SIP (2018), vol. 7, e18, page 1 of 9 © The Authors, 2018.

doi:10.1017/ATSIP.2018.22

ORIGINAL PAPER

Estimation accuracy of covariance matrices when their eigenvalues are almost duplicated

KANTARO SHIMOMURA AND KAZUSHI IKEDA

The covariance matrix of signals is one of the most essential information in multivariate analysis and other signal processing techniques. The estimation accuracy of a covariance matrix is degraded when some eigenvalues of the matrix are almost duplicated. Although the degradation is theoretically analyzed in the asymptotic case of infinite variables and observations, the degradation in finite cases are still open. This paper tackles the problem using the Bayesian approach, where the learning coefficient represents the generalization error. The learning coefficient is derived in a special case, i.e., the covariance matrix is spiked (all eigenvalues take the same value except one) and a shrinkage estimation method is employed. Our theoretical analysis shows a non-monotonic property that the learning coefficient increases as the difference of eigenvalues increases until a critical point and then decreases from the point and converged to the distinct case. The result is validated by numerical experiments.

Keywords: Covariance matrix, Shrinkage estimation, Learning coefficient

Received 29 June 2018; Revised 4 November 2018

I. INTRODUCTION

The problem of estimating covariance matrices often appears in signal processing and multivariate analysis. Thus, the estimation accuracy of the covariance matrices is important in signal processing, in general. The estimation accuracy depends on the estimator. The most popular one is the sample covariance matrix and it is equivalent to the maximum likelihood estimator (MLE) under the Gaussian assumption. The MLE has good properties such as asymptotic unbiasedness, consistency, and asymptotic efficiency, but these apply only for big data. The MLE does not perform well when the dataset is small. Another case of failure in MLE is when some eigenvalues take the same or almost same values. For the identity matrix, e.g., the eigenvalues of the sample covariance matrix are spread out [1]. This problem is not rare. In fact, the subspace methods for system identification assume the covariance matrix is the sum of a low-rank matrix for signals and the identity matrix for noises, which leads to this kind of matrix [2]. Another case is the spike-and-slab prior distribution in Bayesian model selection [3]. Thus, it is important to elucidate the mechanism of the degradation in estimating duplicated eigenvalues especially for a small dataset.

One approach to tackle the problem of duplicated eigenvalues is shrinkage estimation methods [4–7]. Donoho et al. proposed an estimator that is optimal for the spiked covariance matrix, i.e., all the eigenvalues of the population matrix are equal except the maximum one [8]. They gave the risk of estimation called Stein's loss [9] as a function of the maximum eigenvalue using random matrix theory [10]. Their result shows the optimal estimator $\hat{\Sigma} = \mathrm{diag}(\hat{\lambda}_1, 1, 1, \ldots)$ for the population covariance is $\Sigma = \mathrm{diag}(\lambda_1, 1, 1, \ldots)$ is expressed as

$$\hat{\lambda}_1 = \begin{cases} \dfrac{\ell_1}{\alpha + (1-\alpha)\,\ell_1} & \text{if } \ell_1 > \left(1+\sqrt{\gamma}\right)^2, \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$

where

$$\alpha = \frac{1 - (\gamma)/((\ell_1 - 1)^2)}{1 + (\gamma)/(\ell_1 - 1)}, \tag{2}$$

γ is the ratio of the number of variables to that of observations, and ℓ_1 is the maximum eigenvalue of the sample covariance matrix expressed as

$$\ell_1 = \begin{cases} \lambda_1\left(1 + \gamma/\left(\lambda_1 - 1\right)\right) & \text{if } \lambda_1 > 1 + \sqrt{\gamma}, \\ \left(1+\sqrt{\gamma}\right)^2 & \text{otherwise.} \end{cases} \tag{3}$$

The result shows the risk of estimation, $L(\Sigma, \hat{\Sigma})$, increases as λ_1 increases, until the maximum eigenvalue reaches the transition point and then it decreases to zero thereafter. This means the risk is a non-monotonic function of λ_1 (Fig. 1).

Although this property of non-monotonicity is interesting, the theoretical results hold only in the asymptotic that both the numbers of variables and observations go to

Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0192, Japan

Corresponding author:
Kazushi Ikeda
Email: kazushi@is.naist.jp

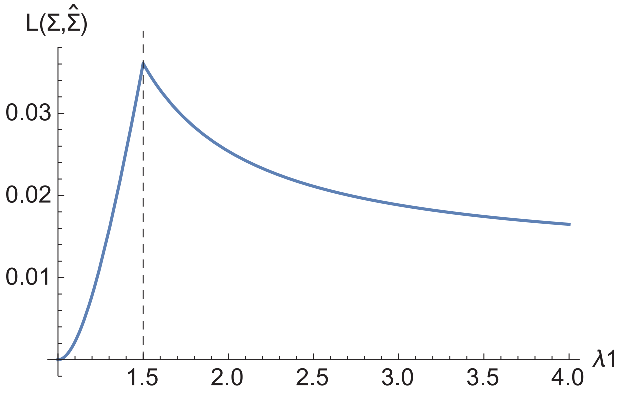

Fig. 1. Risk of estimation vs. the maximum eigenvalue ($\gamma = 1/4$).

infinity. As seen in the MLE case, finite variables and/or observations change the performance. Thus, an analysis of the risk in more realistic cases is necessary.

The problem of finite observations is formulated as the learning curve in machine learning, which is the averaged generalization/prediction error as a function of the number of observations [11]. In Bayesian statistics, the learning curve is expressed as the learning coefficient divided by the number of observations. That is, the learning coefficient is defined as the coefficient of the leading term of the mean Kullback–Leibler divergence from the true to predicted distributions [12, 13]. When the probability model is parametric and regular, i.e., uniquely identifiable, the learning coefficient is represented as a half of the number of the parameters. When the model is singular, however, the learning coefficient takes a smaller value than a half of the number of parameters, depending on the probability model. This idea is applicable to the estimation of covariance matrices with duplicated eigenvalues since this is a singular model due to the duplication of eigenvalues and unidentifiable eigenvectors associated with them [14].

The learning coefficient of a singular model is, however, difficult to derive in general and, only several special cases have been solved so far [15–18]. Thus, to give an exact analysis to a finite case, we considered a specific algorithm, a shrinkage method based on the empirical Bayes method, and derived its learning coefficient in a simple case of two dimensions. As a result, the learning coefficient of the algorithm has a non-monotonic property with respect to the maximum eigenvalue of the population covariance matrices in the same way as the infinite case, although the methods for these analyses are totally different. Finally, we confirmed the theoretical value through numerical experiments.

