{b_h}^2) \in \mathbb{R}_{++}^2, (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}_{++}^2, \end{aligned} \tag{105}$$

where

$$\begin{aligned} \mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) = M \log \frac{c_{a_h}^2}{\sigma_{a_h}^2} + \frac{\mu_{a_h}^2 + M\sigma_{a_h}^2}{c_{a_h}^2} + L \log \frac{c_{b_h}^2}{\sigma_{b_h}^2} + \frac{\mu_{b_h}^2 + L\sigma_{b_h}^2}{c_{b_h}^2} \\ - \frac{2}{\sigma^2}\gamma_h \mu_{a_h}\mu_{b_h} + \frac{1}{\sigma^2}\left(\mu_{a_h}^2 + M\sigma_{a_h}^2\right)\left(\mu_{b_h}^2 + L\sigma_{b_h}^2\right). \end{aligned} \tag{106}$$

Let

$$\kappa = \begin{cases} \sigma^2 \left(\sqrt{\left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h \right)^{-1} & \text{if } \gamma_h > \sqrt{\sigma^2 M}, \\ \infty & \textit{otherwise}. \end{cases}$$

We divide the domain (105) into two regions (see Figure 12):

$$\mathring{\mathcal{R}} = \left\{ (\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}^2 \times \mathbb{R}_{++}^2 \times \mathbb{R}_{++}^2 ; c_{a_h} c_{b_h} \le \kappa \right\}, \tag{107}$$

$$\check{\mathcal{R}} = \left\{ (\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}^2 \times \mathbb{R}_{++}^2 \times \mathbb{R}_{++}^2 ; c_{a_h} c_{b_h} > \kappa \right\}. \tag{108}$$

Below, we will separately investigate the infimum of $\mathcal{L}_h^{\mathrm{EVB}}$ over $\mathring{\mathcal{R}}$,

$$\mathring{\underline{\mathcal{L}}}_h^{\mathrm{EVB}} = \inf_{(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \in \mathring{\mathcal{R}}} \mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2), \tag{109}$$

and the infimum over $\check{\mathcal{R}}$,

$$\check{\underline{\mathcal{L}}}_h^{\mathrm{EVB}} = \inf_{(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \in \check{\mathcal{R}}} \mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2).$$

Rigorously speaking, no minimizer over $\mathring{\mathcal{R}}$ exists. To make discussion simple, we approximate $\mathring{\mathcal{R}}$ by its subregion with an arbitrary accuracy; for any ε ($0 < \varepsilon < \kappa$), we define an ε-margin subregion of $\mathring{\mathcal{R}}$:

$$\mathring{\mathcal{R}}_\varepsilon = \left\{ (\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2, c_{a_h}^2, c_{b_h}^2) \in \mathring{\mathcal{R}} ; c_{a_h} c_{b_h} \ge \varepsilon \right\}.$$

Then the following lemma holds (its proof is given in Appendix G.11):

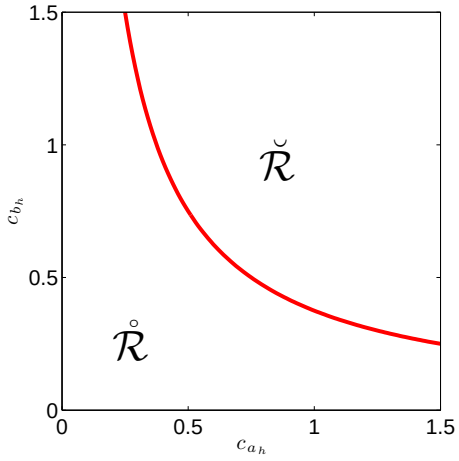

Figure 12: Division of the domain, defined by Equations (107) and (108), when $\gamma = 3, M = L = \sigma^2 = 1$. The hyperbolic boundary belongs to $\mathring{\mathcal{R}}$.

Lemma 19 *The minimizer over $\mathring{\mathcal{R}}_\varepsilon$ is given by*

$$\mathring{\mu}_{a_h} = 0, \tag{110}$$

$$\mathring{\mu}_{b_h} = 0, \tag{111}$$

$$\mathring{\sigma}^2_{a_h} = \frac{1}{2M}\left\{-\left(\frac{\sigma^2}{\varepsilon} - \varepsilon(M-L)\right) + \sqrt{\left(\frac{\sigma^2}{\varepsilon} - \varepsilon(M-L)\right)^2 + 4M\sigma^2}\right\}, \tag{112}$$

$$\mathring{\sigma}^2_{b_h} = \frac{1}{2L}\left\{-\left(\frac{\sigma^2}{\varepsilon} + \varepsilon(M-L)\right) + \sqrt{\left(\frac{\sigma^2}{\varepsilon} + \varepsilon(M-L)\right)^2 + 4L\sigma^2}\right\}, \tag{113}$$

$$\mathring{c}^2_{a_h} = \varepsilon, \tag{114}$$

$$\mathring{c}^2_{b_h} = \varepsilon, \tag{115}$$

and the infimum (109) *over $\mathring{\mathcal{R}}$ is given by*

$$\underline{\mathring{L}}_h^{\mathrm{EVB}} = L + M. \tag{116}$$

Note that Equations (110) and (111) result in the null output ($\widehat{\gamma}_h = \mathring{\mu}_{a_h}\mathring{\mu}_{b_h} = 0$). Accordingly, we call the minimizer (110)–(115) over $\mathring{\mathcal{R}}_\varepsilon$ the *null* (approximated) local minimizer.

On the other hand, we call any stationary point resulting in a *positive* output ($\widehat{\gamma}_h = \breve{\mu}_{a_h}\breve{\mu}_{b_h} > 0$) a *positive* stationary point. The following lemma holds (its proof is given in Appendix G.12):

Lemma 20 *Any* positive *stationary point lies in $\breve{\mathcal{R}}$.*

If

$$\underline{\mathring{L}}_h^{\mathrm{EVB}} < \underline{\breve{L}}_h^{\mathrm{EVB}}, \tag{117}$$

the *null* local minimizer is global over the whole domain (105) (more accurately, over $\mathring{\mathcal{R}}_\varepsilon \cup \check{\mathcal{R}}$ for any $0 < \varepsilon < \kappa$). If

$$\mathring{\underline{\mathcal{L}}}_h^{\mathrm{EVB}} \geq \check{\underline{\mathcal{L}}}_h^{\mathrm{EVB}}, \tag{118}$$

the global minimizers consist of *positive* stationary points, as the following lemma states (its proof is given in Appendix G.13):

Lemma 21 *When Equation* (118) *holds, the global minimizers consist of* positive *stationary points.*

Now, we look for the *positive* stationary points. According to Lemma 20, we can assume that Equation (98) holds. Equations (40) and (41) are reduced to

$$c_{a_h}^2 = \frac{\mu_{a_h}^2 + M\sigma_{a_h}^2}{M}, \tag{119}$$

$$c_{b_h}^2 = \frac{\mu_{b_h}^2 + L\sigma_{b_h}^2}{L}. \tag{120}$$

Then, Equations (74)–(77), (119), and (120) form a necessary and sufficient condition to be a stationary point of the objective function (106). Solving these equations, we have the following lemma (its proof is given in Appendix G.14):

Lemma 22 *At least one* positive *stationary point exists if and only if*

$$\gamma_h^2 \geq (\sqrt{L} + \sqrt{M})^2\sigma^2. \tag{121}$$

At any positive *stationary point, $c_{a_h}^2 c_{b_h}^2$ is given either by*

$$c_{a_h}^2 c_{b_h}^2 = \check{c}_{a_h}^2 \check{c}_{b_h}^2 = \frac{\left(\gamma_h^2 - (L+M)\sigma^2\right) + \sqrt{\left(\gamma_h^2 - (L+M)\sigma^2\right)^2 - 4LM\sigma^4}}{2LM}, \tag{122}$$

or by

$$c_{a_h}^2 c_{b_h}^2 = \acute{c}_{a_h}^2 \acute{c}_{b_h}^2 = \frac{\left(\gamma_h^2 - (L+M)\sigma^2\right) - \sqrt{\left(\gamma_h^2 - (L+M)\sigma^2\right)^2 - 4LM\sigma^4}}{2LM}. \tag{123}$$

We categorize the *positive* stationary points into two groups, based on the above two solutions of $c_{a_h}^2 c_{b_h}^2$; we say that a stationary point satisfying Equation (122) is a *large positive* stationary point, and one satisfying Equation (123) is a *small positive* stationary point. Note that, when

$$\gamma_h^2 = (\sqrt{L} + \sqrt{M})^2\sigma^2, \tag{124}$$

it holds that $\check{c}_{a_h}^2 \check{c}_{b_h}^2 = \acute{c}_{a_h}^2 \acute{c}_{b_h}^2$, and therefore, the *large positive* stationary points and the *small positive* stationary points coincide with each other. The following lemma allows us to focus on the *large positive* stationary points (its proof is given in Appendix G.15.):

Lemma 23 *When*

$$\gamma_h^2 > (\sqrt{L} + \sqrt{M})^2\sigma^2, \tag{125}$$

any small positive *stationary point is a saddle point.*

Summarizing Lemmas 19–23, we have the following lemma:

Lemma 24 *When Equation* (121) *holds, there are two possibilities: that the global minimizers consist of* large positive *stationary points (in the case when Equation* (118) *holds); or that the global minimizer is the* null *local minimizer (in the case when Equation* (117) *holds). When Equation* (121) *does not hold, the global minimizer is the* null *local minimizer.*

Hereafter, we assume that Equation (121) holds. We like to clarify when Equation (118) holds, so that *large positive* stationary points become global minimizers. The EVB objective function (106) is substantially more complex (see Appendix H for illustration) than the VB objective function (73) where the *null* stationary point turns from the global minimum to a saddle point no sooner than any *positive* stationary point arises.

Below, we derive a sufficient condition for any *large positive* stationary point to give a lower objective value than $\mathring{\underline{L}}_h^{\mathrm{EVB}}$. We evaluate the difference between the objectives:

$$\Delta_h(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) = \mathcal{L}_h^{\mathrm{EVB}}(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) - \mathring{\underline{L}}_h^{\mathrm{EVB}}. \tag{126}$$

If $\Delta_h(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) \leq 0$, Equation (118) holds. We obtain the following lemma (its proof is given in Appendix G.16.):

Lemma 25 $\Delta_h(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h})$ *is upper-bounded as*

$$\Delta_h(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) < M\psi(\alpha, \beta), \tag{127}$$

where

$$\psi(\alpha, \beta) = \log\beta + \alpha\log\left(\frac{\beta - (1-\alpha)}{\alpha}\right) + (1-\alpha) + \frac{2}{\sqrt{1 - \frac{(\alpha+\sqrt{\alpha}+1)}{\beta}}} - \beta, \tag{128}$$

$$\alpha = \frac{L}{M}, \tag{129}$$

$$\beta = \frac{\gamma_h^2}{M\sigma^2}. \tag{130}$$

Furthermore, the following lemma states that $\psi(\alpha, \beta)$ is negative when β is large enough (its proof is given in Appendix G.17.):

Lemma 26 $\psi(\alpha, \beta) < 0$ *for any* $0 < \alpha \leq 1$ *and* $\beta \geq 7$.

Combining Lemma 24 and Lemma 25, we obtain the following lemma:

Lemma 27 *When the condition* (127) *holds, the global minimizers consist of* large positive *stationary points.*

Combining Lemma 26 and Lemma 27, we obtain the following lemma:

Lemma 28 *When* $\beta \geq 7$*, the global minimizers consist of* large positive *stationary points.*

Finally, we derive bounds of the *large positive* stationary points (its proof is given in Appendix G.18):

Lemma 29 *Equations* (46), (47), *and* (48) *hold for any* large positive *stationary point.*

Combining Lemma 24, Lemma 28, and Lemma 29 completes the proof of Theorem 5. Combining Lemma 24 and Lemma 29 completes the proof of Theorem 6.

