Large-Dimensional Characterization of Robust Linear Discriminant Analysis

N. Auguin*, D. Morales-Jimenez†, M. R. McKay*

***Abstract*—In standard discriminant analysis, data are commonly assumed to follow a Gaussian distribution, a condition which is often violated in practice. In this work, to account for potential spurious or mislabeled observations in the training data, we consider a robust version of regularized linear discriminant analysis (LDA) classifiers. Essential to such robust version of LDA is the design of a robust discriminant rule which relies on a robust estimate of the covariance matrix of the training data. We propose to use a regularized version of M-estimators of covariance matrices belonging to Maronna's class of estimators. In the regime where both the number of variables and the number of training samples are large, building upon recent results from random matrix theory, we show that when the training data are free from outliers, each classifier within the class of proposed robust classifiers is asymptotically equivalent to traditional, non-robust classifiers. Rather surprisingly, this entails that the use of robust estimators does not degrade the performance of LDA, up to a transformation of the regularization parameter that we precisely characterize. We also demonstrate that the proposed robust classifiers lead to a better classification accuracy when the data are corrupted by outliers or random noise. Furthermore, through simulations on the popular MNIST data set and considering different classification tasks, we show that the worse the classification error of traditional methods is, the further gain is to be expected with the use of our proposed method.**

Index Terms: *Robust estimation, covariance matrices, linear discriminant analysis.*

I. Introduction

Discriminant analysis is a common parametric classification method used in statistics, machine learning, and pattern recognition [2, 3]. The objective is to determine the class to which a new data observation belongs. To that end, a discrimination rule is learned from labeled (training) data, based on estimates of the class means and covariances.

When dealing with real data sets, it is often the case that the number of variables is of the same order as (or even larger than) the number of available samples. In such cases, standard discriminant analysis, based on the classical sample covariance matrix (SCM) estimator, typically fails. To solve this issue, regularized versions of discriminant analysis have been proposed [4], based on regularized versions of the SCM. Regularized discriminant analysis has since established itself as a go-to choice in practice. In recent works [5, 6], regularized linear discriminant analysis (LDA) has been studied from a random matrix theory perspective. Specifically, when the numbers of variables and samples grow large at the same rate, an asymptotic equivalent of the classification error has been found, shedding some light on the influence of the data model on the performance of regularized LDA. This result has been leveraged to estimate the optimal regularization parameter that minimizes the testing error [6]. As this estimation procedure has negligible computational cost, it represents a major improvement over traditional, computationally-costly (and sometimes unstable) procedures like cross-validation or bootstrap [7]. Recently, the performance of LDA has also been studied in the context of dimensionality reduction using random projections [8], and under a spiked covariance model [9].

A common problem arising in discriminant analysis is that the data, although assumed to arise from a Gaussian mixture model, is often not Gaussian. The data distribution may instead be heavy-tailed, or contain outliers[1] (see e.g., [10]). This is important in practice, since the discriminant rule learned from training data requires the estimation of the data covariance matrix, and if outliers are present and/or the data are heavy-tailed, a standard estimate like the SCM may not perform well. It is then natural to employ a robust estimator of the covariance matrix. Here, we consider regularized Maronna's M-estimators of covariance, a class of robust hybrid estimators proposed in [11] combining the original Maronna's M-estimators [12] and the popular regularized SCM (RSCM) [13]. In [14], regularized M-estimators were also applied to the discriminant analysis problem, however: i) the regularized estimators of [14] consider different (general) loss functions which are not specific to the LDA context, i.e., the objective is to minimize certain loss functions such as Euclidean covariance distance, rather than optimizing the LDA performance (classification error); and ii) a systematic performance analysis in terms of the

*N. Auguin and M. R. McKay are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. E-mail: nicolas.auguin@connect.ust.hk, m.mckay@ust.hk.

†D. Morales-Jimenez is with the Institute of Electronics, Communications and Information Technology, Queen's University Belfast, Belfast BT3 9DT, United Kingdom. E-mail: d.morales@qub.ac.uk.

This work has been partially presented at ICASSP 2019 [1].

This work was supported by the Hong Kong RGC General Research Fund under grant number 16202918.

[1] Outliers are data points that statistically differ (significantly) from other observations. In the context of LDA classifiers under a Gaussian mixture model, for example, an outlier could be a data point drawn from a different distribution, e.g., Gaussian with different mean and/or covariance than the ones considered in the model, or a non-Gaussian distribution. The origin of outliers can be diverse [10], e.g., due to variability in the measurement, experimental errors, mislabeling of data observations, etc.

classification error is not provided in [14]—consequently, their LDA solutions require a costly cross-validation procedure in order to find the optimal regularization parameter.