II. PRELIMINARIES

Let $X_i (i = 1, \ldots, N)$ be D-dimensional vectors independently drawn from an identical normal distribution with mean 0 and covariance Σ and

$$S = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^T, \tag{4}$$

the sample covariance matrix, respectively. Then, Σ and S are diagonalized using orthogonal matrices V and U as

$$\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_D) = V^T \Sigma V, \tag{5}$$

$$L = \mathrm{diag}\,(\ell_1, \ldots, \ell_D) = U^T S U, \tag{6}$$

where $\lambda_1 \geq \cdots \geq \lambda_D$ and $\ell_1 \geq \cdots \geq \ell_D$ are the eigenvalues of Σ and S and diag$(\cdot)$ denotes the diagonal matrix of $\cdot$. Note since the sample covariance matrix S and its eigenvalues L are sufficient statistics of the covariance matrix Σ and its eigenvalues Λ, estimators of Σ and Λ are written as $\hat{\Sigma}(S)$ and $\hat{\Lambda}(L)$, respectively, where $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)$.

Let us consider a shrinkage method that can explicitly treat duplication of eigenvalues. Since duplication of eigenvalues makes eigenvectors unidentifiable uniquely, we concentrate on the estimation of eigenvalues irrespective of eigenvectors. This is useful even when eigenvectors are desired because they can be calculated taking the duplication into account. For this purpose, we introduce a hierarchical model where V is chosen from the uniform distribution on the orthogonal matrix space with fixed Λ, i.e.,

$$X_i \sim \mathcal{N}(0, V \Lambda V^T), \quad V \sim p\,(V|\Lambda)\,, \tag{7}$$

and consider a Bayesian estimation method. Note it is natural to consider eigenvalues and eigenvectors separately when treat duplicated or repeated eigenvalues since the latter are not identifiable [19]. The joint distribution of S and V for this model is written as

$$\begin{aligned} p\,(S, V|N, \Lambda) &= p\,(L, U, V|N, \Lambda) \\ &= p\,(L, U|N, V, \Lambda)\; p\,(V|\Lambda)\,. \end{aligned} \tag{8}$$

Here, $p(L, U|N, V, \Lambda)$ is the joint probability of the eigenvalues and their eigenvectors of sample covariance matrix, which is obtained by transforming the Wishart distribution with degree of freedom N and the scale matrix $V \Lambda V^T$. An important property of this model is the support of the distribution of covariance matrix, $p(\Sigma) = p(V \Lambda V^T)$, varies depending on the hyperparameters Λ. For example, when Λ is the identity matrix, Σ is also the identity matrix irrespective of V.

A popular method to determine the hyperparameters Λ is the empirical Bayes method. We calculate the marginalized likelihood by integrating $p(L, U|N, V, \Lambda) p(V|\Lambda)$ over the orthogonal matrix space and maximize it, i.e.,

$$\hat{\Lambda}\,(L) = \underset{\Lambda}{\mathrm{argmax}}\; p\,(S|\Lambda) \tag{9}$$

$$= \underset{\Lambda}{\mathrm{argmax}} \int p\,(L, U|N, V, \Lambda)\; p\,(V|\Lambda)\, dV. \tag{10}$$

Since this integration is invariant under any orthogonal transformation $S \mapsto VSV^T$, the distribution of V can be replaced with the Haar measure on the orthogonal matrix space [20]. Then, the marginal likelihood is expressed using hypergeometric function with matrix argument [21]. Since the marginal likelihood is difficult to maximize due to

https://doi.org/10.1017/ATSIP.2018.22 Published online by Cambridge University Press

the complexity of the hypergeometric function, in general, we only consider a special case of the spiked covariance model for Σ, i.e., $\lambda_1 \geq \lambda_2 = \lambda_3 = \cdots = \lambda_D$ [1]. In addition, we use the eigenvectors of the sample covariance matrix for the estimation instead of the posterior $p(V|S)$, i.e., $\hat{\Sigma} = U\hat{\Lambda}U^T$, which is often supposed in the shrinkage estimation.

Note although the learning coefficient is originally defined for Bayesian estimation, it makes sense since shrinkage estimation can be regarded as an approximation of the posterior in hierarchical models [22].

III. MAIN RESULTS

We show some facts on the shrinkage estimation with our hierarchical model before deriving the learning coefficient of the proposed estimation method. See Appendix for the proofs.

Proposition 1. *For the marginal likelihood of the hierarchical model (7) with spiked covariance model, the following statements hold.*

(i) The marginal likelihood is written as

$$\begin{aligned} p(S|\Lambda) &= \int p(L, U|N, V, \Lambda)\, p(V|\Lambda)\, dV \\ &\propto |\Lambda|^{-N/2} \exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad \cdot {}_1F_1\left(\frac{1}{2}; \frac{D}{2}; \frac{N}{2}\left(\lambda_2^{-1} - \lambda_1^{-1}\right) S\right) \end{aligned}$$

where ${}_1F_1$ denotes the hypergeometric function with matrix argument [23].

(ii) $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_2)$ that maximizes $p(S|\Lambda)$ satisfies the following equations,

$$\mathrm{tr}\Lambda = \lambda_1 + (D-1)\lambda_2 = \mathrm{tr}S, \tag{11}$$

$$\lambda_1 = \frac{1}{N}\frac{F'\left(N/2\left(\lambda_2^{-1} - \lambda_1^{-1}\right)\right)}{F\left(N/2\left(\lambda_2^{-1} - \lambda_1^{-1}\right)\right)}, \tag{12}$$

where $F(z) = {}_1F_1(1/2; D/2; zS)$ and $F'(z)$ is its derivative.

(iii) If there exists a non-trivial solution of the equations above, it maximizes the marginal likelihood, where the trivial solution is $\lambda_1 = \lambda_2 = \mathrm{tr}S/D$.

(iv) The equations have a non-trivial solution if and only if

$$ND\sum_{i<j}(\ell_i - \ell_j)^2 - (D^2 + D - 2)(\mathrm{tr}S)^2 > 0. \tag{13}$$

In the two-dimensional case, the marginal likelihood is explicitly calculated.

