Asymptotic Generalization Bound of Fisher's Linear Discriminant Analysis

Wei Bian , *Member, IEEE,* and Dacheng Tao, *Senior Member, IEEE*

***Abstract*—Fisher's linear discriminant analysis (FLDA) is an important dimension reduction method in statistical pattern recognition. It has been shown that FLDA is asymptotically Bayes optimal under the homoscedastic Gaussian assumption. However, this classical result has the following two major limitations: 1) it holds only for a fixed dimensionality D, and thus does not apply when D and the training sample size N are proportionally large; 2) it does not provide a quantitative description on how the generalization ability of FLDA is affected by D and N. In this paper, we present an asymptotic generalization analysis of FLDA based on random matrix theory, in a setting where both D and N increase and $D/N \longrightarrow \gamma \in [0, 1)$. The obtained lower bound of the generalization discrimination power overcomes both limitations of the classical result, i.e., it is applicable when D and N are proportionally large and provides a quantitative description of the generalization ability of FLDA in terms of the ratio $\gamma = D/N$ and the population discrimination power. Besides, the discrimination power bound also leads to an upper bound on the generalization error of binary-classification with FLDA.**

***Index Terms*—Fisher's linear discriminant analysis, asymptotic generalization analysis, random matrix theory**

I. Introduction

Fisher's linear discriminant analysis (FLDA), first developed by Fisher [1] for binary classification and then extended by Rao [2] to the multiclass scenario, is one of the most representative dimension reduction techniques in statistical pattern recognition. It selects a low dimensional subspace by simultaneously maximizing the between-class scatter and minimizing the within-class scatter. By projecting samples into the low dimensional subspace with the maximum discrimination power, FLDA helps improve the accuracy and the robustness of a decision system [3] [4] [5] [6]. During the past decades, FLDA has been applied to a wide range of areas, from speech/music classification [7] [8], face recognition [9] [10] [11] to financial data analysis [12] [13].

An important property of FLDA is its asymptotic Bayes optimality under the homoscedastic Gaussian assumption [14] [15] [16] , which is a corollary of classical results from multivariate statistics [17]. Actually, as training sample size N goes to infinity, both the within-class scatter matrix $\widehat{\mathbf{\Sigma}}$ (sample covariance) and the between-class scatter matrix $\widehat{\mathbf{S}}$ converge to their population counterparts $\mathbf{\Sigma}$ and $\mathbf{S}$. Therefore, the empirically optimal projection matrix $\widehat{\mathbf{W}}$ of FLDA, obtained by generalized eigendecomposition over $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{S}}$, also converges to its population counterpart $\mathbf{W}$. Thanks to the asymptotic Bayes optimality, we can expect an acceptable performance of FLDA as long as N is sufficiently large. However, this classical result, i.e., the asymptotic Bayes optimality, suffers from two major limitations:

1) It is obtained by fixing the dimensionality D and letting only N increase to infinity. But in practice, D and N can be proportionally large, which makes the classical result inapplicable.
2) It does not provide quantitative description on the performance of FLDA, especially, how the generalization ability of FLDA is affected by D and N.

A. The Contribution of this Paper

To address aforementioned limitations of the classical result, in this paper, we present an asymptotic generalization analysis of FLDA. Our analysis is superior from two aspects. First, we modify the setting of analysis by allowing both D and N to increase and assuming the dimensionality to training sample size ratio $\gamma = D/N$ has a limit in $[0, 1)$. This makes our result applicable in the case where D and N are proportionally large. Second, we quantitatively examine the generalization ability of FLDA. Denoting by $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ the generalization discrimination power of FLDA, we intend to bound it from the lower side in terms of D and N, with respect to the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W})$. Taking a binary-class problem, for example: suppose $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}) = \boldsymbol{\lambda}$ and $\gamma = D/N$, then our asymptotic generalization bound shows that $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ is almost surely larger than

$$\cos^2(\arccos(\sqrt{\boldsymbol{\lambda}/(\boldsymbol{\lambda}+\gamma)}) + \arccos(\sqrt{1-\gamma}))\boldsymbol{\lambda},$$

under mild conditions. Further, as a corollary of the discrimination power bound, we also obtain an asymptotic generalization error bound for binary classification with FLDA.

Based on the obtained asymptotic generalization bound, we can get better insight of FLDA. It is commonly known that the performance of covariance estimation has a severe influence to the generalization ability of FLDA. By assuming a sufficient population discrimination power so as to eliminate the influence from between-class matrix estimation, we show that the mere influence from covariance estimation is proportional to the ratio $\gamma = D/N < 1$, i.e., due to the imperfection of covariance estimation, $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ is about $1 - \gamma$ times of $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W})$. It is worth noticing that such result holds independent of the covariance $\mathbf{\Sigma}$. Besides, the bound shows that the performance of FLDA is substantially determined by the ratio $\gamma = D/N$, given a fixed population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W})$. Therefore, N only needs to scale linearly

Bian and Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia (email: wei.bian@uts.edu.au, dacheng.tao@uts.edu.au).

with respect to D for an acceptable generalization ability of FLDA, although a quadratic number of parameters are to be estimated in the sample covariance.

B. Tools

The technical tools used in our asymptotic generalization analysis are from random matrix theory (RMT) [18] [19] [20] [21], the main goal of which is to provide understanding of the statistics of eigenvalues of matrices with entries drawn randomly from various probability distributions. RMT was originally motivated by applications in nuclear physics in 1950's, and then it was intensively studied in mathematics and statistics. It also found successful applications in engineering fields, e.g., wireless communications [22], recently. In this paper, we make use of two important results from RMT. The first one is the Marčenko-Pastur Law [20], which states that the empirical spectral distribution of a Wishart random matrix converges almost surely to a deterministic distribution $F_\gamma(\lambda)$ as $\lim \gamma = D/N \in [0,1)$. The second one is the almost sure convergence of the extreme singular values of a large Gaussian random matrix [21]. We formulate these two results in following propositions.

Proposition 1: Given $\mathbf{G} \in \mathbb{R}^{D\times N}$, whose entries are independently sampled from standard Gaussian distribution $\mathcal{N}(0,1)$, then as both D and $N \longrightarrow \infty$ and $D/N \longrightarrow \gamma \in [0,1)$, the empirical distribution of the eigenvalues of $\frac{1}{N}\mathbf{G}\mathbf{G}^T$, i.e.,

$$F_N(\lambda) = \frac{1}{D}\sum_{i=1}^{D} 1\{\lambda_i\big(\frac{1}{N}\mathbf{G}\mathbf{G}^T\big) \le \lambda\},\ \lambda \ge 0, \tag{1}$$

where $1\{\cdot\}$ is the indicator function, converges almost surely to a deterministic limit distribution $F_\gamma(\lambda)$ with density

$$dF_\gamma(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda} d\lambda, \tag{2}$$

where

$$\lambda_+ = (1+\sqrt{\gamma})^2 \text{ and } \lambda_- = (1-\sqrt{\gamma})^2. \tag{3}$$

Proposition 2: Letting $\mathbf{G} \in \mathbb{R}^{D\times m}$ with i.i.d. entries sampled from $\mathcal{N}(0,1)$, then as $m/D \longrightarrow \gamma \in [0,1)$,

$$\frac{1}{\sqrt{D}}\sigma_{max}(\mathbf{G}) \xrightarrow{a.s.} 1+\sqrt{\gamma}, \tag{4}$$

and

$$\frac{1}{\sqrt{D}}\sigma_{min}(\mathbf{G}) \xrightarrow{a.s.} 1-\sqrt{\gamma}. \tag{5}$$

C. Notations

Throughout this paper, we will use the following notations. Bold lower case letter $\mathbf{a}$ denotes a vector. Bold upper case letter $\mathbf{A}$ denotes a matrix. $\mathbb{R}^D$ denotes a D-dimensional vector space. $\mathbb{R}^{D_1\times D_2}$ denotes the set of all D_1 by D_2 matrices. $\mathbf{A}_{ii}$ or $\{\mathbf{A}\}_{ii}$ denotes the i-th diagonal entry of a symmetric matrix $\mathbf{A}$. $\mathbf{A}_i$ denotes the i-th column of $\mathbf{A}$. $\mathbf{A}_{1:c}$ denotes the matrix composed by the first c columns of $\mathbf{A}$. $\mathbb{S}^{D-1}$ denotes the D-dimensional unit sphere located on the original point. $\mathbb{S}^{D\times D}_{++}$ denotes the set of all D by D positive definite matrices. $\|\mathbf{a}\|$ denotes the ℓ_2 norm of $\mathbf{a}$. $\sigma_{max}(\mathbf{A})$ and $\sigma_{min}(\mathbf{A})$ are the extreme singular values of $\mathbf{A}$. $\|\mathbf{A}\| = \sigma_{max}(\mathbf{A})$ denotes the operator norm of $\mathbf{A}$. $\lambda_i(\mathbf{A})$ denotes the i-th eigenvalue of $\mathbf{A}$, sorted in a descent order. $\Lambda(\mathbf{A})$ denotes the diagonal matrix composed of the eigenvalues of $\mathbf{A}$, with the eigenvalues sorted in a descent order. $\mathcal{R}(\mathbf{A})$ denotes an orthogonal basis of the range or the column space of $\mathbf{A}$. $[\mathbf{e}_1, ..., \mathbf{e}_D]$ is the canonical basis of $\mathbf{R}^D$. $1\{\cdot\}$ is the indicator function, i.e., $1\{x_0 \le x\} = 1$ if $x \ge x_0$ and $1\{x_0 \le x\} = 0$ if $x < x_0$.

