Journal of Machine Learning Research 21 (2020) 1-24 Submitted 5/19, Revised 12/19; Published 5/20

High-dimensional Linear Discriminant Analysis Classifier for Spiked Covariance Model [*]

Houssem Sifaou HOUSSEM.SIFAOU@KAUST.EDU.SA
Abla Kammoun ABLA.KAMMOUN@KAUST.EDU.SA
Mohamed-Slim Alouini SLIM.ALOUINI@KAUST.EDU.SA
Computer, Electrical and Mathematical Science and Engineering Division,
King Abdullah University of Science and Technology (KAUST), Thuwal, KSA

Editor: Rina Foygel Barber

Abstract

Linear discriminant analysis (LDA) is a popular classifier that is built on the assumption of common population covariance matrix across classes. The performance of LDA depends heavily on the quality of estimating the mean vectors and the population covariance matrix. This issue becomes more challenging in high-dimensional settings where the number of features is of the same order as the number of training samples. Several techniques for estimating the covariance matrix can be found in the literature. One of the most popular approaches are estimators based on using a regularized sample covariance matrix, giving the name regularized LDA (R-LDA) to the corresponding classifier. These estimators are known to be more resilient to the sample noise than the traditional sample covariance matrix estimator. However, the main challenge of the regularization approach is the choice of the optimal regularization parameter, as an arbitrary choice could lead to severe degradation of the classifier performance. In this work, we propose an improved LDA classifier based on the assumption that the covariance matrix follows a spiked covariance model. The main principle of our proposed technique is the design of a parametrized inverse covariance matrix estimator, the parameters of which are shown to be easily optimized. Numerical simulations, using both real and synthetic data, show that the proposed classifier yields better classification performance than the classical R-LDA while requiring lower computational complexity.

Keywords: Linear Discriminant Analysis, Spiked Covariance Models, High-Dimensional Data, Random Matrix Theory.

1. Introduction

Linear Discriminant Analysis (LDA) based classifiers, originally proposed by R. A. Fisher (Fisher, 1936), are among the simplest algorithms used for classification tasks. Grounded in the use of the maximum likelihood principle, LDA turns out to be the optimal classifier that achieves the lowest misclassification rate under the assumption that the data is Gaussian with perfectly known statistics, and common covariance matrix across classes (Hastie et al., 2001). However, in practice, the population covariance matrix and means associated with

[*] This work was funded by the office of sponsored research at KAUST. Part of the results in this paper has been presented at the IEEE International Workshop on Signal Processing Advances in Wireless Communications, 2018.

each class could not be perfectly known. It is common practice to replace them in the discriminant score of the LDA by the sample estimates computed based on the training data. This should not affect severely the performance if the number of training samples is sufficiently large compared to the number of features. However, in high-dimensional settings, the estimation of the covariance matrix and the means associated with each class is highly inaccurate, causing a severe degradation in the classification performance. The impact of the estimation noise has been discussed in (Fan and Fan, 2008) and a solution based on feature selection has been proposed. In some extreme situations in which the sample size is lower than the number of features, the use of the sample covariance matrix as a plug-in estimator is not allowed, as the discriminant score of the LDA involves computing the inverse of the covariance matrix.

One popular attempt to solve all these issues is to use the regularized sample covariance matrices as a plug-in estimator of the population covariance matrix. The resulting classifier is known as Regularized LDA (R-LDA) in reference to the regularization parameter in use. However, the choice of the regularization parameter is critical to achieving good performance. Several works propose to choose the regularization parameter as the optimal value that minimizes the misclassification rate, an approximation of which can be derived using recent results from random matrix theory (Zollanvari and Dougherty, 2015; Bakirov et al., 2016; Elkhalil et al., 2017a). However, this approach presents two major drawbacks. First, the estimation of the optimal regularization parameter is computationally expensive. Second, it does not take advantage of available prior information on the structure of the covariance matrix.

In this paper, we propose, a novel approach based on the assumption that the population covariance matrix is isotropic except for a finite number of symmetry breaking directions (Hoyle and Rattray, 2003; Reimann et al., 1996). Such a model, known as spiked-model covariance matrix, has been used in many real applications including detection (Zhao et al., 1986), EEG signals (Davidson, 2009; Fazli et al., 2011) and financial econometrics (Passemier et al., 2017; Kritchman and Nadler, 2008). Based on this model, we propose to consider a class of sample covariance matrix estimators that follow the same spiked model, that is they can be written as a finite rank perturbation of a scaled identity matrix. The directions of the low-rank perturbation are given by the principal eigenvectors of the sample covariance matrix, while its eigenvalues are some design parameters that are chosen so that an approximation of the misclassification rate rate is minimized. Such an approximation is computed based on results from random matrix theory.

Interestingly, the optimal parameters are obtained in closed form, avoiding the need for using the standard cross-validation approach. Compared to the classical R-LDA, the proposed classifier constitutes a novel improved LDA classifier, which we refer to as ”I-LDA”, that presents a lower complexity and a higher classification performance. The reduction in the computational cost is achieved since, unlike the R-LDA which requires the use of a grid search to optimize the regularization parameter (Hastie et al., 2001; Bishop, 2006; Zollanvari and Dougherty, 2015; Elkhalil et al., 2017a) the optimal hyper-parameters of our proposed classifier admit closed-form expressions that depend only on the training data. We also show that not only the proposed classifier outperforms R-LDA but also other popular classification techniques such as support vector machine (SVM) and k-nearest neighbors (KNN).

We also show that further improvement in the case of imbalanced classes can be obtained by optimizing the intercept of the proposed classifier. This improvement is shown to be significant in such cases, as shown by a set of numerical simulations. Moreover, extension of the multi-class classification is discussed in section 3.4.

The main literature in relation to this work is represented by the works of Donoho et al. in (Donoho and Ghorbani, 2018; Donoho et al., 2018) which consider the problem of covariance matrix estimation under the assumption of a spiked covariance matrix model. However, in contrast to our work, the aforementioned papers consider generic loss functions that are not directly related to the objective of interest which is herein the misclassification rate. Specifically, (Donoho and Ghorbani, 2018) considers the design of the optimal shrinkage that optimizes a condition number loss function while (Donoho et al., 2018) examines the optimization of 26 loss functions. As evidenced later through numerical simulations in section 4, as far as LDA classification is considered, the proposed classifier yields better performance. Under the setting of incomplete data, the problem of high-dimensional LDA has also been considered through the prism of optimality theory in (Cai and Zhang, 2019) where a classification algorithm based on an adaptive constrained ℓ_1 minimization approach has been proposed and compared with (Shao et al., 2011).

The rest of the paper is organized a follows. In the next section, we give a brief overview of LDA and R-LDA classifiers. In section 3, the steps of designing the proposed classifier are detailed. The performance of our technique is studied in section 4 using numerical simulations before concluding in section 5.

Throughout this work, boldface lower case is used for denoting column vectors, $\mathbf{x}$, and upper case for matrices, $\mathbf{X}$. $\mathbf{X}^T$ denotes the transpose. Moreover, $\mathbf{I}_p$ denotes the $p \times p$ identity matrix and $\|.\|$ is used to denote the ℓ_2-norm for vectors and spectral norm for matrices. The notation $\xrightarrow{a.s.}$ is used for the almost sure convergence of sequence of random variables. The trace of a matrix $\mathbf{A}$ is denoted by $\operatorname{tr}\mathbf{A}$ and $\operatorname{diag}(x_1, \cdots, x_n)$ stands for the diagonal matrix with diagonal entries $x_1, \cdots, x_n$. The notation $f(x) = \mathcal{O}(g(x))$ means that $|\frac{f(x)}{g(x)}|$ is bounded as $x \to \infty$.

2. Linear Discriminant Analysis

Consider a set of n vector observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ in $\mathbb{R}^{p\times 1}$ belonging to two classes $\mathcal{C}_0$ and $\mathcal{C}_1$. For $i \in \{0, 1\}$, we denote by n_i the number of observations in $\mathcal{C}_i$ and by $\mathcal{T}_i$ the set of their indexes. Moreover, all observations are assumed to be independent and drawn from a Gaussian mixture model in which both classes have different means but a common covariance matrix. In particular, for $i \in \{0, 1\}$,

$$\mathbf{x}_i \in \mathcal{C}_i \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are respectively the mean vector and the covariance matrix associated with class $\mathcal{C}_i$. We assume that the class labels associated with vector observations $\{\mathbf{x}_\ell\}_{\ell=1}^n$ are perfectly known. These observations constitute the training sample that is used to build a classifier, the aim of which is to predict the class label of an unseen observation $\mathbf{x}$.

For the reader convenience, we start by presenting the classical LDA classifier. Its corresponding discriminant score is (Hastie et al., 2001; Bishop, 2006; Zollanvari and Dougherty, 2015):

$$W^{\mathrm{LDA}}(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \log \frac{\pi_1}{\pi_0}, \tag{1}$$

where π_i is the prior probability corresponding to class i. The unseen observation $\mathbf{x}$ is assigned to class $\mathcal{C}_0$ if $W^{\mathrm{LDA}}(\mathbf{x}) > 0$ and to class $\mathcal{C}_1$ otherwise. In practice, the class statistics namely its mean vector and covariance matrix are unknown. They are usually estimated from training data using the empirical means $\overline{\mathbf{x}}_i$ and the pooled sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ defined as:

$$\hat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\ell \in \mathcal{T}_i} \mathbf{x}_\ell, \tag{2}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{(n_0 - 1)\hat{\boldsymbol{\Sigma}}_0 + (n_1 - 1)\hat{\boldsymbol{\Sigma}}_1}{n - 2}, \tag{3}$$

where

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \sum_{\ell \in \mathcal{T}_i} (\mathbf{x}_\ell - \overline{\mathbf{x}}_i)(\mathbf{x}_\ell - \overline{\mathbf{x}}_i)^T, \quad i = 0, 1.$$

with n_i is the number of observations belonging to class $\mathcal{C}_i$ and $\mathcal{T}_i$ is the set of indexes of observations belonging to class $\mathcal{C}_i$. Replacing $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ by their estimates yields the following LDA discriminant rule:

$$\hat{W}^{\mathrm{LDA}}(\mathbf{x}) = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) - \log \frac{\pi_1}{\pi_0}. \tag{4}$$