Corollary 2. *When $D = 2$, the marginal likelihood is given by*

$$\begin{aligned} p(S|\Lambda) = &\frac{N^N(\ell_1 - \ell_2)(\ell_1\ell_2)^{(N-3)/2}(\lambda_1\lambda_2)^{-N/2}}{4\Gamma(N-1)} \\ &\cdot \exp\left[-\frac{N(\ell_1 + \ell_2)(\lambda_1 + \lambda_2)}{4\lambda_1\lambda_2}\right] \\ &\cdot I_0\left(\frac{N(\ell_1 - \ell_2)(\lambda_1 - \lambda_2)}{4\lambda_1\lambda_2}\right), \end{aligned} \tag{14}$$

where $I_k(z)$ denotes the modified Bessel function of the first kind with order k.

Since $\int p(S|\Lambda)d\ell_1 d\ell_2 = 1$ holds, (14) can be regarded as the joint probability density function of the eigenvalues of the sample covariance matrix with respect to ℓ_1, ℓ_2. By substituting $D = 2$ into Proposition 1, the estimator of eigenvalues is given explicitly.

Corollary 3. *The estimator that maximizes (14) is given by*

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \begin{cases} \left(\dfrac{\ell_1 + \ell_2}{2}, \dfrac{\ell_1 + \ell_2}{2}\right), & \text{if } \left(\dfrac{\ell_1}{\ell_2} - 1\right) N^{1/2} \\ & < 2^{3/2}\left(1 + \dfrac{\sqrt{2}N^{1/2} + 2}{N - 2}\right), \\ \left(\dfrac{\ell_1 + \ell_2 + t}{2}, \dfrac{\ell_1 + \ell_2 - t}{2}\right), & \text{otherwise,} \end{cases} \tag{15}$$

where t is the solution of

$$t = (\ell_1 - \ell_2)\, A\left(\frac{N(\ell_1 - \ell_2)\, t}{(\ell_1 + \ell_2)^2 - t^2}\right) \tag{16}$$

and $A(z) = I_1(z)/I_0(z)$.

Corollary 3 gives the relation between the eigenvalues of the sample covariance matrix and the estimated eigenvalues by our shrinkage estimation (Fig. 2). When the two eigenvalues l_1 and l_2 are closer together than a threshold, the estimators of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ take the same value.

A) Derivation of learning coefficient

The learning coefficient is defined based on the Kullback–Leibler divergence [12]. The mean of the Kullback–Leibler divergence from the true distribution p to the predicted one q is called the Bayesian generalization error in asymptotics, which is written as

$$E\left[\mathrm{KL}\left(p||q\right)\right] = \frac{\kappa}{N} + o\left(\frac{1}{N}\right), \tag{17}$$

where $E[\cdot]$ denotes the expectation with respect to the observations. The coefficient of leading term κ is called the learning coefficient. In case that the distributions are D-dimensional multivariate normal distribution with mean

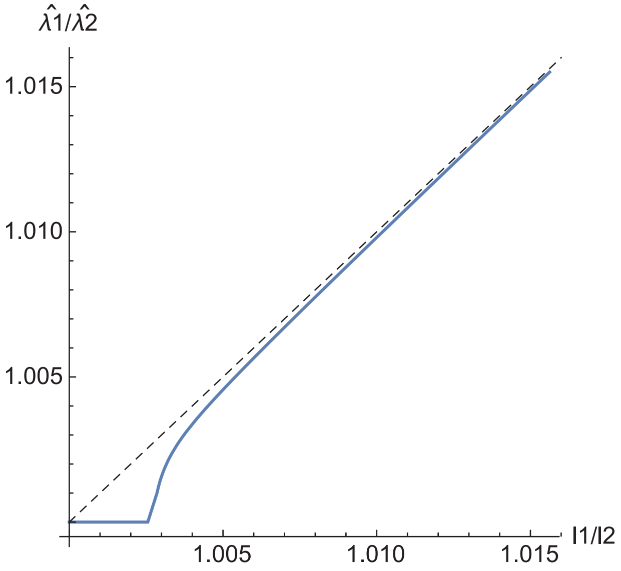

Fig. 2. The relation between the eigenvalues of the sample covariance matrix and those of the estimated eigenvalues. $N = 2^{20}$.

0 and covariance matrices Σ_1 and Σ_2, respectively, the Kullback–Leibler divergence is described as

$$\mathrm{KL}\left(\Sigma_1 || \Sigma_2\right) = \frac{1}{2}\left(\mathrm{tr}\Sigma_1^{-1}\Sigma_2 - \log \Sigma_1^{-1}\Sigma_2 - D\right),$$

which is equivalent to the Stein's loss of covariance matrices [9]. When a statistical model is regular, as a special case, its learning coefficient of maximum likelihood estimation becomes the half of the number of parameters [12], i.e.,

$$E\left[\mathrm{KL}\left(\Sigma || S\right)\right] N \simeq \frac{D\left(D+1\right)}{2},$$

when the estimator of the covariance matrix is the sample covariance matrix.

In our case, the expectation with respect to the observations $X_i (i = 1, \dots, N)$ is equivalent to the expectation with respect to the sample covariance matrix S. Thus, the learning coefficient for $\hat{\Sigma}(S)$ is given by

$$\begin{aligned} & E[\mathrm{KL}(\Sigma || \hat{\Sigma}(S))]N \\ & \simeq \frac{D(D+1)}{2} + E[\mathrm{tr}\Sigma^{-1}(\hat{\Sigma}(S) - S)]N \\ & \quad - E[\log|S^{-1}\hat{\Sigma}(S)|]N. \end{aligned} \tag{18}$$

Using the proposed shrinkage estimator of the covariance matrix,

$$\hat{\Sigma}\left(S\right) = U \begin{bmatrix} \hat{\lambda}_1 & \\ & \hat{\lambda}_2 \end{bmatrix} U^T, \tag{19}$$

and the reparametrization as

$$\begin{aligned} \lambda &= \lambda_2 & c &= \left(\lambda_1/\lambda_2 - 1\right) N^{1/2} \\ \ell &= \ell_2 d & &= \left(\ell_1/\ell_2 - 1\right) N^{1/2} \\ e &= t/\ell_2 N^{1/2}, & & \end{aligned} \tag{20}$$

the second and the third terms of (18) are given by

$$\begin{aligned} & E[\mathrm{tr}\Sigma^{-1}(\hat{\Sigma} - S)]N = \frac{c}{2}\lambda^{-1}(1 + cN^{-1/2})^{-1} \\ & \quad \cdot E\left[(d-e)\ell\left(1 - 2\left(V_{11}^2 + U_{11}^2\right)\right.\right. \\ & \qquad \left.\left. + 4V_{11}U_{11}\left(V_{11}U_{11} + V_{12}U_{12}\right)\right)\right] \end{aligned} \tag{21}$$