II. Main Result

A. Bounding Generalization Discrimination Power

Suppose we have $c+1$ classes, represented by homoscedastic Gaussian distributions in a high-dimensional space $\mathbb{R}^D$, $\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2, ..., c+1$, with class means $\boldsymbol{\mu}_i \in \mathbb{R}^D$ and the common covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}^{D\times D}_{++}$. Assuming the classes have equal prior probability $\frac{1}{c+1}$[1], the following matrix $\mathbf{S}$, which is referred to as the between-class scatter matrix, gives a measure of class separation,

$$\mathbf{S} = \frac{1}{c+1}\sum_{i=1}^{c+1}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \text{ with } \boldsymbol{\mu} = \frac{1}{c+1}\sum_{i=1}^{c+1}\boldsymbol{\mu}_i. \tag{6}$$

Suppose the eigendecomposition of $\boldsymbol{\Sigma}^{-1}\mathbf{S}$ has (at most) c nonzero eigenvalues $\boldsymbol{\lambda}_i$, $i = 1, 2, ..., c$, and associated eigenvectors $\mathbf{W} = [\mathbf{w}_1, .., \mathbf{w}_c]$. FLDA uses $\mathbf{W}$ as a projection matrix to obtain a low-dimensional data representation, and according to Fisher's criterion, the discrimination power in the dimension reduced space is given by [23]

$$\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\mathbf{W}) = \mathrm{Tr}\left((\mathbf{W}^T\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}\mathbf{W}\right) = \sum_{i=1}^{c}\boldsymbol{\lambda}_i. \tag{7}$$

In practice, we do not have access to population parameters $\boldsymbol{\Sigma}$ and $\mathbf{S}$, but their estimates, i.e., the sample covariance $\widehat{\boldsymbol{\Sigma}}$ and the sample between-class scatter matrix $\widehat{\mathbf{S}}$ via sample class means $\widehat{\boldsymbol{\mu}}_i$. Denoting by $\widehat{\mathbf{W}}$ the empirical projection matrix obtained from generalized eigendecomposition of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{S}}$, the generalization discrimination power of FLDA is given by

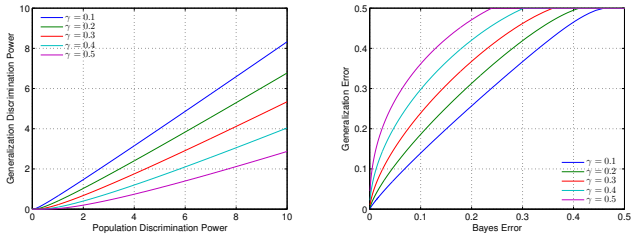
$$\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) = \mathrm{Tr}\left((\widehat{\mathbf{W}}^T\boldsymbol{\Sigma}\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^T\mathbf{S}\widehat{\mathbf{W}}\right), \tag{8}$$

which measures how the classes are separated in the dimension reduced space. When data dimensionality D is fixed and training sample size N goes to infinity, the generalization discrimination power (8) will converge to its population counterpart (7), since $\widehat{\mathbf{W}}$ converges to $\mathbf{W}$. However, such classical result is invalid when D increases proportionally with N. Regarding this, the following theorem gives a new asymptotic result on FLDA's generalization ability, in a setting where D and N increase to infinity proportionally.

Theorem 1: Suppose the population discrimination power $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\mathbf{W}) = \sum_{i=1}^{c}\boldsymbol{\lambda}_i$. The generalization discrimination power $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ can be factorized as

$$\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) = \sum_{i=1}^{c}\boldsymbol{\delta}_i\boldsymbol{\lambda}_i \tag{9}$$

[1]For the convenience of expression, we assume an equal prior probability. This does not substantially change the analysis throughout this paper.



(a) Lower Bound of Discrimination Power (b) Upper Bound of Binary Classification Error

Fig. 1. Asymptotic Generalization Bound of Fisher's Linear Discriminant Analysis.

where $0 \leq \boldsymbol{\delta}_i \leq 1$. Further, as both the dimensionality D and the training sample size N increase ($N > D$) and $D/N \longrightarrow \gamma \in [0, 1)$, it holds asymptotically

$$\boldsymbol{\delta}_i \boldsymbol{\lambda}_i \geq \max^2\Big\{ \cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)}) + \arccos(\sqrt{1-\gamma})), 0\Big\}\boldsymbol{\lambda}_i, \ a.s. \tag{10}$$

Theorem 1 gives an asymptotically lower bound on the generalization ability of FLDA, in terms of the population discrimination power $\boldsymbol{\lambda}_i$ and the dimensionality to training sample size ratio $\gamma = D/N$. An important feature of the bound is that it is determined by the ratio $\gamma = D/N$ rather than the dimensionality D. In other words, a good generalization performance of FLDA only requires a training sample size that scales linearly with respect the dimensionality, although there are a quadratic number of parameters to be estimated in the sample covariance. Figure 1 (a) gives an illustration of the bound under different values of the ratio $\gamma = D/N$.

Besides, according to (10), the influence of the ratio $\gamma = D/N$ to the lower bound comes from two aspects, each through the term $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)}$ and the term $\sqrt{1-\gamma}$. Note that $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)}$ allows a tradeoff between $\boldsymbol{\lambda}_i$ and γ, i.e., when $\boldsymbol{\lambda}_i$ is sufficiently large, $\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)})$ approaches 0 and thus vanishes from the lower bound (10). The second term $\sqrt{1-\gamma}$ only depends on γ, and later proofs reveal that it measures how covariance estimation influences the generalization of FLDA. Assuming a sufficient large $\boldsymbol{\lambda}_i$ such that $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)} \approx 1$, we have

$$\boldsymbol{\delta}_i \boldsymbol{\lambda}_i \approx (1-\gamma)\boldsymbol{\lambda}_i, \tag{11}$$

which shows that the loss of discrimination power due to the imperfection of covariance estimation is approximately proportion to γ. To the best of our knowledge, this is the simplest quantitative result on the influence of covariance estimation to FLDA, compared with related studies in the literature [15] [24] [25]. It is worth noticing that, as long as $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^{D\times D}$, the result is independent of the spectrum of the population covariance $\boldsymbol{\Sigma}$, e.g., the extreme eigenvalues $\lambda_{min}(\boldsymbol{\Sigma})$ and $\lambda_{max}(\boldsymbol{\Sigma})$, or the conditional number $\lambda_{max}(\boldsymbol{\Sigma})/\lambda_{min}(\boldsymbol{\Sigma})$.

B. Bounding Generalization Error of Binary Classification

In binary-class case, FLDA can also be regarded as a linear classifier, where the hyperplane of the linear classifier is perpendicular to the one-dimensional projection vector $\widehat{\mathbf{w}}_1$ of dimension reduction. Without loss of generality, suppose $\widehat{\mathbf{w}}_1^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0$, the generalization error P of binary classification with FLDA can be calculated analytically by [26]

$$P = 0.5\Phi\left\{-\frac{\widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_1 - 0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\right\} + 0.5\Phi\left\{-\frac{0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2) - \widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_2}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\right\}, \tag{12}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard Gaussian. If we replace $\widehat{\mathbf{w}}_1$ and $\widehat{\boldsymbol{\mu}_i}$ by its population counterpart $\mathbf{w}_1$ and $\boldsymbol{\mu}_i$, then (12) gives the Bayes error P_{Bayes}, i.e.,

$$P_{Bayes} = \Phi\left\{-\frac{0.5\mathbf{w}_1^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{\mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1}}\right\} = \Phi\left\{-\sqrt{\frac{\mathbf{w}_1^T\mathbf{S}\mathbf{w}_1}{\mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1}}\right\} = \Phi\left(-\sqrt{\boldsymbol{\lambda}_1}\right). \tag{13}$$

Below, we present a corollary of Theorem 1, which gives an asymptotic upper bound of P in terms of P_{Bayes} and $\gamma = D/N$.

Corollary 1: For binary classification with equal prior probabilities, suppose the population discrimination power $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\mathbf{w}_1) = \boldsymbol{\lambda}_1$, then if both dimensionality D and training sample size N increase ($N > D$) and $D/N \longrightarrow \gamma \in [0, 1)$, the generalization error P of FLDA can be upper bounded asymptotically by

$$P \leq \Phi\left(-\varrho\sqrt{\boldsymbol{\lambda}_1}\right), \ a.s. \tag{14}$$

where

$$\varrho = \max\Big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_1/(\boldsymbol{\lambda}_1+\gamma)})+\arccos(\sqrt{1-\gamma})), 0\Big\}. \tag{15}$$

Further since the Bayes error $P_{Bayes} = \Phi\left(-\sqrt{\boldsymbol{\lambda}_1}\right)$, it holds asymptotically

$$P \leq \Phi\big(\varrho\Phi^{-1}\left(P_{Bayes}\right)\big), \ a.s. \tag{16}$$

with

$$\varrho = \max\left\{\cos\left(\arccos\left(\sqrt{\frac{(\Phi^{-1}(P_{Bayes}))^2}{((\Phi^{-1}(P_{Bayes}))^2+\gamma}}\right) + \arccos(\sqrt{1-\gamma})\right), 0\right\}. \tag{17}$$

Similar to the discrimination power bound, Corollary 1 shows that, given a binary classification problem with Bayes error P_{Bayes}, the generalization error of FLDA is also determined by the dimensionality to training sample size ratio $\gamma = D/N$. Figure 1 (b) gives an illustration of the generalization error bound under different values of γ.

C. Related Work

In recent years, asymptotic analysis on FLDA have also been performed in the case where $D > N$. For example, [15] found that when D increases faster than N the the pseudo-inverse based FLDA approaches to a random guess and therefore suggested a "naive Bayes" approach in this situation. A more detailed analysis on pseudo-inverse FLDA was given in

[25] by investigating the estimation error of pseudo-inverse of the sample covariance. Random matrix theory, e.g., Marčenko-Pastur Law, was also utilized in [25], so as to bound the expected estimation error in the asymptotic case. The result in this paper provides a complementary theory of FLDA in the setting of $D < N$, which shows that the generalization ability of FLDA in such situation is mainly determined by the ratio $\gamma = D/N$.

In contrast to asymptotic analysis, generalization bounds in finite sample case were derived most recently in both linear and kernel spaces, and by using random projection as regularization if $D > N$ [24] [27] [28]. The advantage of these results is they provide explicit probability bounds for finite N and D, while asymptotic results inherently require sufficient large N and D. However, we would like to emphasize that the bounds obtained in this paper have their own merit, by linking the generalization discrimination power (or generalization error) to the population discrimination power (or Bayes error) directly in terms of the ratio $\gamma = D/N$. Besides, as shown by empirical evaluation in later section IV, the bounds hold with high probability (in the empirical sense) for moderate D and N, though they are obtained asymptotically.

III. Proof of Main Result

In this section, we present the proof of Theorem 1, which are mainly based upon the asymptotic results on eigensystems of the sample covariance and the sample between-class scatter matrix.