In the case where the size of the observations p is higher than the number of available training samples n, $\hat{\boldsymbol{\Sigma}}$ is singular. One popular solution consists in using ridge estimators of the inverse covariance matrix (Hastie et al., 2001; Zollanvari and Dougherty, 2015):

$$\mathbf{H} = \left(\mathbf{I}_p + \gamma \hat{\boldsymbol{\Sigma}}\right)^{-1}, \quad \gamma > 0. \tag{5}$$

as a plug-in estimator of the inverse of the covariance matrix. The corresponding discriminant score is known as regularized LDA (R-LDA) in reference to the regularization parameter γ and is given by:

$$\hat{W}^{\mathrm{R-LDA}}(\mathbf{x}) = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \mathbf{H}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) - \log \frac{\pi_1}{\pi_0}. \tag{6}$$

Conditioning on the training samples, the discriminant score $\hat{W}^{\mathrm{LDA}}(\mathbf{x})$ is Gaussian in $\mathbf{x}$. In light of this observation, the conditional misclassification rate associated with class $\mathcal{C}_i$ can be expressed as:

$$\epsilon_i^{\mathrm{LDA}} = \boldsymbol{\Phi}\left(\frac{(-1)^{i+1} G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\boldsymbol{\Sigma}}) + (-1)^i \log \frac{\pi_1}{\pi_0}}{\sqrt{D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma})}}\right), \tag{7}$$

where $\mathbf{\Phi}(.)$ is the cumulative distribution function of a standard normal random variable and

$$G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}}) = \left(\boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right) \hat{\mathbf{\Sigma}}^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1),$$
$$D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}) = (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{\Sigma} \hat{\mathbf{\Sigma}}^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1).$$

The total misclassification rate can be expressed as:

$$\epsilon^{\text{LDA}} = \pi_0 \epsilon_0^{\text{LDA}} + \pi_1 \epsilon_1^{\text{LDA}}. \tag{8}$$

Along the same arguments, the misclassification rate associated with R-LDA takes the same expression with $\hat{\mathbf{\Sigma}}^{-1}$ being replaced by $\mathbf{H}$ in (8). The resulting expression cannot provide insights into how the theoretical mean and covariance of each class impact the classification performances. Such information is critical to properly choose the regularization parameter.

One approach to get around this issue is to use asymptotic results from random matrix theory which lead to approximate the quantities $G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}})$ and $D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma})$ by deterministic equivalents that solely involve $\{\boldsymbol{\mu}_i\}_{i=1}^2$ and $\mathbf{\Sigma}$ (Zollanvari and Dougherty, 2015; Elkhalil et al., 2017b; Dobriban and Wager, 2018). An approximation of the classification performance is thus obtained by replacing $G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}})$ and $D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma})$ with their deterministic equivalents in (8). Such an approximation cannot still be directly used to optimize the regularization parameter since it depends on the unknown covariance matrix $\mathbf{\Sigma}$ and $\{\boldsymbol{\mu}_i\}_{i=0}^1$. The works in (Zollanvari and Dougherty, 2015; Elkhalil et al., 2017b) proposes consistent estimates of the misclassification rates, which were then optimized through a grid search.

This approach has been shown through simulations to outperform the classical cross-validation technique used often for the setting of the regularization parameter. It however presents two major drawbacks. First, the optimization of the regularization parameter involves a grid-search procedure which can lead to prohibitively high computational costs when high dimension settings are considered. Second, it uses the sample covariance matrix as a plug-in estimator of the covariance matrix. Such an estimator is no longer consistent when the number of samples is comparable with that of features, which leads to high estimation noises and in turn to low classification performances.

In this paper, we propose an improved LDA classifier that overcomes both drawbacks and that is particularly suitable in scenarios wherein $\mathbf{\Sigma}$ takes the following particular form (Hoyle and Rattray, 2003; Reimann et al., 1996):

$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p + \sigma^2 \sum_{j=1}^{r} \lambda_j \mathbf{v}_j \mathbf{v}_j^T, \tag{9}$$

where $\sigma^2 > 0$, $\lambda_1 \geq \cdots, \geq \lambda_r > 0$ and $\mathbf{v}_1, \cdots, \mathbf{v}_r$ are orthonormal. The above model is encountered in many real applications, among which detection (Zhao et al., 1986), EEG signals (Davidson, 2009; Fazli et al., 2011) and financial econometrics (Passemier et al., 2017; Kritchman and Nadler, 2008) are the best representatives. The design of the improved classifier will be detailed in the next section.

3. Improved LDA

3.1. Proposed estimation approach

In this section, we present our improved LDA classifier, which unlike the traditional R-LDA, leverages the particular finite rank perturbation property of the true covariance matrix. For the sake of simplicity, we assume that σ^2 and r are perfectly known. In practice, one can resort to the existing efficient algorithms available in the literature for the estimation of these parameters. We refer the reader to the following works and the references therein (Kritchman and Nadler, 2008; Johnstone and Lu, 2009; Ulfarsson and Solo, 2008; Passemier et al., 2017). In our numerical simulations, we have used the method of (Ulfarsson and Solo, 2008). Starting from the eigen decomposition of the pooled covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \sum_{j=1}^{p} s_j \mathbf{u}_j \mathbf{u}_j^T,$$

with s_j being the j-th largest eigenvalue and $\mathbf{u}_j$ its corresponding eigenvector, and noting that:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\left[\mathbf{I}_p - \sum_{j=1}^{r} \frac{\lambda_j}{1+\lambda_j}\mathbf{v}_j\mathbf{v}_j^T\right], \tag{10}$$

it is sensible to seek for estimators of $\boldsymbol{\Sigma}^{-1}$ that takes the form:

$$\hat{\mathbf{C}}^{-1} = \frac{1}{\sigma^2}\left[\mathbf{I}_p + \sum_{j=1}^{r} w_j \mathbf{u}_j\mathbf{u}_j^T\right], \tag{11}$$

where $\{w_j\}_{j=1}^r$ are some design parameters to be optimized. We assume that $w_j \in \mathcal{R} = [-1+\zeta, \chi)$ for some $0 < \zeta < 1$ and $\chi > 1$. From a practical point of view, we only need to assume that $w_j > -1$ to ensure that $\hat{\mathbf{C}}^{-1}$ is positive definite. However, the restriction to the range $\mathcal{R}$ is needed later for the proof of the uniform convergence results. Plugging (11) in place of $\hat{\boldsymbol{\Sigma}}^{-1}$ into (1), yields the following discriminant score:

$$\hat{W}^{\text{I-LDA}} = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \hat{\mathbf{C}}^{-1}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) - \log\frac{\pi_1}{\pi_0}. \tag{12}$$

It is worth mentioning that the estimator $\hat{\mathbf{C}}^{-1}$ can be seen as an instance of the wider class of estimators taking the form $\sum_{j=1}^{p} \eta_j \mathbf{u}_j\mathbf{u}_j^T$, where $\mathbf{u}_j$ being the eigenvector of the sample covariance matrix associated with its j-th largest eigenvalue and scalars η_j are referred to as shrinkage functions that need to be designed (Daniels and Kass, 2001; Ledoit and Wolf, 2004; Karoui, 2018; Chen et al., 2010; Ledoit and Wolf, 2017). Indeed, $\hat{\mathbf{C}}^{-1}$ is obtained by setting $\eta_{r+1} = \ldots = \eta_p = \frac{1}{\sigma^2}$ and $w_j = \sigma^2\eta_j - 1$. The main challenge is how to select the appropriate values of the design parameters w_j.

3.2. Parameter optimization

Given the application into consideration, it is sensible to select the w_j so that they minimize the total misclassification rate:

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}}\ \epsilon^{\text{I-LDA}}(\mathbf{w}), \tag{13}$$

where $\mathbf{w} = [w_1, \cdots, w_r]^T$ and $\epsilon^{\mathrm{I-LDA}}(\mathbf{w})$ is obtained by replacing $\hat{\mathbf{\Sigma}}^{-1}$ by $\hat{\mathbf{C}}^{-1}$ in (8).

Finding the optimal $\mathbf{w}$ that exactly solves (13) could not be in general obtained. To get around this problem, we invoke results from random matrix theory that approximate the total misclassification rate under the asymptotic growth regime defined in the following assumption.

Assumption 1 *Throughout this work, we assume that*

(i) $n, p \longrightarrow \infty$, with fixed ratio $c = p/n$.

(ii) $n_0, n_1, \longrightarrow \infty$, with $p/n_0 \triangleq c_0$, $p/n_1 \triangleq c_1$ where c_0 and c_1 are fixed constants.

(iii) r is fixed and $\lambda_1 > \cdots > \lambda_r > \sqrt{c}$, independently of p and n.

(iv) The mean difference vector $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ has a bounded Euclidean norm, that is $\|\boldsymbol{\mu}\| = \mathcal{O}(1)$.

(v) The spectral norm of $\mathbf{\Sigma}$ is bounded, that is $\|\mathbf{\Sigma}\| = \mathcal{O}(1)$.

Remark 1

- *Assumptions (i), (ii) and (iii) are key assumptions that are generally used in the framework of the theory of large random matrices.*
- *Assumption (iii) is fundamental in our analysis since it guarantees, as per standard results from random matrix theory, the one-to-one mapping between the sample eigenvalues s_j and the unknown λ_j. In fact, when $\lambda_j > \sqrt{c}$, λ_j can be consistently estimated using its corresponding s_j as we will see later. In the case where $\lambda_j \leq \sqrt{c}$, the relation between s_j and λ_j no longer holds and λ_j cannot be estimated (Couillet and Debbah, 2011; Baik et al., 2005).*
- *Assumption (iv) controls the order of the euclidean norm of $\boldsymbol{\mu}$ that ensures asymptotically non-trivial classification.*

It should be noted that from a methodological perspective, the approach undertaken in this work that consists in optimizing a loss function for a spiked covariance model is similar to the one introduced earlier in (Donoho et al., 2018). It basically consists in computing an asymptotic approximation for the considered loss function based on the asymptotics of the eigenvalues and eigenvectors of the spiked model characterized in (Baik et al., 2005) and (Paul, 2007) and then optimizing the parameters that optimize this approximation. However, the main difference with the approach of (Donoho et al., 2018) is that the considered loss function is relevant to the underlying application. In doing so, better performances are expected compared to the approach that considers generic loss functions. Depending on the scenario into consideration, different loss functions can be considered, including the Sharp Ratio in the context of Markowitz portfolio allocation or the SNR in the context the design of MVDR beamforming, as examined in (Yang et al., 2018).

