Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8992841, 11 pages
https://doi.org/10.1155/2021/8992841

Research Article

Degrees of Freedom in Functional Principal Components Analysis

Zixin Lin

Department of Finance and Statistic, University of Science and Technology of China, Hefei, Anhui 230000, China

Correspondence should be addressed to Zixin Lin; linzixin@mail.ustc.edu.cn

Received 16 July 2021; Revised 31 August 2021; Accepted 7 September 2021; Published 8 October 2021

Academic Editor: Sang-Bing Tsai

This paper develops the analytical form of the degrees of freedom in functional principal components analysis. Under the framework of unbiased risk estimation, we derive an unbiased estimator with a clear analytical formula for the degrees of freedom in the one-way penalized functional principal components analysis paradigm. Specifically, a new analytical formula incorporating binary smoothing parameters is also derived based on the singular value decomposition and half-smoothed method regarding the two-way penalized functional principal components analysis framework. The performance of our procedures is demonstrated by simulation studies.

1. Introduction

Functional principal component analysis (FPCA) is a key method for analyzing principal component from smoothing data, such as the data observed from temperature curves with sinusoidal nature. Additionally, FPCA has become a crucial research focus in many statistical fields. The author in [1] proposed the principle of the roughness penalty to deal with data curves using the smoothing spline method. The authors in [1, 2] proposed the roughness penalty to analyze functional data by decomposing variation in a two-way data table. Utilizing the rank-one approximation to the data matrix and penalizing only the right eigenvectors, the authors in [3] considered that a sample **X** was observed from a linear model consisting of the combination of original left singular vectors, right eigenvectors, and observed noise. Moreover, by singular value decomposition (SVD), the authors in [4] analyzed two-way functional data by penalizing left and right singular vectors of a generated covariance matrix. With functional data observed discretely, the initial model in the research of [3, 4] can be written as

$$X_{n\times m} = \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_k^T + E, \tag{1}$$

where $m \leq n$, E is the noise matrix following $N(0, \sigma^2 \mathrm{I}_n \otimes \mathrm{I}_m)$, and $\mathbf{u}_k$ and $\mathbf{v}_k$ denote two fixed and nonobservable vectors, which have the sizes of $n \times 1$ and $m \times 1$, respectively. Based on $K = 1$, the authors in [1, 3] considered the problem of finding the best rank-one approximation of X and estimating $(\hat{u}_1, \hat{v}_1)$ by

$$\arg\min_{(\mathbf{u}_1,\mathbf{v}_1)} \left(\left\| X - \mathbf{u}_1 \mathbf{v}_1^T \right\|^2 + \alpha \mathbf{v}_1^T \Omega \mathbf{v}_1 \right), \tag{2}$$

where α and Ω denote the nonnegative penalized parameter and the penalized matrix, respectively, and $\hat{u}_1 = X\mathbf{v}_1/\mathbf{v}_1^T\mathbf{v}_1$ for any fixed $\mathbf{v}_1$. The authors in [4] further proposed a two-way penalizing approach:

$$\arg\min_{(\mathbf{u}_1,\mathbf{v}_1)} \left\{ \left\| X - \mathbf{u}_1 \mathbf{v}_1^T \right\|^2 + \mathscr{P}(\mathbf{u}_1, \mathbf{v}_1) \right\}, \tag{3}$$

where $\mathrm{P}(\mathbf{u}_1, \mathbf{v}_1) = \mathbf{u}_1^T \alpha_u \Omega_u \quad \mathbf{u}_1 \cdot \|\mathbf{v}_1\|^2 + \|\mathbf{u}_1\|\|^2 \cdot \mathbf{v}_1^T \alpha_v \Omega_v \mathbf{v}_1 + \mathbf{u}_1^T \alpha_u \Omega_u \mathbf{u}_1 \cdot \mathbf{v}_1^T \alpha_v \Omega_v \mathbf{v}_1$, α_u and α_v are penalized parameters, and Ω_u and Ω_v are penalized matrices.

The degrees of freedom (DoF), a conception used to quantitatively measure the complexity of a given model, is of importance in the field of FPCA and for many learning tasks. For example, the authors in [5] derived the soft thresholding DoF using Stein's unbiased risk estimation theory [6] and demonstrated that it leads to*SureShrink*, a procedure of adaptive wavelet shrinkage. Besides, the author in [7] revealed that C_p, as an unbiased estimate that is used to measure the deviation from ture prediction, in some cases provides concretely more excellent accuracy than the coefficient of variation (CV) or related nonparametric methods when the correct DoF was used. See more details about the Bayesian information criterion (BIC) [8], the Generalized Cross Validation (GCV) [9], the Akaike Information Criterion (AIC) [10], the residual information criterion (RIC) [11], and the 'hat' matrix [12] in their respective papers. The author in [7] proposed that let δ be a specific fitting approach, $\hat{\mu} = \delta(\mathbf{y})$ denote its fit, and $\mathbf{y}$ be generated according to $\mathbf{y} \sim (\mu, \sigma^2 \mathbf{I})$, where μ is the true mean vector associated with the common variance σ^2. It was shown that the DoF of δ is

$$\mathrm{d}f(\hat{\boldsymbol{\mu}}) = \frac{\sum_{i=1}^{n} \operatorname{cov}(\hat{\mu}_i, x_i)}{\sigma^2}. \tag{4}$$

Nonetheless, the research of [13, 14] suggested that using the method of perturbation for data to quantitatively calculate an unbiased estimator (whether the estimator is approximate or not) for $\mathrm{d}f(\hat{\mu})$ can be computationally expensive when the analytical formula of $\hat{\mu}$ is unavailable. It is an intriguing research issue based on theoretical and practical aspects to study strict analytical formula of the DoF of FPCA. The main purpose of this paper is to analyze the DoF using SURE for one-way and two-way penalized FPCA. The SURE theorem [15] offers a rigorous definition for the DoF under any fitting process. Particularly, under the background of FPCA, the definition of DoF is

$$\mathrm{d}f(\hat{\mu}) = E\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial \hat{\mu}_{ij}}{\partial x_{ij}}\right). \tag{5}$$

This definition implies that an unbiased estimator for the DoF is

$$\widehat{\mathrm{d}f}(\hat{\mu}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial \hat{\mu}_{ij}}{\partial x_{ij}}. \tag{6}$$

The structure of the paper is organized as follows. We show our results with respect to the setting proposed by [3, 4] and the implications of the results in section of Main Results. Numerical experiments are presented in section of Simulation. The appendix involves all technical issues and details. For ease of presentation, except in Appendix 1 and Appendix 2, we use bold letters to denote vectors.