■

Appendix F. Proof of Corollary 2

Assume that $L = M$. When $\gamma_h \geq 2\sqrt{M}$, Lemma 22 guarantees that at least one *large positive* stationary point exists. In this case, Equation (122) leads to

$$\check{c}_{a_h}\check{c}_{b_h} = \frac{\gamma_h}{M}\rho_+. \tag{131}$$

Its inverse can be written as

$$\frac{1}{\check{c}_{a_h}\check{c}_{b_h}} = \frac{\gamma_h}{\sigma^2}\rho_-.$$

Corollary 1 provides the exact values for the *positive* stationary points $(\check{\mu}_{a_h}, \check{\mu}_{b_h}, \check{\sigma}^2_{a_h}, \check{\sigma}^2_{b_h})$, given $(\check{c}^2_{a_h}, \check{c}^2_{b_h}) = (\check{c}_{a_h}\check{c}_{b_h}, \check{c}_{a_h}\check{c}_{b_h})$. Therefore, we can compute the exact value of the difference (126) of the objective values between the *large positive* stationary points and the *null* local minimizer:

$$\begin{aligned}
\Delta_h &= 2M\log\left(\frac{\gamma_h}{M\sigma^2}\check{\mu}_{a_h}\check{\mu}_{b_h}+1\right) + \frac{1}{\sigma^2}\left(-2\gamma_h\check{\mu}_{a_h}\check{\mu}_{b_h} + M^2\check{c}^2_{a_h}\check{c}^2_{b_h}\right) \\
&= 2M\left\{\log\left(\frac{\gamma_h^2}{M\sigma^2} - \frac{\gamma_h}{M\check{c}_{a_h}\check{c}_{b_h}}\right) - \left(\frac{\gamma_h^2}{M\sigma^2} - \frac{\gamma_h}{M\check{c}_{a_h}\check{c}_{b_h}}\right) + \left(1 + \frac{nM}{2\sigma^2}\check{c}^2_{a_h}\check{c}^2_{b_h}\right)\right\} \\
&= 2M\varphi(\gamma_h).
\end{aligned}$$

Here, the first equation directly comes from Equation (172), and the last equation is obtained by substituting Equation (131) into the second equation.

According to Lemma 24, when $\gamma_h \geq 2\sqrt{M}$ and $\Delta_h \leq 0$, the EVB solutions consist of *large positive* stationary points; otherwise, the EVB solution is the *null* local minimizer. Using Equations (114), (115), and (131), we obtain Equation (51). Equation (52) follows Lemma 26, because $\varphi(\gamma_h) = \Delta_h/(2M) < \psi(\alpha,\beta)/2$ for $\alpha = 1, \beta = \gamma_h^2/(M\sigma^2)$. ■

Appendix G. Proof of Lemmas

In this appendix, the proofs of all the lemmas are given.

G.1 Proof of Lemma 7

We minimize the left-hand side of Equation (61) with respect to A and B:

$$\begin{aligned}
&\min_{A,B}\left\{\mathrm{tr}(AC_A^{-1}A^\top) + \mathrm{tr}(BC_B^{-1}B^\top)\right\} \\
&\text{s.t. } BA^\top = \Omega_L\Gamma\Omega_R^\top.
\end{aligned} \tag{132}$$

We can remove the constraint by changing the variables as follows:

$$A \to \Omega_R\Gamma T^\top C_A^{1/2}, \qquad\qquad B \to \Omega_L T^{-1}C_A^{-1/2},$$

where T is a $H \times H$ non-singular matrix. Then, the problem (132) is rewritten as

$$\min_T\left\{\mathrm{tr}\left(T^\top T\Gamma^2\right) + \mathrm{tr}\left((TT^\top)^{-1}(C_AC_B)^{-1}\right)\right\}. \tag{133}$$

Let

$$T^{-1} = U_T D_T V_T^\top$$

be the singular value decomposition of T^{-1}, where $D_T = \mathrm{diag}(d_1, \ldots, d_H)$ ($\{d_h\}$ are in non-increasing order). Then, the problem (133) is written as

$$\min_{U_T, D_T, V_T} \left\{ \mathrm{tr}\left(U_T D_T^{-2} U_T^\top \Gamma^2 \right) + \mathrm{tr}\left(V_T D_T^2 V_T^\top (C_A C_B)^{-1} \right) \right\}. \tag{134}$$

The objective function in Equation (134) can be written with the doubly stochastic matrices

$$Q_U = U_T \bullet U_T,$$
$$Q_V = V_T \bullet V_T,$$

where $\bullet$ denotes the Hadamard product, as follows (Marshall et al., 2009):

$$(d_1^{-2}, \ldots, d_H^{-2}) Q_U (\widehat{\gamma}_1^2, \ldots, \widehat{\gamma}_H^2)^\top + (d_1^2, \ldots, d_H^2) Q_V ((c_{a_1} c_{b_1})^{-1}, \ldots, (c_{a_H} c_{b_H})^{-1})^\top.$$

Since $\{\widehat{\gamma}_h^2\}$ and $\{d_h^2\}$ are in non-increasing order, and $\{d_h^{-2}\}$ and $(c_{a_h} c_{b_h})^{-1}$ are in non-decreasing order, this is minimized when $Q_U = Q_V = I_H$ (which is attained with $U_T = V_T = I_H$) for any D_T.

Thus, the problem (134) is reduced to

$$\min_{\{d_h\}} \sum_{h=1}^H \left(\frac{\widehat{\gamma}_h^2}{d_h^2} + \frac{d_h^2}{(c_{a_h} c_{b_h})^2} \right).$$

This is minimized when $d_h^2 = \widehat{\gamma}_h c_{a_h} c_{b_h}$,[5] and the minimum coincides to the right-hand side of Equation (61), which completes the proof.

■

G.2 Proof of Lemma 8

It is known that the second term of Equation (60) is minimized when

$$A = (\sqrt{\gamma_1} \boldsymbol{\omega}_{a_1}, \ldots, \sqrt{\gamma_H} \boldsymbol{\omega}_{a_H}) T^\top,$$
$$B = (\sqrt{\gamma_1} \boldsymbol{\omega}_{b_1}, \ldots, \sqrt{\gamma_H} \boldsymbol{\omega}_{b_H}) T^{-1},$$

where T is any $H \times H$ non-singular matrix. Since the first term of Equation (60) does not depend on the directions of $\{\boldsymbol{a}_h, \boldsymbol{b}_h\}$, any minimizer can be written in the form of Equation (62) with $\{\widehat{\gamma}_h \geq 0\}$.

The degeneracy with respect to T is partly resolved by the first term of Equation (60). Suppose that we have obtained the best set of $\{\widehat{\gamma}_h\}$. Then, minimizing Equation (60) is equivalent to the following problem:

$$\begin{aligned} &\text{Given } \{\widehat{\gamma}_h \geq 0\}, \\ &\min_{A,B} \quad \left\{ \mathrm{tr}(A C_A^{-1} A^\top) + \mathrm{tr}(B C_B^{-1} B^\top) \right\} \\ &\text{s.t. } B A^\top = \sum_{h=1}^H \widehat{\gamma}_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top. \end{aligned} \tag{135}$$

5. If $\widehat{\gamma}_h = 0$, the minimum is attained by simply setting the corresponding column vectors of A and B to $(\boldsymbol{a}_h, \boldsymbol{b}_h) = (\boldsymbol{0}, \boldsymbol{0})$.

Lemma 7 guarantees that

$$\boldsymbol{a}_h = \sqrt{\frac{c_{a_h}}{c_{b_h}}\widehat{\gamma}_h}\boldsymbol{\omega}_{a_h},$$
$$\boldsymbol{b}_h = \sqrt{\frac{c_{b_h}}{c_{a_h}}\widehat{\gamma}_h}\boldsymbol{\omega}_{b_h},$$

give a solution for the problem (135) for any (so far unknown) set of $\{\widehat{\gamma}_h\}$, which completes the proof. ■

G.3 Proof of Lemma 9

Equation (60) can be written as

$$\mathcal{L}^{\mathrm{MAP}}(A,B) = \mathrm{tr}(AC_A^{-1}A^\top) + \mathrm{tr}(BC_B^{-1}B^\top) + \frac{1}{\sigma^2}\left\|V - BA^\top\right\|_{\mathrm{Fro}}^2.$$

This is invariant with respect to the transform

$$A \to A\Theta^\top,$$
$$B \to B\Theta^{-1},$$

since

$$\mathrm{tr}(A\Theta^\top C_A^{-1}\Theta A^\top) = \mathrm{tr}(AC_A^{-1/2}\Xi^\top C_A^{1/2}C_A^{-1}C_A^{1/2}\Xi C_A^{-1/2}A^\top) = \mathrm{tr}(AC_A^{-1}A^\top),$$
$$\mathrm{tr}(B\Theta^{-1}C_B^{-1}(\Theta^{-1})^\top B^\top) = \mathrm{tr}(BC_B^{-1/2}\Xi^\top C_B^{1/2}C_B^{-1}C_B^{1/2}\Xi C_B^{-1/2}B^\top) = \mathrm{tr}(BC_B^{-1}B^\top),$$
$$B\Theta^{-1}\Theta A = BA.$$

This completes the proof. ■

G.4 Proof of Lemma 10

Let

$$\Sigma_{a_h} = \sum_{m=1}^{M} \tau_m^{(a_h)} \boldsymbol{t}_m^{(a_h)} \boldsymbol{t}_m^{(a_h)\top},$$
$$\Sigma_{b_h} = \sum_{l=1}^{L} \tau_l^{(b_h)} \boldsymbol{t}_l^{(b_h)} \boldsymbol{t}_l^{(b_h)\top},$$

be the eigenvalue decompositions of Σ_{a_h} and Σ_{b_h} , where

$$\left(\tau_1^{(a_h)},\ldots,\tau_M^{(a_h)}\right) \in \mathbb{R}_{++}^M, \qquad \left(\tau_1^{(b_h)},\ldots,\tau_L^{(b_h)}\right) \in \mathbb{R}_{++}^L.$$

are the eigenvalues. Then, the objective function (67) is written as

$$\begin{aligned}
&\mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \tau_m^{(a_h)}, \tau_l^{(b_h)}\}) \\
&= \sum_{h=1}^{H} \left(- \sum_{m=1}^{M} \log \tau_m^{(a_h)} + \frac{\|\boldsymbol{\mu}_{a_h}\|^2 + \sum_{m=1}^{M} \tau_m^{(a_h)}}{c_{a_h}^2} - \sum_{l=1}^{L} \log \tau_l^{(b_h)} + \frac{\|\boldsymbol{\mu}_{b_h}\|^2 + \sum_{l=1}^{L} \tau_l^{(b_h)}}{c_{b_h}^2} \right) \\
&\quad + \frac{1}{\sigma^2} \left\| V - \sum_{h=1}^{H} \boldsymbol{\mu}_{b_h} \boldsymbol{\mu}_{a_h}^{\top} \right\|_{\mathrm{Fro}}^2 \\
&\quad + \frac{1}{\sigma^2} \sum_{h=1}^{H} \left(\|\boldsymbol{\mu}_{a_h}\|^2 \sum_{l=1}^{L} \tau_l^{(b_h)} + \sum_{m=1}^{M} \tau_m^{(a_h)} \|\boldsymbol{\mu}_{b_h}\|^2 + \left(\sum_{m=1}^{M} \tau_m^{(a_h)} \right) \left(\sum_{l=1}^{L} \tau_l^{(b_h)} \right) \right).
\end{aligned}$$

Since the second and the third terms are positive, this is lower-bounded as

$$\begin{aligned}
\mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \tau_m^{(a_h)}, \tau_l^{(b_h)}\}) &> \sum_{h=1}^{H} \left(\frac{\|\boldsymbol{\mu}_{a_h}\|^2}{c_{a_h}^2} + \sum_{m=1}^{M} \left(\frac{\tau_m^{(a_h)}}{c_{a_h}^2} - \log \frac{\tau_m^{(a_h)}}{c_{a_h}^2} \right) \right) \\
&\quad + \sum_{h=1}^{H} \left(\frac{\|\boldsymbol{\mu}_{b_h}\|^2}{c_{b_h}^2} + \sum_{l=1}^{L} \left(\frac{\tau_l^{(b_h)}}{c_{b_h}^2} - \log \frac{\tau_l^{(b_h)}}{c_{b_h}^2} \right) \right) - \sum_{h=1}^{H} \left(M \log c_{a_h}^2 + L \log c_{b_h}^2 \right). \qquad (136)
\end{aligned}$$