Theorem 1 *Under the settings of Assumption 1, we have*

$$G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}) - \frac{1}{2}\left[\frac{(-1)^i\|\boldsymbol{\mu}\|^2}{\sigma^2}\overline{G}(\mathbf{w}) - \frac{p}{n_0} + \frac{p}{n_1}\right] \xrightarrow{a.s.} 0, \tag{14}$$

$$D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma}) - \left[\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\overline{D}(\mathbf{w}) + \frac{p}{n_0} + \frac{p}{n_1}\right] \xrightarrow{a.s.} 0, \tag{15}$$

where

$$\overline{G}(\mathbf{w}) = 1 + \sum_{j=1}^{r} a_j b_j w_j,$$
$$\overline{D}(\mathbf{w}) = 1 + \sum_{j=1}^{r} \left[\lambda_j b_j + 2a_j b_j(\lambda_j + 1)w_j\right] + \sum_{j=1}^{r} \left[a_j b_j (1 + \lambda_j a_j) w_j^2\right],$$

with

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1, \quad a_j = \frac{1 - c/\lambda_j^2}{1 + c/\lambda_j}, \quad b_j = \frac{\boldsymbol{\mu}^T \mathbf{v}_j \mathbf{v}_j^T \boldsymbol{\mu}}{\|\boldsymbol{\mu}\|^2}, \quad j = 1, \cdots, r \tag{16}$$

Proof See Appendix A. ■

Using these deterministic equivalents and after some manipulations, a deterministic equivalent of the global misclassification rate can be obtained as:

$$\epsilon^{\mathrm{I-LDA}}(\mathbf{w}) - \bar{\epsilon}^{\mathrm{I-LDA}}(\mathbf{w}) \xrightarrow{a.s.} 0.$$

where

$$\bar{\epsilon}^{\mathrm{I-LDA}}(\mathbf{w}) = \pi_0 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}(\overline{G}(\mathbf{w}) - \eta)}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}}\right] + \pi_1 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}(\overline{G}(\mathbf{w}) + \eta)}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}}\right] \tag{17}$$

with $\alpha = \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$, $\eta = \frac{1}{\alpha}[c/\pi_0 - c/\pi_1 + 2\log\frac{\pi_1}{\pi_0}]$ and $\kappa = \frac{1}{\alpha}[p/n_0 + p/n_1]$. In order to ensure that the this convergence result is guaranteed at optimality, we need to establish the uniform convergence of $\epsilon^{\mathrm{I-LDA}}(\mathbf{w})$. This is the objective of the following theorem.

Theorem 2 *Under the settings of Assumption 1, we have*

$$\sup_{\mathbf{w} \in \mathcal{R}^r} |\epsilon^{\mathrm{I-LDA}}(\mathbf{w}) - \bar{\epsilon}^{\mathrm{I-LDA}}(\mathbf{w})| \xrightarrow{a.s.} 0.$$

where $\bar{\epsilon}^{\mathrm{I-LDA}}(\mathbf{w})$ *is given in* (17).

Proof See Appendix D ■

Now, with the asymptotic equivalent of the misclassification rate on hand, we can determine the optimal parameters $w_j^\star$.

Theorem 3 *Under the settings of Assumption 1, the optimal parameters* $\{w_j^\star\}$ *that minimize* $\bar{\epsilon}^{\mathrm{I-LDA}}(\mathbf{w})$ *are given by:*

$$w_j^\star = \frac{u^\star}{\beta}\frac{\gamma_j}{\alpha_j} - \beta_j, \quad j = 1, \cdots, r \tag{18}$$

where

$$\alpha_j = \lambda_j a_j^2 b_j + a_j b_j, \quad \beta_j = \frac{\lambda_j + 1}{\lambda_j a_j + 1}, \quad \gamma_j = a_j b_j, \quad j = 1, ..., r,$$

$\beta = \sqrt{\sum_{j=1}^r \gamma_j^2 / \alpha_j}$, *and* $u^\star$ *is the minimizer of the scalar function* $\tilde{g}(u)$ *given by*

$$\tilde{g}(u) = \pi_0 \mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\beta u + d_0}{\sqrt{u^2 + b}}\right) + \pi_1 \mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\beta u + d_1}{\sqrt{u^2 + b}}\right),$$

with

$$b = 1 + \kappa + \sum_{j=1}^r \left[\lambda_j b_j - \frac{(\lambda_j a_j b_j + a_j b_j)^2}{\lambda_j a_j^2 b_j + a_j b_j}\right], \tag{19}$$

$$d_0 = 1 - \eta - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}, \tag{20}$$

$$d_1 = 1 + \eta - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}. \tag{21}$$

Proof See Appendix B. ■

> **Remark 2** *In the case of equiprobable classes,* $u^\star$ *can be computed in closed form expression as* $u^\star = \frac{\beta b}{d}$, *where* $d = \frac{d_0 + d_1}{2} = 1 - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}$.

3.3. Improved LDA with optimal intercept

Bias correction is a general procedure that aims to optimize the bias in the discriminant score of a given classifier to minimize the misclassification rate. It has been considered in several previous works; see for instance (Chan and Peter, 2009) and (Huang et al., 2010) and the references therein, and has been recently applied to LDA and R-LDA in (Cheng and Binyan, 2018). Generally, bias correction allows to bring a gain in performance especially in high-dimensional settings and imbalanced classes. This is because the original bias in LDA, introduced essentially to calibrate the case of unequal sample sizes across classes, is devised under the assumption that the sample size is sufficiently high. There is thus room for further improvement to adapt this bias to high-dimensional settings. Following this line of ideas, we apply the bias correction procedure to the improved LDA classifier and show that not only the optimization of the parameters get simplified but also a significant gain is obtained once the bias is optimally set. The resulting classifier will be named as "OII-LDA" in reference to optimal-intercept-improved-LDA. Starting off from our proposed classifier classifier and replacing the constant term with a parameter θ, the score function can be written as:

$$\hat{W}^{\text{OII-LDA}} = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \hat{\mathbf{C}}^{-1}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) + \theta,$$

where θ is a parameter that will be optimized. The corresponding misclassification rate can be expressed in this case as,

$$\epsilon^{\mathrm{OII-LDA}} = \pi_0 \mathbf{\Phi}\left[\frac{-G(\boldsymbol{\mu}_0, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}) - \theta}{\sqrt{D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma})}}\right] + \pi_1 \mathbf{\Phi}\left[\frac{G(\boldsymbol{\mu}_1, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}) + \theta}{\sqrt{D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma})}}\right].$$

Following similar steps as in the previous section, the asymptotic equivalent of $\epsilon^{\mathrm{OII-LDA}}$ can be obtained as for $\epsilon^{\mathrm{I-LDA}}$:

$$\epsilon^{\mathrm{OII-LDA}} - \overline{\epsilon}^{\mathrm{OII-LDA}} \xrightarrow{a.s.} 0.$$

where

$$\overline{\epsilon}^{\mathrm{OII-LDA}} = \pi_0 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}\left[\overline{G}(\mathbf{w}) - \frac{1}{\alpha}\left(\frac{p}{n_0} - \frac{p}{n_1} - 2\theta\right)\right]}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}}\right] + \pi_1 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}\left[\overline{G}(\mathbf{w}) + \frac{1}{\alpha}\left(\frac{p}{n_0} - \frac{p}{n_1} - 2\theta\right)\right]}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}}\right]$$

Taking the derivative of $\overline{\epsilon}^{\mathrm{OII-LDA}}$ and equating it to zero yields the optimal θ:

$$\theta^{\star} = \frac{1}{2}\left[\frac{p}{n_0} - \frac{p}{n_1}\right] - \frac{\overline{D}(\mathbf{w}) + \kappa}{\overline{G}(\mathbf{w})} \log \frac{\pi_1}{\pi_0}$$

Replacing $\theta^{\star}$ by its expression, the asymptotic misclassification rate becomes

$$\overline{\epsilon}^{\mathrm{OII-LDA}} = \pi_0 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}\overline{G}(\mathbf{w})}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}} + \frac{\sqrt{\overline{D}(\mathbf{w}) + \kappa}}{\sqrt{\alpha}\overline{G}(\mathbf{w})} \log \frac{\pi_1}{\pi_0}\right] + \pi_1 \mathbf{\Phi}\left[\frac{-\sqrt{\alpha}\overline{G}(\mathbf{w})}{2\sqrt{\overline{D}(\mathbf{w}) + \kappa}} - \frac{\sqrt{\overline{D}(\mathbf{w}) + \kappa}}{\sqrt{\alpha}\overline{G}(\mathbf{w})} \log \frac{\pi_1}{\pi_0}\right] \tag{22}$$

Now, we are able to find the new optimal parameter vector $\mathbf{w}$ that minimizes the asymptotic misclassification rate $\overline{\epsilon}^{\mathrm{OII-LDA}}$.

Theorem 4 *The optimal parameter vector* $\mathbf{w}$ *that minimizes* $\overline{\epsilon}^{\mathrm{OII-LDA}}$ *is given by*

$$w_j^{\star} = \frac{b}{d}\frac{\gamma_j}{\alpha_j} - \beta_j, \quad j = 1, \cdots, r. \tag{23}$$

where b, α_j, β_j *and* γ_j *are defined in Theorem 3 and* d *is given by:*

$$d = 1 - \sum_{j=1}^{r} \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}. \tag{24}$$

Proof See Appendix C. ■

Remark 3 *Interestingly, the optimal parameters of OII-LDA are obtained in closed-form even in the case of imbalanced classes. Thus, in practice OII-LDA should be used instead I-LDA proposed in the previous section, since it yields better performance and the optimization of its parameters are less computationally demanding.*

The optimal design parameters in Theorems 3 and 4 could not be directly used in practice, since they depend on the unobservable quantities λ_j and b_j. To solve this issue, consistent estimators for these quantities need to be retrieved. This is the objective of the following result:

Proposition 5 *Under the settings of Assumption 1, we have*

$$|\alpha - \hat{\alpha}| \xrightarrow{a.s.} 0, \qquad |\lambda_j - \hat{\lambda}_j| \xrightarrow{a.s.} 0, \qquad \text{and} \qquad |b_j - \hat{b}_j| \xrightarrow{a.s.} 0,$$

where

$$\begin{aligned}
\hat{\alpha} &= \frac{\|\hat{\boldsymbol{\mu}}\|^2}{\sigma^2} - c_1 - c_0 \\
\hat{\lambda}_j &= \frac{s_j/\sigma^2 + 1 - c + \sqrt{(s_j/\sigma^2 + 1 - c)^2 - 4s_j/\sigma^2}}{2} - 1, \\
\hat{b}_j &= \frac{1 + c/\hat{\lambda}_j}{1 - c/\hat{\lambda}_j^2} \frac{\hat{\boldsymbol{\mu}}^T \mathbf{u}_j \mathbf{u}_j^T \hat{\boldsymbol{\mu}}}{\|\hat{\boldsymbol{\mu}}\|^2 - c_1\sigma^2 - c_0\sigma^2}, \quad j = 1, \cdots, r.
\end{aligned}$$

with $c_0 = \frac{p}{n0}$, $c_1 = \frac{p}{n1}$, $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1$, *and* s_j *is the* j*-th largest eigenvalue of the pooled covariance matrix.*

Proof The proof is a direct application of results from (Couillet and Debbah, 2011, Theorem 9.1 and Theorem 9.9) and it is thus omitted. ■

Replacing α, λ_j and b_j by their estimates yields a consistent estimator $\hat{w}_j^\star$ of $w_j^\star$. Even though we have assumed that the data follow a Gaussian distribution, our results hold for non-Gaussian data under some moment assumptions on the entries. This is due to the fact that all the results regarding the covariance spiked model were proven for non-Gaussian data in (Benaych-Georges and Nadakuditi, 2011). Moreover, the optimal parameters of our proposed classifier OII-LDA, are independent of the specific distribution, that is, replacing $\boldsymbol{\Phi}$ with other cumulative distribution function will not affect the expressions of the optimal parameters in (23).