2. Main Results

For a fixed $\mathbf{v}_1$, plugging this $\hat{u}_1 = X\mathbf{v}_1/\mathbf{v}_1^T\mathbf{v}_1$ into (2) and transforming it as

$$\arg\min_{\mathbf{v}_1} \frac{\left(\mathbf{v}_1^T\left(\alpha\Omega - X^TX\right)\mathbf{v}_1\right)}{\mathbf{v}_1^T\mathbf{v}_1}, \tag{7}$$

it is easy to conclude that the solution to (7) is the eigenvector that corresponds to the minimum eigenvalue of the matrix $\alpha\Omega - X^TX$. Without loss of generality, we make the following assumptions: firstly, Ω is a symmetric matrix; secondly, the eigenvalues of $\alpha\Omega - X^TX$ are ordered by $\lambda_1 > \cdots > \lambda_m$, where $|\lambda_j| \neq |\lambda_i|$ for any $i \neq j$; and thirdly, $\dot{v}_k$ is the eigenvector of $\alpha\Omega - X^TX$ corresponding to the eigenvalue λ_k and is orthonormal with other eigenvectors. The same assumptions apply to more general situations, which are only more trivial in terms of notions. The rank-one approximation to X is written as

$$\hat{X} = X\dot{v}_m\dot{v}_m^T. \tag{8}$$

Theorem 1. *For* $1 \le k \le m$, *let* $\dot{v}_k^T = (v_{ik}, \ldots, v_{mk})$, $X = (\mathbf{x}_1, \ldots, \mathbf{x_m})$ *denote the sample matrix, and* $\Omega > 0$. *An unbiased estimate of the DoF for (8) is*

$$\widehat{\mathrm{d}f}(\hat{X}) = n - \sum_{j=1}^{m}\sum_{k=1}^{m-1} 2\mathbf{x}_j^T\mathbf{x}_j\left(\lambda_m^2 - \lambda_k^2\right)^{-1}\left(\lambda_k v_{jk}^2 + \lambda_m v_{jm}^2\right) - \sum_{j=1}^{m}\sum_{k=1}^{m-1} \mathbf{x}_j^T\mathbf{x}_j\left(\lambda_m - \lambda_k\right)^{-1} v_{jk}^2 v_{jm}^2. \tag{9}$$

Theorem 1 implies that if $\alpha\Omega - X^TX$ remains positive definite, the first principal component $\mathbf{u}_{n\times 1} = X\mathbf{v}$ intuitively should have n free elements, but $\widehat{\mathrm{d}f}(\hat{X}) > n$ in this case. Hence, the number of free elements does not measure the complexity of the model correctly. In other words, the number of free elements always underestimates the true DoF and should not be considered an accurate estimate for the DoF.

Let $S_u^{1/2} = (I + \alpha_u\Omega_u)^{-1/2}$ and $S_v^{1/2} = (I + \alpha_v\Omega_v)^{-1/2}$. Let $\tilde{X} = S_u^{1/2}XS_v^{1/2}$ be a data matrix with X whose rows and columns are half-smoothed. Take $\overline{u}_1 = S_u^{-1/2}\mathbf{u}_1$ and $\overline{v}_1 = S_v^{-1/2}\mathbf{v}_1$; in this case, (3) can be transformed into

$$\arg\min_{(\overline{u}_1,\overline{v}_1)}\left(\|X\|^2 - \|\tilde{X}\|^2 + \left\|\tilde{X} - \overline{u}_1\overline{v}_1\right\|^2\right). \tag{10}$$

Without loss of generality, we make following notions and assumptions: firstly, $\tilde{X}_{n\times m}$ is a half-smoothed data matrix with singular values $\sigma_1 > \cdots > \sigma_m > \sigma_{m+1} = \cdots = \sigma_n = 0$; secondly, the SVD of $\tilde{X}_{n\times m}$ is $\tilde{X}_{n\times m} = \tilde{U}\nabla\tilde{V}^T$ where $\tilde{V}_{m\times m} = (\tilde{v}_1, \ldots, \tilde{v}_m)$ and $\tilde{U}_{n\times n} = (\tilde{u}_1, \ldots, \tilde{u}_n)$, and the elements of diagonal of ∇ are σ_k with nondiagonal ones taken as zero; thirdly, let T_v be a diagonal matrix such that its diagonal elements are the same as those of $S_v^{1/2}$ with nondiagonal ones taken as zero, and T_u corresponds to $S_u^{1/2}$ under the same settings. Then, for any combination of $\hat{u}_1\hat{v}_1^T$ maximizing $\mathscr{C}(\overline{u}_1, \overline{v}_1)$, the Eckart–Young theorem suggests that next combination of $\hat{u}_2\hat{v}_2^T$ could be driven from minimizing

$$\|X\|^2 - \|\tilde{X}\|^2 + \left\|\tilde{X} - \hat{u}_1\hat{v}_1^T - \overline{u}_2\overline{v}_2\right\|^2. \tag{11}$$

Therefore, the approximation to the first-K-th principal component is

$$\hat{X} = \sum_{l=1}^{K} \tilde{u}_l\tilde{v}_l^T. \tag{12}$$

Theorem 2. *An unbiased estimate for the DoF of (12) is*

$$\begin{aligned}\widehat{\mathrm{d}}f(\widehat{X}) &= \sigma_l^{-1}\mathrm{tr}\left(S_v^{1/2}\right)\widetilde{u}_l^T T_u \widetilde{u}_l + \sigma_1^{-1}\mathrm{tr}\left(S_u^{1/2}\right)\widetilde{v}_l^T T_v \widetilde{v}_l \\ &- \sigma_l^{-1} 2\widetilde{v}_l^T T_v \widetilde{v}_l \widetilde{u}_l^T T_u \widetilde{u}_l \\ &+ \sum_{l=1}^{K}\sum_{k\neq l}\left(\sigma_l^2 - \sigma_k^2\right)^{-1}\sigma_k \widetilde{v}_l^T T_v \widetilde{v}_k \widetilde{u}_l^T T_u \widetilde{u}_k \\ &+ \sum_{l=1}^{K}\sum_{k\neq l}\left(\sigma_l^2 - \sigma_k^2\right)^{-1}\sigma_l^{-1}\sigma_k^2\left(\widetilde{v}_k^T T_v \widetilde{v}_k \widetilde{u}_l^T T_u \widetilde{u}_l \right. \\ &\left. + \widetilde{v}_l^T T_v \widetilde{v}_l \widetilde{u}_k^T T_u \widetilde{u}_k\right).\end{aligned} \tag{13}$$