$$\simeq \frac{c}{2}A\left(\frac{c^2}{4}\right)E[d-e], \tag{22}$$

$$\begin{aligned} & E[\log|S^{-1}\hat{\Sigma}|]N \\ & = E\left[\log(1 + dN^{-1/2})^{-1}\left(1 + \frac{d+e}{2}N^{-1/2}\right)\right. \\ & \qquad \left. \cdot \left(1 + \frac{d-e}{2}N^{-1/2}\right)\right] \end{aligned}$$

$$\simeq \frac{1}{4}E[d^2 - e^2], \tag{23}$$

respectively. Note that this kind of reparameterization is widely used in evaluating learning coefficients [13, 22]. In (22) and (23), the difference between the eigenvalues of shrinkage estimator, e, and the probability density function of the distance of eigenvalues of the sample covariance matrices, d, can respectively be approximated to the solution of $dI_1(de/4) - eI_0(de/4) = 0$ and the solution of $p(d) = \frac{d}{4}\exp[-(c^2 + d^2)/8]I_0(cd/4)$. See Appendix for the validity of these approximations.

B) Numerical experiments

To validate the derived learning coefficient, some numerical experiments were carried out. The experimental learning coefficients were calculated as the Kullback–Leibler divergence from the true distribution to the predicted distribution averaged over 10^4 sample covariance matrices, multiplied by $2N$, while the theoretical learning coefficients were numerically calculated using (18), (22), and (23). The two learning coefficients coincide well, including their non-monotonicity (Fig. 3).

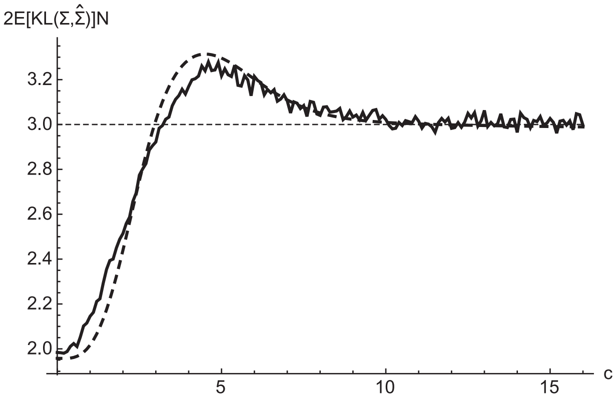

Fig. 3. Learning coefficients versus c, the normalized eigenvalue ratio. Experiments (solid) and theory (dashed).

https://doi.org/10.1017/ATSIP.2018.22 Published online by Cambridge University Press

IV. DISCUSSION

How duplication of eigenvalues in covariance matrices affects estimation was clarified here. We derived the learning coefficient of a shrinkage estimator based on the empirical Bayes method where both the numbers of variables and observations are finite. Our shrinkage estimation method has two phases depending on the sample eigenvalues, i.e., in one phase the estimation error increases when the population covariance matrix has eigenvalues closer to each other than a threshold while in the other phase the learning coefficient varies smoothly with respect to the difference of eigenvalues. This phenomenon is consistent with an asymptotic case where the numbers of variables and observations are infinite. However, the influence of duplication is stochastic in the finite case and the state fluctuates between the two phases while the influence of duplication is deterministic in the infinite case. The analysis of the fluctuation is still open.

Our analysis treated the simple case of two-dimensional. To expand our results, we need to calculate the marginal likelihood in (8) for general cases. Although this is written using the generalized Bingham distribution or its product, the derivation of the estimators or the conditions is difficult due to the complexity of the hypergeometric function. We need a new method to solve this problem. Nonetheless, this work has given a new insight to this problem because the Bayesian approach is shown to be hopeful.

FINANCIAL SUPPORT

This work was supported by JSPS KAKENHI Grant Numbers 25280083, 17H01817, 17H05863.

STATEMENT OF INTEREST

None.

REFERENCES

[1] Johnstone, I.M.: On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, **29** (2001), 475–501.

[2] Katayama, T.: Subspace Methods for System Identification, Springer, London, 2005.

[3] Ishwaran, H.; Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.*, **33** (2005), 730–773.

[4] Bickel, P.J.; Levina, E.: Regularized estimation of large covariance matrices. *Ann. Stat.*, **36** (2008), 199–227.

[5] Karoui, N.E.: Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Stat.*, **36** (2008), 2757–2790.

[6] Rao, N.R.; Mingo, J.A.; Speicher, R.; Edelman, A.: Statistical eigen-inference from large Wishart matrices. *Ann. Stat.*, **36** (2008), 2850–2885.

[7] Ledoit, O. et al.: Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Stat.*, **40** (2012), 1024–1060.

[8] Donoho, D.L.; Gavish, M.; Johnstone, I.M.: Optimal shrinkage of eigenvalues in the spiked covariance model. In Technical Report 2013-10, Department of Statistics, Stanford University, Stanford, 2013.

[9] James, W.; Stein, C.: Estimation with Quadratic Loss, in *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability*, 1, 1961, 361–379.

[10] Baik, J.; Ben, Arousa, G., Peche, S.: Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann Probab.*, **33** (2005), 1643–1697.

[11] Amari, S.; Fujita, N.; Shinomoto, S.: Four types of learning curves. *Neural. Comput.*, **4** (1992), 605–618.

[12] Watanabe, S.: Algebraic analysis for nonidentifiable learning machines. *Neural. Comput.*, **13** (2001), 899–933.

[13] Watanabe, S.; Amari, S.: Learning coefficients of layered models when the true distribution mismatches the singularities. *Neural. Comput.*, **15** (2003), 1013–1033.

[14] Sheena, Y.: Inference on the eigenvalues of the covariance matrix of a multivariate normal distribution—Geometrical view. *J. Stat. Plan. Inference.*, **150** (2014), 66–83.

[15] Yamazaki, K.; Watanabe, S.: Singularities in mixture models and upper bounds of stochastic complexity. *Neural. Netw.*, **16** (2003), 1029–1038.

[16] Yamazaki, K.; Watanabe, S.: Singularities in complete bipartile graph-type Boltzmann machines and upper bounds of stochastic complexities. *IEEE Trans. Neural Netw.*, **16** (2005), 312–324.

[17] Aoyagi, M.; Watanabe, S.: Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural. Netw.*, **18** (2005), 914–933.

[18] Watanabe, K.; Watanabe, S.: Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *J. Mach. Learn. Res.*, **7** (2006), 625–643.