3.4. Extension to multi-class classification

The proposed binary classifier can be easily extended to the case of multi-class classification. In the machine learning literature, this can be performed by combining the use of multiple binary classifiers to devise a multi-class classifier. Two popular methods have been proposed towards this aim. These are commonly known as One vs. Rest and One vs. One approaches Assume that the total number of classes is C. The one vs the rest approach consists in training one binary classifier for each class by considering the rest of the classes as forming one class. Then, the output of these C binary classifiers are combined to decide the class of the unseen observation $\mathbf{x}$. On the other hand, the one vs one consists in training $C(C-1)/2$ binary classifiers corresponding to each possible pair of classes. Then, the output of these binary classifiers are combined to predict the class of the unseen observation $\mathbf{x}$. In our case and this also holds for R-LDA, the one vs the rest approach is not applicable since the class formed by the remaining observations from the $C-1$ classes is not homogeneous in the sense that it is formed by observations drawn from a mixture of Gaussian distribution rather than from a Gaussian distribution as required by our results. However, the second approach (one vs. one) is applicable in our case and should allow extending our binary classifier to the multi-class classification case. As already known, the one vs one approach may lead to ambiguities (Bishop, 2006) as some classes may receive the same number of votes. This issue can be solved by giving priority to the decision with the highest scores.

4. Numerical Simulations

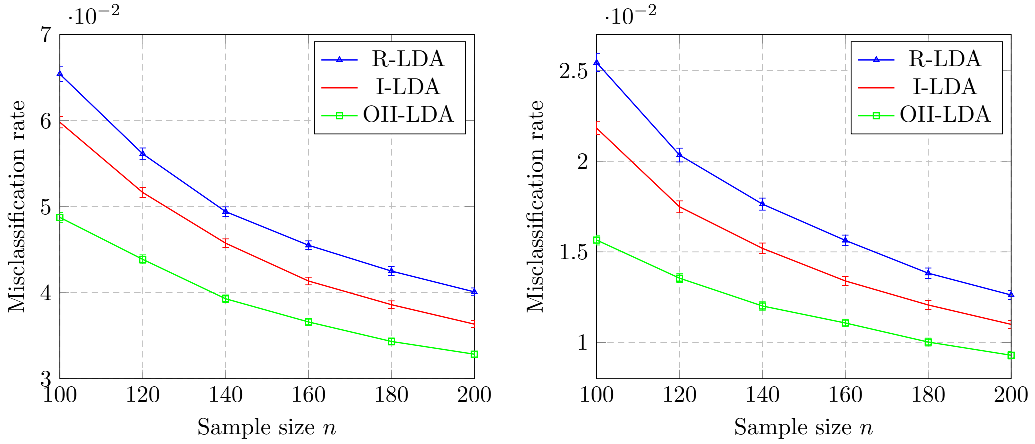
In this section, we study the performance of the proposed improved LDA classifier. We compare its performance with R-LDA classifier and other classical classifiers based on both synthetic and real data.

4.1. Synthetic data

In the synthetic data simulations, we use the following Monte Carlo protocol to estimate the true misclassification rate:

- Step 1: Set $\sigma^2 = 1$ and choose $r = 3$ orthogonal symmetry breaking directions as follows : $\mathbf{v}_1 = [1, 0, \cdots, 0]^T$, $\mathbf{v}_2 = [0, 1, 0, \cdots, 0]^T$, $\mathbf{v}_3 = [0, 0, 1, 0 \cdots, 0]^T$ and their corresponding weights $\lambda_1 = 8$, $\lambda_2 = 7$, $\lambda_3 = 6$. Set $\boldsymbol{\mu}_0 = \frac{1}{\sqrt{p}}[a, a, \cdots, a]^T$ and $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_0$ where a is a finite constant. We choose $a = 2$ and $a = 2.5$.
- Step 2: Generate n_i training samples for class i.
- Step 3: Using the training set, design the improved LDA classifier as explained in section 3 and determine the optimal parameter γ^* of R-LDA using grid search over $\gamma \in \{10^{i/10}, i = -10 : 1 : 10\}$.
- Step 4: Estimate the misclassification rate of both classifiers using a set of 1000 testing samples.
- Step 5: Repeat Step 2–4, 500 times and determine the average classification true error of both classifiers. In all figures, 95% confidence intervals are plotted along with the Monte Carlo estimate of the misclassification rate.

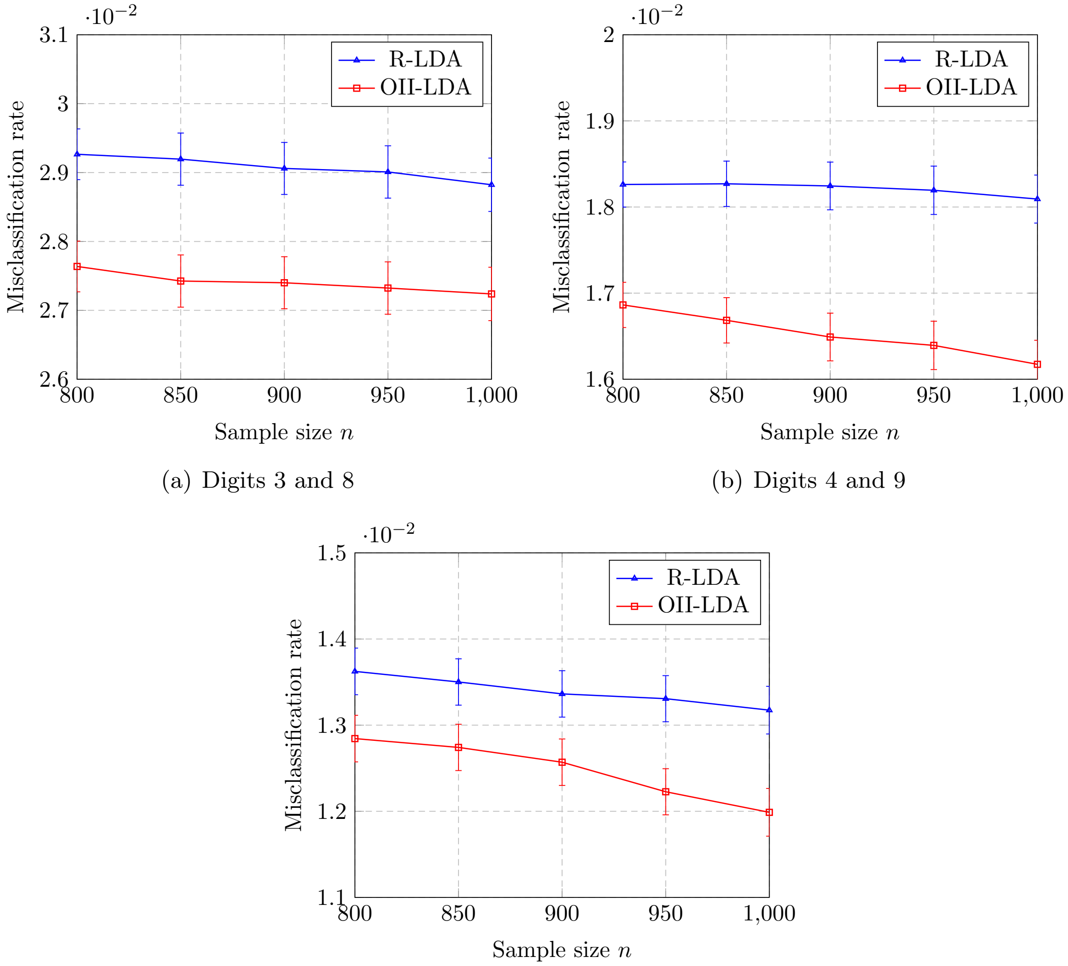
In Fig. 1, we plot the misclassification rate vs. training sample size n when $p = 150$ and $\pi_0 = 0.2$ for the proposed improved LDA and the classical R-LDA using synthetic data. It is observed that the improved LDA outperforms the classical R-LDA and the gap between the two schemes is significant.



(a) $a = 2$

(b) $a = 2.5$

Figure 1: Misclassification rate vs. sample size n for $p = 150$ and $\pi_0 = 0.2$. Comparison between Improved LDA and R-LDA with synthetic data.



(a) Digits 3 and 8

(b) Digits 4 and 9

(c) Digits 8 and 9

Figure 2: Misclassification rate vs. sample size n. Comparison between Improved LDA and R-LDA using USPS dataset.

4.2. Real data

For real data simulation, we use two datasets. The first one is the ”USPS” dataset which is one of the standard datasets for handwritten digit recognition. The dataset is publicly available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`. For this dataset, the number of features (observation size) is $p = 256$ and the total number of samples for all digits including test samples is 9298. The second dataset is the ”Phoneme” dataset which is composed of 4509 speech frames corresponding to five phonemes. Each frame is represented by log-periodogram of length 256. We used in our simulations the phonemes ’aa’ and ’ao’, which are the most confusing phonemes, as class 0 and class 1.

We use the following protocol for the real dataset:

- Step 1: Let q_0 be the ratio between the total number of samples in class $\mathcal{C}_0$ to the total number of samples available in the full dataset. Denote by n_{Full} the total number of samples in the full dataset. Choose $n < n_{\text{Full}}$ the number of training samples; set $n_0 = \lfloor q_0 n \rfloor$, where $\lfloor . \rfloor$ is the floor function and $n_1 = n - n_0$. Take n_i training samples belonging to class $\mathcal{C}_i$ randomly from the full dataset. The remaining samples are used as a test dataset in order to estimate the misclassification rate.