A. On $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$

We begin the proof by bounding the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ in terms of eigenvalues and/or eigenvectors of a normalized version of the sample covariance and sample between-class scatter matrix.

Lemma 1: Given a problem with population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, there is a nonsingular matrix $\mathbf{X}$ that simultaneously diagonalizes $\mathbf{\Sigma}$ and $\mathbf{S}$, i.e.,

$$\mathbf{X}^T \mathbf{\Sigma} \mathbf{X} = \mathbf{I} \text{ and } \mathbf{X}^T \mathbf{S} \mathbf{X} = \mathbf{\Lambda}_0, \tag{18}$$

where $\mathbf{\Lambda}_0 = \text{diag}(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_c, 0, ..., 0)$.

Lemma 2: Given the normalized estimates $\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}^T \widehat{\mathbf{\Sigma}} \mathbf{X}$ and $\widehat{\mathbf{S}}_0 = \mathbf{X}^T \widehat{\mathbf{S}} \mathbf{X}$, and their eigendecompositions $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U} \Lambda(\widehat{\mathbf{\Sigma}}_0) \mathbf{U}^T$ and $\widehat{\mathbf{S}}_0 = \mathbf{V} \Lambda(\widehat{\mathbf{S}}_0) \mathbf{V}^T$, the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}})$ can be expressed as

$$\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i, \tag{19}$$

where

$$\boldsymbol{\delta}_i = \left\| \mathcal{R}^T \left(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0) \mathbf{U}^T \mathbf{V}_{1:c} \right) \mathbf{U}^T \mathbf{e}_i \right\|^2. \tag{20}$$

Lemma 3: Given $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ and $\mathbf{V}_{1:c}$ from Lemma 2, it holds

$$\boldsymbol{\delta}_i \geq \max^2 \Big\{ \cos \big(\arccos(\|\mathbf{V}_{1:c}^T \mathbf{e}_i\|) + \arccos \left(\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0) \xi \Big/ \sqrt{\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0) \xi} \right) \big), 0 \Big\}. \tag{21}$$

where ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$.

Lemma 2 and Lemma 3 show that the generalization discrimination power of FLDA are determined by the eigensystems of the normalized estimates $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$. Since $\widehat{\mathbf{\Sigma}}_0$ is actually an estimate of the identity covariance matrix $\mathbf{I}$, we have that given the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, the generalization ability of FLDA, i.e., $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, is independent of the population covariance $\mathbf{\Sigma}$. Next, we present properties on the eigensymstems of $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$, which are necessary for evaluating the lower bound of $\boldsymbol{\delta}_i$ in (21).

B. Properties of $\widehat{\mathbf{\Sigma}}_0$

We have the following lemma on the eigensystem of the normalized sample covariance $\widehat{\mathbf{\Sigma}}_0$.

Lemma 4: Given the eigendecomposition $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U} \Lambda(\widehat{\mathbf{\Sigma}}_0) \mathbf{U}^T$, it holds

1) $\mathbf{U}$ and $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ are independent random variables;
2) $\mathbf{U}$ follows the Haar distribution, i.e., it is uniformly distributed on the set of all orthonormal matrices in $\mathbb{R}^{D \times D}$;
3) denoting by $F_N(\lambda)$ the empirical spectral distribution of the eigenvalues of $\widehat{\mathbf{\Sigma}}_0$, i.e.,

$$F_N(\lambda) = \frac{1}{D} \sum_{i=1}^{D} 1\{\lambda_i(\widehat{\mathbf{\Sigma}}_0) \leq \lambda\}, \ \lambda \geq 0, \tag{22}$$

then, as $D/N \longrightarrow \gamma \in [0, 1)$,

$$F_N(\lambda) \xrightarrow{a.s.} F_\gamma(\lambda), \tag{23}$$

where the limit distribution $F_\gamma(\lambda)$ has the density

$$dF_\gamma(\lambda) = \frac{1}{2\pi\gamma} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} d\lambda, \tag{24}$$

with

$$\lambda_+ = (1 + \sqrt{\gamma})^2 \text{ and } \lambda_- = (1 - \sqrt{\gamma})^2. \tag{25}$$

The first and the second statements in Lemma 4 can be understood by the fact that $\widehat{\mathbf{\Sigma}}_0$ is an empirical estimate of $\mathbf{I}$, whose probability density is invariant to any orthogonal transformation. The last statement is a corollary of the Marčenko-Pastur law, i.e., Proposition 1, which says that the empirical spectral distribution of the matrix $\frac{1}{N} \mathbf{G} \mathbf{G}^T$, wherein $\mathbf{G} \in \mathbb{R}^{D \times N}$ has i.i.d entries sampled from $\mathcal{N}(0, 1)$, converges almost surely to the deterministic distribution $F_\gamma(\lambda)$ as $D/N \longrightarrow \gamma \in [0, 1)$.

Further, we need the following lemma on the inverse of the eigenvalues $\Lambda(\widehat{\mathbf{\Sigma}}_0)$, which says that the energy of $\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)$ and $\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)$ projected onto a random direction is almost surely deterministic in the limit. It is worth noticing that the results in Lemma 5 generalize the results on the expectations $\mathbb{E}[\sum_i \lambda_i^{-1}(\widehat{\mathbf{\Sigma}}_0)]$ and $\mathbb{E}[\sum_i \lambda_i^{-2}(\widehat{\mathbf{\Sigma}}_0)]$ in [25].

Lemma 5: Suppose ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$ and it is indepen-

dent of $\widehat{\mathbf{\Sigma}}_0$, then as $D/N \longrightarrow \gamma \in [0,1)$,

$$\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-1} dF_\gamma(\lambda) = \frac{1}{1-\gamma}, \tag{26}$$

and

$$\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-2} dF_\gamma(\lambda) = \frac{1}{(1-\gamma)^3}. \tag{27}$$

C. Properties of $\widehat{\mathbf{S}}_0$

We have the following lemma on the eigenvectors of $\widehat{\mathbf{S}}_0$.

Lemma 6: Given the eigendecomposition $\widehat{\mathbf{S}}_0 = \mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T$, then as $D/N \longrightarrow \gamma \in [0,1)$,

$$\lim_{D/N \longrightarrow \gamma} \|\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2 \geq \frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i + \gamma}, \ a.s., \ i = 1, 2, ..., c, \tag{28}$$

where $\boldsymbol{\lambda}_i$ is from the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}) = \sum_{i=1}^c \boldsymbol{\lambda}_i$.

Recalling Lemma 1, the population counterpart of $\widehat{\mathbf{S}}_0$ is actually the diagonal matrix $\mathbf{\Lambda}_0 = \mathbf{X}^T\mathbf{S}\mathbf{X}$. Therefore, we expect the first c eigenvectors $\mathbf{V}_{1:c}$ of $\widehat{\mathbf{S}}_0$ to be close to $\mathbf{I}_{1:c} = [\mathbf{e}_1, ..., \mathbf{e}_c]$. Lemma 6 shows that the performance of eigenvector estimation is determined by the $\boldsymbol{\lambda}_i$ and γ, and in particular, as γ approaches 0 the estimation becomes consistent.

D. Proof of Theorem 1

Now, we are ready to prove our main result Theorem 1, which is a conclusion out of the combination of Lemmas 2, 3, 5 and 6.

Proof: By Lemma 5, we have

$$\lim_{D/N \longrightarrow \gamma} \frac{\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi}{\sqrt{\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi}} = \frac{\frac{1}{1-\gamma}}{\frac{1}{(1-\gamma)^{1.5}}} = \sqrt{1-\gamma}, \ \text{a.s.} \tag{29}$$

By Lemma 6, we have

$$\lim_{D/N \longrightarrow \gamma} \|\mathbf{V}_{1:c}^T \mathbf{e}_i\| \geq \sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}, \ \text{a.s.} \tag{30}$$

Then the proof is completed by substituting (29) and (30) into Lemma 2 and Lemma 3. ■

IV. Empirical Evaluations

A. On the Bound of Generalization Discrimination Power

According to Theorem 1, the generalization discrimination power of FLDA for dimension reduction can be factorized as $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) = \sum_{i=1}^c \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, where $\boldsymbol{\lambda}_i$ measures the population discrimination power, and each component $\boldsymbol{\delta}_i \boldsymbol{\lambda}_i$ of the generalization discrimination power can be lower bounded by

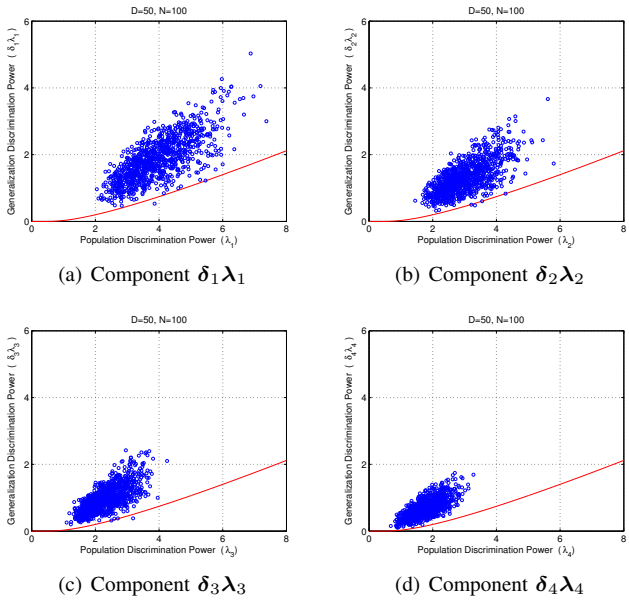
$$\boldsymbol{\delta}_i \boldsymbol{\lambda}_i \geq \max^2\big\{ \cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}) + \arccos(\sqrt{1-\gamma})), 0\big\}\boldsymbol{\lambda}_i.$$

We evaluate this result on both simulated and real datasets by comparing $\boldsymbol{\delta}_i \boldsymbol{\lambda}_i$ with the lower bound above.