Formula (8) categorically measures the shared weights of (α_u, α_v). Let the norm and orthogonality of the vector space be as follows:

$$\begin{aligned}\overline{u}_k^T T_u \overline{u}_k &= \langle \overline{u}_k, \overline{u}_k \rangle_T = 1, \\ \overline{u}_l^T T_u \overline{u}_k &= \langle \overline{u}_l, \overline{u}_k \rangle_T = 0, \\ \overline{v}_k^T T_v \overline{v}_k &= \langle \overline{v}_k, \overline{v}_k \rangle_T = 1, \\ \overline{v}_l^T T_v \overline{v}_k &= \langle \overline{\overline{v}}_l, \overline{v}_k \rangle_T = 0.\end{aligned} \tag{14}$$

Theorem 3. *Under transformation of (14), formula (13) in Theorem 2 can be rewritten as*

$$\widehat{d}f(\widehat{X}) = \sum_{l=1}^{K}\frac{tr\left(S_u^{1/2}\right) + tr\left(S_v^{1/2}\right) - 2}{\sigma_l} + \sum_{l=1}^{K}\sum_{k\neq l}\frac{2\sigma_k^2}{\sigma_l\left(\sigma_l^2 - \sigma_k^2\right)}. \tag{15}$$

As Theorem 3 shows, if $\alpha_u = 0$, under this situation, the original model (13) degrades to a one-way penalized problem but the left eigenvectors must still be considered. Combining (14) with (15), it is easy to show that under the transformation, an unbiased estimate for the DoF of a one-rank approximation with $K = 1$ is

$$\widehat{\mathrm{d}}f(\widehat{X}) = \frac{\mathrm{tr}\left(S_v^{1/2}\right) - 2}{\sigma_1} + \sum_{k=2}^{m}\frac{2\sigma_k^2}{\sigma_1\left(\sigma_1^2 - \sigma_k^2\right)}. \tag{16}$$

3. Simulation

The settings of the simulations for Theorem 1 can be clarified with the following detail; firstly, the definition for the true DoF proposed by [15] under the circumstance of (6) is

$$\mathrm{d}f(\widehat{X}) = \tau^{-2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{cov}\left(\widehat{X}_{ij}, X_{ij}\right) = \tau^{-2}\mathrm{Etr}\left(\widehat{X}^T X\right), \tag{17}$$

and we use the average of 1000 sample values to estimate the $\mathrm{Etr}(\widehat{X}^T X)$ of the true DoF with a trivial value $\tau = 1$; secondly, compared with using 1000 samples to estimate the true DoF, we instead use the average of 50 sample values to measure the unbiased estimate suggested in Theorem 1 because when facing practical applications, we may not have a sufficient sample size. Thirdly, a naive estimate for the DoF is a constant number of free parameters which is shown in Figure 1 with the other parameters.

Taking

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ 0 & 1 & -2 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -2 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{18}$$

a popular choice for the penalty standard is $\mathbf{v}^T\Omega\mathbf{v} = \sum_{j=2}^{m-1}(v_{j+1} - 2v_j + v_{j-1})^2$, whose principal philosophy is based on second differences of $\mathbf{v}$, which means that the matrix Ω is as follows:

$$\Omega = A_{m\times(m-2)}^T A_{m\times(m-2)}. \tag{19}$$

Figure 1 shows the estimated differences in various situations by adjusting the size of the matrix.

There are several pieces of information that can be obtained from Figure 1. The naive estimator of the DoF does not correctly assess the true DoF of the model; furthermore, the true complexity of the model is generally greater than the number of free elements in a given eigenvector; that is, the free elements of this eigenvector underestimate the true complexity. Regardless of whether n is much greater than m, although there is a deviation due to the number of samples utilized, unbiased estimation (7) generally has excellent performance with regard to the true DoF.

Many conditions that were argued for the first simulation of model (3) will continue to be utilized in the second simulation including the argument for the sample size and $\tau = 1$. A few special conditions should be clear. Under the coordinate transformation of (14), several of the crucial taken values stated in the previous part are

$$\begin{aligned}\Omega_v &= A_{m\times(m-2)}^T A_{m\times(m-2)}, \\ \Omega_u &= A_{n\times(n-2)}^T A_{n\times(n-2)}, \\ \mathrm{d}f(\widehat{X}) &= \tau^{-2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{cov}\left(\widehat{X}_{ij}, Y_{ij}\right) = \tau^{-2}\mathrm{Etr}\left(\widehat{X}^T Y\right),\end{aligned} \tag{20}$$

where $Y = S_u^{1/2} X S_v^{1/2}$ [15]. Figure 2 shows that given the mutual constraint for the two penalty parameters, the aforementioned unbiased estimator (15) is an excellent fit for determining the true DoF.