[19] Kenny, S.P.; Hou, G.J.W.: Approximate analysis for repeated eigenvalue problems with applications to controls structures integrated design, NASA Technical Report, 3439, 1994.

[20] Braun, D.: Invariant integration over the orthogonal group. *J. Phys. A: Gen. Phys.*, **39** (2006), 14581.

[21] James, A.T.: Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Stat.*, **35** (1964), 475–501.

[22] Nakajima, S.; Watanabe, S.: Generalization error of linear neural networks in an empirical Bayes approach, in *Proc. IJCAI2015* 2005, 804–810.

[23] Magnus, W.; Oberhettinger, F.: Formulas and Theorems for the Special Functions of Mathematical Physics, Chelsea Publishing Company, New York, 1949.

[24] Jupp, P.E.; Mardia, K.V.: Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *Ann. Stat.*, **7** (1979), 599–606.

[25] Mardia, K.V.; Zemroch, P.J.: Algorithm AS 86: the von Mises distribution function. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **24** (1975), 268–272.

APPENDIX

Proof of Proposition 1

Since $p(L, U|N, V, \Lambda)$ in (8) is the joint distribution of eigenvalues and eigenvectors transformed from the Wishart distribution with degree of freedom N and scale matrix

https://doi.org/10.1017/ATSIP.2018.22 Published online by Cambridge University Press

$V\Lambda V^T$, the probability density function is given by

$$\begin{aligned} &p\left(L, U|N, V, \Lambda\right) \\ &= \frac{N^{ND/2}\left|S\right|^{(N-D-1)/2}\exp\left(-\left(N/2\right)\mathrm{tr}\left(SV\Lambda^{-1}V^T\right)\right)}{2^{ND/2}\left|\Lambda\right|^{N/2}\Gamma_D\left(N/2\right)} \\ &\quad\cdot\prod_{i<j}\left|\ell_i-\ell_j\right| \qquad\qquad (\mathrm{A.1}) \\ &= \frac{N^{ND/2}\left|L\right|^{(N-D-1)/2}\exp\!\left(-\left(N/2\right)\mathrm{tr}\!\left(ULU^TV\Lambda^{-1}V^T\right)\right)}{2^{ND/2}\left|\Lambda\right|^{N/2}\Gamma_D\left(N/2\right)} \\ &\quad\cdot\prod_{i<j}\left|\ell_i-\ell_j\right|, \qquad\qquad (\mathrm{A.2}) \end{aligned}$$

where Γ_D denotes the multivariate gamma function defined as

$$\Gamma_D\left(N/2\right) = \pi^{D(D-1)/2}\prod_{i=1}^{D}\Gamma\left(\frac{1}{2}\left(N-i+1\right)\right). \qquad (\mathrm{A.3})$$

Here, the last term $\prod_{i<j}\left|\ell_i-\ell_j\right|$ is the Jacobian to transform variables from the elements of matrix to eigenvalues and eigenvectors.

The marginal likelihood in (8) is calculated as

$$\begin{aligned} &p\left(S|\Lambda\right) \\ &= \int p\left(L, U|N, V, \Lambda\right)p\left(V|\Lambda\right)dV \\ &\propto \left|\Lambda\right|^{-N/2}\int\exp\left[-\frac{N}{2}\mathrm{tr}SV\Lambda^{-1}V^T\right]dV \\ &= \left|\Lambda\right|^{-N/2}\int\exp\left[-\frac{N}{2}\mathrm{tr}SV\left(\begin{bmatrix}\lambda_1^{-1}-\lambda_2^{-1} & & \\ & 0 & \\ & & \ddots & \\ & & & 0\end{bmatrix}\right.\right. \\ &\quad\left.\left.+\begin{bmatrix}\lambda_2^{-1} & & & \\ & \lambda_2^{-1} & & \\ & & \ddots & \\ & & & \lambda_2^{-1}\end{bmatrix}\right)V^T\right]dV \\ &= \left|\Lambda\right|^{-N/2}\exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad\cdot\int\exp\left[-\frac{N}{2}\mathrm{tr}SV\left(\begin{bmatrix}\lambda_1^{-1}-\lambda_2^{-1} & & \\ & 0 & \\ & & \ddots & \\ & & & 0\end{bmatrix}\right)V^T\right]dV \\ &= \left|\Lambda\right|^{-N/2}\exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad\cdot\int\exp\left[-\frac{N}{2}\left(\lambda_1^{-1}-\lambda_2^{-1}\right)\mathrm{tr}SV\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}^TV^T\right]dV \\ &\propto \left|\Lambda\right|^{-N/2}\exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad\cdot\int\exp\left[-\frac{N}{2}\left(\lambda_1^{-1}-\lambda_2^{-1}\right)\mathrm{tr}Svv^T\right]dS^{D-1} \\ &\propto \left|\Lambda\right|^{-N/2}\exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad\cdot\int\mathrm{Bingham}\left(v|\frac{N}{2}\left(\lambda_1^{-1}-\lambda_2^{-1}\right)S\right)dS^{D-1} \\ &= \left|\Lambda\right|^{-N/2}\exp\left[-\frac{N}{2}\lambda_2^{-1}\mathrm{tr}S\right] \\ &\quad\cdot{}_1F_1\left(\frac{1}{2};\frac{D}{2};\frac{N}{2}\left(\lambda_2^{-1}-\lambda_1^{-1}\right)S\right), \end{aligned}$$

where $\mathrm{Bingham}(v|N/2(\lambda_1^{-1}-\lambda_2^{-1})S)$ denotes the probability density function of the Bingham distribution without normalization and dS^{D-1} means the integration over the D-dimensional unit sphere. The normal constant of the Bingham distribution is written as hypergeometric function with matrix argument. Note that we can extend this calculation to the case of

$$\Lambda = \begin{bmatrix}\lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_1 & & & \\ & & & \lambda_2 & & \\ & & & & \ddots & \\ & & & & & \lambda_2\end{bmatrix}$$

by using the generalized Bingham distribution [24]. Let $F(z)$ denote ${}_1F_1(1/2; D/2; zS)$ hereafter for simplicity.