Focusing on the first term in Equation (136), we find that

$$\lim_{\|\mu_{a_h}\| \to \infty} \mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \tau_m^{(a_h)}, \tau_l^{(b_h)}\}) = \infty$$

for any h. Further,

$$\begin{aligned}
\lim_{\tau_m^{(a_h)} \to 0} \mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \tau_m^{(a_h)}, \tau_l^{(b_h)}\}) &= \infty, \\
\lim_{\tau_m^{(a_h)} \to \infty} \mathcal{L}^{\mathrm{VB}}(\{\boldsymbol{a}_h, \boldsymbol{b}_h, \tau_m^{(a_h)}, \tau_l^{(b_h)}\}) &= \infty,
\end{aligned}$$

for any (h,m), because $(x - \log x) \geq 1$ for any $x > 0$, $\lim_{x \to +0}(x - \log x) = \infty$, and $\lim_{x \to \infty}(x - \log x) = \infty$. The same holds for $\{\boldsymbol{\mu}_{b_h}\}$ and $\{\tau_l^{(b_h)}\}$ because of the second term in Equation (136). Consequently, the objective function (67) goes to infinity when approaching to any point on the boundary of the domain (69). Since the objective function (67) is differentiable in the domain, any minimizer is a stationary point. For any observation V, the objective function (67) can be finite, for example, when $\|\boldsymbol{\mu}_{a_h}\| = \|\boldsymbol{\mu}_{b_h}\| = 0, \Sigma_{a_h} = I_M, \Sigma_{b_h} = I_L$. Therefore, at least one minimizer always exists. ■

G.5 Proof of Lemma 11

Combining Equations (76) and (77) and eliminating $\sigma_{b_h}^2$, we obtain

$$M \left(\mu_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right) \sigma_{a_h}^4 + \left(\widehat{\eta}_h^2 - \sigma^2 (M - L) \right) \sigma_{a_h}^2 - \sigma^2 \left(\mu_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2} \right) = 0.$$

This has one positive and one negative solutions. Neglecting the negative one, we obtain

$$\sigma_{a_h}^2 = \frac{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 - \sigma^2(M-L))^2 + 4M\sigma^2\widehat{\eta}_h^2}}{2M(\mu_{b_h}^2 + \sigma^2 c_{a_h}^{-2})}. \tag{137}$$

Similarly, combining Equations (76) and (77) and eliminating $\sigma_{a_h}^2$, we obtain

$$\sigma_{b_h}^2 = \frac{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(M-L))^2 + 4L\sigma^2\widehat{\eta}_h^2}}{2L(\mu_{a_h}^2 + \sigma^2 c_{b_h}^{-2})}. \tag{138}$$

Note that Equations (137) and (138) are real and positive for any $(\mu_{a_h}, \mu_{b_h}) \in \mathbb{R}^2$ and $\widehat{\eta}_h \in \mathbb{R}_{++}$.

Let us focus on the *null* stationary points. Apparently, Equations (80) and (81) are necessary to satisfy Equations (74) and (75) and result in the *null* output $\widehat{\gamma}_h = \mathring{\mu}_{a_h}\mathring{\mu}_{b_h} = 0$. Substituting Equations (80) and (81) into Equations (137) and (138) leads to Equations (82) and (83).

■

G.6 Proof of Lemma 12

To prove the lemma, we transform the set of variables $(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2)$ to $(\widehat{\gamma}_h, \widehat{\delta}_h, \sigma_{a_h}^2, \sigma_{b_h}^2, \widehat{\eta}_h)$, and the necessary and sufficient condition (74)–(77) to (86)–(90). The transform (92) is obtained from the definitions (78) and (84), which we use in the following when necessary.

First we show that Equation (91) is necessary for any *positive* stationary point. $\widehat{\gamma}_h$ and $\widehat{\delta}_h$ must be positive because Equations (74) and (75) imply that μ_{a_h} and μ_{b_h} have the same sign. $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$ must be positive because of their original domain (72). $\widehat{\eta}_h$ must be positive by its definition (79).

Next, we obtain Equations (86)–(90) from Equations (74)–(77). Equation (86) simply comes from the definition (79) of the additional variable $\widehat{\eta}_h$, which we have introduced for convenience. Equations (89) and (90) are equivalent to Equations (137) and (138), which were derived from Equations (76) and (77) in Appendix G.5. Equations (87) and (88) are derived from Equations (74) and (75), as shown below.

Equations (137) and (138) can be rewritten as

$$\sigma_{a_h}^2 = \frac{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(L+M))^2 - 4\sigma^4 LM}}{2M(\mu_{b_h}^2 + \sigma^2 c_{a_h}^{-2})}, \tag{139}$$

$$\sigma_{b_h}^2 = \frac{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{(\widehat{\eta}_h^2 + \sigma^2(L+M))^2 - 4\sigma^4 LM}}{2L(\mu_{a_h}^2 + \sigma^2 c_{b_h}^{-2})}. \tag{140}$$

Substituting Equations (139) and (140) into Equations (74) and (75), respectively, we have

$$2\sigma^2 M\left(\mu_{b_h}^2+\frac{\sigma^2}{c_{a_h}^2}\right)\frac{\mu_{a_h}}{\mu_{b_h}} = \gamma_h\left\{-\left(\widehat{\eta}_h^2-\sigma^2(M-L)\right)+\sqrt{\left(\eta_h^2+\sigma^2(L+M)\right)^2-4\sigma^4 LM}\right\}, \tag{141}$$

$$2\sigma^2 L\left(\mu_{a_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right)\frac{\mu_{b_h}}{\mu_{a_h}} = \gamma_h\left\{-\left(\widehat{\eta}_h^2+\sigma^2(M-L)\right)+\sqrt{\left(\eta_h^2+\sigma^2(L+M)\right)^2-4\sigma^4 LM}\right\}. \tag{142}$$

Subtraction of Equation (142) from Equation (141) gives

$$2\sigma^2(M-L)\mu_{a_h}\mu_{b_h}+2\sigma^4\left(\frac{M\mu_{a_h}}{c_{a_h}^2\mu_{b_h}}-\frac{L\mu_{b_h}}{c_{b_h}^2\mu_{a_h}}\right)=2\sigma^2(M-L)\gamma_h,$$

which is equivalent to Equation (88).

The last condition (87) is derived by multiplying Equations (141) and (142) (of which the both sides are positive):

$$4\sigma^4 LM\widehat{\eta}_h^2=\gamma_h^2\left(2\widehat{\eta}_h^4+2\widehat{\eta}_h^2\sigma^2(L+M)-2\widehat{\eta}_h^2\sqrt{(\widehat{\eta}_h^2+\sigma^2(L+M))^2-4\sigma^4 LM}\right).$$

Dividing both sides by $2\widehat{\eta}_h^2\gamma_h^2$ (>0), we have

$$\sqrt{(\widehat{\eta}_h^2+\sigma^2(L+M))^2-4\sigma^4 LM}=\widehat{\eta}_h^2+\sigma^2(L+M)-\frac{2\sigma^4 LM}{\gamma_h^2}. \tag{143}$$

Note that the left-hand side of Equation (143) is always real and positive since

$$\begin{aligned}(\widehat{\eta}_h^2+\sigma^2(L+M))^2-4\sigma^4 LM &= (\widehat{\eta}_h^2-\sigma^2(M-L))^2+4M\sigma^2\widehat{\eta}_h^2\\ &>0.\end{aligned}$$

Therefore, the right-hand side of Equation (143) is non-negative when Equation (143) holds:

$$\widehat{\eta}_h^2+\sigma^2(L+M)-\frac{2\sigma^4 LM}{\gamma_h^2}\geq 0. \tag{144}$$

To obtain Equation (87) from Equation (143), we square Equation (143):

$$(\widehat{\eta}_h^2+\sigma^2(L+M))^2-4\sigma^4 LM=\left(\widehat{\eta}_h^2+\sigma^2(L+M)-\frac{2\sigma^4 LM}{\gamma_h^2}\right)^2. \tag{145}$$

Note that this is equivalent to Equation (143) only when Equation (144) holds. Equation (145) leads to

$$\frac{\sigma^4 LM}{\gamma_h^2}-(\widehat{\eta}_h^2+\sigma^2(L+M))+\gamma_h^2=0.$$

Solving this with respect to $\widehat{\eta}_h^2$ results in Equation (87). Equation (87) cannot hold with any real and positive value of $\widehat{\eta}_h$ when $\sigma^2 L \leq \gamma_h^2 \leq \sigma^2 M$. Further, substituting Equation (87) into Equation (144) gives

$$\gamma_h^2 - \frac{\sigma^4 LM}{\gamma_h^2} \geq 0.$$

Therefore, Equation (87) satisfies Equation (144) only when $\gamma_h^2 \geq \sigma^2\sqrt{LM}$. Accordingly, when Equation (85) holds, Equation (87) is equivalent to Equation (143). Otherwise, Equation (143) cannot hold, and no *positive* stationary point exists. ■

G.7 Proof of Lemma 13

Squaring both sides of Equation (86) (which are positive) and substituting Equation (87) into it, we have

$$\begin{aligned}\widehat{\gamma}_h^2 + \frac{\sigma^2}{c_{a_h} c_{b_h}} \left(\frac{c_{b_h} \widehat{\delta}_h}{c_{a_h}} + \frac{c_{a_h}}{c_{b_h} \widehat{\delta}_h} \right) \widehat{\gamma}_h \\ + \left(\frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} - \left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right) \left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right) \gamma_h^2 \right) = 0. \end{aligned} \tag{146}$$

Multiplying both sides of Equation (88) by $\widehat{\delta}_h$ (> 0) and solving it with respect to $\widehat{\delta}_h$, we obtain

$$\widehat{\delta}_h = \frac{(M-L)(\gamma_h - \widehat{\gamma}_h) + \sqrt{(M-L)^2(\gamma_h - \widehat{\gamma}_h)^2 + \frac{4\sigma^4 LM}{c_{a_h}^2 c_{b_h}^2}}}{2\sigma^2 M c_{a_h}^{-2}} \tag{147}$$

as a positive solution. We neglect the other solution, since it is negative. Substituting Equation (147) into Equation (146) gives Equation (93). Thus, we have transformed the necessary and sufficient condition Equations (86)–(90) to (93), (87), (147), (89), and (90). This proves the necessity.

Assume that Equation (85) holds and a positive real solution $\widehat{\gamma}_h$ of Equation (93) exists. Then, a positive real $\widehat{\eta}_h$ satisfying Equation (87) exists. For any existing $(\widehat{\gamma}_h, \widehat{\eta}_h) \in \mathbb{R}_{++}^2$, a positive real $\widehat{\delta}_h$ satisfying Equation (147) exists. For any existing $(\widehat{\gamma}_h, \widehat{\delta}_h, \widehat{\eta}_h) \in \mathbb{R}_{++}^3$, positive real $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$ satisfying Equations (89) and (90) exist. Thus, whenever a positive real solution $\widehat{\gamma}_h$ of Equation (93) exists, the corresponding point $(\widehat{\gamma}_h, \widehat{\delta}_h, \sigma_{a_h}^2, \sigma_{b_h}^2, \widehat{\eta}_h) \in \mathbb{R}_{++}^5$ satisfying the necessary and sufficient condition (93), (87), (147), (89), and (90) exists. This proves the sufficiency.