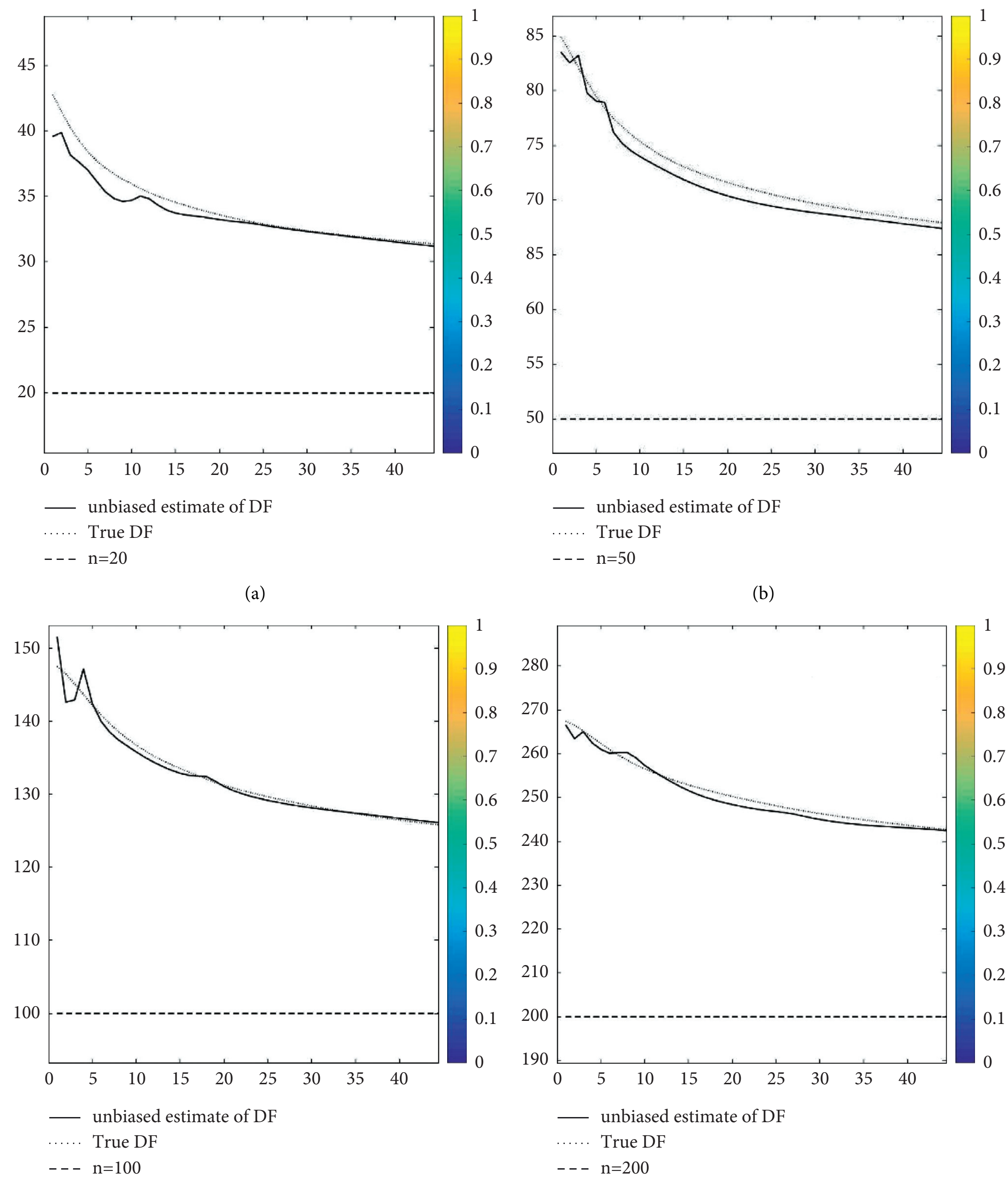
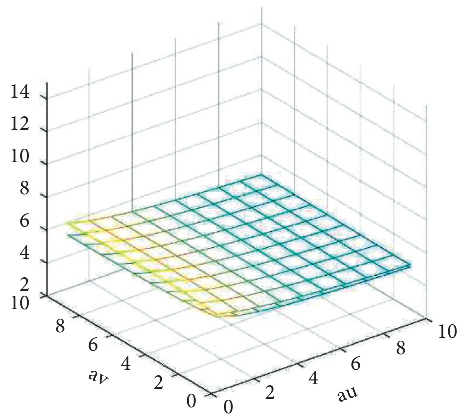


Figure 1: The sizes of the matrix are (a) 20×8, (b) 50×8, (c) 100×8, (d) 200×8. The horizontal axis represents the penalty parameter α, and the vertical axis corresponds the DoF.



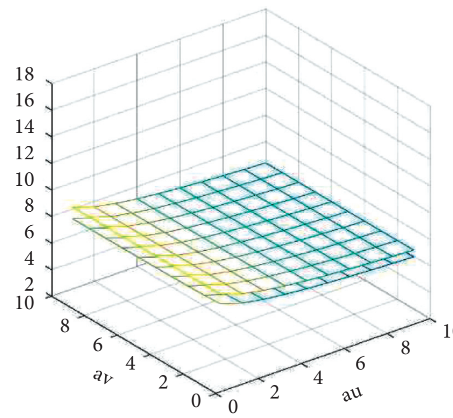


Figure 2: The sizes of the matrix are (a) 50×8 and (b) 100×8. The plane coordinate axis denotes the selection of the parameters (α_u, α_v). The z axis denotes the DoF.

Appendix

(A) Proof for Theorem 1

For ease of discussion, we no longer make use of bold letters to represent a certain matrix. The central problem of proof of Theorem 1 can be focused on finding a way to calculate the following:

$$\lim_{\varepsilon \longrightarrow 0} \frac{v_m v_m^T(\alpha\Omega - X^T X + \varepsilon A) - v_m v_m^T(\alpha\Omega - X^T X)}{\varepsilon}, \tag{21}$$

where v_m is the eigenvector of $a\Omega - X^T X$ corresponding to minimum eigenvalue λ_m and A stands for an arbitrary matrix with a coefficient $\varepsilon > 0$. As formula (12) shows, every elements of matrix $v_m v_m^T$ could be viewed as the first derivative of variable $a\Omega - X^T X$, and once we take $A = I_{ij}$, the $(i, j)^{th}$ unit in matrix $v_m v_m^T$ is

$$\frac{\partial(v_m v_m^T)_{ij}}{\partial(\alpha\Omega - X^T X)_{ij}}. \tag{22}$$

Recall that Stein's unbiased estimator (2) for degrees of freedom in our situation is

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\partial \widehat{X}_{ij}}{\partial X_{ij}} = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\partial(X v_m v_m^T)_{ij}}{\partial X_{ij}} = n + \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{p=1}^{m} X_{ip}\frac{\partial(v_m v_m^T)_{pj}}{\partial X_{ij}}. \tag{23}$$

The last term of the above formula can be rewritten as

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{p=1}^{m} X_{ip}\frac{\partial(v_m v_m^T)_{pj}}{\partial X_{ij}} = -\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{p=1}^{m} X_{ip}^2\frac{\partial(v_m v_m^T)_{pj}}{\partial(\alpha\Omega - X^T X)_{pj}} - \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2\frac{\partial(v_m v_m^T)_{jj}}{\partial(\alpha\Omega - X^T X)_{jj}}. \tag{24}$$

Therefore, the original formula is

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\partial \widehat{X}_{ij}}{\partial X_{ij}} = n - \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{p=1}^{m} X_{ip}^2\frac{\partial(v_m v_m^T)_{pj}}{\partial(\alpha\Omega - X^T X)_{pj}} - \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2\frac{\partial(v_m v_m^T)_{jj}}{\partial(\alpha\Omega - X^T X)_{jj}}. \tag{25}$$

Write

$$J = \begin{pmatrix} 0 & \alpha\Omega - X^T X \\ \alpha\Omega - X^T X & 0 \end{pmatrix}. \tag{26}$$