The conditions of Λ to maximize $p(S|\Lambda)$ is given by making the gradients of the likelihood function null, i.e.,

$$\begin{aligned} &\frac{\partial}{\partial\lambda_1}\log p\left(S|\Lambda\right) \\ &\quad= -\frac{N}{2}\lambda_1^{-1}+\frac{N}{2}\lambda_1^{-2}\frac{F'\left(N/2\left(\lambda_2^{-1}-\lambda_1^{-1}\right)\right)}{F\left(N/2\left(\lambda_2^{-1}-\lambda_1^{-1}\right)\right)} = 0, \\ &\frac{\partial}{\partial\lambda_2}\log p\left(S|\Lambda\right) = -\frac{N\left(D-1\right)}{2}\lambda_2^{-1}+\frac{N\mathrm{tr}S}{2}\lambda_2^{-1} \\ &\quad-\frac{N}{2}\lambda_2^{-2}\frac{F'\left(N/2\left(\lambda_2^{-1}-\lambda_1^{-1}\right)\right)}{F\left(N/2\left(\lambda_2^{-1}-\lambda_1^{-1}\right)\right)} \\ &\quad= 0. \end{aligned}$$

The second claim in Proposition 1 is straightforward from the above. The third claim is shown by the uniqueness of the maximum likelihood estimator of the Bingham distribution [24] and the monotonicity of $z^{-N/2}$, $\exp(-z)$ and $F(z)$. From (11), it is sufficient to consider only the case $\lambda_2 = (\mathrm{tr}S-\lambda_1)/(D-1)$ to derive the condition a non-trivial solution exists. To do this, we rewrite $\log p(S|\Lambda)$ as a function of λ_1,

$$\begin{aligned} g\left(\lambda_1\right) &= \log p\left(S|\mathrm{diag}\left(\lambda_1, \left(\mathrm{tr}S-\lambda_1\right)/\left(D-1\right)\right)\right) \\ &= \log\left[\lambda_1^{-N/2}\left(\frac{\mathrm{tr}S-\lambda_1}{D-1}\right)^{-N/2+D-1}\right. \end{aligned}$$

$$\exp\left[-\frac{N(D-1)\,\mathrm{tr}S}{2(\mathrm{tr}S-\lambda_1)}\right.$$
$$\left.\cdot\, F\left(\frac{N}{2}\left(\left(\frac{\mathrm{tr}S-\lambda_1}{D-1}\right)^{-1}-\lambda_1^{-1}\right)\right)\right],$$

and consider the condition where the trivial solution is not the maximizer, i.e., $g''(\lambda_1) > 0$ at $\lambda_1 = \mathrm{tr}S/D$. To calculate $g''(\mathrm{tr}S/D)$, we need the series expansion of $F(z)$ at $z = 0$. The series expansion of hypergeometric function with matrix argument can be written using zonal polynomial $\mathcal{C}_\kappa$ [21],

$$\begin{aligned}F(z) &= {}_1F_1\left(\frac{1}{2};\frac{D}{2};zS\right)\\ &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{(1/2)_\kappa}{(D/2)_\kappa}\frac{\mathcal{C}_\kappa(zS)}{k!}\\ &= \sum_{k=0}^{\infty}\left(\sum_{\kappa\vdash k}\frac{(1/2)_\kappa}{(D/2)_\kappa}\mathcal{C}_\kappa(S)\right)\frac{z^k}{k!}\end{aligned}$$

where $\kappa = (k_1,\ldots,k_l) \vdash k$ denotes partitions of an integer and $(a)_\kappa$ is the generalized Pochhammer symbol defined as

$$\begin{aligned}(a)_\kappa &= \prod_{i=1}^{l}\left(a-\frac{i-1}{2}\right)_{k_i},\\ (a)_{k_i} &= \frac{\Gamma(a+k_i)}{\Gamma(a)} = a(a+1)\cdots(a+k_i-1).\end{aligned}$$

From the above, we have

$$\begin{aligned}F(0) &= 1,\\ F'(0) &= \frac{\mathrm{tr}S}{D},\\ F''(0) &= \frac{(\mathrm{tr}S)^2+2\sum_{i=1}^{D}\ell_i^2}{D(D+2)}\\ &= \frac{(D+2)(\mathrm{tr}S)^2+2\sum_{i<j}\left(\ell_i-\ell_j\right)^2}{D^2(D+2)},\end{aligned}$$

and then

$$g''\left(\frac{\mathrm{tr}S}{D}\right)$$
$$= -\frac{D^3N((D^2+D-2)(\mathrm{tr}S)^2-ND\sum_{i<j}(\ell_i-\ell_j)^2)}{2(D-1)^2(D+2)(\mathrm{tr}S)^4}.$$

Since $g''(\mathrm{tr}S/D) > 0$, we get the fourth claim in Proposition 1.

Proof of Corollary 2

Using the formula

$$I_0(1) = \frac{1}{\pi}\int_0^{\pi}\exp[\cos x]\,dx$$

and the invariance of integration against the transformation $V \mapsto UV$, (16) is calculated as

$$\int p(L,U|N,V,\Lambda)\,p(V|\Lambda)\,dV$$
$$\propto (\ell_1-\ell_2)(\ell_1\ell_2)^{(N-3)/2}\int\exp\left[-\frac{N}{2}ULU^TV\Lambda^{-1}V^T\right]dV$$
$$= (\ell_1-\ell_2)(\ell_1\ell_2)^{(N-3)/2}$$
$$\cdot\int_0^{\pi}\exp\left[-\frac{N}{2}\mathrm{tr}\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}L\right.$$
$$\left.\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\Lambda^{-1}\right]d\theta$$
$$= (\ell_1-\ell_2)(\ell_1\ell_2)^{(N-3)/2}$$
$$\cdot\int_0^{\pi}\exp\left[\frac{N(\ell_1-\ell_2)(\lambda_1-\lambda_2)}{4\lambda_1\lambda_2}\cos 2\theta\right.$$
$$\left.-\frac{N(\ell_1+\ell_2)(\lambda_1+\lambda_2)}{4\lambda_1\lambda_2}\right]d\theta$$
$$\propto (\ell_1-\ell_2)(\ell_1\ell_2)^{(N-3)/2}\exp\left[-\frac{N(\ell_1+\ell_2)(\lambda_1+\lambda_2)}{4\lambda_1\lambda_2}\right]$$
$$\cdot\, I_0\left(\frac{N(\ell_1-\ell_2)(\lambda_1-\lambda_2)}{4\lambda_1\lambda_2}\right).$$

The remainder term is the normalization constant.