- Step 2: Using the training dataset, design the improved LDA classifier as explained in section 3 and determine the optimal parameter γ^* of R-LDA using grid search over $\gamma \in \{10^{i/10}, i = -10 : 1 : 10\}$
- Step 3: Using the test dataset, estimate the true misclassification rate for both classifiers.
- Step 4: Repeat steps 1–4 500 times and determine the average misclassification rate of both classifiers.

In Fig. 2, the misclassification rate vs. training sample size is plotted for the R-LDA and our proposed classifier. Three pairs of the most confusing digits are chosen for simulation; (3,8), (4,9) and (8,9). As seen, the performance of the proposed classifier is better than R-LDA classifier and the gain is significant. For example, more that 10% gain in terms of misclassification rate is obtained for digits (4,9) when $n = 1000$.

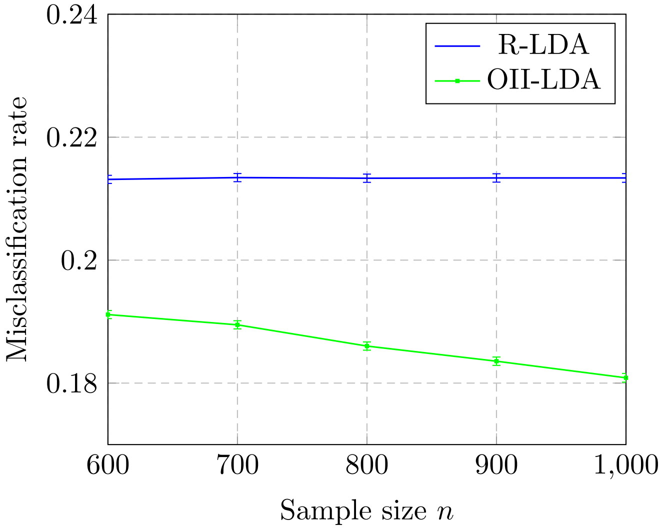


Figure 3: Misclassification rate vs. sample size n. Comparison between Improved LDA and R-LDA using Phoneme dataset.

In Fig. 3, we compare the misclassification rate of our proposed classifier and the classical R-LDA. The most confusing phonemes 'aa and 'ao' are chosen for binary classification. The proposed classifier exhibits significantly lower misclassification rate.

In the last experiment, we compare the proposed classifier with classical classifiers SVM and KNN. For SVM, we used the linear and the polynomial kernels with other parameters fixed. We also compare with the SVM with all the hyper-parameters optimized. The performance of LDA with the covariance matrix estimator proposed in (Donoho et al., 2018) is also assessed in Table 1. The resulting classifier is denoted by DG-LDA for notational convenience. In (Donoho et al., 2018), 26 loss functions have been used, we report here the one with the lowest misclassification rate. Using the phonemes 'aa and 'ao' again, we report in Table 1 the misclassification rate for different values of the number of training samples. As observed, the proposed classifier yields the best performance

Table 1: Comparison between the proposed classifier, R-LDA, DG-LDA, SVM, and KNN, using Phoneme dataset.

	$n = 800$	$n = 900$	$n = 1000$
OII-LDA	**0.177**	**0.175**	**0.174**
R-LDA	0.213	0.213	0.212
DG-LDA	0.186	0.185	0.184
SVM (linear kernel)	0.206	0.204	0.201
SVM (All hyper-parameters optimized)	0.179	0.178	0.176
SVM (polynomial kernel)	0.255	0.254	0.250
KNN ($k = 1$)	0.226	0.225	0.224
KNN ($k = 5$)	0.229	0.228	0.227

while presenting the lowest computationally complexity among all classifiers, the complexity of SVM and KNN being shown to be higher than LDA (Li et al., 2006).

5. Conclusion

This work presents an improved LDA classifier to perform binary classification of observations drawn from Gaussian distribution. The population covariance matrix is assumed to be common for both classes and to follow a spiked model. Leveraging this particular structure, the proposed classifier uses an estimate of the covariance matrix that follows a parametrized spiked model, the parameters of which corresponds to its largest eigenvalues. Using standard results from random matrix theory, asymptotic characterization of the misclassification rate is provided, and the parameters are selected so that a consistent estimate of the misclassification rate is minimized. Interestingly, we show that the proposed classifier provides better performance while presenting a much lower complexity as compared to other state-of-the-art classification techniques. As a further extension, the same ideas underlying the proposed classification method can be extended to other classification methods, including for instance, quadratic discriminant analysis classifier or other spiked models in which the population covariance matrix is a low-rank perturbation of a diagonal matrix not necessarily equal to identity. The same approach could also be applied to Sharpe ratio maximization in the context of Markowitz's portfolio allocation.

Appendix A. Proof of Theorem 1

For notational convenience, we define $\mathbf{X}_i \in \mathbb{R}^{p\times n_i}$ the matrix of training data associated with class $\mathcal{C}_i$. Hence $\mathbf{X}_i$ can be written as $\mathbf{X}_i = \mathbf{Y}_i + \boldsymbol{\mu}_i \mathbf{1}_{n_i}^T$ such that the vectors of $\mathbf{Y}_i$ are independent Gaussian vectors with zero mean and covariance $\boldsymbol{\Sigma}$. Using these notations, it can be easily shown that the covariance matrix associated with class $\mathcal{C}_i$ writes as

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\left(\mathbf{Y}_i\mathbf{Y}_i^T - \mathbf{Y}_i\frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T}{n_i}\mathbf{Y}_i^T\right).$$

Let $\frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T}{n_i}=\mathbf{O}_i\mathbf{E}_i\mathbf{O}_i^T$, the eigenvalue decomposition of $\frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T}{n_i}$, with $\mathbf{E}_i=\operatorname{diag}\left([1,\mathbf{0}_{(n_i-1)\times 1}]\right)$ and $\mathbf{O}_i\in\mathbb{R}^{n_i\times n_i}$ orthogonal matrix whose first column is $\frac{1}{\sqrt{n_i}}\mathbf{1}_{n_i}$. Define $\tilde{\mathbf{Y}}_i=\mathbf{Y}_i\mathbf{O}_i$. Thus, we have

$$\hat{\mathbf{\Sigma}}_i=\frac{1}{n_i-1}\left(\mathbf{Y}_i\mathbf{O}_i\mathbf{O}_i^T\mathbf{Y}_i^T-\mathbf{Y}_i\mathbf{O}_i\mathbf{E}_i\mathbf{O}_i^T\mathbf{Y}_i^T\right),=\frac{1}{n_i-1}\left(\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i^T-\tilde{\mathbf{y}}_{i,1}\tilde{\mathbf{y}}_{i,1}^T\right),$$

where $\tilde{\mathbf{y}}_{i,1}$ is the first column of $\tilde{\mathbf{Y}}_i$. Let $\overline{\mathbf{Y}}_i$ be $\tilde{\mathbf{Y}}_i$ with the first column removed. Due to the invariance of Gaussian distribution under orthogonal transformation, the columns of $\tilde{\mathbf{Y}}_i$ follow a Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. The pooled covariance matrix given by

$$\hat{\mathbf{\Sigma}}=\frac{1}{n-2}\left((n_0-1)\overline{\mathbf{Y}}_0\overline{\mathbf{Y}}_0^T+(n_1-1)\overline{\mathbf{Y}}_1\overline{\mathbf{Y}}_1\right), \tag{25}$$

follow a standard spiked model for which standard results from random matrix theory apply. In this respect, we have the following results from (Couillet and Debbah, 2011)

$$\begin{aligned}&\mathbf{v}_j^T\mathbf{u}_k\mathbf{u}_k^T\mathbf{v}_j-a_j\delta_{j,k}\xrightarrow{a.s.}0,\quad k=1,\ldots,r\\&\frac{1}{\|\boldsymbol{\mu}\|^2}\boldsymbol{\mu}^T\mathbf{u}_j\mathbf{u}_j^T\boldsymbol{\mu}-a_jb_j\xrightarrow{a.s.}0,\quad j=1,\cdots,r\end{aligned} \tag{26}$$

where $\delta_{j,k}$ denotes the Kronecker delta and a_j and b_j are given in (16). With these results at hand, we are now ready to find deterministic equivalents for $G(\boldsymbol{\mu}_i,\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\hat{\mathbf{C}})$ and $D(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\hat{\mathbf{C}},\mathbf{\Sigma})$. We start by the term $G(\boldsymbol{\mu}_i,\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\hat{\mathbf{C}})$ which can be expressed as

$$\begin{aligned}G(\boldsymbol{\mu}_i,\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\hat{\mathbf{C}})&=\left(\frac{(-1)^i}{2}\boldsymbol{\mu}^T-\frac{1}{2n_0}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T-\frac{1}{2n_1}\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\right)\hat{\mathbf{C}}^{-1}\left(\frac{1}{n_0}\mathbf{Y}_0\mathbf{1}_{n_0}-\frac{1}{n_1}\mathbf{Y}_1\mathbf{1}_{n_1}+\boldsymbol{\mu}\right)\\&=\frac{(-1)^i}{2}\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\boldsymbol{\mu}-\frac{1}{2n_0}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\hat{\mathbf{C}}^{-1}\boldsymbol{\mu}-\frac{1}{2n_1}\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\hat{\mathbf{C}}^{-1}\boldsymbol{\mu}\\&+\frac{(-1)^i}{2n_0}\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_0\mathbf{1}_{n_0}-\frac{(-1)^i}{2n_1}\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_1\mathbf{1}_{n_1}-\frac{1}{2n_0^2}\mathbf{1}_{n_0}^T\mathbf{Y}_0\hat{\mathbf{C}}^{-1}\mathbf{Y}_0\mathbf{1}_{n_0}\\&+\frac{1}{2n_1^2}\mathbf{1}_{n_1}^T\mathbf{Y}_1\hat{\mathbf{C}}^{-1}\mathbf{Y}_1\mathbf{1}_{n_1}+\frac{1}{2n_0n_1}\left[\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_1\mathbf{1}_{n_1}-\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_0\mathbf{1}_{n_0}\right]\end{aligned}$$