Finally, suppose that we obtain a solution satisfying Equations (86)–(90) in the domain (91). Then, Equation (87) implies that

$$\gamma_h > \widehat{\eta}_h.$$

Moreover, ignoring the positive terms $\sigma^2/c_{b_h}^2$ and $\sigma^2/c_{a_h}^2$ in Equation (86), we have

$$\widehat{\eta}_h > \widehat{\gamma}_h.$$

Therefore, Equation (96) holds. ■

G.8 Proof of Lemma 15

Assume that $\gamma_h^2 > \sigma^2 M$. Then, the second inequality in Equation (98) holds if and only if

$$\left(1 - \frac{\sigma^2 L}{\gamma_h^2}\right)\left(1 - \frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} > 0.$$

The left-hand side can be factorized as

$$\gamma_h^{-2}\left(\gamma_h^2 - \left(\kappa + \sqrt{\kappa^2 - LM\sigma^4}\right)\right)\left(\gamma_h^2 - \left(\kappa - \sqrt{\kappa^2 - LM\sigma^4}\right)\right) > 0, \tag{148}$$

where

$$\kappa = \frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2}.$$

Since

$$\kappa - \sqrt{\kappa^2 - LM\sigma^4} < M\sigma^2 < \kappa + \sqrt{\kappa^2 - LM\sigma^4},$$

Equation (148) holds if and only if

$$\gamma_h^2 > \kappa + \sqrt{\kappa^2 - LM\sigma^4},$$

which leads to Equation (30). ■

G.9 Proof of Lemma 16

We show that the Hessian of the objective function (73) has at least one negative and one positive eigenvalues at the *null* stationary point, when any *positive* stationary point exists. We only focus on the 2-dimensional subspace spanned by (μ_{a_h}, μ_{b_h}). The partial derivatives of Equation (73) are given by

$$\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{VB}}}{\partial \mu_{a_h}} = \frac{\mu_{a_h}}{c_{a_h}^2} + \left(\frac{-\gamma_h \mu_{b_h} + (\mu_{b_h}^2 + L\sigma_{b_h}^2)\mu_{a_h}}{\sigma^2}\right),$$
$$\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{VB}}}{\partial \mu_{b_h}} = \frac{\mu_{b_h}}{c_{b_h}^2} + \left(\frac{-\gamma_h \mu_{a_h} + (\mu_{a_h}^2 + M\sigma_{a_h}^2)\mu_{b_h}}{\sigma^2}\right).$$

Then, the Hessian is given by

$$\begin{aligned}\frac{1}{2}\mathcal{H}^{\mathrm{VB}} &= \begin{pmatrix} \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{VB}}}{(\partial \mu_{a_h})^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{VB}}}{\partial \mu_{a_h}\partial \mu_{b_h}} \\ \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{VB}}}{\partial \mu_{a_h}\partial \mu_{b_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{VB}}}{(\partial \mu_{b_h})^2} \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \frac{\sigma^2}{c_{a_h}^2} + (\mu_{b_h}^2 + L\sigma_{b_h}^2) & -\gamma_h + 2\mu_{a_h}\mu_{b_h} \\ -\gamma_h + 2\mu_{a_h}\mu_{b_h} & \frac{\sigma^2}{c_{b_h}^2} + (\mu_{a_h}^2 + M\sigma_{a_h}^2) \end{pmatrix}. \end{aligned} \tag{149}$$

The determinant of Equation (149) is written as

$$\left|\frac{1}{2}\mathcal{H}^{\mathrm{VB}}\right| = \frac{1}{\sigma^4}\left(\frac{\sigma^2}{c_{a_h}^2} + (\mu_{b_h}^2 + L\sigma_{b_h}^2)\right)\left(\frac{\sigma^2}{c_{b_h}^2} + (\mu_{a_h}^2 + M\sigma_{a_h}^2)\right) - \frac{1}{\sigma^4}(2\mu_{a_h}\mu_{b_h} - \gamma_h)^2$$
$$= \frac{1}{\sigma_{a_h}^2\sigma_{b_h}^2} - \frac{1}{\sigma^4}(2\mu_{a_h}\mu_{b_h} - \gamma_h)^2, \tag{150}$$

where Equations (76) and (77) are used in the second equation.

The determinant (150) of the Hessian at the *null* stationary point, given by Equations (80)–(83), is written as

$$\left|\frac{1}{2}\mathring{\mathcal{H}}^{\mathrm{VB}}\right| = \frac{1}{\mathring{\sigma}_{a_h}^2\mathring{\sigma}_{b_h}^2} - \frac{1}{\sigma^4}\gamma_h^2. \tag{151}$$

Assume the existence of any *positive* stationary point, for which it holds that

$$\gamma_h^2 = \frac{\sigma^4}{\breve{\sigma}_{a_h}^2\breve{\sigma}_{b_h}^2}. \tag{152}$$

This is obtained by substituting Equation (75) into Equation (74) and dividing both sides by $\breve{\mu}_{a_h}\breve{\sigma}_{a_h}^2\breve{\sigma}_{b_h}^2/\sigma^4$ (>0). Note that Equation (152) is not required for the *null* stationary point where $\mathring{\mu}_{a_h} = 0$. Substituting Equation (152) into Equation (151), we have

$$\left|\frac{1}{2}\mathring{\mathcal{H}}^{\mathrm{VB}}\right| = \frac{1}{\mathring{\sigma}_{a_h}^2\mathring{\sigma}_{b_h}^2} - \frac{1}{\breve{\sigma}_{a_h}^2\breve{\sigma}_{b_h}^2}. \tag{153}$$

Multiplying Equations (139) and (140) leads to

$$\sigma_{a_h}^2\sigma_{b_h}^2 = \frac{1}{4LM\widehat{\eta}_h^2}\left\{-\left(\widehat{\eta}_h^2 - \sigma^2(M-L)\right) + \sqrt{\left(\eta_h^2 + \sigma^2(L+M)\right)^2 - 4\sigma^4 LM}\right\}$$
$$\times\left\{-\left(\widehat{\eta}_h^2 + \sigma^2(M-L)\right) + \sqrt{\left(\eta_h^2 + \sigma^2(L+M)\right)^2 - 4\sigma^4 LM}\right\}$$
$$= \frac{1}{2LM}\left\{\widehat{\eta}_h^2 + \sigma^2(L+M) - \sqrt{\left(\widehat{\eta}_h^2 + \sigma^2(L+M)\right)^2 - 4\sigma^4 LM}\right\},$$

which is decreasing with respect to $\widehat{\eta}_h$. Equation (79) implies that $\widehat{\eta}_h$ is larger at any *positive* stationary point than at the *null* stationary point. Therefore, it holds that $\mathring{\sigma}_{a_h}^2\mathring{\sigma}_{b_h}^2 > \breve{\sigma}_{a_h}^2\breve{\sigma}_{b_h}^2$, and Equation (153) is negative. This means that the Hessian $\mathring{\mathcal{H}}^{\mathrm{VB}}$ has one negative and one positive eigenvalues.

Consequently, the Hessian of the objective function (73) with respect to $(\mu_{a_h}, \mu_{b_h}, \sigma_{a_h}^2, \sigma_{b_h}^2)$ has at least one negative and one positive eigenvalues at the *null* stationary point, which proves the lemma. ■

G.10 Proof of Lemma 18

We rely on the monotonicity of the positive solution of the quadratic equation (97) with respect to q_1 and q_0; the positive solution $\widehat{\gamma}$ of (97) is a monotone decreasing function of q_1 and q_0 (see

Figure 11). Although Equation (93) is not really quadratic with respect to $\widehat{\gamma}_h$ because Equation (94) depends on $\widehat{\gamma}_h$, we can bound the positive solutions of Equation (93) by replacing the coefficients q_1 and q_0 with their bounds. Equation (93) might have multiple positive solutions if the left-hand side oscillates when crossing the horizontal axis in Fig.11. However, our approach bounds all the positive solutions, and Lemma 17 guarantees that the minimizers consist of some of them when Equation (98) holds.

First we derive an upper-bound of $\widehat{\gamma}_h^2$. Let us lower-bound Equation (94) by ignoring the positive term $4\sigma^4 LM/(c_{a_h}^2 c_{b_h}^2)$:

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

We also lower-bound Equation (95) by ignoring the positive term $\sigma^4/(c_{a_h}^2 c_{b_h}^2)$. Then we can obtain an upper-bound of $\widehat{\gamma}_h$:

$$\widehat{\gamma}_h < \widehat{\gamma}_h^{\mathrm{up}},$$

where $\widehat{\gamma}_h^{\mathrm{up}}$ is the larger solution of the following equation:

$$(\widehat{\gamma}_h^{\mathrm{up}})^2 + \left(\frac{M}{L}-1\right)\gamma_h\widehat{\gamma}_h^{\mathrm{up}} - \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h^2 = 0.$$

This can be factorized as

$$\left(\widehat{\gamma}_h^{\mathrm{up}} - \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h\right)\left(\widehat{\gamma}_h^{\mathrm{up}} + \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\gamma_h\right) = 0.$$

Thus, the larger solution of this equation,

$$\widehat{\gamma}_h^{\mathrm{up}} = \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h,$$

gives the upper-bound in Equation (28).

Similarly, we derive a lower-bound of $\widehat{\gamma}_h^2$. Let us upper-bound Equation (94) by using the relation $\sqrt{x^2+y^2} \leq \sqrt{x^2+y^2+2xy} \leq x+y$ for $x,y \geq 0$:

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

We also upper-bound Equation (95) by adding a non-negative term

$$\frac{(M-L)\sigma^2}{Lc_{a_h}c_{b_h}}\left(\frac{1}{c_{a_h}c_{b_h}}+\frac{\sigma^2\sqrt{LM}}{\gamma_h}\right).$$

Then we can obtain a lower-bound of $\widehat{\gamma}_h$:

$$\widehat{\gamma}_h \geq \widehat{\gamma}_h^{\mathrm{lo}},$$

where $\widehat{\gamma}_h^{\mathrm{lo}}$ is the larger solution of the following equation:

$$L(\widehat{\gamma}_h^{\mathrm{lo}})^2+\left((M-L)\gamma_h+\frac{\sigma^2(L+M)\sqrt{M/L}}{c_{a_h}c_{b_h}}\right)\widehat{\gamma}_h^{\mathrm{lo}} + \frac{M^2\sigma^4}{Lc_{a_h}^2c_{b_h}^2}+\frac{\sigma^4M(M-L)\sqrt{M/L}}{\gamma_h c_{a_h}c_{b_h}}-M\left(1-\frac{\sigma^2L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2M}{\gamma_h^2}\right)\gamma_h^2=0.$$

This can be factorized as

$$\left(\widehat{\gamma}_h^{\mathrm{lo}}-\left(1-\frac{\sigma^2M}{\gamma_h^2}\right)\gamma_h+\frac{\sigma^2\sqrt{M/L}}{c_{a_h}c_{b_h}}\right)\left(L\widehat{\gamma}_h^{\mathrm{lo}}+M\left(1-\frac{\sigma^2L}{\gamma_h^2}\right)\gamma_h+\frac{\sigma^2M\sqrt{M/L}}{c_{a_h}c_{b_h}}\right)=0.$$

Thus, the larger solution of this equation,

$$\widehat{\gamma}_h^{\mathrm{lo}}=\left(1-\frac{\sigma^2M}{\gamma_h^2}\right)\gamma_h-\frac{\sigma^2\sqrt{M/L}}{c_{a_h}c_{b_h}},$$

gives the lower-bound in Equation (28).