By leveraging tools from random matrix theory, we study the asymptotic performance of LDA when employing these robust covariance estimators (regularized Maronna's M-estimators) in lieu of the RSCM. To do so, we build on a series of recent works concerned with the performance analysis of regularized LDA [5, 6] and the asymptotic behavior of Maronna's M estimators [15], which consider the regime where the number of variables and the number of samples grow large at the same rate. As a key technical finding, we demonstrate that when no outliers are present in the data, there is no performance loss when using regularized Maronna's M-estimators rather than the RSCM. This departs from the fact that, in classical statistical settings (small number of variables, large number of samples) robust estimation methods are sometimes accompanied with a loss in accuracy in uncontaminated data, as compared with classical non-robust methods [10]. The tradeoff between robustness and efficiency is indeed a well known issue in robust statistics [10, 16, 17]—while offering some form of protection against outliers, robust estimators should ideally differ as little as possible from standard, non-robust alternatives under uncontaminated data [16]. This is also known as "premium", i.e., the cost of using the robust estimator when the data is clean[2]. Our analysis shows that the considered robust estimators have a low premium in high-dimensional settings, and that the associated cost of using robust LDA classifiers vanishes asymptotically as both the number of variables and the number of samples grow large.

As a key theoretical result to characterizing the large-dimensional behaviour of the robust LDA classifiers, we show that Maronna's M-estimators tend to behave analogously to the RSCM. This is consistent with previous results [15, 18], derived for settings different to LDA. For LDA, a notable technical challenge is that we have to consider a mixture model, with data samples that typically exhibit different, non-zero means. Our results necessitate a centering of the data prior to estimation, which complicates the analysis. Similar technical difficulties were also encountered in [19], in the context of portfolio optimization. The family of Maronna's M-estimators that we consider is quite broad, encompassing in particular Tyler-type [20] and Huber-type estimators [21]. A large-dimensional analysis of these estimators was presented in [22] and [15]; however, the analysis was not done in the context of LDA and considered different data models.

Our theoretical results are validated through simulations, on synthetic data and on the MNIST dataset. We show that when the data are corrupted by outliers represented by salt-and-pepper noise, there is a clear benefit in using Maronna's M-estimators, relative to RSCM. That is, when departing from clean-data models, the proposed robust classification methods provide protection against outliers and lead to enhanced performance compared to traditional non-robust methods. Generally, our results argue in favor of the use of robust covariance estimators for LDA-based classification.

Notation: The superscript $^{\mathrm{T}}$ means transpose, $\mathrm{Tr}[\mathbf{A}]$ represents the trace of the Hermitian matrix $\mathbf{A}$ and $\lambda_1(\mathbf{A}) \leq \cdots \leq \lambda_N(\mathbf{A})$ represent its ordered eigenvalues, $||.||$ denotes the Euclidean norm for vectors and the spectral norm for matrices. The notation $\mathbf{A} = \mathrm{diag}(a_1, \ \cdots, a_n)$ indicates that the matrix $\mathbf{A}$ is an $n \times n$ diagonal matrix with diagonal entries $a_1, \cdots, a_n$. We use $\mathbf{1}_n$ to denote the $n \times 1$ vector of all ones, and $\mathbf{I}_N$ the $N \times N$ identity matrix. $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The arrow $\xrightarrow{\text{a.s.}}$ designates the almost sure convergence of a random variable, while δ_x denotes the Dirac measure at point x. If $\mathbf{h}$ is a Gaussian random vector with mean μ and covariance matrix $\mathbf{C}$, we write $\mathbf{h} \sim \mathcal{N}(\mu, \mathbf{C})$. For a functional f, we say that $f = \mathcal{O}(1)$ if there exists $M > 0$ such that $|f| \leq M$.