From (25), it follows that $\frac{1}{\sqrt{n_i}}\mathbf{Y}_i\mathbf{1}_{n_i}=\tilde{\mathbf{y}}_{i,1}$ is independent of $\hat{\mathbf{\Sigma}}$. Since $\hat{\mathbf{C}}^{-1}$ is fully constructed from the eigenvectors of $\hat{\mathbf{\Sigma}}$, $\frac{1}{\sqrt{n_i}}\mathbf{Y}_i\mathbf{1}_{n_i}$ is also independent of $\hat{\mathbf{C}}^{-1}$. This yields the following convergences

$$\begin{aligned}&\frac{1}{n_i}\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_i\mathbf{1}_{n_i}\xrightarrow{a.s.}0,\\&\frac{1}{n_0n_1}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_1\mathbf{1}_{n_1}\xrightarrow{a.s.}0.\end{aligned}$$

Using (26), it follows that

$$\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\boldsymbol{\mu}-\frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\left(1+\sum_{j=1}^r w_ja_jb_j\right)\xrightarrow{a.s.}0.$$

It remains to deal with the term $\frac{1}{n_i^2}\mathbf{1}_{n_i}^T\mathbf{Y}_i^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_i\mathbf{1}_{n_i}=\frac{1}{n_i}\tilde{\mathbf{y}}_{i,1}^T\hat{\mathbf{C}}^{-1}\tilde{\mathbf{y}}_{i,1}$. Using the independence of $\tilde{\mathbf{y}}_{i,1}$ and $\hat{\mathbf{C}}^{-1}$ and applying the trace Lemma (Couillet and Debbah, 2011, Theorem 3.7), we have

$$\frac{1}{n_i^2}\mathbf{1}_{n_i}^T\mathbf{Y}_i^T\hat{\mathbf{C}}^{-1}\mathbf{Y}_i\mathbf{1}_{n_i}-\frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\xrightarrow{a.s.}0.$$

Finally, since $\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}$ is equal to identity plus a low rank perturbation matrix, we have that

$$\frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1} = \frac{p}{n_i} + o(1).$$

Putting all the above results together yields the result in (14). Let us now deal with the term $D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma})$. Using the notations defined in this proof, this term can be rewritten as:

$$D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma}) = \left(\frac{1}{n_0}\mathbf{Y}_0\mathbf{1}_{n_0} - \frac{1}{n_1}\mathbf{Y}_1\mathbf{1}_{n_1} + \boldsymbol{\mu}\right)^T \hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\left(\frac{1}{n_0}\mathbf{Y}_0\mathbf{1}_{n_0} - \frac{1}{n_1}\mathbf{Y}_1\mathbf{1}_{n_1} + \boldsymbol{\mu}\right). \quad (27)$$

Again due to the independence between $\frac{1}{\sqrt{n_i}}\mathbf{Y}_i\mathbf{1}_{n_i}$ and $\hat{\mathbf{C}}^{-1}$, the cross-products in (27) will converge to zero almost surely. Hence,

$$\begin{aligned} D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \mathbf{\Sigma}) = &\frac{1}{n_0^2}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{Y}_0\mathbf{1}_{n_0} + \frac{1}{n_1^2}\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{Y}_1\mathbf{1}_{n_1} \\ &+ \boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\boldsymbol{\mu} + o(1). \end{aligned}$$

Replacing $\mathbf{\Sigma}$ and $\mathbf{C}^{-1}$ by their expressions and using the results in (26), it can be easily shown that

$$\boldsymbol{\mu}^T\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\boldsymbol{\mu} - \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\left(1 + \sum_{j=1}^r \lambda_j b_j + 2\sum_{j=1}^r a_j b_j[\lambda_j + 1]w_j + \sum_{j=1}^r a_j b_j[1 + \lambda_j a_j]w_j^2\right) \xrightarrow{a.s.} 0$$

Moreover, using the independence of $\tilde{\mathbf{y}}_{i,1}$ and $\hat{\mathbf{C}}^{-1}$ and applying the trace Lemma (Couillet and Debbah, 2011, Theorem 3.7), we have

$$\frac{1}{n_i^2}\mathbf{1}_{n_i}^T\mathbf{Y}_i^T\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{Y}_i\mathbf{1}_{n_i} - \frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1} \xrightarrow{a.s.} 0.$$

Finally, $\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}$ is identity matrix plus a low rank matrix. Thus, we have $\frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1}\mathbf{\Sigma}\hat{\mathbf{C}}^{-1} = \frac{p}{n_i} + o(1)$. Putting these results together yields the convergence result in (15).

Appendix B. Proof of Theorem 3

The objective function is given by

$$f(\mathbf{w}) = \pi_0\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}f_0(\mathbf{w})\right) + \pi_1\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}f_1(\mathbf{w})\right),$$

where

$$f_0(\mathbf{w}) = \frac{\overline{G}(\mathbf{w}) - \eta}{\sqrt{\overline{D}(\mathbf{w}) + \kappa}} \quad \text{and} \quad f_1(\mathbf{w}) = \frac{\overline{G}(\mathbf{w}) + \eta}{\sqrt{\overline{D}(\mathbf{w}) + \kappa}}.$$

The numerator of $f_0(\mathbf{w})$ can be rewritten as

$$\sum_{j=1}^r\left[a_j b_j\left(w_j + \frac{\lambda_j + 1}{\lambda_j a_j + 1}\right) - \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}\right] + 1 - \eta.$$

And the square of the denominator of $f_0(\mathbf{w})$ can be expressed as

$$1 + \kappa + \sum_{j=1}^r\left[(\lambda_j a_j^2 b_j + a_j b_j)\left(w_j + \frac{\lambda_j + 1}{\lambda_j a_j + 1}\right)^2 + \lambda_j b_j - \frac{(\lambda_j a_j b_j + a_j b_j)^2}{\lambda_j a_j^2 b_j + a_j b_j}\right].$$

Thus, $f_0(\mathbf{w})$ can be rewritten as

$$f_0(\mathbf{w}) = \frac{\sum_{j=1}^r \gamma_j(w_j+\beta_j)+d_0}{\sqrt{\sum_{j=1}^r \alpha_j(w_j+\beta_j)^2+b}},$$

where

$$\alpha_j = \lambda_j a_j^2 b_j + a_j b_j, \quad \beta_j = \frac{\lambda_j+1}{\lambda_j a_j+1}, \quad \gamma_j = a_j b_j, \quad j=1,...,r$$

$$b = 1+\kappa+\sum_{j=1}^r\left[\lambda_j b_j - \frac{(\lambda_j a_j b_j + a_j b_j)^2}{\lambda_j a_j^2 b_j + a_j b_j}\right],$$

$$d_0 = 1-\eta-\sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j+1}.$$

Then, we have

$$f_0(\mathbf{w}) = \frac{\mathbf{c}^T\tilde{\mathbf{w}}+d_0}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}}+b}},$$

where the elements of $\tilde{\mathbf{w}}$ are $\tilde{w}_j = w_j+\beta_j$, $\mathbf{c} = [\gamma_1,\cdots,\gamma_r]^T$ and $\mathbf{D} = \text{diag}(\alpha_1,\cdots,\alpha_r)$. Similarly, it can be shown that

$$f_1(\mathbf{w}) = \frac{\mathbf{c}^T\tilde{\mathbf{w}}+d_1}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}}+b}},$$

where

$$d_1 = 1+\eta-\sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j+1}.$$

Thus, the objective function can be rewritten as

$$f(\tilde{\mathbf{w}}) = \pi_0\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\mathbf{c}^T\tilde{\mathbf{w}}+d_0}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}}+b}}\right)+\pi_1\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\mathbf{c}^T\tilde{\mathbf{w}}+d_1}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}}+b}}\right).$$

Letting $u = \|\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}\|$ and $\bar{\mathbf{w}} = \frac{\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}}{\|\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}\|}$, we have

$$\min_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}}) = \min_{\substack{(\bar{\mathbf{w}},u)\\ \|\bar{\mathbf{w}}\|=1,u>0}} g(\bar{\mathbf{w}},u) = \min_{\substack{u\\u>0}}\min_{\substack{\bar{\mathbf{w}}\\ \|\bar{\mathbf{w}}\|=1}} g(\bar{\mathbf{w}},u),$$

where

$$g(\bar{\mathbf{w}},u) = \pi_0\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\mathbf{c}^T\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{w}}u+d_0}{\sqrt{u^2+b}}\right)+\pi_1\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\mathbf{c}^T\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{w}}u+d_1}{\sqrt{u^2+b}}\right).$$

Since $u>0$ and $\mathbf{\Phi}(.)$ is an increasing function, $\bar{\mathbf{w}}^\star$ that minimizes $g(\bar{\mathbf{w}},u)$ subject to $\|\bar{\mathbf{w}}\|=1$ is the minimizer of $-\mathbf{c}^T\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{w}}$ subject to $\|\bar{\mathbf{w}}\|=1$. Thus, $\bar{\mathbf{w}}^\star = \frac{1}{\sqrt{\mathbf{c}^T\mathbf{D}^{-1}\mathbf{c}}}\mathbf{D}^{-\frac{1}{2}}\mathbf{c}$. Replacing $\bar{\mathbf{w}}^\star$ in $g(\bar{\mathbf{w}},u)$ yields,

$$\tilde{g}(u) = \pi_0\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\beta u+d_0}{\sqrt{u^2+b}}\right)+\pi_1\mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}\frac{\beta u+d_1}{\sqrt{u^2+b}}\right),$$

where $\beta = \sqrt{\mathbf{c}^T\mathbf{D}^{-1}\mathbf{c}} = \sqrt{\sum_{j=1}^r \gamma_j^2/\alpha_j}$. Finally, computing the minimizer $u^\star$ of the function $\tilde{g}(u)$ yields the optimal parameters vector $\tilde{\mathbf{w}}^\star = u^\star\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{w}}^\star = \frac{u^\star}{\sqrt{\mathbf{c}^T\mathbf{D}^{-1}\mathbf{c}}}\mathbf{D}^{-1}\mathbf{c}$.