The coefficient of the second term of Equation (146),

$$\frac{\sigma^2}{c_{a_h}c_{b_h}}\left(\frac{c_{b_h}\widehat{\delta}_h}{c_{a_h}}+\frac{c_{a_h}}{c_{b_h}\widehat{\delta}_h}\right),$$

is minimized when

$$\widehat{\delta}_h=\frac{c_{a_h}}{c_{b_h}}.$$

Then we can obtain another upper-bound of $\widehat{\gamma}_h$:

$$\widehat{\gamma}_h \leq \widehat{\gamma}_h^{\mathrm{up}},$$

where $\widehat{\gamma}_h^{\mathrm{up}}$ is the larger solution of the following equation:

$$(\widehat{\gamma}_h^{\mathrm{up}})^2+\left(\frac{2\sigma^2}{c_{a_h}c_{b_h}}\right)\widehat{\gamma}_h^{\mathrm{up}}+\frac{\sigma^4}{c_{a_h}^2c_{b_h}^2}-\left(1-\frac{\sigma^2L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2M}{\gamma_h^2}\right)\gamma_h^2=0.$$

This can be factorized as

$$\left(\widehat{\gamma}_h^{\mathrm{up}} - \sqrt{\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right) \times \left(\widehat{\gamma}_h^{\mathrm{up}} + \sqrt{\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right) = 0.$$

Thus, the larger solution of this equation,

$$\widehat{\gamma}_h^{\mathrm{up}} = \sqrt{\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h - \frac{\sigma^2}{c_{a_h}c_{b_h}},$$

gives the upper-bound in Equation (31). ■

G.11 Proof of Lemma 19

Consider the two-step minimization, (103) and (104). Lemma 17 implies that the minimizer of Equation (103) is the *null* stationary point for any given $(c_{a_h}^2, c_{b_h}^2)$ in $\mathring{\mathcal{R}}$. The *null* stationary point is explicitly given by Lemma 11. Substituting Equations (80)–(83) into Equation (106) gives

$$\mathring{\mathcal{L}}_h^{\mathrm{EVB}}(c_{a_h}^2, c_{b_h}^2) = M(-\log\lambda_{a,1} + \lambda_{a,1}) + L(-\log\lambda_{b,1} + \lambda_{b,1}) + \frac{LM\lambda_{a,0}\lambda_{b,0}}{\sigma^2}. \tag{154}$$

where

$$\lambda_{a,k}(c_{a_h}c_{b_h}) = \frac{1}{2M(c_{a_h}c_{b_h})^k}\left\{-\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} - c_{a_h}c_{b_h}(M-L)\right) + \sqrt{\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} - c_{a_h}c_{b_h}(M-L)\right)^2 + 4M\sigma^2}\right\},$$

$$\lambda_{b,k}(c_{a_h}c_{b_h}) = \frac{1}{2L(c_{a_h}c_{b_h})^k}\left\{-\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} + c_{a_h}c_{b_h}(M-L)\right) + \sqrt{\left(\frac{\sigma^2}{c_{a_h}c_{b_h}} + c_{a_h}c_{b_h}(M-L)\right)^2 + 4L\sigma^2}\right\}.$$

Note that $\lambda_{a,k} > 0$, $\lambda_{b,k} > 0$ for any k, and that Equation (154) depends on $c_{a_h}^2$ and $c_{b_h}^2$ only through their product $c_{a_h}c_{b_h}$.

Consider a decreasing mapping $x = \sigma^2/(c_{a_h}^2 c_{b_h}^2)$ (> 0). Then, $\lambda_{a,1}$ and $\lambda_{b,1}$ are written as

$$\lambda'_{a,1}(x) = 1 - \frac{(x+(L+M)) - \sqrt{(x+(L+M))^2 - 4ML}}{2M},$$

$$\lambda'_{b,1}(x) = 1 - \frac{(x+(L+M)) - \sqrt{(x+(L+M))^2 - 4ML}}{2L}.$$

Since they are increasing with respect to x, $\lambda_{a,1}$ and $\lambda_{b,1}$ are decreasing with respect to $c_{a_h}c_{b_h}$. Further, $\lambda_{a,1}$ and $\lambda_{b,1}$ are upper-bounded as

$$\lambda_{a,1}(c_{a_h}c_{b_h}) < \lim_{c_{a_h}c_{b_h}\to+0} \lambda_{a,1}(c_{a_h}c_{b_h}) = \lim_{x\to\infty} \lambda'_{a,1}(x) = 1,$$
$$\lambda_{b,1}(c_{a_h}c_{b_h}) < \lim_{c_{a_h}c_{b_h}\to+0} \lambda_{b,1}(c_{a_h}c_{b_h}) = \lim_{x\to\infty} \lambda'_{b,1}(x) = 1.$$

Since $(-\log\lambda+\lambda)$ is decreasing in the range $0<\lambda<1$, the first two terms in Equation (154) are increasing with respect to $c_{a_h}c_{b_h}$, and lower-bounded as

$$M(-\log\lambda_{a,1}+\lambda_{a,1}) > \lim_{c_{a_h}c_{b_h}\to+0} M(-\log\lambda_{a,1}+\lambda_{a,1}) = M, \tag{155}$$
$$L(-\log\lambda_{b,1}+\lambda_{b,1}) > \lim_{c_{a_h}c_{b_h}\to+0} L(-\log\lambda_{b,1}+\lambda_{b,1}) = L. \tag{156}$$

Similarly, using the same decreasing mapping, we have

$$\lambda'_{a,0}(x)\cdot\lambda'_{b,0}(x) = \frac{\sigma^2}{2LM}\left((x+(L+M)) - \sqrt{(x+(L+M))^2-4LM}\right).$$

Since this is decreasing with respect to x and lower-bounded by zero, $\lambda_{a,0}\lambda_{b,0}$ is increasing with respect to $c_{a_h}c_{b_h}$ and lower-bounded as

$$\lambda_{a,0}(c_{a_h}c_{b_h})\cdot\lambda_{b,0}(c_{a_h}c_{b_h}) > \lim_{c_{a_h}c_{b_h}\to+0} \lambda_{a,0}(c_{a_h}c_{b_h})\cdot\lambda_{b,0}(c_{a_h}c_{b_h}) = \lim_{x\to\infty} \lambda'_{a,0}(x)\cdot\lambda'_{b,0}(x) = 0.$$

Therefore, the third term in Equation (154) is increasing with respect to $c_{a_h}c_{b_h}$, and lower-bounded as

$$\frac{LM\lambda_{a,0}\lambda_{b,0}}{\sigma^2} > \lim_{c_{a_h}c_{b_h}\to+0} \frac{LM\lambda_{a,0}\lambda_{b,0}}{\sigma^2} = 0. \tag{157}$$

Now we have found that Equation (154) is increasing with respect to $c_{a_h}c_{b_h}$, because it consists of the increasing terms. Equations (114) and (115) minimize $c_{a_h}c_{b_h}$ over $\mathring{\mathcal{R}}_\varepsilon$ when Equation (43) is adopted. Therefore, they minimize Equation (154). Equations (110)–(113) are obtained by substituting Equations (114) and (115) into Equations (80)–(83). Since the infima (155)–(157) of the three terms of Equation (154) are obtained at the same time with the minimizer in the limit when $\varepsilon\to+0$, we have Equation (116). ■

G.12 Proof of Lemma 20

Existence of any *positive* stationary point lying in $\mathring{R}$ contradicts with Lemma 14. ■

G.13 Proof of Lemma 21

Assume that Equation (118) holds. Then, any global minimizer or point sequence giving the global infimum $\underline{\check{\mathcal{L}}}_h^{\mathrm{EVB}}$ exists in $\check{R}$. Let us investigate the objective function (106). It is differentiable in the

domain (102), and lower-bounded as

$$\mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2,c_{a_h}^2,c_{b_h}^2) \geq \mu_{a_h}^2\left(\frac{1}{c_{a_h}^2}+\frac{1}{\sigma^2}L\sigma_{b_h}^2\right)+\mu_{b_h}^2\left(\frac{1}{c_{b_h}^2}+\frac{1}{\sigma^2}M\sigma_{a_h}^2\right)$$
$$+M\left(\frac{\sigma_{a_h}^2}{c_{a_h}^2}-\log\frac{\sigma_{a_h}^2}{c_{a_h}^2}\right)+L\left(\frac{\sigma_{b_h}^2}{c_{b_h}^2}-\log\frac{\sigma_{b_h}^2}{c_{b_h}^2}\right)+\frac{1}{\sigma^2}\left(LM\sigma_{a_h}^2\sigma_{b_h}^2-\gamma_h^2\right). \tag{158}$$

Note that each term is lower-bounded by a finite value, since $(x-\log x)\geq 1$ for any $x>0$.

Since any sequence such that $c_{a_h}^2\to 0$ or $c_{b_h}^2\to 0$ goes into $\mathring{R}$, it cannot give $\underline{\check{\mathcal{L}}}_h^{\mathrm{EVB}}$. Accordingly, we neglect such sequences. Then, we find that the lower-bound (158) goes to infinity when $\sigma_{a_h}^2\to 0$ or $\sigma_{b_h}^2\to 0$, because of the third and the fourth terms (note that $\lim_{x\to+0}(x-\log x)=\infty$). Further, it goes to infinity when $\sigma_{a_h}^2\to\infty$ or $\sigma_{b_h}^2\to\infty$, because of the fifth term. It also goes to infinity when $|\mu_{a_h}|\to\infty$ or $|\mu_{b_h}|\to\infty$, because of the first and the second terms. Finally, it goes to infinity when $c_{a_h}^2\to\infty$ or $c_{b_h}^2\to\infty$, because of the third and the fourth terms.

The above mean that the objective function (106) goes to infinity when approaching to any point on the domain boundary included in $\check{R}$. Consequently, the minimizers consist of stationary points in $\check{R}$. According to Lemma 14 and Lemma 16, the *null* stationary points in $\check{R}$ are saddle points. Therefore, the minimizers consist of *positive* stationary points. ■

G.14 Proof of Lemma 22

Substituting Equation (75) into Equation (74) gives

$$\gamma_h^2=\frac{\sigma^4}{\sigma_{a_h}^2\sigma_{b_h}^2}. \tag{159}$$

Substituting Equations (76) and (77) into Equation (159), we have

$$\gamma_h^2=\left(\mu_{a_h}^2+M\sigma_{a_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right)\left(\mu_{b_h}^2+L\sigma_{b_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right). \tag{160}$$

Substituting Equations (119) and (120) into Equation (160) gives

$$\gamma_h^2=\left(Mc_{a_h}^2+\frac{\sigma^2}{c_{b_h}^2}\right)\left(Lc_{b_h}^2+\frac{\sigma^2}{c_{a_h}^2}\right).$$

From this, we have

$$LMc_{a_h}^4c_{b_h}^4-\left(\gamma_h^2-(L+M)\sigma^2\right)c_{a_h}^2c_{b_h}^2+\sigma^4=0. \tag{161}$$

Solving Equation (161) with respect to $c_{a_h}^2c_{b_h}^2$, we obtain two solutions:

$$c_{a_h}^2c_{b_h}^2=\frac{\left(\gamma_h^2-(L+M)\sigma^2\right)\pm\sqrt{\left(\gamma_h^2-(L+M)\sigma^2\right)^2-4LM\sigma^4}}{2LM}. \tag{162}$$

On the other hand, because of the redundancy with respect to the transform (42), we can fix the ratio of the hyperparameters as in Equation (43). Thus, we have transformed the necessary and sufficient condition (74)–(77), (119), and (120) to (74)–(77), and (162). Since

$$\begin{aligned}&\sqrt{\left(\gamma_h^2-(L+M)\sigma^2\right)^2-4LM\sigma^4}\\&\qquad=\sqrt{\left(\gamma_h^2-(\sqrt{L}+\sqrt{M})^2\sigma^2\right)\left(\gamma_h^2-(\sqrt{M}-\sqrt{L})^2\sigma^2\right)}\end{aligned}$$

and

$$\sqrt{(\sqrt{M}-\sqrt{L})^2\sigma^2}<\sqrt{M\sigma^2},$$

the two solutions (162) are real and positive if and only if Equation (121) holds. This proves the necessity.