The authors in [16] suggest that after transformating any given matrix to the Jordan–Wielandt matrix form, especially for a symmetric matrix, its new eigenvalues will be $\pm\sigma_1, \pm\sigma_2, \dots, \pm\sigma_m$ with no zero number corresponding to original eigenvalue of $\alpha\Omega - X^T X$. Furthermore, the eigenvectors that correspond to singular values, σ_k and $-\sigma_k$, are $(u_k^T, v_k^T)^T/\sqrt{2} = \beta_k$ and $(u_k^T, -v_k^T)^T/\sqrt{2} = \gamma_k$, respectively. Without loss of generality, the following discussion shall assume that there is no overlap among $\sigma_1, \dots, \sigma_m$, and $\sigma_m > 0$ in the rest of the proof. The same process and conditions apply to the more general situation where a given matrix has equal eigenvalues but only produce more only a little difference in notation. Note that

$$J = \begin{pmatrix} 0 & a\Omega - X^T X \\ \alpha\Omega - X^T X & 0 \end{pmatrix} = \sum_{k=1}^{m}\sigma_k\beta_k\beta_k^T + \sum_{k=1}^{m}(-\sigma_k)\gamma_k\gamma_k^T, \tag{27}$$

$$\frac{1}{\theta I - J} = \sum_{k=1}^{m}\frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m}\frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T.$$

By the Cauchy residue formula, it is easy to find a closed curve C_m in the complex space to form an open set that contains only a single eigenvalue σ_m of J as the only singularity. From mathematical perspective, there exists a $c > 0$ which allows the circle $O(\sigma_m, c)$ to belong to open set of C_m. Therefore, we shall have

$$\frac{1}{2\pi i}\oint_{C_m}\frac{1}{\theta I - J}\mathrm{d}\theta = \frac{1}{2\pi i}\oint_{C_m}\sum_{k=1}^{m}\frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m}\frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T\mathrm{d}\theta. \quad (28)$$

It is easy to have

$$\beta_m\beta_m^T = \frac{1}{2\pi i}\oint_{C_m}\frac{1}{\theta I - J}\mathrm{d}\theta = \frac{1}{2\pi i}\oint_{C_m}\sum_{k=1}^{m}\frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m}\frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T\mathrm{d}\theta. \quad (29)$$

Similarly, we write

$$J + \varepsilon B = \begin{pmatrix} 0 & \alpha\Omega - X^T X \\ \alpha\Omega - X^T X & 0 \end{pmatrix} + \varepsilon\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}. \quad (30)$$

Then,

$$J + \varepsilon B = \sum_{k=1}^{m}\widehat{\sigma}_k\widehat{\beta}_k\widehat{\beta}_k^T - \sum_{k=1}^{m}\widehat{\sigma}_k\widehat{\gamma}_k\widehat{\gamma}_k^T. \quad (31)$$

We take

$$M = u_m v_m^T, \quad \widehat{M} = \widehat{u}_m\widehat{v}_m^{\mathrm{T}}, \quad (32)$$

where $\widehat{u}_m$ and $\widehat{v}_m$ correspond to $\alpha\Omega - X^T X + \varepsilon A$ with similar definitions, and we make for J. From previous discussion, write

$$\varepsilon^{-1}\begin{pmatrix} 0 & \widehat{M} \\ \widehat{M}^T & 0 \end{pmatrix} - \varepsilon^{-1}\begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix} = \varepsilon^{-1}\left(\widehat{\beta}_m\widehat{\beta}_m^T - \beta_m\beta_m^T\right) + \varepsilon^{-1}\left(\widehat{\gamma}_m\widehat{\gamma}_m^T - \gamma_m\gamma_m^T\right), \quad (33)$$

where the notations $\widehat{\sigma}_k$, $\widehat{\gamma}_k$, and $\widehat{\beta}_k$ are all similar definitions made for J. When ε is small enough, it is natural that $\widehat{\sigma}_k$ is close to $\widehat{\sigma}_k$; moreover, C_m defined above will contain not only $\widehat{\sigma}_m$, an eigenvalues of $J + \varepsilon B$, but also σ_m as only two singularities [17]. Write

$$\widehat{\beta}_m\widehat{\beta}_m^T = \frac{1}{2\pi i}\oint_{C_m}\frac{1}{\theta I - J - \varepsilon B}\mathrm{d}\theta = \frac{1}{2\pi i}\oint_{C_m}\sum_{k=1}^{m}\frac{1}{\theta - \widehat{\sigma}_k}\widehat{\beta}_k\widehat{\beta}_k^T + \sum_{k=1}^{m}\frac{1}{\theta + \widehat{\sigma}_k}\widehat{\gamma}_k\widehat{\gamma}_k^T\mathrm{d}\theta. \quad (34)$$

Therefore,

$$\begin{aligned}
\widehat{\beta}_m\widehat{\beta}_m^T - \beta_m\beta_m^T &= \frac{1}{2\pi i}\oint_{C_m}\frac{1}{\theta I - J}(\varepsilon B)\frac{1}{\theta I - J}d\theta + O\left(\varepsilon^2\right), \\
\beta_k\beta_k^T B\beta_m\beta &= \frac{\left(u_k^T, v_k^T\right)^T\left(u_k^T, v_k^T\right)B\left(u_m^T, v_m^T\right)^T\left(u_m^T, v_m^T\right)}{4} \\
&= \frac{\left(u_k^T, v_k^T\right)^T\left(v_k^T A^T, u_k^T A\right)\left(u_m^T, v_m^T\right)^T\left(u_m^T, v_m^T\right)}{4} \\
&= \frac{u_k^T A v_m + v_k^T A^T u_m}{4}\begin{pmatrix} \sim & \sim \\ \sim & v_k v_m^T \end{pmatrix}, \\
\sum_{k=1}^{m-1}\frac{1}{\sigma_m - \sigma_k}\beta_k\beta_k^T B\beta_m\beta_m^T &= \sum_{k=1}^{m-1}\frac{1}{\sigma_m - \sigma_k}\frac{u_k^T A v_m + v_k^T A^T u_m}{4}\begin{pmatrix} \sim & \sim \\ \sim & v_k v_m^T \end{pmatrix} \\
&= \left(\sim \sum_{k=1}^{m-1}\frac{u_k^T A v_m + v_k^T A^T u_m}{4\left(\sigma_m - \sigma_k\right)}v_k v_m^T\right).
\end{aligned} \quad (35)$$

Similarly,

$$\gamma_k\gamma_k^T B\beta_m\beta_m^T = \frac{\left(u_k^T, -v_k^T\right)^T \left(u_k^T - v_k^T\right) B \left(u_m^T, v_m^T\right)^T \left(u_m^T, v_m^T\right)}{4}$$
$$= \frac{v_k^T A^T u_m - u_k^T A v_m}{4} \begin{pmatrix} \sim & \sim \\ \sim & v_k v_m^T \end{pmatrix}, \tag{36}$$
$$\sum_{k=1}^{m} \frac{1}{\sigma_m + \sigma_k} \gamma_k\gamma_k^T B\beta_m\beta_m^T = \sum_{k=1}^{m} \frac{1}{\sigma_m + \sigma_k} \frac{v_k^T A^T u_m - u_k^T A v_m}{4} \begin{pmatrix} \sim & \sim \\ \sim & v_k v_m^T \end{pmatrix}.$$