For simulated data, we fix the ratio $\gamma = D/N = 0.5$, with $D = 50$ and $N = 100$. Note the settings give moderate size problems; however, due to the asymptotic characteristic of the bound, which inherently fits to large size problem, the evaluation on moderate size problems is more critical. We generate 1,000 experiments, each having 5 classes with randomly generated population covariance $\mathbf{\Sigma}$ and class means $\boldsymbol{\mu}_i$, $i = 1, ..., 5$. The population discrimination power $\boldsymbol{\lambda}_i$, $i = 1, ..., 4$, are calculated via eigendecomposition of $\mathbf{\Sigma}^{-1}\mathbf{S}$, where $\mathbf{S}$ is the between-class scatter matrix. For the generalization discrimination power $\boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, the factor $\boldsymbol{\delta}_i$ has a close form formulation as shown by Lemma 2, i.e.,

$$\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2,$$

where $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ and $\mathbf{U}$ are the eigensystems of $\widehat{\mathbf{\Sigma}}_0$ and $\mathbf{V}_{1:c}$ are the first c eigenvectors of $\widehat{\mathbf{S}}_0$, with $\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}^T\widehat{\mathbf{\Sigma}}\mathbf{X}$ and $\widehat{\mathbf{S}}_0 = \mathbf{X}^T\widehat{\mathbf{S}}\mathbf{X}$ being the normalized sample covariance and between-class scatter matrix and $\mathbf{X}$ simultaneously diagonalizing $\mathbf{\Sigma}$ and $\mathbf{S}$. Since a larger discrimination power means a better separation between classes, we expect that on most of the experiments the generalization discrimination power of FLDA can be bounded from the lower side by the generalization bound. Indeed, as shown by Figure 2, the bound holds with an overwhelming probability in the empirical sense (i.e., on more than 990 out of the 1,000 experiments).



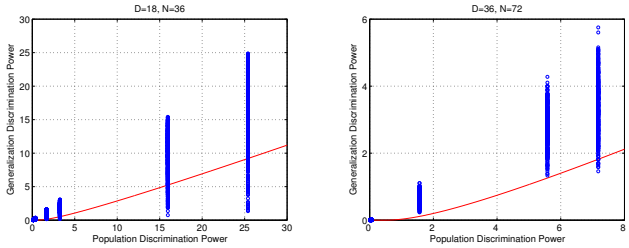
(a) Component $\boldsymbol{\delta}_1\boldsymbol{\lambda}_1$ (b) Component $\boldsymbol{\delta}_2\boldsymbol{\lambda}_2$

(c) Component $\boldsymbol{\delta}_3\boldsymbol{\lambda}_3$ (d) Component $\boldsymbol{\delta}_4\boldsymbol{\lambda}_4$

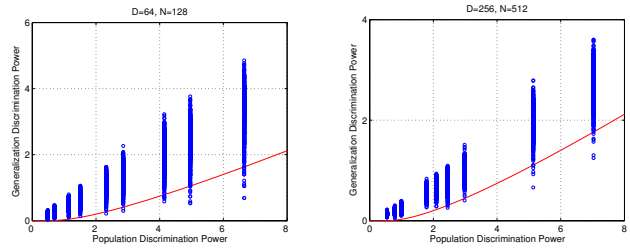
Fig. 2. Evaluation of the Generalization Discrimination Power Bound with Simulated Data.

We further evaluate the bound of generalization discrimination power on four benchmark datesets from the UCI machine learning repository [29]: 1) the image segmentation (ImageSeg) dataset [2], which contains 7 classes and in total 2,310 examples from $\mathbb{R}^{18}$; 2) the Landsat dataset, which constants 6 classes and in total 6,435 examples from $\mathbb{R}^{36}$; 3) the optical recognition of handwritten digits (Optdigits) dataset, which contains 10 classes and in total 5,620 examples from $\mathbb{R}^{60}$; and 4) the USPS handwritten digits dataset, which contains 10 classes and in total 9,298 examples from $\mathbb{R}^{256}$. Note that for real dataset, the population parameters $\mathbf{\Sigma}$ and

[2]The original dataset has 19 features; however the 3rd feature is a constant for all examples, and therefore is discarded in the experiments.



(a) ImageSeg (b) LandSat



(c) Optdigits (d) USPS

Fig. 3. Evaluation of the Generalization Discrimination Power Bound with Real Data.

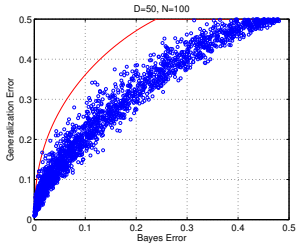
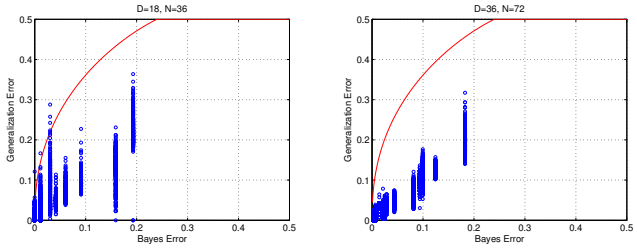
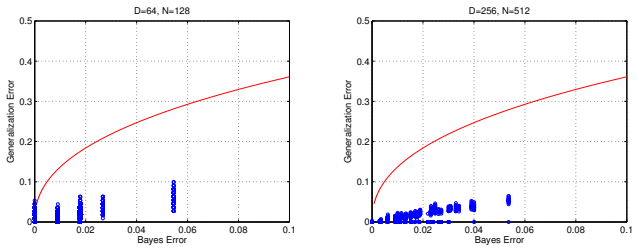


Fig. 4. Evaluation of the Generalization Error Bound with Simulated Data.



(a) ImageSeg (b) LandSat



(c) Optdigits (d) USPS

Fig. 5. Evaluation of the Generalization Error Bound with Real Data.

$\mathbf{S}$ are unknown. Thus, we use the entire dataset to get their estimates and treat them as population parameters. Again, we fix the ratio $\gamma = D/N = 0.5$, i.e., we randomly select examples twice of the dimensionality as the training data. The generalization discrimination powers over 1,000 random experiments are shown in Figure 3. On the panel for each dataset, the columns of the scatters correspond to different components of the generalization discrimination power $\boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, and the horizontal axis location of each column equals the population discrimination power $\boldsymbol{\lambda}_i$ (the column number is class number minus 1). On three out of the four datasets, including LandSat, Optdigits and USPS, the generalization discrimination power is properly bounded by the lower bound, with a high probability in the empirical sense. On the ImageSeg dataset, the bound does not hold with high probability as on the other three datasets. The major reason is that the size of the problem is considerably small, with $D = 18$ and $N = 36$, while the bound favors large or moderate size problems.

B. On the Bound of Generalization Errors

According to Corollary 1, suppose the Bayes error of a binary classification problem is P_{Bayes}, then the generalization error P of FLDA can be boudned by

$$P \leq \Phi(\varrho \Phi^{-1}(P_{Bayes})),$$

where $\Phi(\cdot)$ is the CDF of the standard Gaussian distribution and

$$\varrho = \max\left\{\cos\left(\arccos\left(\sqrt{\frac{(\Phi^{-1}(P_{Bayes}))^2}{((\Phi^{-1}(P_{Bayes}))^2+\gamma}}\right) + \arccos(\sqrt{1-\gamma})\right), 0\right\}. \quad (31)$$

To evaluate this result, we perform binary classification with FLDA on 1,000 experiments, with randomly generated covariance matrix and class means. The same as in previous simulation, we fix the ratio $\gamma = D/N = 0.5$, with $D = 50$ and $N = 100$. Figure 4 shows the result, where the generalization error of FLDA is properly bounded by the upper bound.

In addition, we run experiments on the previous four real datasets to evaluate the generalization error bound. We randomly select class pairs from each dataset to perform binary classification. We hold out 10% data as the evaluation set, which is used to estimate the "Bayes" error and generalization error. The "Bayes" classifier is obtained by training FLDA on the rest 90% data, and the empirical classifier is trained with a subset of the rest data, such that $N = 2D$, namely fixing the ratio $\gamma = D/N = 0.5$. On each dataset, 1,000 random experiments are performed, with the result shown in Figure 5. Similar to the result in Figure 3, on three out of the four datasets, the generalization error can be bounded by the upper bound, while the bound does not dominate all the experiment on the ImageSeg dataset due to the small size of the problem.

V. Conclusion

FLDA is an important statistical model in pattern recognition. The result obtain in this paper enriches the existing theory of FLDA, by showing that the generalization ability of FLDA is mainly determined by the dimensionality to training sample size ratio $\gamma = D/N$, given D and N are reasonably large and

$N > D$. Important conclusions from this result include: 1) to ensure FLDA performing well, training sample size only needs to scale linearly with respect to data dimensionality, although a quadratic number of parameters are to be estimated in the sample covariance; and 2) the generalization ability of FLDA (with respect to the Bayes optimum) is independent of the spectral structure of the population covariance, given its nonsingularity and above conditions.

VI. Proofs

We provide below the detailed proofs of Lemmas in Section III and Corollary 1 in Section II.

A. Proof of Lemma 1

It is a direct result of the simultaneous diagonalization theorem for a pair of semidefinite matrices [23].