Suppose that Equation (121) holds. Then, the two solutions (162) exist. The inverse of the smaller solution (123) is written as

$$\frac{1}{\acute{c}_{a_h}^2\acute{c}_{b_h}^2}=\frac{\left(\gamma_h^2-(L+M)\sigma^2\right)+\sqrt{\left(\gamma_h^2-(L+M)\sigma^2\right)^2-4LM\sigma^4}}{2\sigma^4}. \tag{163}$$

This is upper-bounded as

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

Using this bound, we have

$$\begin{aligned}&\sqrt{\left(1-\frac{\sigma^2L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2M}{\gamma_h^2}\right)}\gamma_h-\frac{\sigma^2}{\acute{c}_{a_h}\acute{c}_{b_h}}\\&\qquad>\sqrt{\gamma_h^2-(L+M)\sigma^2+\frac{LM\sigma^4}{\gamma_h^2}}-\sqrt{\gamma_h^2-(L+M)\sigma^2}\\&\qquad>0.\end{aligned}$$

This means that Equation (98) holds. The same holds for the larger solution (122), since

$$\frac{1}{\check{c}_{a_h}\check{c}_{b_h}}\leq\frac{1}{\acute{c}_{a_h}\acute{c}_{b_h}}.$$

Consequently, Lemma 14 guarantees the existence of at least one *positive* stationary point $(\check{\mu}_{a_h},\check{\mu}_{b_h},\check{\sigma}_{a_h}^2,\check{\sigma}_{b_h}^2)\in\mathbb{R}^2\times\mathbb{R}_{++}^2$ satisfying Equations (74)–(77), given any $(c_{a_h}^2,c_{b_h}^2)\in\mathbb{R}_{++}^2$ constructed from Equation (43) and either of the two solutions (162). Thus, we have shown the existence of at least one *positive* stationary point satisfying the necessary and sufficient condition (74)–(77), and (162) when Equation (121) holds. This proves the sufficiency. ■

G.15 Proof of Lemma 23

We show that, when Equation (125) holds, the Hessian of the objective function (106) has at least one negative and one positive eigenvalues at any *small positive* stationary point. We only focus on the 4-dimensional subspace spanned by $(\mu_{a_h}, \mu_{b_h}, c_{a_h}^2, c_{b_h}^2)$. The partial derivatives of the objective function (106) are

$$\begin{aligned}
\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{a_h}} &= \frac{\mu_{a_h}}{c_{a_h}^2} + \frac{-\gamma_h \mu_{b_h} + (\mu_{b_h}^2 + L\sigma_{b_h}^2)\mu_{a_h}}{\sigma^2}, \\
\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{b_h}} &= \frac{\mu_{b_h}}{c_{b_h}^2} + \frac{-\gamma_h \mu_{a_h} + (\mu_{a_h}^2 + M\sigma_{a_h}^2)\mu_{b_h}}{\sigma^2}, \\
\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{a_h}^2} &= \frac{1}{2}\left(\frac{M}{c_{a_h}^2} - \frac{(\mu_{a_h}^2 + M\sigma_{a_h}^2)}{c_{a_h}^4}\right), \\
\frac{1}{2}\frac{\partial \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{b_h}^2} &= \frac{1}{2}\left(\frac{L}{c_{b_h}^2} - \frac{(\mu_{b_h}^2 + L\sigma_{b_h}^2)}{c_{b_h}^4}\right).
\end{aligned}$$

Then, the Hessian is given by

$$\begin{aligned}
\frac{1}{2}\mathcal{H}^{\mathrm{EVB}} &= \begin{pmatrix}
\frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{(\partial \mu_{a_h})^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{a_h}\partial \mu_{b_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{a_h}\partial c_{a_h}^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{a_h}\partial c_{b_h}^2} \\
\frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{b_h}\partial \mu_{a_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{(\partial \mu_{b_h})^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{b_h}\partial c_{a_h}^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial \mu_{b_h}\partial c_{b_h}^2} \\
\frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{a_h}^2\partial \mu_{a_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{a_h}^2\partial \mu_{b_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{(\partial c_{a_h}^2)^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{a_h}^2\partial c_{b_h}^2} \\
\frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{b_h}^2\partial \mu_{a_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{b_h}^2\partial \mu_{b_h}} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{\partial c_{b_h}^2\partial c_{a_h}^2} & \frac{1}{2}\frac{\partial^2 \mathcal{L}_h^{\mathrm{EVB}}}{(\partial c_{b_h}^2)^2}
\end{pmatrix} \\
&= \begin{pmatrix}
\frac{1}{c_{a_h}^2} + \frac{\mu_{b_h}^2 + L\sigma_{b_h}^2}{\sigma^2} & \frac{2\mu_{a_h}\mu_{b_h} - \gamma_h}{\sigma^2} & -\frac{\mu_{a_h}}{c_{a_h}^4} & 0 \\
\frac{2\mu_{a_h}\mu_{b_h} - \gamma_h}{\sigma^2} & \frac{1}{c_{b_h}^2} + \frac{\mu_{a_h}^2 + M\sigma_{a_h}^2}{\sigma^2} & 0 & -\frac{\mu_{b_h}}{c_{b_h}^4} \\
-\frac{\mu_{a_h}}{c_{a_h}^4} & 0 & \frac{2(\mu_{a_h}^2 + M\sigma_{a_h}^2) - Mc_{a_h}^2}{2c_{a_h}^6} & 0 \\
0 & -\frac{\mu_{b_h}}{c_{b_h}^4} & 0 & \frac{2(\mu_{b_h}^2 + L\sigma_{b_h}^2) - Lc_{b_h}^2}{2c_{b_h}^6}
\end{pmatrix}.
\end{aligned} \tag{164}$$

At any *positive* stationary point, Equations (74)–(77), (119), and (120) hold. Substituting Equations (76), (77), (119), and (120) into (164), we have

$$\frac{1}{2}\mathcal{H}^{\mathrm{EVB}} = \begin{pmatrix}
\frac{1}{\sigma_{a_h}^2} & \frac{\gamma_h - 2\mu_{a_h}\mu_{b_h}}{\sigma^2} & -\frac{\mu_{a_h}}{c_{a_h}^4} & 0 \\
\frac{\gamma_h - 2\mu_{a_h}\mu_{b_h}}{\sigma^2} & \frac{1}{\sigma_{b_h}^2} & 0 & -\frac{\mu_{b_h}}{c_{b_h}^4} \\
-\frac{\mu_{a_h}}{c_{a_h}^4} & 0 & \frac{M}{2c_{a_h}^4} & 0 \\
0 & -\frac{\mu_{b_h}}{c_{b_h}^4} & 0 & \frac{L}{2c_{b_h}^4}
\end{pmatrix}.$$

Its determinant is calculated as

$$\left|\frac{1}{2}\mathcal{H}^{\mathrm{EVB}}\right| = -\frac{\mu_{b_h}}{c_{b_h}^4}\begin{vmatrix} -\frac{\mu_{a_h}}{c_{a_h}^4} & 0 & \frac{M}{2c_{a_h}^4} \\ 0 & -\frac{\mu_{b_h}}{c_{b_h}^4} & 0 \\ \frac{1}{\sigma_{a_h}^2} & \frac{\gamma_h - 2\mu_{a_h}\mu_{b_h}}{\sigma^2} & -\frac{\mu_{a_h}}{c_{a_h}^4} \end{vmatrix} + \frac{L}{2c_{b_h}^4}\begin{vmatrix} \frac{1}{\sigma_{a_h}^2} & \frac{\gamma_h - 2\mu_{a_h}\mu_{b_h}}{\sigma^2} & -\frac{\mu_{a_h}}{c_{a_h}^4} \\ \frac{\gamma_h - 2\mu_{a_h}\mu_{b_h}}{\sigma^2} & \frac{1}{\sigma_{b_h}^2} & 0 \\ -\frac{\mu_{a_h}}{c_{a_h}^4} & 0 & \frac{M}{2c_{a_h}^4} \end{vmatrix}$$
$$= \frac{1}{c_{a_h}^4 c_{b_h}^4}\left(\frac{\mu_{a_h}^2\mu_{b_h}^2}{c_{a_h}^4 c_{b_h}^4} - \frac{M\mu_{b_h}^2}{2\sigma_{a_h}^2 c_{b_h}^4} - \frac{L\mu_{a_h}^2}{2\sigma_{b_h}^2 c_{a_h}^4} + \frac{LM}{4\sigma^4}\left(\frac{\sigma^4}{\sigma_{a_h}^2\sigma_{b_h}^2} - (\gamma_h - 2\mu_{a_h}\mu_{b_h})^2\right)\right).$$

Multiplying both sides of Equation (74) by μ_{a_h} gives

$$\mu_{a_h}^2 = \frac{\sigma_{a_h}^2}{\sigma^2}\gamma_h\widehat{\gamma}_h,$$

and therefore

$$\frac{\mu_{a_h}^2}{\sigma_{a_h}^2} = \frac{\gamma_h\widehat{\gamma}_h}{\sigma^2}. \tag{165}$$

Similarly from Equation (75), we obtain

$$\frac{\mu_{b_h}^2}{\sigma_{b_h}^2} = \frac{\gamma_h\widehat{\gamma}_h}{\sigma^2}. \tag{166}$$

By using Equations (78), (84), (159), (165), and (166), we obtain

$$\left|\frac{1}{2}\mathcal{H}^{\mathrm{EVB}}\right| = \frac{1}{c_{a_h}^4 c_{b_h}^4}\left(\frac{\widehat{\gamma}_h^2}{c_{a_h}^4 c_{b_h}^4} - \frac{\gamma_h\widehat{\gamma}_h}{2\sigma^2}\left(\frac{M\widehat{\delta}^{-2}}{c_{b_h}^4} + \frac{L\widehat{\delta}^2}{c_{a_h}^4}\right) + \frac{LM}{\sigma^4}\left(\widehat{\gamma}_h\gamma_h - \widehat{\gamma}_h^2\right)\right). \tag{167}$$

Since

$$\frac{M\widehat{\delta}^{-2}}{c_{b_h}^4} + \frac{L\widehat{\delta}^2}{c_{a_h}^4} \geq \frac{2\sqrt{LM}}{c_{a_h}^2 c_{b_h}^2}$$

for any $\widehat{\delta}^2 > 0$, Equation (167) is upper-bounded by

$$\left|\frac{1}{2}\mathcal{H}^{\mathrm{EVB}}\right| \leq \frac{1}{c_{a_h}^4 c_{b_h}^4}\left(\frac{\widehat{\gamma}_h^2}{c_{a_h}^4 c_{b_h}^4} - \frac{\gamma_h\widehat{\gamma}_h\sqrt{LM}}{\sigma^2 c_{a_h}^2 c_{b_h}^2} + \frac{LM}{\sigma^4}\left(\widehat{\gamma}_h\gamma_h - \widehat{\gamma}_h^2\right)\right)$$
$$= \frac{\widehat{\gamma}_h}{c_{a_h}^4 c_{b_h}^4}\left(\frac{1}{c_{a_h}^2 c_{b_h}^2} - \frac{\sqrt{LM}}{\sigma^2}\right)\left\{\left(\frac{1}{c_{a_h}^2 c_{b_h}^2} + \frac{\sqrt{LM}}{\sigma^2}\right)\widehat{\gamma}_h - \frac{\sqrt{LM}}{\sigma^2}\gamma_h\right\}. \tag{168}$$

At any *small positive* stationary point, Equation (123) is upper-bounded as

$$\acute{c}_{a_h}^2\acute{c}_{b_h}^2 < \frac{\sigma^2}{\sqrt{LM}}$$

when Equation (125) holds. Therefore, Equation (168) is written as

$$\left|\frac{1}{2}\acute{\mathcal{H}}^{\mathrm{EVB}}\right| \leq C\left\{\left(\frac{1}{\acute{c}_{a_h}^2\acute{c}_{b_h}^2} + \frac{\sqrt{LM}}{\sigma^2}\right)\widehat{\gamma}_h - \frac{\sqrt{LM}}{\sigma^2}\gamma_h\right\},$$

with a positive factor

$$C = \frac{\widehat{\gamma}_h}{\acute{c}_{a_h}^4 \acute{c}_{b_h}^4} \left(\frac{1}{\acute{c}_{a_h}^2 \acute{c}_{b_h}^2} - \frac{\sqrt{LM}}{\sigma^2} \right).$$

Using Equation (31), we have

$$\begin{aligned}
\left| \frac{1}{2} \acute{\mathcal{H}}^{\mathrm{EVB}} \right| &\leq C \left\{ \left(\frac{1}{\acute{c}_{a_h}^2 \acute{c}_{b_h}^2} + \frac{\sqrt{LM}}{\sigma^2} \right) \left(\sqrt{\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)} \gamma_h - \frac{\sigma^2}{\acute{c}_{a_h} \acute{c}_{b_h}} \right) \right. \\
&\qquad \left. - \frac{\sqrt{LM}}{\sigma^2} \gamma_h \right\} \\
&= C \left\{ - \frac{\sigma^2}{\acute{c}_{a_h}^3 \acute{c}_{b_h}^3} + \sqrt{\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)} \frac{\gamma_h}{\acute{c}_{a_h}^2 \acute{c}_{b_h}^2} - \frac{\sqrt{LM}}{\acute{c}_{a_h} \acute{c}_{b_h}} \right. \\
&\qquad \left. - \frac{\sqrt{LM}}{\sigma^2} \left(1 - \sqrt{\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)} \right) \gamma_h \right\} \\
&< \frac{C}{\acute{c}_{a_h} \acute{c}_{b_h}} \left(- \frac{\sigma^2}{\acute{c}_{a_h}^2 \acute{c}_{b_h}^2} + \sqrt{\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)} \frac{\gamma_h}{\acute{c}_{a_h} \acute{c}_{b_h}} - \sqrt{LM} \right).
\end{aligned}$$

At the last inequality, we neglected the negative last term in the curly braces.