II. Linear discriminant analysis

A. Model

In discriminant analysis, a discriminant rule is applied to decide the class that a given (unseen) data observation belongs to. Such rule is built based on a training data set composed of n samples, known (labeled) to belong to one of the classes. Consider 2 classes C_0, C_1, and assume that the $n_i > 0$ observations from class C_i are independent samples from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i \in \mathbb{R}^{N \times 1}$ and covariance matrix $\mathbf{C}_N \in \mathbb{R}^{N \times N}$, with $\mathbf{C}_N \succeq \mathbf{0}$. Thus, a training sample $\mathbf{y}^{(i)}$ from class C_i $(i = 0, 1)$ is such that

$$\mathbf{y}^{(i)} = \boldsymbol{\mu}_i + \mathbf{C}_N^{1/2}\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N). \tag{1}$$

The linear discriminant rule assigns a new (test) measurement $\mathbf{y}$ to class C_k if

$$k = \underset{i \in \{0,1\}}{\mathrm{argmin}} \left\{ (\mathbf{y} - \boldsymbol{\mu}_i)^{\mathrm{T}} \mathbf{C}_N^{-1} (\mathbf{y} - \boldsymbol{\mu}_i) - \log \pi_i \right\}, \tag{2}$$

with π_i the *a priori* probability of class C_i. Throughout the paper the prior probabilities π_0, π_1 will be assumed known. It can be seen that the LDA rule assigns the label 0 to observation $\mathbf{y}$ if $\mathbb{P}(\mathbf{y}|\mathbf{y} \in C_0) > \mathbb{P}(\mathbf{y}|\mathbf{y} \in C_1)$, and the label 1 otherwise.

Since the true means $\boldsymbol{\mu}_i$ and the population matrix $\mathbf{C}_N$ appearing in the LDA rule are unknown, in practice they need to be estimated based on the training data $\{\mathbf{y}_j^{(i)} \in C_i, \quad i = 0, 1, \quad j = 1, \cdots, n_i\}$. Common estimates for $\boldsymbol{\mu}_i$, $\mathbf{C}_N$ are the sample estimates

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_j^{(i)}, \quad i \in \{0, 1\}$$

$$\hat{\mathbf{R}} = \frac{1}{n-2} \left((n_0 - 1)\hat{\mathbf{R}}_0 + (n_1 - 1)\hat{\mathbf{R}}_1 \right),$$

[2] For example, a robust estimate of location like the median incurs a high premium: under uncontaminated data, it requires a higher number of samples to reach the same performance as the (non-robust) mean [17].

where $n = n_0 + n_1$ and where

$$\hat{\mathbf{R}}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} \tilde{\mathbf{y}}_j^{(i)}(\tilde{\mathbf{y}}_j^{(i)})^{\mathrm{T}}, \quad i \in \{0, 1\}. \tag{3}$$

Here $\tilde{\mathbf{y}}_j^{(i)}$ designates the re-centered version of sample $\mathbf{y}_j^{(i)} \in C_i$, $i = 0, 1$, $j \in \{1, \cdots, n_i\}$ obtained after subtracting the sample mean $\hat{\boldsymbol{\mu}}_i$ from $\mathbf{y}_j^{(i)}$, i.e.,

$$\tilde{\mathbf{y}}_j^{(i)} \triangleq \mathbf{y}_j^{(i)} - \hat{\boldsymbol{\mu}}_i.$$

The estimator $\hat{\mathbf{R}}$ is usually referred to as the "pooled SCM". A main issue with the pooled SCM estimator is that it performs poorly when the number of variables is of the same order as the number of samples (or possibly larger). In practice, to alleviate this issue (and avoid possible ill-conditioning of the SCM), the regularized estimator [5, 23]

$$\hat{\mathbf{R}}(\kappa, \beta) = \kappa\left(\mathbf{I}_N + \beta\hat{\mathbf{R}}\right) \tag{4}$$

is typically used, where $\kappa, \beta \geq 0$ are regularization parameters. Hereafter, this estimator will be referred to as the RSCM. When the RSCM and the sample means are used as "plug-in" estimators in (2), the corresponding LDA classifier will be referred to as RLDA.

Next, we recall some known results concerning the performance of LDA for a general estimator of $\mathbf{C}_N$ [5], and the asymptotic performance of LDA when the estimator of $\mathbf{C}_N$ is chosen to be $\hat{\mathbf{R}}(\kappa, \beta)$ [5, 6]. These results will be important when studying the asymptotic performance of our LDA solutions, introduced in Section III.