B. Proof of Lemma 2

The proof is divided into two steps.

i) Since $\mathbf{X}$ in Lemma 1 is nonsingular, there exists some $\mathbf{Q} \in \mathbb{R}^{D\times c}$ such that $\widehat{\mathbf{W}} = \mathbf{X}\mathbf{Q}$. Then,

$$\begin{aligned}\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}) &= \mathrm{Tr}((\widehat{\mathbf{W}}^T\mathbf{\Sigma}\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^T\mathbf{S}\widehat{\mathbf{W}}) \\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{X}^T\mathbf{S}\mathbf{X}\mathbf{Q}) \\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{X}^T\mathbf{\Lambda}\mathbf{Q}) \\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\mathbf{\Lambda}_1\mathbf{Q}_1) \\ &= \mathrm{Tr}(\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\mathbf{\Lambda}_1) \\ &= \sum_{i=1}^{c}\boldsymbol{\delta}_i\boldsymbol{\lambda}_i,\end{aligned} \tag{32}$$

where $\mathbf{Q}_1$ contains the first c rows of $\mathbf{Q}$ and $\mathbf{\Lambda}_1$ is the upper-left $c \times c$ submatrix of $\mathbf{\Lambda}$, and clearly,

$$\boldsymbol{\delta}_i = \{\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\}_{ii}. \tag{33}$$

ii) In FLDA, $\widehat{\mathbf{W}}$ are the eigenvectors of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{S}}$, and we can restrict the scale of $\widehat{\mathbf{W}}$ such that

$$\widehat{\mathbf{W}}^T\widehat{\mathbf{\Sigma}}\widehat{\mathbf{W}} = \mathbf{I}_c \text{ and } \widehat{\mathbf{W}}^T\widehat{\mathbf{S}}\widehat{\mathbf{W}} = \widehat{\mathbf{\Lambda}}_1, \tag{34}$$

where $\widehat{\mathbf{\Lambda}}_1$ is some $c \times c$ diagonal matrix. Substituting $\widehat{\mathbf{W}} = \mathbf{X}\mathbf{Q}$ into (34) and recalling $\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}^T\widehat{\mathbf{\Sigma}}\mathbf{X}$ and $\widehat{\mathbf{S}}_0 = \mathbf{X}^T\widehat{\mathbf{S}}\mathbf{X}$, we get

$$\mathbf{Q}^T\widehat{\mathbf{\Sigma}}_0\mathbf{Q} = \mathbf{I}_c \text{ and } \mathbf{Q}^T\widehat{\mathbf{S}}_0\mathbf{Q} = \widehat{\mathbf{\Lambda}}_1. \tag{35}$$

Given the eigendecomposition $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$, we have from the first equation in (35) that there must exist some orthogonal matrix $\mathbf{O} \in \mathbb{R}^{D\times c}$, $\mathbf{O}^T\mathbf{O} = \mathbf{I}_c$, such that

$$\mathbf{Q} = \mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{O}. \tag{36}$$

Further, given the eigendecomposition $\widehat{\mathbf{S}}_0 = \mathbf{V}^T\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}$, we get from the second equation in (35) that

$$\mathbf{O}^T\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T\mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{O} = \widehat{\mathbf{\Lambda}}_1. \tag{37}$$

In addition, since $\widehat{\mathbf{S}}_0$ has rank c, we can rewrite (37) as

$$\mathbf{O}^T\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{O} = \widehat{\mathbf{\Lambda}}_1, \tag{38}$$

where $\Lambda_1(\widehat{\mathbf{\Sigma}}_0)$ is the upper-left $c\times c$ submatrix of $\Lambda(\widehat{\mathbf{\Sigma}}_0)$. (38) implies the columns of $\mathbf{O}$ must be the left singular vectors of $\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)$. Thus, $\mathbf{O}$ spans the range space of $\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)$ and therefore the range space of $\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}$. Then, there must exist some matrix $\mathbf{A} \in \mathbb{R}^{c\times c}$ such that $\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c} = \mathbf{O}\mathbf{A}$, and thus

$$\mathbf{O} = \Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}^{-1}, \tag{39}$$

where the nonsingularity of $\mathbf{A}$ is implied by the nonsingularity of $\Lambda^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$.

By (36) and (39), we have

$$\mathbf{Q} = \mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}, \tag{40}$$

and

$$\mathbf{Q}_1 = \mathbf{I}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}. \tag{41}$$

Therefore,

$$\begin{aligned}&\{\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1\}_{ii} = \\ &\mathbf{e}_i^T\mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}(\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})^{-1} \\ &\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{e}_i.\end{aligned} \tag{42}$$

Letting $\mathbf{R} = \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})$, then

$$\begin{aligned}&\mathbf{R}\mathbf{R}^T = \\ &\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}(\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})^{-1}\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0),\end{aligned} \tag{43}$$

which together with (42) gives

$$\begin{aligned}\{\mathbf{Q}_1(\mathbf{Q}_\ell^T\mathbf{Q}_\ell)^{-1}\mathbf{Q}_1\}_{ii} &= \mathbf{e}_i^T\mathbf{U}\mathbf{R}\mathbf{R}^T\mathbf{U}^T\mathbf{e}_i = \|\mathbf{R}^T\mathbf{U}^T\mathbf{e}_i\|^2 \\ &= \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2.\end{aligned} \tag{44}$$

This completes the proof.

C. Proof of Lemma 3

Recall Lemma 2 that $\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2$. Denote by $\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}))$ the angle between vector $\mathbf{U}^T\mathbf{e}_i$ and subspace $\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})$, we have

$$\boldsymbol{\delta}_i = \cos^2(\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}))). \tag{45}$$

Two basic facts that hold for arbitrary vector $\mathbf{a}_1$, $\mathbf{a}_2$ and subspace $\mathbf{A}$ are

$$\measuredangle(\mathbf{a}_1, \mathbf{A}) \le \measuredangle(\mathbf{a}_1, \mathbf{a}_2) + \measuredangle(\mathbf{a}_2, \mathbf{A}). \tag{46}$$

and

$$\measuredangle(\mathbf{a}_1, \mathbf{A}) \le \measuredangle(\mathbf{a}_1, \mathbf{a}), \text{ if } \mathbf{a} \in \mathbf{A}. \tag{47}$$

Then, by using (46) and (47), we get

$$\begin{aligned}&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_i)) \\ \le&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i) \\ &+ \measuredangle(\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})) \\ \le&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i) \\ &+ \measuredangle(\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i, \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i) \\ =&\theta_1 + \theta_2.\end{aligned} \tag{48}$$

Denoting $\theta = \theta_1 + \theta_2$, since $\cos(x)$ is positive and decreasing on $[0, \pi/2]$, x^2 is increasing on $[0, 1]$, and $\boldsymbol{\delta}_i$ is nonnegative, we have

$$\begin{aligned}\boldsymbol{\delta}_i &\geq \begin{cases} \cos^2(\theta), & \theta \leq \frac{\pi}{2} \\ 0, & \text{else} \end{cases} \\ &= \max^2\{\cos(\theta), 0\}.\end{aligned} \tag{49}$$

It remains to calculate θ_1 and θ_2. For θ_1, We have

$$\begin{aligned}\cos^2(\theta_1) &= \frac{|\mathbf{e}_i \mathbf{V}_{1:c}^T \mathbf{U}\mathbf{U}^T \mathbf{V}_{1:c}\mathbf{e}_i|^2}{\|\mathbf{U}^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2} \\ &= \frac{|\mathbf{e}_i^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i|^2}{\mathbf{e}_i^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i} = \|\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2,\end{aligned} \tag{50}$$

which gives

$$\theta_1 = \arccos(\|\mathbf{V}_{1:c}^T \mathbf{e}_i\|). \tag{51}$$

For θ_2, as rescaling does not change the direction of a vector, we can rewrite θ_2 as

$$\theta_2 = \measuredangle(\mathbf{U}^T\zeta, \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\zeta), \tag{52}$$

where

$$\zeta = \frac{\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i}{\|\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i\|}. \tag{53}$$

Note that ζ is a unit-length random vector and is independent of $\mathbf{U}$ due to the independency between $\mathbf{V}_{1:c}$ and $\mathbf{U}$. Then, we have

$$\begin{aligned}\cos^2(\theta_2) &= \frac{|\zeta^T \mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\zeta|^2}{\|\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\zeta\|^2} \\ &= \frac{(\zeta^T \mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\zeta)^2}{\zeta^T \mathbf{U}\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\zeta}.\end{aligned} \tag{54}$$

We have known, from Lemma 4, $\mathbf{U}$ is uniformly distributed on the set of all orthonormal matrices in $\mathbb{R}^{D\times D}$, and ζ is a unit-length random vector independent of $\mathbf{U}$. Thus, $\xi = \mathbf{U}^T\zeta$ must be a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$. Finally, (54) gives

$$\theta_2 = \arccos\left(\xi^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi \Big/ \sqrt{\xi^T\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\xi}\right). \tag{55}$$

This completes the proof.

D. Proof of Lemma 4

Since $\widehat{\boldsymbol{\Sigma}}_0 = \mathbf{X}^T\widehat{\boldsymbol{\Sigma}}\mathbf{X}$ is a normalized sample covariance, wherein $\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X} = \mathbf{I}$, we have

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}(\mathbf{x}_j^i - \bar{\mathbf{x}}_i)(\mathbf{x}_j^i - \bar{\mathbf{x}}_i)^T, \tag{56}$$

where $\mathbf{x}_j^i$ is sampled from some $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I})$ and $\bar{\mathbf{x}}_i$ is the sample mean. Letting $\mathbf{z}_j^i = \mathbf{x}_j^i - \boldsymbol{\mu}_i$, which implies $\mathbf{z}_j^i$ is sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, and $\bar{\mathbf{z}}^i = \bar{\mathbf{x}}^i - \boldsymbol{\mu}_i$, then $\widehat{\boldsymbol{\Sigma}}_0$ can be rewritten as

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}(\mathbf{z}_j^i - \bar{\mathbf{z}}^i)(\mathbf{z}_j^i - \bar{\mathbf{z}}^i)^T, \tag{57}$$

One property of $\widehat{\boldsymbol{\Sigma}}_0$ in (57) is that, as a random variable, its distribution is invariant to orthogonal similarity transformation, i.e., $\widehat{\boldsymbol{\Sigma}}_0$ and $\mathbf{O}\widehat{\boldsymbol{\Sigma}}_0\mathbf{O}^T$, wherein $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, have the same distribution. This is due to the fact that $\mathbf{O}^T\widehat{\boldsymbol{\Sigma}}_0\mathbf{O}$ corresponds to (57) in the case of replacing $\mathbf{z}_j^i$ by $\mathbf{O}\mathbf{z}_j^i$ while $\mathbf{O}\mathbf{z}_j^i$ has the same distribution with $\mathbf{z}_j^i$, i.e., the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. Then, according to Theorem 3.2 in [30], the invariant property to orthogonal similarity transformation implies that the distribution of $\widehat{\boldsymbol{\Sigma}}_0$ is independent of its eigenvectors $\mathbf{U}$ but only depends on its eigenvalues $\Lambda(\widehat{\boldsymbol{\Sigma}}_0)$, and $\mathbf{U}$ is a random matrix uniformly distributed on the set of all possible orthonormal matrices in $\mathbb{R}^{D\times D}$. This completes the statements 1) and 2) in Lemma 4.