Using Equation (163), we have

$$\left| \frac{1}{2} \acute{\mathcal{H}}^{\mathrm{EVB}} \right| < -C'(f(\gamma_h) - g(\gamma_h)), \tag{169}$$

where

$$\begin{aligned}
C' &= \frac{\gamma_h^2 C}{2\sigma^2 \acute{c}_{a_h} \acute{c}_{b_h}}, \\
f(\gamma_h) &= \left(1 - \frac{(\sqrt{M} - \sqrt{L})^2 \sigma^2}{\gamma_h^2}\right) + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}, \\
g(\gamma_h) &= \sqrt{2\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)} \\
&\qquad \times \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right) + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}}}.
\end{aligned}$$

Since C', $f(\gamma_h)$, and $g(\gamma_h)$ are positive, the right-hand side of Equation (169) is negative if $f^2(\gamma_h) - g^2(\gamma_h) > 0$. This is shown below.

$$
\begin{aligned}
f^2(\gamma_h) - g^2(\gamma_h) &= \left(\left(1 - \frac{(\sqrt{M}-\sqrt{L})^2\sigma^2}{\gamma_h^2}\right) + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}} \right)^2 \\
&- 2\left(1 - \frac{L\sigma^2}{\gamma_h^2}\right)\left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)\left(\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right) + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}} \right) \\
&= 2\frac{\sqrt{LM}\sigma^2}{\gamma_h^2}\left(2 - \frac{\sqrt{LM}\sigma^2}{\gamma_h^2}\right) \\
&\quad \times \left(\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right) + \sqrt{\left(1 - \frac{(L+M)\sigma^2}{\gamma_h^2}\right)^2 - \frac{4LM\sigma^4}{\gamma_h^4}} \right) \\
&> 0.
\end{aligned}
$$

Consequently, it holds that $|\acute{\mathcal{H}}^{\mathrm{EVB}}| < 0$. This means that $\acute{\mathcal{H}}^{\mathrm{EVB}}$ has at least one negative and one positive eigenvalues. Therefore, the Hessian of the objective function (106) with respect to $(\mu_{a_h}, \mu_{b_h}, \sigma^2_{a_h}, \sigma^2_{b_h}, c^2_{a_h}, c^2_{b_h})$ has at least one negative and one positive eigenvalues at any *small positive* stationary point, when Equation (125) holds. This proves the lemma. ■

G.16 Proof of Lemma 25

Substituting Equations (106) and (116) into Equation (126), we have

$$
\begin{aligned}
\Delta_h(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) &= \mathcal{L}_h^{\mathrm{EVB}}(\breve{\mu}_{a_h}, \breve{\mu}_{b_h}, \breve{\sigma}^2_{a_h}, \breve{\sigma}^2_{b_h}, \breve{c}^2_{a_h}, \breve{c}^2_{b_h}) - (L+M) \\
&= M\log\frac{\breve{c}^2_{a_h}}{\breve{\sigma}^2_{a_h}} + L\log\frac{\breve{c}^2_{b_h}}{\breve{\sigma}^2_{b_h}} + \frac{\breve{\mu}^2_{a_h} + M\breve{\sigma}^2_{a_h}}{\breve{c}^2_{a_h}} + \frac{\breve{\mu}^2_{b_h} + L\breve{\sigma}^2_{b_h}}{\breve{c}^2_{b_h}} \\
&\quad + \frac{1}{\sigma^2}\left(-2\gamma_h\breve{\mu}_{a_h}\breve{\mu}_{b_h} + \left(\breve{\mu}^2_{a_h} + M\breve{\sigma}^2_{a_h}\right)\left(\breve{\mu}^2_{b_h} + L\breve{\sigma}^2_{b_h}\right)\right) - (L+M). \qquad (170)
\end{aligned}
$$

Substituting Equations (119) and (120) into Equation (170), we have

$$
\Delta_h = M\log\left(\frac{\breve{\mu}^2_{a_h}}{M\breve{\sigma}^2_{a_h}} + 1\right) + L\log\left(\frac{\breve{\mu}^2_{b_h}}{L\breve{\sigma}^2_{b_h}} + 1\right) + \frac{1}{\sigma^2}\left(-2\gamma_h\breve{\mu}_{a_h}\breve{\mu}_{b_h} + LM\breve{c}^2_{a_h}\breve{c}^2_{b_h}\right). \qquad (171)
$$

Substituting Equations (165) and (166) into Equation (171) and using Equation (78), we have

$$
\Delta_h = M\log\left(\frac{\gamma_h}{M\sigma^2}\widehat{\gamma}_h + 1\right) + L\log\left(\frac{\gamma_h}{L\sigma^2}\widehat{\gamma}_h + 1\right) + \frac{1}{\sigma^2}\left(-2\gamma_h\widehat{\gamma}_h + LM\breve{c}^2_{a_h}\breve{c}^2_{b_h}\right). \qquad (172)
$$

Using the bounds (28), Equation (172) is upper-bounded as

$$\begin{aligned}\Delta_h &< M\log\left(\frac{\gamma_h^2}{M\sigma^2}\left(1-\frac{M\sigma^2}{\gamma_h^2}\right)+1\right)+L\log\left(\frac{\gamma_h^2}{L\sigma^2}\left(1-\frac{M\sigma^2}{\gamma_h^2}\right)+1\right)\\ &\qquad+\frac{1}{\sigma^2}\left(-2\gamma_h\left(\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h-\frac{\sigma^2\sqrt{M/L}}{\check{c}_{a_h}\check{c}_{b_h}}\right)+LM\check{c}_{a_h}^2\check{c}_{b_h}^2\right)\\ &= M\log\left(\frac{\gamma_h^2}{M\sigma^2}\right)+L\log\left(\frac{\gamma_h^2}{L\sigma^2}-\frac{M}{L}+1\right)\\ &\qquad+\frac{1}{\sigma^2}\left(-2\gamma_h\left(\gamma_h-\frac{\sigma^2 M}{\gamma_h}-\frac{\sigma^2\sqrt{M/L}}{\check{c}_{a_h}\check{c}_{b_h}}\right)+LM\check{c}_{a_h}^2\check{c}_{b_h}^2\right)\\ &= M\log\left(\frac{\gamma_h^2}{M\sigma^2}\right)+L\log\left(\frac{\gamma_h^2}{L\sigma^2}-\frac{M}{L}+1\right)+2M+\frac{2\sqrt{M/L}}{\check{c}_{a_h}\check{c}_{b_h}}\gamma_h-\frac{2\gamma_h^2}{\sigma^2}+\frac{LM\check{c}_{a_h}^2\check{c}_{b_h}^2}{\sigma^2}.\end{aligned}$$

Since $\sqrt{x^2-y^2} > x-y$ for $x > y > 0$, Equation (122) yields

$$\check{c}_{a_h}^2\check{c}_{b_h}^2 \geq \frac{\gamma_h^2-(L+M+\sqrt{LM})\sigma^2}{LM}. \tag{173}$$

Ignoring the positive term $4LM\sigma^4$ in Equation (122), we obtain

$$\check{c}_{a_h}^2\check{c}_{b_h}^2 < \frac{\gamma_h^2-(L+M)\sigma^2}{LM}. \tag{174}$$

Equations (173) and (174) result in

$$\sqrt{\frac{\gamma_h^2-(L+M+\sqrt{LM})\sigma^2}{LM}} \leq \check{c}_{a_h}\check{c}_{b_h} < \sqrt{\frac{\gamma_h^2-(L+M)\sigma^2}{LM}}.$$

Using these bounds, we obtain

$$\begin{aligned}\Delta_h &< M\log\left(\frac{\gamma_h^2}{M\sigma^2}\right)+L\log\left(\frac{\gamma_h^2}{L\sigma^2}-\frac{M}{L}+1\right)+2M+\frac{2\sqrt{M/L}}{\sqrt{\frac{\gamma_h^2-(L+M+\sqrt{LM})\sigma^2}{LM}}}\gamma_h\\ &\qquad-\frac{2\gamma_h^2}{\sigma^2}+\gamma_h^2-(L+M)\\ &= M\log\left(\frac{\gamma_h^2}{M\sigma^2}\right)+L\log\left(\frac{\gamma_h^2}{L\sigma^2}-\frac{M}{L}+1\right)+M-L+\frac{2M}{\sqrt{1-\frac{(L+M+\sqrt{LM})\sigma^2}{\gamma_h^2}}}-\frac{\gamma_h^2}{\sigma^2}.\end{aligned}$$

Using Equations (128), (129), and (130), we obtain Equation (127). ■

G.17 Proof of Lemma 26

For $0 < \alpha \leq 1$ and $\beta \geq 7$, Equation (128) is increasing with respect to α, because

$$\begin{aligned}
\frac{\partial \psi(\alpha,\beta)}{\partial \alpha} &= \log\left(\frac{\beta-1+\alpha}{\alpha}\right) - \left(\frac{\beta-1}{\beta-1+\alpha}\right) - 1 + \frac{(\sqrt{\alpha}+1/2)}{\beta\sqrt{\alpha}\left(1-\frac{(\alpha+\sqrt{\alpha}+1)}{\beta}\right)^{3/2}} \\
&> \log\left(\frac{\beta-1}{\alpha}+1\right) - 2 + \frac{1}{\beta} \\
&\geq \log(\beta) - 2 + \frac{1}{\beta} \\
&> 0.
\end{aligned}$$

Here, we used the numerical estimation that $\log(\beta) - 2 + 1/\beta \approx 0.0888$ when $\beta = 7$, and the fact that $\log(\beta) - 2 + 1/\beta$ is increasing with respect to β when $\beta > 1$.

For $0 < \alpha \leq 1$ and $\beta > 3$, Equation (128) is decreasing with respect to β, because

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

Consequently, if $\psi(1,\widetilde{\beta}) < 0$, it holds that $\psi(\alpha,\beta) < 0$ for any $0 < \alpha \leq 1$ and $\beta \geq \widetilde{\beta}$. The fact that $\psi(1,7) \approx -0.462 < 0$ completes the proof. ■

G.18 Proof of Lemma 29

Since the upper-bound in Equation (28) does not depend on $(c_{a_h}^2, c_{b_h}^2)$, Equation (46) holds.

Since the lower-bound in Equation (28) is nondecreasing with respect to $c_{a_h}c_{b_h}$, substituting Equation (173) into Equation (28) yields

$$\widehat{\gamma}_h \geq \max\left\{0, \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h - \frac{\sigma^2 M}{\sqrt{\gamma_h^2 - (L+M+\sqrt{LM})\sigma^2}}\right\}.$$

It holds that

$$-\frac{\sigma^2 M}{\gamma_h} > -\frac{\sigma^2 M}{\sqrt{\gamma_h^2 - (L+M+\sqrt{LM})\sigma^2}} > -\frac{\sigma^2 M}{\gamma_h - \sqrt{(L+M+\sqrt{LM})\sigma^2}},$$

where the positive term $(L+M+\sqrt{LM})\sigma^2$ is subtracted in the first inequality and the relation $\sqrt{x^2-y^2} > x-y$ for $x > y > 0$ is used in the second inequality. Then we have

$$\widehat{\gamma}_h > \max\left\{0, \gamma_h - \frac{2\sigma^2 M}{\gamma_h - \sqrt{(L+M+\sqrt{LM})\sigma^2}}\right\},$$

which leads to Equation (47).