Approximations

By the reparametrization in (20), Σ, S, and $\hat{\Sigma}(S)$ are rewritten as

$$\begin{aligned}\Sigma &= V\begin{bmatrix}\lambda\left(1+cN^{-1/2}\right) & \\ & \lambda\end{bmatrix}V^T,\\ S &= U\begin{bmatrix}\ell\left(1+dN^{-1/2}\right) & \\ & \ell\end{bmatrix}U^T,\\ \hat{\Sigma}(S) &= U\begin{bmatrix}\ell\left(1+\frac{d+e}{2}N^{-1/2}\right) & \\ & \ell\left(1+\frac{d-e}{2}N^{-1/2}\right)\end{bmatrix}U^T,\end{aligned}$$

respectively. Then,

$$E[\mathrm{tr}\Sigma^{-1}(\hat{\Sigma}-S)]N$$
$$= \frac{1}{2}\lambda^{-1}E\left[(d-e)\,\ell\,\mathrm{tr}V\begin{bmatrix}(1+cN^{-1/2})^{-1} & \\ & 1\end{bmatrix}\right.$$
$$\left.V^TU\begin{bmatrix}-1 & \\ & 1\end{bmatrix}U^T\right]$$
$$= \frac{c}{2}\lambda^{-1}(1+cN^{-1/2})^{-1}$$
$$\cdot\, E\left[(d-e)\,\ell\left(1-2\left(V_{11}^2+U_{11}^2\right)\right.\right.$$
$$+4V_{11}U_{11}\left(V_{11}U_{11}+V_{12}U_{12}\right))].$$

By substituting $V_{11} = \cos\theta_0, U_{11} = \cos\theta$, we get

$$\begin{aligned}
&E\left[(d-e)\,\ell\left(1-2\left(V_{11}^2+U_{11}^2\right)\right.\right. \\
&\qquad \left.\left. +4V_{11}U_{11}\left(V_{11}U_{11}+V_{12}U_{12}\right)\right)\right] \\
&= \frac{c}{2}\lambda^{-1}(1+cN^{-1/2})^{-1}E\left[(d-e)\,\ell\cos\left(2\left(\theta_0-\theta\right)\right)\right] \\
&\simeq \frac{c}{2}\lambda^{-1}(1+cN^{-1/2})^{-1}E\left[(d-e)\,\ell\left(1-2\left(\theta_0-\theta\right)^2\right)\right].
\end{aligned}$$

The conditional distribution of θ for expectation is

$$\begin{aligned}
&p\left(L,U|N,V,\Lambda\right) \\
&\propto \exp\left[-\frac{N}{2}ULU^TV\Lambda^{-1}V^T\right] \\
&\propto \exp\left[-\frac{N}{4}\left(\ell_1-\ell_2\right)\left(\lambda_2^{-1}-\lambda_1^{-1}\right)\cos\left(2\left(\theta_0-\theta\right)\right)\right] \\
&= \exp\left[-\frac{1}{4}cd\ell\lambda^{-1}\left(1+cN^{-1/2}\right)^{-1}\cos\left(2\left(\theta_0-\theta\right)\right)\right]
\end{aligned}$$

from (A.2), which is the probability density function of the von Mises distribution [25]. Thus, we have

$$\begin{aligned}
&E\left[(d-e)\ell\left(1-2\left(\theta_0-\theta\right)^2\right)\right] \\
&= E\left[(d-e)\ell A\left(\frac{1}{4}cd\ell\lambda^{-1}(1+cN^{-1/2})^{-1}\right)\right]
\end{aligned}$$

by considering the variance of von Mises distribution, where

$$\begin{aligned}
E\left[\ell\right] &\simeq \lambda, \\
A\left(\frac{1}{4}cd\ell\lambda^{-1}(1+cN^{-1/2})^{-1}\right) &\simeq A(c^2/4)
\end{aligned}$$

and $A(z) = I_1(z)/I_0(z)$. Assuming the independence of ℓ and d, we get

$$\begin{aligned}
E\left[\mathrm{tr}\Sigma^{-1}(\hat{\Sigma}-S)\right]N &\simeq \frac{c}{2}(1+cN^{-1/2})^{-1}A(c^2/4)E\left[d-e\right] \\
&\to \frac{c}{2}A(c^2/4)E\left[d-e\right] \quad (N\to\infty).
\end{aligned}$$

In a similar fashion, we get

$$\begin{aligned}
&E\left[\log\left|S^{-1}\hat{\Sigma}\right|\right]N \\
&= E\left[\log(1+dN^{-1/2})^{-1}\left(1+\frac{d+e}{2}N^{-1/2}\right)\right. \\
&\qquad \left.\cdot\left(1+\frac{d-e}{2}N^{-1/2}\right)\right]N \\
&= \left(E\left[-\log(1+dN^{-1/2})\right.\right. \\
&\qquad \left.\left.+\log\left(1+dN^{-1/2}+\frac{d^2-e^2}{4}N^{-1}\right)\right]\right)N \\
&= \left(E\left[-dN^{-1/2}+\frac{d^2}{2}N^{-1}+dN^{-1/2}+\frac{d^2-e^2}{4}N^{-1}\right.\right. \\
&\qquad \left.\left.-\frac{1}{2}\left(dN^{-1/2}+\frac{d^2-e^2}{4}N^{-1}\right)^2+O(N^{-3/2})\right]\right)N \\
&= E\left[\frac{d^2-e^2}{4}+O(N^{-1/2})\right] \\
&\to \frac{1}{4}E\left[d^2-e^2\right] \quad (N\to\infty).
\end{aligned}$$

As for the approximation on e and d, it is necessary to evaluate $E[d-e]$ and $E[d^2-e^2]$ numerically, where

$$e = \begin{cases} 0 & \text{if } d < 2^{3/2}\left(1+\frac{\sqrt{2}N^{1/2}+2}{N-2}\right), \\ t/\ell N^{1/2} & \text{otherwise,} \end{cases}$$

with t in Corollary 3. Rewriting (16), we get the approximation of e as

$$\begin{aligned}
e\ell N^{-1/2} &= d\ell N^{-1/2}A\left(\frac{de\ell^2}{\ell^2(2+dN^{-1/2})^2-e^2\ell^2N^{-1}}\right), \\
e &= dA\left(\frac{de}{(2+dN^{-1/2})^2-e^2N^{-1}}\right) \\
&\to dA\left(\frac{de}{4}\right) \quad (N\to\infty).
\end{aligned}$$