Further, (57) can be rewritten as

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_0 &= \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}\mathbf{z}_j^i\mathbf{z}_j^{iT} - \frac{1}{c+1}\sum_{i=1}^{c+1}\bar{\mathbf{z}}^i\bar{\mathbf{z}}^{iT} \\ &= \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}\mathbf{z}_j^i\mathbf{z}_j^{iT} - \frac{1}{(c+1)n}\sum_{i=1}^{c+1}\sqrt{n}\bar{\mathbf{z}}^i\sqrt{n}\bar{\mathbf{z}}^{iT} \\ &= \frac{1}{N}\mathbf{G}_1\mathbf{G}_1^T - \frac{1}{N}\mathbf{G}_2\mathbf{G}_2^T = T_1 + T_2.\end{aligned} \tag{58}$$

where $\mathbf{G}_1 \in \mathbb{R}^{D\times N}$ and $\mathbf{G}_2 \in \mathbb{R}^{D\times(c+1)}$. For the first term $T_1 = \frac{1}{N}\mathbf{G}_1\mathbf{G}_1^T$, by Proposition 1, we know that the empirical distribution of its eigenvalues converges almost surely to $F_\gamma(\lambda)$ with density,

$$dF_\gamma(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda}d\lambda, \tag{59}$$

where $\gamma = \lim D/N$ and

$$\lambda_+ = (1 + \sqrt{\gamma})^2 \text{ and } \lambda_- = (1 - \sqrt{\gamma})^2. \tag{60}$$

For the second term $T_2 = \frac{1}{N}\mathbf{G}_2\mathbf{G}_2^T$, clearly it has finite rank $c + 1$. According to [31], a finite rank perturbation does not effect the convergence of the empirical spectral distribution, i.e., $\lim F_N(\lambda(T_1 + T2)) = \lim F_N(\lambda(T_1)) = F_\gamma(\lambda)$. This completes the proof.

E. Proof of Lemma 5

The condition that ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$ can be replaced by $\xi \in \mathbb{R}^D$ with entries independently sampled from $\mathcal{N}(0, 1/D)$. This is because, in the later case, $\xi/\|\xi\|$ is uniformly distributed on $\mathbb{S}^{D-1}$, and $\|\xi\|^2 \xrightarrow{a.s.} 1$ due to the Strong Law of Large Numbers.

For (26), we divide the proof into two steps. First, we show that $\xi^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-1}dF_\gamma(\lambda)$, and then we calculate the integral.

i) Recall $\lambda_- = (1 - \sqrt{\gamma})^2$, and let $\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0) = \operatorname{diag}(\min\{\lambda_-, \lambda_i^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\})$, i.e., a truncated version of $\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)$ by clamping $\lambda_i^{-1}(\widehat{\boldsymbol{\Sigma}}_0)$ to be λ_-^{-1} if $\lambda_i^{-1}(\widehat{\boldsymbol{\Sigma}}_0) \geq \lambda_-^{-1}$. Then, we divide the left-hand side of (26) into three terms

$$\xi^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \xi^T\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi, \tag{61}$$

$$\xi^T\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \frac{1}{D}\operatorname{Tr}(\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)), \tag{62}$$

and

$$\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0)) - \int \lambda^{-1} dF_\gamma(\lambda). \tag{63}$$

We show that all the three terms converge almost surely to zero.

For the first term (61), we have

$$\begin{aligned} 0 \leq & \xi^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0) - \overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0))\xi \\ \leq & \|\xi\|^2 \max\{0, \lambda_{\min}^{-1}(\widehat{\mathbf{\Sigma}}_0) - \lambda_-^{-1}\}. \end{aligned} \tag{64}$$

By the same argument in the proof of Lemma 4, we know that

$$\begin{aligned} \lim \lambda_{min}(\widehat{\mathbf{\Sigma}}_0) &= \lim \lambda_{min}\left(\frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}\mathbf{z}_j^i\mathbf{z}_j^{iT}\right) \\ &= \left(\lim \frac{1}{\sqrt{N}}\sigma_{min}(\mathbf{Z})\right)^2, \end{aligned} \tag{65}$$

where $\mathbf{Z} = [\mathbf{z}_1^1, ..., \mathbf{z}_n^{c+1}] \in \mathbb{R}^{D\times N}$, with entries independently sampled from $\mathcal{N}(0,1)$. By Proposition 2, we have $\lim \frac{1}{\sqrt{N}}\sigma_{min}(\mathbf{Z}) = 1 - \sqrt{\gamma}$, and thus $\lambda_{min}(\widehat{\mathbf{\Sigma}}_0) \xrightarrow{a.s.} (1 - \sqrt{\gamma})^2 = \lambda_-$. Accordingly,

$$\max\{0, \lambda_{\min}^{-1}(\widehat{\mathbf{\Sigma}}_0) - \lambda_-^{-1}\} \xrightarrow{a.s.} 0. \tag{66}$$

Then, by $\|\xi\|^2 \xrightarrow{a.s.} 1$, (64) and (66), we have

$$\xi^T\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi - \xi^T\overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} 0. \tag{67}$$

For the second term (62), since $\|\overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0)\| \leq \lambda_-$ for all D, i.e., it is uniformly bounded, we apply Theorem 3.4 in [22] and get

$$\xi^T\overline{\Lambda}_\alpha^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi - \frac{1}{D}\mathrm{Tr}(\overline{\Lambda}_\alpha^{-1}(\widehat{\mathbf{\Sigma}}_0)) \xrightarrow{a.s.} 0. \tag{68}$$

For the third term (63), since $dF_\gamma(\lambda)$ is nonzero only on $[\lambda_-, \lambda_+]$, it is sufficient to examine

$$\begin{aligned} &\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0)) - \int \lambda^{-1} dF_\gamma(\lambda) \\ =& \int_0^\infty \min(\lambda_-, \lambda^{-1}) dF_N(\lambda) - \int_{\lambda_-}^{\lambda_+} \lambda^{-1} dF_\gamma(\lambda) \\ =& \int_{\lambda_-}^{\lambda_+} \lambda^{-1} d(F_N(\lambda) - F_\gamma(\lambda)) + \lambda_-^{-1}\int_0^{\lambda_-} dF_N(\lambda) \\ &+ \int_{\lambda_+}^\infty \lambda^{-1} dF_N(\lambda). \end{aligned} \tag{69}$$

Sine $F_N(\lambda) \xrightarrow{a.s.} F_\gamma(\lambda)$ and λ^{-1} is bounded on $[\lambda_-, \lambda_+]$, it holds [32]

$$\int_{\lambda_-}^{\lambda_+} \lambda^{-1} d(F_N(\lambda) - F_\gamma(\lambda)) \xrightarrow{a.s.} 0. \tag{70}$$

Further, sine $F_\gamma(\lambda_-) = 0$ and $F_\gamma(\lambda_+) = 1$, it holds

$$\int_0^{\lambda_-} dF_N(\lambda) = F_N(\lambda_-) \xrightarrow{a.s.} F_\gamma(\lambda_-) = 0, \tag{71}$$

and

$$\begin{aligned} 0 \leq & \int_{\lambda_+}^\infty \lambda^{-1} dF_N(\lambda) \leq \lambda_+^{-1}(1 - F_N(\lambda_+)) \\ & \xrightarrow{a.s.} \lambda_+^{-1}(1 - F_\gamma(\lambda_+)) = 0. \end{aligned} \tag{72}$$

Thus,

$$\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}_\alpha^{-1}(\widehat{\mathbf{\Sigma}}_0)) - \int \lambda^{-1} dF_\gamma(\lambda) \xrightarrow{a.s.} 0. \tag{73}$$

ii) We now calculate the integral

$$I = \int \lambda^{-1} dF_\gamma(\lambda) = \int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda^2} d\lambda \tag{74}$$

where $\lambda_+ = (1 + \sqrt{\gamma})^2$ and $\lambda_- = (1 - \sqrt{\gamma})^2$.

Letting $\lambda = 1 + \gamma - 2\sqrt{\gamma}\cos x$, $x \in [0, \pi]$ and substituting it into (74), we have

$$I = \frac{2}{\pi}\int_0^\pi \frac{\sin^2 x}{(1 + \gamma - 2\sqrt{\gamma}\cos x)^2} dx. \tag{75}$$

Further, letting $t = \tan\frac{x}{2}$, we have

$$\begin{aligned} I &= \frac{2}{\pi}\int_0^\infty \frac{\left(\frac{2t}{1+t^2}\right)^2}{\left(1 + \gamma - 2\sqrt{\gamma}\frac{1-t^2}{1+t^2}\right)^2}\frac{2}{1+t^2} dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\gamma)(t^2+1) - 2\sqrt{\gamma}(1-t^2)\right)^2}\frac{1}{1+t^2} dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\sqrt{\gamma})^2 t^2 + (1-\sqrt{\gamma})^2\right)^2}\frac{1}{1+t^2} dt \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\int_0^\infty \frac{t^2}{\left(t^2 + \left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^2\right)^2}\frac{1}{1+t^2} dt. \end{aligned} \tag{76}$$

Letting $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$ and by partial fraction, we have

$$\begin{aligned} &\int_0^\infty \frac{t^2}{(t^2+\alpha^2)^2}\frac{1}{1+t^2} dt = \\ &\int_0^\infty \frac{-\frac{1}{(1-\alpha^2)^2}}{t^2+1} dt + \int_0^\infty \frac{\frac{1}{(1-\alpha^2)^2}}{t^2+\alpha^2} dt + \int_0^\infty \frac{-\frac{\alpha^2}{(1-\alpha^2)}}{(t^2+\alpha^2)^2} dt. \end{aligned} \tag{77}$$

Denoting by I_1, I_2 and I_3 the terms in the righthand side of (77), we have

$$\begin{aligned} I_1 &= \int_0^\infty \frac{-\frac{1}{(1-\alpha^2)^2}}{t^2+1} dt = \frac{-1}{(1-\alpha^2)^2}\int_0^\infty d\arctan t \\ &= \frac{-\pi}{2(1-\alpha^2)^2}, \end{aligned} \tag{78}$$

$$\begin{aligned} I_2 &= \int_0^\infty \frac{\frac{1}{(1-\alpha^2)^2}}{t^2+\alpha^2} dt = \frac{1}{\alpha(1-\alpha^2)^2}\int_0^\infty d\arctan\frac{t}{\alpha} \\ &= \frac{\pi}{2\alpha(1-\alpha^2)^2}, \end{aligned} \tag{79}$$

$$\begin{aligned} I_3 &= \int_0^\infty \frac{-\frac{\alpha^2}{(1-\alpha^2)}}{(t^2+\alpha^2)^2} dt \\ &= \frac{-1}{2(1-\alpha^2)}\int_0^\infty d\frac{t}{t^2+\alpha^2} + \frac{-1}{2(1-\alpha^2)}\int_0^\infty \frac{1}{t^2+\alpha^2} dt \\ &= 0 + \frac{-\pi}{4\alpha(1-\alpha^2)} = \frac{-\pi}{4\alpha(1-\alpha^2)}. \end{aligned} \tag{80}$$

Combining (76) to (80) and noticing $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we get

$$\begin{aligned} I &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\left(\frac{-\pi}{2(1-\alpha^2)^2}+\frac{\pi}{2\alpha(1-\alpha^2)^2}+\frac{-\pi}{4\alpha(1-\alpha^2)}\right) \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\frac{\pi}{4\alpha(1+\alpha)^2} = \frac{1}{1-\gamma}. \end{aligned} \tag{81}$$

This completes the proof of (26).