Substituting Equation (174) into Equation (31), we obtain

$$\begin{aligned}\widehat{\gamma}_h &< \sqrt{\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h - \frac{\sigma^2\sqrt{LM}}{\sqrt{\gamma_h^2-(L+M)\sigma^2}} \\ &< \sqrt{\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)}\gamma_h - \frac{\sigma^2\sqrt{LM}}{\gamma_h},\end{aligned}$$

where the positive term $(L+M)\sigma^2$ is ignored in the second inequality. This gives Equation (48), and completes the proof. ■

Appendix H. Illustration of EVB Objective Function

Here we illustrate the EVB objective function (106). Let us consider a partially minimized objective function:

$$\tilde{\mathcal{L}}_h^{\mathrm{EVB}}(c_{a_h}c_{b_h}) = \min_{(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2)} \mathcal{L}_h^{\mathrm{EVB}}(\mu_{a_h},\mu_{b_h},\sigma_{a_h}^2,\sigma_{b_h}^2,c_{a_h}c_{b_h},c_{a_h}c_{b_h}). \tag{175}$$

According to Lemma 19, the infimum at the *null* local minimizer is given by

$$\lim_{c_{a_h}c_{b_h}\to 0} \tilde{\mathcal{L}}_h^{\mathrm{EVB}}(c_{a_h}c_{b_h}) = \mathring{\underline{\mathcal{L}}}_h^{\mathrm{EVB}} = L+M. \tag{176}$$

Figure 13 depicts the partially minimized objective function (175) when $L=M=H=1$, $\sigma^2=1$, and $V = 1.5, 2.0, 2.1, 2.7$. Corollary 1 provides the exact values for drawing these graphs. The *large* and the *small positive* stationary points, specified by Equations (122) and (123), respectively, are also plotted in the graphs if they exist. When

$$V = 1.5 \left(< 2 = (\sqrt{L}+\sqrt{M})\sigma\right),$$

Equation (121) does not hold. In this case, the objective function (175) has no stationary point as Lemma 22 states (the upper-left graph of Figure 13). The curve is identical for $0 \le V < 2.0$.

When $V = 2.0$ (the upper-right graph), Equation (124) holds. In this case, the objective function (175) has a stationary point at $c_{a_h}c_{b_h} = 1$. This corresponds to the coincident *large* and *small positive* stationary point. Still no local minimum exists.

When $V = 2.1$ (the lower-left graph), Equation (125) holds. In this case, there exists a *large positive* stationary point (which is a local minimum) at $c_{a_h}c_{b_h} \approx 1.37$, as well as a *small positive* stationary point (which is a local maximum) at $c_{a_h}c_{b_h} \approx 0.73$. However, we see that

$$\tilde{\mathcal{L}}_h^{\mathrm{EVB}}(1.37) \approx 2.24 > 2 = \mathring{\underline{\mathcal{L}}}_h^{\mathrm{EVB}}.$$

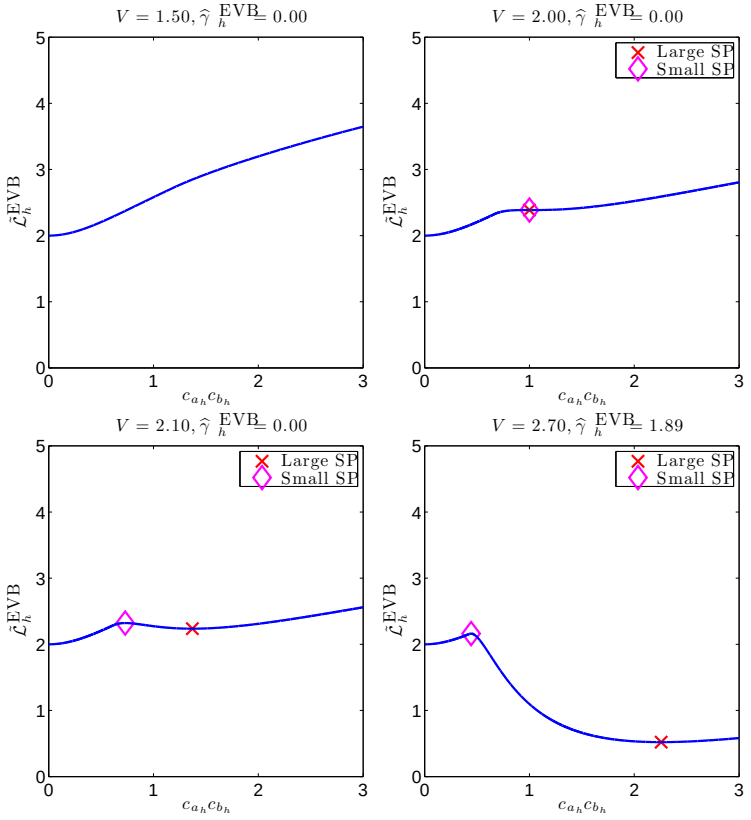

Figure 13: Illustration of the partially minimized objective function (175) when $L = M = H = 1$, $\sigma^2 = 1$, and $V = 1.5, 2.0, 2.1, 2.7$. The convergence $\tilde{\mathcal{L}}_h^{\mathrm{EVB}}(c_{a_h} c_{b_h}) \to L + M \ (= 2)$ as $c_{a_h} c_{b_h} \to 0$ is observed (see Equation (176)). 'Large SP' and 'Small SP' indicate the *large* and the *small positive* stationary points, respectively.

Therefore, the *null* local minimizer ($c_{a_h} c_{b_h} \to 0$) is still global, resulting in $\widehat{\gamma}_h^{\mathrm{EVB}} = 0$.

When $V = 2.7$ (the lower-right graph), $\gamma_h \geq \sqrt{7M} \cdot \sigma$ holds. As Lemma 28 states, a *large positive* stationary point at $c_{a_h} c_{b_h} \approx 2.26$ gives the global minimum:

$$\tilde{\mathcal{L}}_h^{\mathrm{EVB}}(2.26) \approx 0.52 < 2 = \mathring{\underline{\mathcal{L}}}_h^{\mathrm{EVB}},$$

resulting in a *positive* output $\widehat{\gamma}_h^{\mathrm{EVB}} \approx 1.89$.

Appendix I. Derivation of Equations (57) and (58)

Let $p(\boldsymbol{v}|\boldsymbol{\theta})$ be a model distribution, where $\boldsymbol{v}$ is a random variable and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a d-dimensional parameter vector. The *Jeffreys non-informative prior* (Jeffreys, 1946) is defined as

$$\phi^{\mathrm{Jef}}(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{F}|}, \tag{177}$$

where $\mathcal{F} \in \mathbb{R}^{d \times d}$ is the Fisher information matrix defined by

$$\mathcal{F}_{jk} = \int \frac{\partial \log p(\boldsymbol{v}|\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log p(\boldsymbol{v}|\boldsymbol{\theta})}{\partial \theta_k} p(\boldsymbol{v}|\boldsymbol{\theta}) d\boldsymbol{v}. \tag{178}$$

Let us first derive the Jeffreys prior for the non-factorizing model:

$$p_U(V|U) \propto \exp\left(-\frac{1}{2\sigma^2}(V-U)^2\right). \tag{179}$$

In this model, the parameter vector is one-dimensional: $\theta = U$. Since

$$\frac{\partial \log p_U(V|U)}{\partial U} = \frac{V-U}{\sigma^2},$$

the Fisher information (178) is given by

$$\mathcal{F}_U = \frac{1}{\sigma^2}.$$

This is constant over the parameter space. Therefore, the Jeffreys prior (177) for the model (179) is given by Equation (57).

Let us move on to the MF model:

$$p_{A,B}(V|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}(V-AB)^2\right). \tag{180}$$

In this model, the parameter vector is $\boldsymbol{\theta} = (A,B)$. Since

$$\frac{\partial \log p_{A,B}(Y|A,B)}{\partial A} = \frac{1}{\sigma^2}(Y-AB)B,$$
$$\frac{\partial \log p_{A,B}(Y|A,B)}{\partial B} = \frac{1}{\sigma^2}(Y-AB)A,$$

the Fisher information matrix is given by

$$\mathcal{F}_{A,B} = \frac{1}{\sigma^2}\begin{pmatrix} B^2 & AB \\ AB & A^2 \end{pmatrix},$$

whose eigenvalues are $\sigma^{-2}\sqrt{A^2+B^2}$ and 0.

The common (over the parameter space) zero-eigenvalue comes from the invariance of the MF model (180) under the transform $(A,B) \to (sA, s^{-1}B)$ for any $s > 0$. Neglecting it, we re-define the Jeffreys prior by

$$\phi^{\mathrm{Jef}}(\boldsymbol{\theta}) \propto \sqrt{\textstyle\prod_{j=1}^{d-1} \lambda_j},$$

where λ_j is the j-th largest eigenvalue of the Fisher information matrix. Thus, we obtain Equation (58).

■

References

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

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

P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, USA, 1998.

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

A. J. Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Department of Statistics, Stanford University, Stanford, CA, USA, 1964.

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

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

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

O. Chapelle and Z. Harchaoui. A machine learning approach to conjoint analysis. In *Advances in Neural Information Processing Systems*, volume 17, pages 257–264, 2005.

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

A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, West Sussex, UK, 2009.

M. J. Daniels and R. E. Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4): 1173–1184, 2001.

B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130, 1973.

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

A. Gelman. Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99:537–545, 2004.

Y. Y. Guo and N. Pal. A sequence of improvements over the James-Stein estimator. *Journal of Multivariate Analysis*, 42(2):302–317, 1992.

K. Hayashi, J. Hirayama, and S. Ishii. Dynamic exponential family matrix factorization. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 452–462, Berlin, 2009. Springer.

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

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

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

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

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

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

O. Ledoit and M. Wolf. A well-conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis*, pages 365–411, 2004.

D. D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

E. L. Lehmann. *Theory of Point Estimation*. Wiley, New York, 1983.

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

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(2):415–447, 1992.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

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

S. Nakajima and M. Sugiyama. Implicit regularization in variational Bayesian matrix factorization. In A. T. Joachims and J. Fürnkranz, editors, *Proceedings of 27th International Conference on Machine Learning (ICML2010)*, Haifa, Israel, Jun. 21–25 2010.

S. Nakajima and S. Watanabe. Variational Bayes solution of linear neural networks and its generalization performance. *Neural Computation*, 19(4):1112–1153, 2007.

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

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

T. Raiko, A. Ilin, and J. Karhunen. Principal component analysis for large scale problems with lots of missing values. In J. Kok, J. Koronacki, R. Lopez de Mantras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Proceedings of the 18th European Conference on Machine Learning*, volume 4701 of *Lecture Notes in Computer Science*, pages 691–698, Berlin, 2007. Springer-Verlag.

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

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

J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14(3):1080–1100, 1986.

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

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.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

P. Y.-S. Shao and W. E. Strawderman. Improving on the James-Stein positive-part estimator. *The Annals of Statistics*, 22:1517–1538, 1994.

N. Srebro and T. Jaakkola. Weighted low rank approximation. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning*. AAAI Press, 2003.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum margin matrix factorization. In *Advances in NIPS*, volume 17, 2005.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the 3rd Berkeley Symp. on Math. Stat. and Prob.*, pages 197–206, 1956.

C. Stein. Estimation of a covariance matrix. In *Rietz Lecture, 39th Annual Meeting IMS*, 1975.

W. E. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, 42:385–388, 1971.

D. Tao, M. Song, X. Li, J. Shen, J. Sun, X. Wu, C. Faloutsos, and S. J. Maybank. Tensor approach for 3-D face modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(10): 1397–1410, 2008.

K. Watanabe and S. Watanabe. Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, 7:625–644, 2006.

S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001.

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

D. Wipf and S. Nagarajan. A new view of automatric relevance determination. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1625–1632, Cambridge, MA, 2008. MIT Press.

H. Wold. Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah, editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York, NY, USA, 1966.

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

K. Yamazaki and S. Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, 2003.

K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. In *Proceedings of the Twenty-Second International Conference on Machine learning*, pages 1012–1019, 2005.

S. Yu, J. Bi, and J. Ye. Probabilistic interpretations and extensions for a family of 2D PCA-style algorithms. In *KDD Workshop on Data Mining using Matrices and Tensors*, 2008.