Furthermore,

$$\beta_m\beta_m^T B\gamma_k\gamma_k^T = \left(\gamma_k\gamma_k^T B\beta_m\beta_m^T\right)^T = \frac{v_k^T A^T u_m - u_k^T A v_m}{4} \begin{pmatrix} \sim & \sim \\ \sim & v_m v_k^T \end{pmatrix},$$
$$\beta_m\beta_m^T B\beta_k\beta_k^T = \left(\beta_k\beta_k^T B\beta_m\beta_m^T\right)^T = \frac{u_k^T A v_m + v_k^T A^T u_m}{4} \begin{pmatrix} \sim & \sim \\ \sim & v_m v_k^T \end{pmatrix}. \tag{37}$$

Note $\widehat{\beta}_m\widehat{\beta}_m^T - \beta_m\beta_m^T$ can be decomposed as

$$\frac{\varepsilon}{2\pi i}\oint_{C_m} \left(\sum_{k=1}^{m-1} \frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m-1} \frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T\right) B \left(\frac{1}{\theta - \sigma_m}\beta_m\beta_m^T\right) d\theta,$$
$$\frac{\varepsilon}{2\pi i}\oint_{C_m} \left(\frac{1}{\theta - \sigma_m}\beta_m\beta_m^T\right) B \left(\sum_{k=1}^{m-1} \frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m} \frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T\right) d\theta, \tag{38}$$
$$\frac{\varepsilon}{2\pi i}\oint_{C_m} \left(\frac{1}{\theta - \sigma_m}\beta_m\beta_m^T\right) B \left(\frac{1}{\theta - \sigma_m}\beta_m\beta_m^T\right) d\theta = 0.$$

Thus, we will have

$$\lim_{\varepsilon \longrightarrow 0} \frac{\widehat{\beta}_m\widehat{\beta}_m^T - \beta_m\beta_m^T}{\varepsilon} = \left(\sum_{k=1}^{m-1} \frac{1}{\sigma_m - \sigma_k}\beta_k\beta_k^T A\beta_m\beta_m^T + \sum_{k=1}^{m} \frac{1}{\sigma_m + \sigma_k}\gamma_k\gamma_k^T A\beta_m\beta_m^T\right)$$
$$+ \left(\sum_{k=1}^{m-1} \frac{1}{\sigma_m - \sigma_k}\beta_m\beta_m^T A\beta_k\beta_k^T + \sum_{k=1}^{m} \frac{1}{\sigma_m + \sigma_k}\beta_m\beta_m^T A\gamma_k\gamma_k^T\right) \tag{39}$$
$$+ O(\varepsilon).$$

Then, the lower right block of the matrix can be transformed as

$$\begin{aligned}(\text{LRB}) = &\sum_{k=1}^{m-1} \frac{u_k^T A v_m + v_k^T A^T u_m}{4(\sigma_m - \sigma_k)} v_k v_m^T + \sum_{k=1}^{m} \frac{v_k^T A^T u_m - u_k^T A v_m}{4(\sigma_m + \sigma_k)} v_k v_m^T \\ &+ \sum_{k=1}^{m-1} \frac{u_k^T A v_m + v_k^T A^T u_m}{4(\sigma_m - \sigma_k)} v_m v_k^T + \sum_{k=1}^{m} \frac{v_k^T A^T u_m - u_k^T A v_m}{4(\sigma_m + \sigma_k)} v_m v_k^T.\end{aligned} \quad (40)$$

Note that the original matrix is a symmetric matrix; the argument we make for and result we get regarding $\widehat{\beta}_m \widehat{\beta}_m^T - \beta_m \beta_m^T$ are same for $\widehat{\gamma}_m \widehat{\gamma}_m^T - \gamma_m \gamma_m^T$ since their lower right block is equal; moreover, we take $A = I_{ij}$ and take our focus just on $(i, j)^{th}$ elements; we can conclude that

$$\begin{aligned}\sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{p=1}^{m} X_{ip}^2 \frac{\partial\left(v_m v_m^T\right)_{pi}}{\partial\left(\alpha\Omega - X^T X\right)_{pi}} &= \sum_{k=1}^{m-1} \sum_{p=1}^{m} 2x_p^T x_p \left\{ \frac{\sigma_k \operatorname{sgn}(y_k) v_{pk}^2 + \sigma_m \operatorname{sgn}(y_m) v_{pm}^2}{\left(\sigma_m^2 - \sigma_k^2\right)} \right\} \\ &= \sum_{k=1}^{m-1} \sum_{p=1}^{m} 2x_p^T x_p \left\{ \frac{\lambda_k v_{pk}^2 + \lambda_m v_{pm}^2}{\left(\lambda_m^2 - \lambda_k^2\right)} \right\}, \\ \sum_{j=1}^{m} \sum_{j=1}^{n} \sum_{i=1}^{n} X_{ij}^2 \frac{\partial\left(v_m v_m^T\right)_{ij}}{\partial\left(\alpha\Omega - X^T X\right)_{ij}} &= \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij}^2 \sum_{k=1}^{m-1} \frac{\sigma_m v_{jk}^2 u_{jm} v_{jm} + \sigma_k v_{jm}^2 u_{jk} v_{jk}}{\left(\sigma_m^2 - \sigma_k^2\right)} \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij}^2 \sum_{k=1}^{m-1} \frac{\sigma_m v_{jk}^2 u_{jm} v_{jm} + \sigma_k v_{jm}^2 u_{jk} v_{jk}}{\left(\sigma_m^2 - \sigma_k^2\right)} \\ &= \sum_{j=1}^{m} \sum_{k=1}^{m-1} x_j^T x_j \frac{\lambda_m v_{jk}^2 v_{jm}^2 + \lambda_k v_{jk}^2 v_{jm}^2}{\left(\lambda_m^2 - \lambda_k^2\right)}.\end{aligned} \quad (41)$$

Finally, we apply our conclusion to equation (21).

$$\begin{aligned}\sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\partial \widehat{X}_{ij}}{\partial X_{ij}} = n - \sum_{j=1}^{m} \sum_{k=1}^{m-1} 2x_j^T x_j \frac{\lambda_k v_{jk}^2 + \lambda_m v_{jm}^2}{\left(\lambda_m^2 - \lambda_k^2\right)} \\ - \sum_{j=1}^{m} \sum_{k=1}^{m-1} x_j^T x_j \frac{v_{jk}^2 v_{jm}^2}{\left(\lambda_m - \lambda_k\right)}.\end{aligned} \quad (42)$$