For (27), by the same strategy as used in the proof of (26), we have $\xi^T\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-2}dF_\gamma(\lambda)$. Below, we calculate the integral.

$$I = \int \lambda^{-2}dF_\gamma(\lambda) = \int_{\lambda_-}^{\lambda_+}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{2\pi\gamma\lambda^3}d\lambda, \tag{82}$$

where $\lambda_+ = (1+\sqrt{\gamma})^2$ and $\lambda_- = (1-\sqrt{\gamma})^2$. Letting $\lambda = 1+\gamma-2\sqrt{\gamma}\cos x$, $x\in[0,\pi]$ and substituting it into (74), we have

$$I = \frac{2}{\pi}\int_0^\pi \frac{\sin^2 x}{(1+\gamma-2\sqrt{\gamma}\cos x)^3}dx. \tag{83}$$

Further, letting $t = \tan\frac{x}{2}$, we have

$$\begin{aligned} I &= \frac{2}{\pi}\int_0^\infty \frac{\left(\frac{2t}{1+t^2}\right)^2}{\left(1+\gamma-2\sqrt{\gamma}\frac{1-t^2}{1+t^2}\right)^3}\frac{2}{1+t^2}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\gamma)(t^2+1)-2\sqrt{\gamma}(1-t^2)\right)^3}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\sqrt{\gamma})^2t^2+(1-\sqrt{\gamma})^2\right)^3}dt \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^6}\int_0^\infty \frac{t^2}{\left(t^2+\left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^2\right)^3}dt. \end{aligned} \tag{84}$$

Letting $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we have

$$\begin{aligned} &\int_0^\infty \frac{t^2}{(t^2+\alpha^2)^3}dt \\ =&-\frac{1}{4}\int_0^\infty d\frac{t}{(t^2+\alpha^2)^2}+\frac{1}{4}\int_0^\infty\frac{1}{(t^2+\alpha^2)^2}dt \\ =&\frac{\pi}{16\alpha^3}. \end{aligned} \tag{85}$$

Thus, by $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we get $I = \frac{16}{\pi(1+\sqrt{\gamma})^6}\frac{\pi}{16\alpha^3} = \frac{1}{(1-\gamma)^3}$. This completes the proof of (27).

F. Proof of Lemma 6

By Lemmas 1 and 2, $\widehat{\mathbf{S}}_0$ is an estimate of $\mathbf{X}^T\mathbf{S}\mathbf{X} = \mathbf{\Lambda}_0 = \text{diag}(\boldsymbol{\lambda}_1,...,\boldsymbol{\lambda}_c,0,...,0)$. Suppose the original distributions of the $c+1$ classes are $\mathcal{N}(\boldsymbol{\mu}_i,\mathbf{\Sigma})$ and the between-class scatter matrix is $\mathbf{S}$. Then, $\mathbf{\Lambda}_0$ should be the between-class scatter matrix of an equivalent problem with distributions $\mathcal{N}(\boldsymbol{\mu}_i',\mathbf{I})$, wherein $\boldsymbol{\mu}_i' = \mathbf{X}^T\boldsymbol{\mu}_i$. Therefore, $\mathbf{\Lambda}_0 = \frac{1}{c+1}\sum_{i=1}^{c+1}(\boldsymbol{\mu}_i'-\boldsymbol{\mu}')(\boldsymbol{\mu}_i'-\boldsymbol{\mu}')^T$, with $\boldsymbol{\mu}' = \frac{1}{c+1}\sum_{i=1}^{c+1}\boldsymbol{\mu}_i'$. Letting $\mathbf{M} = [\boldsymbol{\mu}_1',...,\boldsymbol{\mu}_{c+1}']$ and $\mathbf{E}\in\mathbb{R}^{(c+1)\times(c+1)}$ with all entries equal to $\frac{1}{c+1}$, we have $\mathbf{\Lambda}_0 = \frac{1}{c+1}\mathbf{M}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\mathbf{M}^T$. Similarly, we have $\widehat{\mathbf{S}}_0 = \frac{1}{c+1}\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\widehat{\mathbf{M}}^T$, where $\widehat{\mathbf{M}} = [\widehat{\boldsymbol{\mu}}_1',...,\widehat{\boldsymbol{\mu}}_{c+1}']$ and $\widehat{\boldsymbol{\mu}}_1'$ is an estimate of $\boldsymbol{\mu}_1'$. As there are n training examples per class, we have $\widehat{\mathbf{M}} = \mathbf{M}+\mathbf{X}$, where the entries of $\mathbf{X}\in\mathbb{R}^{D\times(c+1)}$ are i.i.d. samples from $\mathcal{N}(0,1/n)$.

Note that the nonzero diagonal entries of $\mathbf{\Lambda}_0$ are $\boldsymbol{\lambda}_i$, $i=1,2,...,c$, which are actually eigenvalues of $\mathbf{\Lambda}_0$, associated with eigenvectors $\mathbf{e}_i$, $i=1,2,...,c$. Thus, $\mathbf{\Lambda}_0 = \frac{1}{c+1}\mathbf{M}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\mathbf{M}^T$ implies that $\mathbf{M}(\mathbf{I}-\mathbf{E})$ has singular values $\sqrt{(c+1)\boldsymbol{\lambda}_i}$, $i=1,2,...,c$ and left singular vectors $\mathbf{I}_{1:c} = [\mathbf{e}_1,...,\mathbf{e}_c]$. Denoting by $\mathbf{Q}\in\mathbb{R}^{(c+1)\times c}$ the right singular vectors of $\mathbf{M}(\mathbf{I}-\mathbf{E})$, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c$, we have

$$\mathbf{M}(\mathbf{I}-\mathbf{E})\mathbf{Q} = \left[\sqrt{(c+1)\boldsymbol{\lambda}_1}\mathbf{e}_1,...,\sqrt{(c+1)\boldsymbol{\lambda}_c}\mathbf{e}_c\right]. \tag{86}$$

Consequently, by $\widehat{\mathbf{M}} = \mathbf{M}+\mathbf{X}$, we have

$$\begin{aligned} &\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})\mathbf{Q} \\ =&\left[\sqrt{(c+1)\boldsymbol{\lambda}_1}\mathbf{e}_1,...,\sqrt{(c+1)\boldsymbol{\lambda}_c}\mathbf{e}_c\right]+\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q} \\ =&[\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_c], \end{aligned} \tag{87}$$

where

$$\boldsymbol{\xi}_i = \sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i+\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i,\ i=1,2,...,c. \tag{88}$$

Then, by $\widehat{\mathbf{S}}_0 = \frac{1}{c+1}\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\widehat{\mathbf{M}}^T$, we have for the first c eigenvectors $\mathbf{V}_{1:c}$ of $\widehat{\mathbf{S}}_0$ that

$$\begin{aligned} \mathbf{V}_{1:c} &= \mathcal{R}(\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})) = \mathcal{R}(\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})\mathbf{Q}) \\ &= \mathcal{R}([\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_c]). \end{aligned} \tag{89}$$

Accordingly,

$$\begin{aligned} &\|\mathbf{V}_{1:c}^T\mathbf{e}_i\|^2 \\ =&\|\mathcal{R}^T([\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_c])\mathbf{e}_i\|^2 \geq \|\mathcal{R}^T(\boldsymbol{\xi}_i)\mathbf{e}_i\|^2 = \frac{1}{\|\boldsymbol{\xi}_i\|^2}|\boldsymbol{\xi}_i^T\mathbf{e}_i|^2 \\ =&\frac{|\mathbf{e}_i^T\sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i+\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|^2}{\|\sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i+\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|^2} \\ \geq&\frac{(c+1)\boldsymbol{\lambda}_i+|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|^2-2\sqrt{(c+1)\boldsymbol{\lambda}_i}|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{(c+1)\boldsymbol{\lambda}_i+\|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|^2+2\sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i}. \end{aligned} \tag{90}$$

It can be verified that as $N=(c+1)n\longrightarrow\infty$

$$|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i| \leq \|\mathbf{e}_i^T\mathbf{X}\| = \sqrt{\sum_{j=1}^{c+1}\mathbf{X}_{ij}^2}\xrightarrow{a.s.}0, \tag{91}$$

where the inequality is due to $\|(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\| \leq \|(\mathbf{I}-\mathbf{E})\|\|\mathbf{Q}_i\| \leq 1$ and the limit is because $\mathbf{X}_{ij}$ follows the distribution $\mathcal{N}(0,\frac{1}{n})$.

In addition, by Proposition 2 and letting $\mathbf{G} = \sqrt{n}\mathbf{X}$, we have

$$\|\mathbf{X}\| = \frac{1}{\sqrt{n}}\|\mathbf{G}\|\xrightarrow{a.s.}\sqrt{\frac{D}{n}} = \sqrt{\frac{(c+1)D}{N}}\longrightarrow\sqrt{(c+1)\gamma}. \tag{92}$$

Thus,

$$\|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\| \leq \|\mathbf{X}\|\xrightarrow{a.s.}\sqrt{(c+1)\gamma}. \tag{93}$$

Combining (90), (91) and (93), we obtain

$$\lim_{D/N\longrightarrow\gamma}\|\mathbf{V}_{1:c}^T\mathbf{e}_i\|^2\geq\frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i+\gamma},\ a.s. \tag{94}$$

This completes the proof.