Appendix C. Proof of Theorem 4

Letting $R = \frac{\overline{G}(\mathbf{w})}{\sqrt{\overline{D}(\mathbf{w})+\kappa}}$, we can write

$$\overline{\epsilon} = \pi_0 \mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}R + \frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right) + \pi_1 \mathbf{\Phi}\left(-\frac{\sqrt{\alpha}}{2}R - \frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right), \tag{28}$$

We will assume without loss of generality that $\pi_1 > \pi_0$. First, we will prove that $\overline{\epsilon}$ is a strictly decreasing function of R for $R \in]0, +\infty[$. Taking the derivative of $\overline{\epsilon}$ with respect to R, we get

$$\begin{aligned}\frac{d\overline{\epsilon}}{dR} = \frac{1}{\sqrt{2\pi}}\Bigg[&\pi_0\left(-\frac{\sqrt{\alpha}}{2} - \frac{1}{\sqrt{\alpha}R^2}\log\frac{\pi_1}{\pi_0}\right)e^{-\frac{\left[-\frac{\sqrt{\alpha}}{2}R+\frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right]^2}{2}}. \\ &+\pi_1\left(-\frac{\sqrt{\alpha}}{2} + \frac{1}{\sqrt{\alpha}R^2}\log\frac{\pi_1}{\pi_0}\right)e^{-\frac{\left[\frac{\sqrt{\alpha}}{2}R+\frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right]^2}{2}}\Bigg]\end{aligned}$$

Multiplying both sides by $\frac{1}{\pi_0}e^{-\frac{\left[-\frac{\sqrt{\alpha}}{2}R+\frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right]^2}{2}}$, and after simple simplifications, we get

$$\begin{aligned}&\frac{1}{\pi_0}e^{-\frac{\left[-\frac{\sqrt{\alpha}}{2}R+\frac{1}{\sqrt{\alpha}R}\log\frac{\pi_1}{\pi_0}\right]^2}{2}}\frac{d\overline{\epsilon}}{dR} \\ &= \frac{1}{\sqrt{2\pi}}\left[-\frac{\sqrt{\alpha}}{2} - \frac{1}{\sqrt{\alpha}R^2}\log\frac{\pi_1}{\pi_0} + \frac{\pi_1}{\pi_0}\left(-\frac{\sqrt{\alpha}}{2} + \frac{1}{\sqrt{\alpha}R^2}\log\frac{\pi_1}{\pi_0}\right)e^{-\log\frac{\pi_1}{\pi_0}}\right] = -\frac{\sqrt{\alpha}}{\sqrt{2\pi}}\end{aligned}$$

Thus, $\frac{d\overline{\epsilon}}{dR} < 0$ for all $R \in]0, +\infty[$ and consequently $\overline{\epsilon}$ is a strictly decreasing function of R for $R \in]0, +\infty[$. On the other hand, using the notations of Appendix B, we have

$$R = \frac{\mathbf{c}^T\tilde{\mathbf{w}} + d}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}} + b}},$$

where

$$d = 1 - \sum_{j=1}^{r}\frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}.$$

Obviously, $0 < R \leq R_1$ with $R_1 = \max_{\tilde{\mathbf{w}}}\frac{\mathbf{c}^T\tilde{\mathbf{w}}+d}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}}+b}}$. Since $\overline{\epsilon}$ is a strictly decreasing function of R, the optimal R is R_1. It remains now to solve the following optimization problem

$$\max_{\tilde{\mathbf{w}}}\frac{\mathbf{c}^T\tilde{\mathbf{w}} + d}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}} + b}}.$$

Letting $u = \|\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}\|$ and $\bar{\mathbf{w}} = \frac{\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}}{\|\mathbf{D}^{\frac{1}{2}}\tilde{\mathbf{w}}\|}$, we can write

$$\max_{\tilde{\mathbf{w}}}\frac{\mathbf{c}^T\tilde{\mathbf{w}} + d}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}} + b}} = \max_{u}\max_{\|\bar{\mathbf{w}}\|=1}\frac{u\mathbf{c}^T\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{w}} + d}{\sqrt{u^2 + b}}.$$

Clearly, $\bar{\mathbf{w}}^\star = \frac{1}{\sqrt{\mathbf{c}^T\mathbf{D}^{-1}\mathbf{c}}}\mathbf{D}^{-\frac{1}{2}}\mathbf{c}$. Consequently,

$$\max_{\tilde{\mathbf{w}}}\frac{\mathbf{c}^T\tilde{\mathbf{w}} + d}{\sqrt{\tilde{\mathbf{w}}^T\mathbf{D}\tilde{\mathbf{w}} + b}} = \max_{u}\frac{\beta u + d}{\sqrt{u^2 + b}},$$

with $\beta = \sqrt{\mathbf{c}^T\mathbf{D}^{-1}\mathbf{c}}$. Taking the derivative with respect to u and noting that $d > 0$, one can simply obtain $u^\star = \frac{\beta b}{d}$. Thus, $\tilde{\mathbf{w}}^\star = \frac{b}{d}\mathbf{D}^{-1}\mathbf{c}$. Returning back to $\mathbf{w}$ yields the result of Theorem 4.

Appendix D. Proof of Theorem 2

We will establish here the uniform convergence in $\mathbf{w} \in \mathcal{R}^r$ of $\epsilon^{\mathrm{I-LDA}}(\mathbf{w})$ given by

$$\epsilon^{\mathrm{I-LDA}}(\mathbf{w}) = \sum_{i=1}^{2} \pi_i \mathbf{\Phi}\left(\frac{(-1)^{i+1} G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}^{-1}) + (-1)^i \log\frac{\pi_1}{\pi_0}}{\sqrt{D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}^{-1}, \mathbf{\Sigma})}}\right),$$

to its deterministic equivalent given in (17). For simplicity, we will prove the result in the case where the classes are equiprobable and $n_0 = n_1$. The generalization to the case of imbalanced classes would be easy. In the case of equiprobable classes, the misclassification rate can be written as

$$\epsilon^{\mathrm{I-LDA}}(\mathbf{w}) = \frac{1}{2}\mathbf{\Phi}\left(-\sqrt{\frac{N_0(\mathbf{w})}{M(\mathbf{w})}}\right) + \frac{1}{2}\mathbf{\Phi}\left(\sqrt{\frac{N_1(\mathbf{w})}{M(\mathbf{w})}}\right),$$

where for notational convenience, we define

$$N_i(\mathbf{w}) = [G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}(\mathbf{w}))]^2 = \left[\left(\boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)\hat{\mathbf{C}}^{-1}(\mathbf{w})(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)\right]^2, \tag{29}$$

$$M(\mathbf{w}) = D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}(\mathbf{w}), \mathbf{\Sigma}) = (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)^T \hat{\mathbf{C}}^{-1}(\mathbf{w}) \mathbf{\Sigma} \hat{\mathbf{C}}^{-1}(\mathbf{w})(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1), \tag{30}$$

where the dependence of $\hat{\mathbf{C}}$ on $\mathbf{w}$ is made explicit. Since by assumption $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = \mathcal{O}(1)$, then there exist positive constants η_0, η_1 and η_2 such that, almost surely,

$$\eta_0 \leq \|\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1\|^2 \leq \eta_1 \quad \text{and} \quad \left\|\boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right\|^2 \leq \eta_2.$$

Moreover, since the uniform convergence is preserved by composition with continuous function, it suffices to prove the uniform convergence of

$$\vartheta_i(\mathbf{w}) = \frac{N_i(\mathbf{w})}{M(\mathbf{w})},$$

to its deterministic equivalent given by

$$\overline{\vartheta}(\mathbf{w}) = \frac{\overline{N}(\mathbf{w})}{\overline{M}(\mathbf{w})},$$

where $\overline{N}(\mathbf{w}) = \frac{\alpha}{2}\overline{G}(\mathbf{w})$ and $\overline{M}(\mathbf{w}) = \alpha\overline{D}(\mathbf{w}) + 1$ with α, $\overline{G}(\mathbf{w})$ and $\overline{D}(\mathbf{w})$ are defined in Theorem 1. Explicitly, we need to establish that for all $\delta > 0$ there exists K such that

$$\sup_{\mathbf{w} \in \mathcal{R}^r} |\vartheta(\mathbf{w}) - \overline{\vartheta}(\mathbf{w})| < K\delta, \tag{31}$$

for large n almost surely. Since $\mathcal{R}$ is bounded, for any $\delta > 0$, we can always construct a lattice of $\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta \in \mathcal{R}^r$ with J finite, such that for each $\mathbf{w} \in \mathcal{R}^r$, there exists $\mathbf{w}' \in \{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\}$ verifying $\max_{i \in \{1, \cdots, r\}} |w_i - w_i'| < \delta$. Thus, for such $\mathbf{w}'$, we can write

$$\begin{aligned}
&\sup_{\mathbf{w} \in \mathcal{R}^r} |\vartheta_i(\mathbf{w}) - \overline{\vartheta}(\mathbf{w})| && (32)\\
&\leq \sup_{\mathbf{w} \in \mathcal{R}^r} \left[|\vartheta_i(\mathbf{w}) - \vartheta_i(\mathbf{w}')| + |\overline{\vartheta}(\mathbf{w}') - \overline{\vartheta}(\mathbf{w})| + |\vartheta_i(\mathbf{w}') - \overline{\vartheta}(\mathbf{w}')|\right] \\
&\leq \sup_{\mathbf{w} \in \mathcal{R}^r} |\vartheta_i(\mathbf{w}) - \vartheta_i(\mathbf{w}')| + \sup_{\mathbf{w} \in \mathcal{R}^r} |\overline{\vartheta}(\mathbf{w}') - \overline{\vartheta}(\mathbf{w})| + \max_{\mathbf{w}'' \in \{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\}} |\vartheta_i(\mathbf{w}'') - \overline{\vartheta}(\mathbf{w}'')|. && (33)
\end{aligned}$$

Let us begin by the first term, we have

$$|\vartheta_i(\mathbf{w}) - \vartheta_i(\mathbf{w}')| = \left| \frac{M(\mathbf{w}')[N_i(\mathbf{w}) - N_i(\mathbf{w}')] + N_i(\mathbf{w}')[M(\mathbf{w}') - M(\mathbf{w})]}{M(\mathbf{w})M(\mathbf{w}')} \right|.$$

Using the properties of the spectral norm, we have

$$\begin{aligned}|N_i(\mathbf{w}) - N_i(\mathbf{w}')| &\leq \left\| \boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2} \right\|^2 \|\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1\|^2 \left\| \hat{\mathbf{C}}^{-1}(\mathbf{w}) - \hat{\mathbf{C}}^{-1}(\mathbf{w}') \right\| \left\| \hat{\mathbf{C}}^{-1}(\mathbf{w}) + \hat{\mathbf{C}}^{-1}(\mathbf{w}') \right\| \\ &\leq \eta_1 \eta_2 \frac{1}{\sigma^2} \max_{j \in \{1, \cdots, r\}} |w_j - w_j'| \left(2 + \max_{j \in \{1, \cdots, r\}} |w_j + w_j'| \right) \\ &< h_1 \delta,\end{aligned}$$

where $h_1 = \frac{2}{\sigma^2}\eta_1\eta_2(1+\chi)$. The last inequality is obtained by recalling that $w_j \in [-1+\zeta, \chi)$ with $q > 1$. Similarly, it can be shown that

$$|M(\mathbf{w}) - M(\mathbf{w}')| < h_2 \delta,$$

where $h_2 = 2\eta_1(\lambda_1 + 1)(1+\chi)$. Thus,

$$|\vartheta_i(\mathbf{w}) - \vartheta_i(\mathbf{w}')| < h\delta,$$

with

$$h = \frac{M(\mathbf{w})h_1 + N_i(\mathbf{w}')h_2}{M(\mathbf{w})M(\mathbf{w}')}.$$