(B) Proof for Theorem 2

The method and idea we use to proof Theorem 2 are analogous to that in Appendix 1. First, according to our initial definition for degrees of freedom, note that

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \partial \frac{\left(\sum_{l=1}^{K} u_l v_l^T\right)_{ij}}{\partial X_{ij}} = \sum_{j=1}^{m} \left(S_v^{1/2}\right)_{jj} \sum_{i=1}^{n} \left(S_u^{1/2}\right)_{ii} \partial \frac{\left(\sum_{l=1}^{K} u_l v_l^T\right)_{ij}}{\partial \widehat{X}_{ij}}. \quad (43)$$

We continue to use some notations made in Section 2, write

$$\begin{aligned}M &= \widehat{X} = \sum_{l=1}^{K} \widehat{u}_l \widehat{v}_l^T, \\ \widehat{M} &= \widehat{X} = \sum_{l=1}^{K} \widehat{u}_l \widehat{v}_l^T + \varepsilon A,\end{aligned} \quad (44)$$

where A is an arbitrary matrix. The transformation of the Jordan–Wielandt matrix for M and $\widehat{M}$ is as same as what we argue in the last appendix; write

$$\varepsilon^{-1}\begin{pmatrix} 0 & \widehat{M} \\ \widehat{M}^T & 0 \end{pmatrix} - \varepsilon^{-1}\begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix} = \varepsilon^{-1}\sum_{l=1}^{K}\left(\widehat{\beta}_l\widehat{\beta}_l^T - \beta_l\beta_l^T\right) + \varepsilon^{-1}\sum_{l=1}^{K}\left(\widehat{\gamma}_l\widehat{\gamma}_l^T - \gamma_l\gamma_l^T\right). \tag{45}$$

Similarly,

$$(\theta I - J)^{-1} = \sum_{k=1}^{m}\frac{1}{\theta - \sigma_k}\beta_k\beta_k^T + \sum_{k=1}^{m}\frac{1}{\theta + \sigma_k}\gamma_k\gamma_k^T + \frac{1}{\theta}\left(I - \sum_{k=1}^{m}\left(\beta_k\beta_k^T + \gamma_k\gamma_k^T\right)\right). \tag{46}$$

Take

$$\begin{aligned} P_1 &= \sum_{k=2}^{m}(\theta + \sigma_k)^{-1}\beta_k\beta_k^T + \sum_{k=1}^{m}(\theta - \sigma_k)^{-1}\gamma_k\gamma_k^T + \theta^{-1}\left(I - \sum_{k=1}^{m}\left(\beta_k\beta_k^T + \gamma_k\gamma_k^T\right)\right), \\ P_2 &= \sum_{k=1}^{m}(\theta + \sigma_k)^{-1}\beta_k\beta_k^T + \sum_{k=2}^{m}(\theta - \sigma_k)^{-1}\gamma_k\gamma_k^T + \theta^{-1}\left(I - \sum_{k=1}^{m}\left(\beta_k\beta_k^T + \gamma_k\gamma_k^T\right)\right). \end{aligned} \tag{47}$$

Naturally,

$$\begin{aligned} &\widehat{\beta}_1\widehat{\beta}_1^T - \beta_1\beta_1^T = (2\pi i)^{-1}\oint_{C_1}(\theta I - J)^{-1}(\varepsilon B)(\theta I - J)^{-1}\mathrm{d}\theta + O\left(\varepsilon^2\right) \\ &(2\pi i)^{-1}\oint_{C_1}(\theta I - J)^{-1}(\varepsilon B)(\theta I - J)^{-1}\mathrm{d}\theta \\ &= (2\pi i)^{-1}\varepsilon\oint P_1 B\left((\theta - \sigma_1)^{-1}\beta_1\beta_1^T\right)\mathrm{d}\theta + (2\pi i)^{-1}\varepsilon\oint\left((\theta - \sigma_1)^{-1}\beta_1\beta_1^T\right)BP_1\mathrm{d}\theta. \end{aligned} \tag{48}$$

Furthermore,

$$\begin{aligned} &\widehat{\gamma}_1\widehat{\gamma}_1^T - \gamma_1\gamma_1^T = (2\pi i)^{-1}\oint_{C_1}(\theta I + J)^{-1}(\varepsilon B)(\theta I + J)^{-1}\mathrm{d}\theta + O\left(\varepsilon^2\right), \\ &(2\pi i)^{-1}\oint_{C_1}(\theta I + J)^{-1}(\varepsilon B)(\theta I + J)^{-1}\mathrm{d}\theta \\ &= (2\pi i)^{-1}\varepsilon\oint P_2 B\left((\theta + \sigma_1)^{-1}\gamma_1\gamma_1^T\right)\mathrm{d}\theta + (2\pi i)^{-1}\varepsilon\oint\left((\theta + \sigma_1)^{-1}\gamma_1\gamma_1^T\right)BP_2\mathrm{d}\theta. \end{aligned} \tag{49}$$

Therefore,

$$
\begin{aligned}
&(2\pi i)^{-1}\oint_{C_1}(\theta I+J)^{-1}(\varepsilon B)(\theta I+J)^{-1}\mathrm{d}\theta\\
&=\sigma_1^{-1}B\left(\gamma_1\gamma_1^T+\beta_1\beta_1^T\right)-\sigma_1^{-1}\left(\gamma_1\gamma_1^TB\gamma_1\gamma_1^T+\beta_1\beta_1^TB\beta_1\beta_1^T\right)\\
&\quad+\sum_{k=2}^{m}(\sigma_1-\sigma_k)^{-1}\sigma_1^{-1}\sigma_k\left(\gamma_1\gamma_1^TB\gamma_k\gamma_k^T+\beta_1\beta_1^TB\beta_k\beta_k^T\right)\\
&\quad-\sum_{k=1}^{m}(\sigma_1+\sigma_k)^{-1}\sigma_1^{-1}\sigma_k\left(\gamma_1\gamma_1^TB\beta_k\beta_k^T+\beta_1\beta_1^TB\gamma_k\gamma_k^T\right)\\
&=4^{-1}\begin{pmatrix}\sim & \sigma_1^{-1}u_1u_1^TA-\sigma_1^{-1}u_1u_1^TAv_1v_1^T\\ \sim & \sim\end{pmatrix}\\
&\quad+4^{-1}\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_k\begin{pmatrix}\sim & u_1v_1^TAu_kv_k^T\\ \sim & \sim\end{pmatrix}\\
&\quad+4^{-1}\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_1^{-1}\sigma_k^2\begin{pmatrix}\sim & u_1u_1^TAv_kv_k^T\\ \sim & \sim\end{pmatrix}.
\end{aligned}
\tag{50}
$$