B. Classification error of linear discriminant analysis

Let $\hat{\mathbf{H}}$ be an estimator of $\mathbf{C}_N^{-1}$. Then, conditioned on the training data $\mathbf{x}_1, \cdots, \mathbf{x}_n$, the probability of misclassification is given by [5]

$$\epsilon^{\mathrm{LDA}}(\hat{\mathbf{H}}) = \pi_0\epsilon_0^{\mathrm{LDA}}(\hat{\mathbf{H}}) + \pi_1\epsilon_1^{\mathrm{LDA}}(\hat{\mathbf{H}}) \tag{5}$$

with $\epsilon_i^{\mathrm{LDA}}(\hat{\mathbf{H}})$, $i \in \{0, 1\}$, the class-conditional classification error verifying

$$\epsilon_i^{\mathrm{LDA}}(\hat{\mathbf{H}}) = \Phi\left(\frac{(-1)^{i+1}G_i(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{H}}) + (-1)^i \log\frac{\pi_1}{\pi_0}}{\sqrt{D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{H}})}}\right), \tag{6}$$

where

$$G_i(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{H}}) = \left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)^{\mathrm{T}}\hat{\mathbf{H}}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) \tag{7}$$

$$D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{H}}) = (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\hat{\mathbf{H}}\mathbf{C}_N\hat{\mathbf{H}}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1). \tag{8}$$

For RLDA, $\hat{\mathbf{H}}$ is chosen to be $\hat{\mathbf{R}}(\kappa, \beta)^{-1}$, which yields the class-conditional classification error

$$\epsilon_i^{\mathrm{RLDA}}(\kappa, \beta) = \Phi\left(\frac{(-1)^{i+1}G_i(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta)^{-1}) + (-1)^i\kappa\log\frac{\pi_1}{\pi_0}}{\sqrt{D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta)^{-1})}}\right), \tag{9}$$

where $\hat{\mathbf{R}}(\beta) \triangleq \frac{\hat{\mathbf{R}}(\kappa,\beta)}{\kappa} = \mathbf{I}_N + \beta\hat{\mathbf{R}}$. Note that the first regularization parameter of the RSCM κ is inconsequential when the class priors are equal, i.e., when $\pi_0 = \pi_1$. The corresponding asymptotic classification error has been studied in [5, 6], under a double-asymptotic regime defined as follows:

Assumption 1. $c_N \triangleq N/n \to c \in (0, \infty)$ as $N, n \to \infty$, and $n_i/n \to \pi_i \in (0, 1)$, $i = 0, 1$, as $n, n_i \to \infty$.

This specifies the growth regime under consideration, allowing random matrix theory results to be exploited.

Moreover, the following assumption is also required:

Assumption 2. $||\mathbf{C}_N|| = \mathcal{O}(1)$, and $||\boldsymbol{\mu}|| = \mathcal{O}(1)$, where $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$.

This assumption controls the scaling of the covariance and mean of the training data, and ensures that a non-trivial (i.e., neither 0 nor 1) classification error probability is attainable as the dimension of the data N grows large [6, 24].

Under Assumptions 1 and 2, [5, 6] showed that the class-conditional classification error of RLDA in (9) converges to a deterministic quantity $\overline{\epsilon}_i^{\mathrm{RLDA}}(\kappa, \beta)$ as $n, N \to \infty$, with $\overline{\epsilon}_i^{\mathrm{RLDA}}(\kappa, \beta)$ depending only on the true means of each class and the underlying covariance matrix. The result is recalled as follows:

Proposition 1. *[5, Theorem 1] Let Assumptions 1 and 2 hold. As $N, n \to \infty$, $|\epsilon_i^{\mathrm{RLDA}}(\kappa, \beta) - \overline{\epsilon}_i^{\mathrm{RLDA}}(\kappa, \beta)| \xrightarrow{\text{a.s.}} 0$ for each $\kappa, \beta > 0$, with*

$$\overline{\epsilon}_i^{\mathrm{RLDA}}(\kappa, \beta) = \Phi\left(\frac{(-1)^{i+1}\overline{G}_i(\beta) + (-1)^i\kappa\log\left(\frac{\pi_0}{\pi_1}\right)}{\sqrt{\overline{D}(\beta)}}\right).$$