G. Proof of Corollary 1

Recall that

$$P=0.5\Phi\left\{-\frac{\widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_1-0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\right\}+0.5\Phi\left\{-\frac{0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)-\widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_2}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\right\}, \tag{95}$$

assumed $\widehat{\mathbf{w}}_1^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\geq 0$. First, we have

$$\begin{aligned}&-\frac{\widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_1-0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\\&=-0.5\frac{\widehat{\mathbf{w}}_1^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}+0.5\frac{\widehat{\mathbf{w}}_1^T((\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2))}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\\&=-\sqrt{\frac{\widehat{\mathbf{w}}_1^T\mathbf{S}\widehat{\mathbf{w}}_1}{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}+0.5\frac{\widehat{\mathbf{w}}_1^T((\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2))}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\\&=-\sqrt{\boldsymbol{\delta}_1\boldsymbol{\lambda}_1}+0.5T,\end{aligned} \tag{96}$$

and similarly

$$\begin{aligned}&-\frac{0.5\widehat{\mathbf{w}}_1^T(\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)-\widehat{\mathbf{w}}_1^T\boldsymbol{\mu}_2}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\\&=-0.5\frac{\widehat{\mathbf{w}}_1^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}-0.5\frac{\widehat{\mathbf{w}}_1^T((\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2)-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2))}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\\&=-\sqrt{\boldsymbol{\delta}_1\boldsymbol{\lambda}_1}-0.5T,\end{aligned} \tag{97}$$

As long as $T\xrightarrow{a.s.}0$, we have by Theorem 1 that

$$P=\Phi(-\sqrt{\boldsymbol{\delta}_1\boldsymbol{\lambda}_1})\leq\Phi(-\varrho\sqrt{\boldsymbol{\lambda}_1}) \tag{98}$$

with

$$\varrho=\max\big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)})+\arccos(\sqrt{1-\gamma})),0\big\}. \tag{99}$$

Below, we verify that it indeed holds

$$T=\frac{\widehat{\mathbf{w}}_1^T((\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)+(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2))}{\sqrt{\widehat{\mathbf{w}}_1^T\boldsymbol{\Sigma}\widehat{\mathbf{w}}_1}}\xrightarrow{a.s.}0. \tag{100}$$

By using similar strategy in the proof of Lemma 2, in particular (40), we have $\widehat{\mathbf{w}}_1=\mathbf{X}\mathbf{q}$, wherein $\mathbf{X}$ satisfies $\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}=\mathbf{I}$ and

$$\mathbf{q}=a\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2),\ \text{for some } a\neq 0, \tag{101}$$

since $\mathbf{X}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)$ is the first eigenvector of the normalized sample between-scatter matrix $\widehat{\mathbf{S}}_0=\mathbf{X}^T\widehat{\mathbf{S}}\mathbf{X}$. Substituting (101) into T, we have

$$T=\frac{(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T((\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)+(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2))}{\sqrt{(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)}}. \tag{102}$$

For the numerator, we have

$$\begin{aligned}&(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T((\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)+(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2))\\=&(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)^T\mathbf{X}\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)\\&-(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2)\\&+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T((\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)+(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2))\\=&T_1-T_2+T_3.\end{aligned} \tag{103}$$

Due to the normalization, we know that $\xi_1=\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)$ follows the multivariate Gaussian distribution $\mathcal{N}(0,\frac{1}{n}\mathbf{I})$, with $n=N/2$ being the training data number per class. Then, by Lemma 5 and $\|\xi_1\|^2\xrightarrow{a.s.}2\gamma$, we have

$$T_1=\xi_1^T\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi_1=\|\xi_1\|^2\frac{\xi_1^T}{\|\xi_1\|}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\frac{\xi_1}{\|\xi_1\|}\xrightarrow{a.s.}\frac{2\gamma}{1-\gamma}. \tag{104}$$

Similarly, letting $\xi_2=\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2)$, the same argument gives $T_2\xrightarrow{a.s.}\frac{2\gamma}{1-\gamma}$. Denoting $\xi_3=\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ and recalling Lemma 5, we have

$$\begin{aligned}\|\xi_3\|^2&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\&\xrightarrow{a.s.}\frac{\|\mathbf{X}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\|^2}{(1-\gamma)^3}<\infty.\end{aligned} \tag{105}$$

Then, since ξ follows $\mathcal{N}(0,\frac{1}{n}\mathbf{I})$ and ξ_3 has bounded entries due to (105), we have

$$\xi_3^T\xi_1\xrightarrow{a.s.}0. \tag{106}$$

Similarly, $\xi_3^T\xi_2\xrightarrow{a.s.}0$. Thus,

$$T_3=\xi_3^T(\xi_1+\xi_2)\xrightarrow{a.s.}0. \tag{107}$$

Therefore, we have the numerator $T_1-T_2+T_3\xrightarrow{a.s.}0$.

For the dominator, letting $\zeta=\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)$, we have

$$\begin{aligned}&\sqrt{(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)^T\mathbf{X}\mathbf{U}^T\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}\mathbf{X}^T(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2)}\\=&\|\zeta\|\sqrt{\frac{\zeta^T}{\|\zeta\|}\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\frac{\zeta}{\|\zeta\|}}\xrightarrow{a.s.}\frac{\lim\|\zeta\|}{(1-\gamma)^{3/2}}.\end{aligned} \tag{108}$$

Note that $\lim\|\zeta\|>0$, because $\widehat{\boldsymbol{\mu}}_1\neq\widehat{\boldsymbol{\mu}}_2$ almost surely. Thus, the dominator must be positive. Therefore, we have T in (100) has limit 0.

ACKNOWLEDGMENT

This study is supported by the Australian Research Council Projects DP-140102164 and ARC FT-130101457.

REFERENCES

[1] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals Eugen.*, vol. 7, pp. 179–188, 1936.

[2] C. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society series B: Methodological*, vol. 10, pp. 159–203, 1948.

[3] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001.

[4] D. Tao, X. Li, X. Wu, and S. Maybank, "Geometric mean for subspace selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 260–274, 2009.

[5] W. Bian and D. Tao, "Max-min distance analysis by using sequential sdp relaxation for dimension reduction." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 1037–1050, 2011.

[6] F. De la Torre and T. Kanade, "Multimodal oriented discriminant analysis," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 177–184.

[7] G. Potamianos and H. Graf, "Linear discriminant analysis for speechreading," in *Workshop on Multimedia Signal Process*, 1998, pp. 221– 226.

[8] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. Lopez-Ferreras, "Application of Fisher linear discriminant analysis to speech/music classification," in *The International Conference on Computer as a Tool*, 2005, pp. 1666–1669.

[9] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[10] T. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, 2005.

[11] J. Ye and Q. Li, "A two-stage linear discriminant analysis via qr-decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929–941, 2005.

[12] E. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[13] K. Kumar and S. Bhattacharya, "Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances," *Review of Accounting and Finance*, vol. 5, no. 3, pp. 216–227, August 2006.

[14] O. Hamsici and A. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 647–657, 2008.

[15] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[16] J. Fan, Y. Fan, and Y. Wu, "High-dimensional classification," in *High-dimensional Data Analysis*, T. Cai and X. Shen, Eds. New Jersey: World Scientific, 2011, pp. 3–37.

[17] T. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York, NY: Wiley, 1984.

[18] E. Wigner, "Characteristic vectors of bordered matrices with infinite dimensions," *The Annals of Mathematics*, vol. 62, no. 3, pp. 548–564, 1955.

[19] ——, "On the distribution of the roots of certain symmetric matrices," *The Annals of Mathematics*, vol. 67, no. 2, pp. 325–327, 1958.

[20] V. Marčenko and L. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, p. 457, 1967.

[21] A. Edelman and N. Rao, "Random matrix theory," *Acta Numerica*, vol. 14, no. 233-297, p. 139, 2005.

[22] A. Tulino and S. Verdú, *Random matrix theory and wireless communications*. Now Publishers Inc, 2004, vol. 1.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press, September 1990.

[24] R. J. Durrant and A. Kabán, "A bound on the performance of lda in randomly projected data spaces," in *International Conference on Pattern Recognition*, 2010, pp. 4044–4047.

[25] D. C. Hoyle, "Accuracy of pseudo-inverse covariance learning – a random matrix theory analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 33, no. 7, pp. 1470–1481, Jul. 2011.

[26] G. J. Mclachlan, *Discriminant Analysis and Statistical Pattern Recognition (Wiley Series in Probability and Statistics)*. Wiley-Interscience, Aug. 2004.

[27] R. J. Durrant and A. Kaban, "Compressed fisher linear discriminant analysis: classification of randomly projected data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 1119–1128.

[28] R. J. Durrant and A. Kabán, "Error bounds for kernel fisher linear discriminant in gaussian hilbert space," *Journal of Machine Learning Research - Proceedings Track*, vol. 22, pp. 337–345, 2012.

[29] C. Blake and C. Merz, "UCI repository of machine learning databases," Dept. of Information and Computer Sciences, University of California, Irvine, Tech. Rep., 1998.

[30] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, Massachusetts Institute of Technology, 1989.

[31] T. Tao, *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

[32] P. Billingsley, *Convergence of Probability Measures*, ser. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., 1999, vol. 175.

Wei Bian (M'14) received the B.Eng. degree in electronic engineering and the B.Sc degree in applied mathematics in 2005, the M.Eng. degree in electronic engineering in 2007, all from the Harbin Institute of Technology, China, and the PhD degree in computer science in 2012 from the University of Techonology, Sydney, Australia. His research interests are pattern recognition and machine learning.

Dacheng Tao (M'07-SM'12) is Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering & Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored 100+ scientific articles at top venues including IEEE T-PAMI, T-NNLS, T-IP, NIPS, ICML, AISTATS, ICDM, CVPR, ICCV, ECCV; ACM T-KDD, Multimedia and KDD, with the best theory/algorithm paper runner up award in IEEE ICDM07 and best student paper award in IEEE ICDM'13.