Same argument is applicable for $\widehat{\gamma}_1\widehat{\gamma}_1^T-\gamma_1\gamma_1^T$, and we omit this trivial process and write

$$
\begin{aligned}
&\varepsilon^{-1}\left(\widehat{\beta}_1\widehat{\beta}_1^T-\beta_1\beta_1^T\right)+\varepsilon^{-1}\left(\widehat{\gamma}_1\widehat{\gamma}_1^T-\gamma_1\gamma_1^T\right)\\
&=4^{-1}\begin{pmatrix}\sim & \sigma_1^{-1}u_1u_1^TA+\sigma_1^{-1}Av_1v_1^T-2\sigma_1^{-1}u_1u_1^TAv_1v_1^T\\ \sim & \sim\end{pmatrix}\\
&\quad+4^{-1}\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_k\begin{pmatrix}\sim & u_1v_1^TAu_kv_k^T+u_kv_k^TAu_1v_1^T\\ \sim & \sim\end{pmatrix}\\
&\quad+4^{-1}\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_1^{-1}\sigma_k^2\begin{pmatrix}\sim & u_1u_1^TAv_kv_k^T+u_ku_k^TAv_1v_1^T\\ \sim & \sim\end{pmatrix}\\
&\quad+O(\varepsilon).
\end{aligned}
\tag{51}
$$

We take $A=I_{ij}$ and take our focus just on $(i,j)^{th}$ elements, then (51) can be written as

$$
\begin{aligned}
&\frac{\partial\left(u_1v_1^T\right)_{ij}}{\partial\widehat{X}_{ij}}\\
&=\sigma_1^{-1}\left(u_{i1}^2+v_{j1}^2-2u_{i1}^2v_{j1}^2\right)+\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_ku_{i1}u_{ik}v_{j1}v_{jk}\\
&\quad+\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_1^{-1}\sigma_k^2\left(u_{i1}^2v_{jk}^2+u_{ik}^2v_{j1}^2\right).
\end{aligned}
\tag{52}
$$

Combining the above formula with the original one, we complete the proof:

$$
\begin{aligned}
\widehat{\mathrm{d}}f\left(u_1v_1^T\right)&=\sigma_1^{-1}\mathrm{tr}\left(S_v^{1/2}\right)\widehat{u}_1^TT_u\widehat{u}_1+\sigma_1^{-1}\mathrm{tr}\left(S_u^{1/2}\right)\widehat{v}_1^TT_v\widehat{v}_1-\sigma_1^{-1}2\widehat{v}_1^TT_v\widehat{v}_1\widehat{u}_1^TT_u\widehat{u}_1\\
&\quad+\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_k\widehat{v}_1^TT_v\widehat{v}_k\widehat{u}_1^TT_u\widehat{u}_k\\
&\quad+\sum_{k=2}^{m}\left(\sigma_1^2-\sigma_k^2\right)^{-1}\sigma_1^{-1}\sigma_k^2\left(\widehat{v}_k^TT_v\widehat{v}_k\widehat{u}_1^TT_u\widehat{u}_1+\widehat{v}_1^TT_v\widehat{v}_1\widehat{u}_k^TT_u\widehat{u}_k\right).
\end{aligned}
\tag{53}
$$

Note that same computational process for $u_1v_1^T$ can be applied for other sequent combination of right and left singular vectors, incorporating with first-K-th principal component to complete the proof:

$$\widehat{\mathrm{d}}f(\widehat{X}) = \sum_{l=1}^{K} \sigma_l^{-1}\operatorname{tr}\left(S_v^{1/2}\right)\widehat{u}_l^T T_u \widehat{u}_l + \sigma_1^{-1}\operatorname{tr}\left(S_u^{1/2}\right)\widehat{v}_l^T T_v \widehat{v}_l - \sigma_l^{-1} 2\widehat{v}_l^T T_v \widehat{v}_l \widehat{u}_l^T T_u \widehat{u}_l$$
$$+ \sum_{l=1}^{K}\sum_{k\neq l}\left(\sigma_l^2 - \sigma_k^2\right)^{-1}\sigma_k \widehat{v}_l^T T_v \widehat{v}_k \widehat{u}_l^T T_u \widehat{u}_k \tag{54}$$
$$+ \sum_{l=1}^{K}\sum_{k\neq l}\left(\sigma_l^2 - \sigma_k^2\right)^{-1}\sigma_l^{-1}\sigma_k^2\left(\widehat{v}_k^T T_v \widehat{v}_k \widehat{u}_l^T T_u \widehat{u}_l + \widehat{v}_l^T T_v \widehat{v}_l \widehat{u}_k^T T_u \widehat{u}_k\right).$$

Data Availability

No data were used to support this study.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Project: No. 12071452).

References

[1] B. W. Silverman, "Smoothed functional principal components analysis by choice of norm," *Annals of Statistics*, vol. 24, pp. 1–24, 1996.

[2] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer-Verlag, New York, NY, USA, 2nd ed edition, 2005.

[3] J. Huang, H. Shen, and A. Buja, "Functional principal components analysis via penalized rank one approximation," *Electronic Journal of Statistics*, vol. 2, pp. 678–695, 2008.

[4] J. Z. Huang, H. Shen, and A. Buja, "The analysis of two-way functional data using two-way regularized singular value decompositions," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1609–1620, 2009.

[5] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 120–1224, 1995.

[6] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[7] B. Efron, "The estimation of prediction error," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 619–632, 2004.

[8] G. Schwarz, "Cumulative index - volumes I-xvi," *Progress in Optics*, vol. 6, pp. 461–464, 1978.

[9] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, vol. 31, pp. 317–403, 1979.

[10] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 267–281, Springer, New York, NY, USA, 1973.

[11] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *Annals of Statistics*, vol. 22, pp. 1947–1975, 1994.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, Berlin, Germany, 2009.

[13] X. Shen and J. Ye, "Adaptive model selection," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 210–221, 2002.

[14] J. Ye, "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 120–131, 1998.

[15] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, pp. 1135–1151, 1981.

[16] G. H. Golub and C. F. Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, MA, USA, 3rd ed. edition, 1996.

[17] I. Stewart and D. Tall, *Complex Analysis*, Cambridge University Press, Cambridge, UK, 2018.