We have to prove now that h is bounded. To this end, we begin by noting that

$$\lambda_{\min}\left[\hat{\mathbf{C}}^{-1}(\mathbf{w})\boldsymbol{\Sigma}\hat{\mathbf{C}}^{-1}(\mathbf{w})\right] \leq \frac{M(\mathbf{w})}{\|\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1\|^2} \leq \lambda_{\max}\left[\hat{\mathbf{C}}^{-1}(\mathbf{w})\boldsymbol{\Sigma}\hat{\mathbf{C}}^{-1}(\mathbf{w})\right]$$

where $\lambda_{\min}\left[\hat{\mathbf{C}}^{-1}(\mathbf{w})\boldsymbol{\Sigma}\hat{\mathbf{C}}^{-1}(\mathbf{w})\right] \geq \frac{\zeta}{\sigma^2}$ and $\lambda_{\max}\left[\hat{\mathbf{C}}^{-1}(\mathbf{w})\boldsymbol{\Sigma}\hat{\mathbf{C}}^{-1}(\mathbf{w})\right] \leq \frac{1}{\sigma^2}(1+\lambda_1)(1+\chi)^2$ The same inequalities hold for $M(\mathbf{w}')$. As for $N_i(\mathbf{w}')$, it can be bounded as

$$|N_i(\mathbf{w}')| \leq \eta_1\eta_2(1+\chi)^2\frac{1}{\sigma^2}.$$

Thus, $h \leq \frac{\sigma^2(1+\lambda_1)(1+q)^2 h_2 + \sigma^2\eta_1\eta_2(1+q)^2 h_1}{\eta_0^2} \triangleq K_1$. With this, we have bounded the first term in (33) as

$$\sup_{\mathbf{w} \in \mathcal{R}^r} |\vartheta_i(\mathbf{w}) - \vartheta_i(\mathbf{w}')| < K_1\delta. \tag{34}$$

We focus now on bounding the second term in (33), we start by rewriting $|\overline{\vartheta}(\mathbf{w}) - \overline{\vartheta}(\mathbf{w}')|$ as

$$|\overline{\vartheta}(\mathbf{w}) - \overline{\vartheta}(\mathbf{w}')| = \left| \frac{\overline{M}(\mathbf{w}')[\overline{N}(\mathbf{w}) - \overline{N}(\mathbf{w}')] + \overline{N}(\mathbf{w}')[\overline{M}(\mathbf{w}') - \overline{M}(\mathbf{w})]}{\overline{M}(\mathbf{w})\overline{M}(\mathbf{w}')} \right|.$$

Now, we have

$$\begin{aligned}|\overline{M}(\mathbf{w}) - \overline{M}(\mathbf{w}')| &= \alpha \left| \sum_{j=1}^{r} 2a_j b_j(\lambda_j + 1)(w_j - w_j') + \sum_{j=1}^{r} a_j b_j (a_j\lambda_j + 1)(w_j^2 - w_j'^2) \right| \\ &\leq \alpha \max_{j \in \{1, \cdots, r\}} |w_j - w_j'| \left(\sum_{j=1}^{r} 2a_j b_j (\lambda_j + 1) \right) \\ &\quad + \alpha \max_{j \in \{1, \cdots, r\}} |w_j^2 - w_j'^2| \left(\sum_{j=1}^{r} a_j b_j (a_j\lambda_j + 1) \right) \\ &< h_3\delta,\end{aligned}$$

where

$$h_3 = \alpha \left(\sum_{j=1}^{r} 2a_j b_j (\lambda_j + 1) \right) + 2\alpha\chi \left(\sum_{j=1}^{r} a_j b_j (a_j \lambda_j + 1) \right).$$

Replacing a_j and b_j by their expressions, it is easy to see that $\sum_{j=1}^{r} 2a_j b_j (\lambda_j + 1)$ and $\sum_{j=1}^{r} a_j b_j (a_j \lambda_j + 1)$ are positive and finite. Similarly, we have

$$|\overline{N}(\mathbf{w}) - \overline{N}(\mathbf{w}')| < h_4 \delta,$$

with $h_4 = \frac{\alpha}{2} \sum_{j=1}^{r} a_j b_j$. Hence,

$$|\overline{\vartheta}(\mathbf{w}) - \overline{\vartheta}(\mathbf{w}')| < h_5 \delta,$$

where

$$h_5 = \frac{\overline{M}(\mathbf{w}') h_4 + \overline{N}(\mathbf{w}') h_3}{\overline{M}(\mathbf{w}) \overline{M}(\mathbf{w}')}.$$

It remains to show that h_5 is bounded. This can be achieved by noting that

$$\overline{N}(\mathbf{w}') < 1 + \chi \sum_{j=1}^{r} a_j b_j \triangleq h_6,$$

and

$$1 \leq \overline{M}(\mathbf{w}) \leq 2 + \sum_{j=1}^{r} \lambda_j b_j + \chi \sum_{j=1}^{r} 2a_j b_j (\lambda_j + 1) \triangleq h_7,$$

where the inequality $1 \leq \overline{M}(\mathbf{w})$ is obtained by checking that $\overline{D}(\mathbf{w}) \geq 0$ for all $\mathbf{w}$. Again here, replacing a_j and b_j by their expression, one can easily show that h_6 and h_7 are finite. Thus, h_5 is bounded as

$$h_5 < h_7 h_4 + h_6 h_3 \triangleq K_2.$$

With this, we have

$$\sup_{\mathbf{w} \in \mathcal{R}^r} |\overline{\vartheta}(\mathbf{w}') - \overline{\vartheta}(\mathbf{w})| < K_2 \delta. \tag{35}$$

It remains to bound the last term in (33). Since we have established that $|\vartheta_i(\mathbf{w}_k) - \overline{\vartheta}(\mathbf{w}_k)| \xrightarrow{a.s.} 0$, for all $\mathbf{w} \in \mathcal{R}^r$ including all $\mathbf{w}_k^\delta$ in the lattice $\{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\}$. Therefore, for each $\mathbf{w}_k^\delta \in \{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\}$, there exists N_k such that for all $n > N_k$ (and $p = cn$),

$$|\vartheta_i(\mathbf{w}_k^\delta) - \overline{\vartheta}(\mathbf{w}_k^\delta)| < \delta.$$

Letting $N = \max(N_1, \cdots, N_J)$, we have for $n > N$,

$$|\vartheta_i(\mathbf{w}'') - \overline{\vartheta}(\mathbf{w}'')| < \delta, \ \forall \ \ \mathbf{w}'' \in \{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\},$$

which implies that for sufficiently large n,

$$\max_{\mathbf{w}'' \in \{\mathbf{w}_1^\delta, \cdots, \mathbf{w}_J^\delta\}} |\vartheta_i(\mathbf{w}'') - \overline{\vartheta}(\mathbf{w}'')| < \delta. \tag{36}$$

Combining (34), (35) and (36) yields the desired result in (31).

References

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, Sept. 2005.

D. Bakirov, A. P. James, and A. Zollanvari. An efficient method to estimate the optimum regularization parameter in RLDA. *Bioinformatics*, 32(22):3461–3468, 2016.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

T Tony Cai and Linjun Zhang. High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

Y.-B Chan and H. Peter. Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96(2):469–478, 04 2009.

Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, Oct. 2010.

W. Cheng and J. Binyan. On the dimension effect of regularized linear discriminant analysis. *Electronic Journal of Statistics*, 12(2):2709–2742, 2018.

R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications.* U.K., Cambridge: Cambridge Univ. Press, 2011.

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

D. J. Davidson. Functional mixed-effect models for electrophysiological responses. *Neurophysiology*, 41(1):71–79, Feb 2009.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

David L Donoho and Behrooz Ghorbani. Optimal covariance estimation for condition number loss in the spiked model. *arXiv preprint arXiv:1810.07403*, 2018.

David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018.

K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M. S. Alouini. Asymptotic performance of regularized quadratic discriminant analysis based classifiers. In *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sept. 2017a.

Khalil Elkhalil, Abla Kammoun, Romain Couillet, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini. Asymptotic performance of regularized quadratic discriminant analysis based classifiers. In *27th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2017, Tokyo, Japan, September 25-28, 2017*, pages 1–6, 2017b.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

S. Fazli, M. Danöczy, J. Schelldorfer, and K.-R. Müller. l1-penalized linear mixed-effects models for high dimensional data with application to bci. *NeuroImage*, 56(4):2100 – 2108, 2011.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7(2):179–188, 1936.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

D. C. Hoyle and M. Rattray. Limiting form of the sample covariance eigenspectrum in PCA and kernel PCA. *Proceedings of Neural Information Processing Systems*, 2003.

S. Huang, T. Tong, and H. Zhao. Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics*, 66(4), 2010.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Stat. Assoc.*, 104(486):682–693, 2009.

N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.*, 36(6):2757–2790, 2018.

S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19 – 32, 2008.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411, 2004.

O. Ledoit and M. Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, 10(4):453–472, Nov 2006.

D. Passemier, Z. Li, and J. Yao. On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):51–67, 2017.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.

P. Reimann, C. Van den Broeck, and G.J. Bex. A gaussian scenario for unsupervised learning. *J. Phys. A:Math. Gen.*, 1996.

Jun Shao, Yazhen Wang, Xinwei Deng, Sijian Wang, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265, 2011.

M. O. Ulfarsson and V. Solo. Dimension estimation in noisy pca with sure and random matrix theory. *IEEE Transactions on Signal Processing*, 56(12):5804–5816, Dec. 2008.

L. Yang, M. R. McKay, and R. Couillet. High-dimensional mvdr beamforming: Optimized solutions based on spiked random matrix models. *IEEE Trans. Signal Processing*, 66(7):1933–1947, 2018.

L.C. Zhao, P.R. Krishnaiah, and Z.D. Bai. On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20(1):1– 25, 1986. ISSN 0047-259X.

A. Zollanvari and E. R. Dougherty. Generalized consistent error estimator of linear discriminant analysis. *IEEE Transactions on Signal Processing*, 63(11):2804–2814, Jun. 2015.