In addition, when N goes to infinity,

$$2^{3/2}\left(1+\frac{\sqrt{2}N^{1/2}+2}{N-2}\right) \to 2^{3/2}$$

holds. This means e approximates to the solution of $dI_1(de/4) - eI_0(de/4) = 0$ when $d < 2^{3/2}$ and $e = 0$ otherwise. To get the approximation of d, we transform $p(\ell_1, \ell_2)$ into $p(\ell, d)$ and marginalize it as

$$\begin{aligned}
&p\left(\ell_1,\ell_2\right)d\ell_1 d\ell_2 \\
&= \frac{N^N\left(\ell_1-\ell_2\right)\left(\ell_1\ell_2\right)^{(N-3)/2}\left(\lambda_1\lambda_2\right)^{-N/2}}{4\Gamma\left(N-1\right)} \\
&\quad \cdot\exp\left[-\frac{N\left(\ell_1+\ell_2\right)\left(\lambda_1+\lambda_2\right)}{4\lambda_1\lambda_2}\right] \\
&\quad \cdot I_0\left(\frac{N\left(\ell_1-\ell_2\right)\left(\lambda_1-\lambda_2\right)}{4\lambda_1\lambda_2}\right)d\ell_1 d\ell_2 \\
&= \frac{N^N(dN^{-1/2})\ell^{N-2}(1+dN^{-1/2})^{(N-3)/2}\lambda^{-N}(1+cN^{-1/2})^{-N/2}}{4\Gamma\left(N-1\right)} \\
&\quad \cdot\exp\left[-\frac{N\ell(2+dN^{-1/2})(2+cN^{-1/2})}{4\lambda(1+cN^{-1/2})}\right] \\
&\quad \cdot I_0\left(\frac{cd\ell}{4\lambda(1+cN^{-1/2})}\right)\ell N^{-1/2}d\ell dd,
\end{aligned}$$

https://doi.org/10.1017/ATSIP.2018.22 Published online by Cambridge University Press

and

$$\begin{aligned}
p(d) &= \int_0^\infty p(\ell, d) d\ell \\
&= \frac{N^{N-1/2}(dN^{-1/2})(1+dN^{-1/2})^{(N-3)/2}\lambda^{-N}(1+cN^{-1/2})^{-N/2}}{4\Gamma(N-1)} \\
&\quad \cdot \int_0^\infty \ell^{N-1} \exp\left[-\frac{N\ell(2+dN^{-1/2})(2+cN^{-1/2})}{4\lambda(1+cN^{-1/2})}\right] \\
&\quad \cdot I_0\left(\frac{cd\ell}{4\lambda(1+cN^{-1/2})}\right) d\ell \\
&= \frac{N^{N-1/2}(dN^{-1/2})(1+dN^{-1/2})^{(N-3)/2}\lambda^{-N}(1+cN^{-1/2})^{-N/2}}{4\Gamma(N-1)} \\
&\quad \cdot \Gamma(N)\left(\frac{N(2+dN^{-1/2})(2+cN^{-1/2})}{cd}\right)^{-N} \\
&\quad \cdot {}_2F_1\left(\frac{1}{2}+\frac{N}{2}, \frac{N}{2}; 1; \left(\frac{N(2+dN^{-1/2})(2+cN^{-1/2})}{cd}\right)^{-2}\right) \\
&\quad \cdot \left(\frac{cd}{4\lambda(1+cN^{-1/2})}\right)^{-N} \\
&= \frac{d(1-N^{-1})(1+dN^{-1/2})^{-3/2}}{4} \\
&\quad \cdot \left(\frac{4(1+cN^{-1/2})^{1/2}(1+dN^{-1/2})^{1/2}}{(2+dN^{-1/2})(2+cN^{-1/2})}\right)^N \\
&\quad \cdot {}_2F_1\left(\frac{1}{2}+\frac{N}{2}, \frac{N}{2}; 1; \left(\frac{N(2+dN^{-1/2})(2+cN^{-1/2})}{cd}\right)^{-2}\right) \\
&= \frac{d(1-N^{-1})(1+dN^{-1/2})^{-3/2}}{4} \\
&\quad \cdot \left(\frac{4(1+cN^{-1/2})^{1/2}(1+dN^{-1/2})^{1/2}}{(2+dN^{-1/2})(2+cN^{-1/2})}\right)^N \\
&\quad \cdot \Bigg(\left(1-\frac{cd}{N(2+dN^{-1/2})(2+cN^{-1/2})}\right. \\
&\quad \cdot \left.\left(1+\frac{cd}{N(2+dN^{-1/2})(2+cN^{-1/2})}\right)^{-1}\right)^N \\
&\quad \cdot {}_2F_1\Bigg(N, \frac{1}{2}; 1; \frac{2cd}{N(2+dN^{-1/2})(2+cN^{-1/2})} \\
&\quad \cdot \left.\left(1+\frac{cd}{N(2+dN^{-1/2})(2+cN^{-1/2})}\right)^{-1}\right) \\
&\to \frac{d}{4}\exp\left[-\frac{c^2+d^2}{8}\right] I_0\left(\frac{cd}{4}\right) \quad (N \to \infty),
\end{aligned}$$

respectively. In the above derivation, the formulas

$$\begin{aligned}
{}_2F_1(\alpha, \beta; \gamma; z) &= (1-z)^{\gamma-\alpha-\beta}{}_2F_1(\gamma-\alpha, \gamma-\beta; \gamma; z) \\
&= (1-z)^{-\alpha}{}_2F_1\left(\alpha, \gamma-\beta; \gamma; \frac{z}{z-1}\right), \\
&\lim_{\gamma\to\infty} {}_{p+1}F_q(\alpha_1, \ldots, \alpha_p, \gamma; \beta_1, \ldots, \beta_q; z) \\
&= {}_pF_q(\alpha_1, \ldots, \alpha_p; \beta_1, \ldots, \beta_q; z)
\end{aligned}$$

and

$$I_k(z) = \frac{(z/2)^k}{\Gamma(k+1)}{}_1F_1\left(k+\frac{1}{2}; 2k+1; -2z\right)$$

are applied.

Kantaro Shimomura received his B.S. from Tokyo University of Science in 2014 and received his M.S. in Information Science from Nara Institute of Science and Technology in 2016. His research interests include theoretical analyses of machine learning algorithms.

Kazushi Ikeda received his B.E., M.E., and Ph.D. in Mathematical Engineering and Information Physics from the University of Tokyo in 1989, 1991, and 1994, respectively. He joined Kanazawa University in 1994, moved to Kyoto University in 1998, and became a full professor of Nara Institute of Science and Technology in 2008. He is currently serving as a governing board member of Asia-Pacific Neural Network Society and an associate editor of APSIPA-TSIP. His research interests include machine learning, signal processing, and mathematical biology.

https://doi.org/10.1017/ATSIP.2018.22 Published online by Cambridge University Press