Here, $\overline{G}_i(\beta)$ and $\overline{D}(\beta)$ are defined as

$$\overline{G}_i(\beta) = \frac{(-1)^i}{2}\boldsymbol{\mu}^{\mathrm{T}}\left(\mathbf{I}_N + \frac{\beta}{1+\beta\delta}\mathbf{C}_N\right)^{-1}\boldsymbol{\mu} - \frac{n\delta}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)$$

$$\overline{D}(\beta) = \frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}_N\mathbf{A}\boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\mathrm{Tr}[\mathbf{C}_N^2\mathbf{A}]}{1 - \frac{\beta^2}{n(1+\beta\delta)^2}\mathrm{Tr}[\mathbf{C}_N^2\mathbf{A}]},$$

where $\mathbf{A} = \left(\mathbf{I}_N + \frac{\beta}{1+\beta\delta}\mathbf{C}_N\right)^{-2}$, and δ is the unique solution to

$$\delta = \frac{1}{N}\mathrm{Tr}\left[\mathbf{C}_N\left(\mathbf{I}_N + \frac{\beta}{1+\beta\delta}\mathbf{C}_N\right)^{-1}\right].$$

This result links the misclassification probability associated with the RSCM with a deterministic quantity that does not depend on the (random) training data. As will be seen, this can be used to identify the optimal regularization parameter β with minimal classification error.

C. Sensitivity of RLDA to the presence of outliers

An issue with RLDA is that it is not robust if outlying samples are present in the training data. We show this with an example. Consider a covariance matrix with entries $[\mathbf{C}_N]_{ij} = 0.8^{|i-j|}$, and with eigenvalue decomposition $\mathbf{C}_N = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}$. Assume that $\boldsymbol{\mu} \propto \mathbf{V}\mathbf{1}_N$, and that outliers distributed as $\mathcal{N}(5\boldsymbol{\mu}, \mathbf{I}_N)$ are introduced in the training sample. In Fig. 1, we plot the probability of classification error for the RLDA classifier (i.e., based on the RSCM), as the proportion of outliers increases. Here, we assume equal priors (as such, the choice of κ is irrelevant), and β is chosen empirically to minimize the classification error on the testing data set. The performance of the oracle classifier ($\mathbf{C}_N$ assumed known) and the performance of a robust classifier (based on Tyler's estimator, to be introduced next), are also presented.

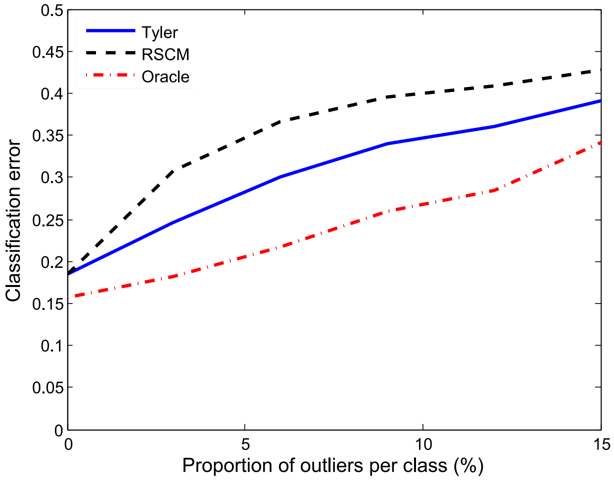

Fig. 1: Classification error probability for the RLDA and oracle classifiers. $N = 100$, $n_0 = n_1 = 200$ ($c_N = 1/4$), averaged over 1,000 realizations. Outliers $\sim \mathcal{N}(5\boldsymbol{\mu}, \mathbf{I}_N)$.

It is evident that the performance of RLDA is seriously affected by the outliers, even when the outlier proportion is reasonably small. This is in contrast to the robust classifier, which can mitigate the effect of the outliers. In the next section, we will introduce a class of robust covariance estimators which can be used in the context of LDA.

III. Robust LDA Using Maronna's M-estimators

We propose a robust version of RLDA by utilizing robust estimators of covariance matrices known as Maronna's M-estimators [12]. A regularized version of these estimators, adapted to data-scarce applications, has been proposed in [11]. The general behavior of these regularized, robust estimators has been studied in [15] under double asymptotics. However it is not known how these estimators perform in the context of LDA. By leveraging and extending some of the results obtained in [15], we will show that the use of these estimators lead to a robust version of RLDA which is guaranteed to perform as well as the traditional (RSCM-based) RLDA when data are not corrupted by outliers.

In the context of the two-mixture model introduced above, robust regularized M-estimators of covariance matrices are then defined as the unique solution to the equation [15]

$$\hat{\mathbf{C}}_N(\rho) = \frac{(1-\rho)}{n-2}\sum_{i=0}^{1}\sum_{j=1}^{n_i} u\left((\tilde{\mathbf{y}}_j^{(i)})^{\mathrm{T}}\hat{\mathbf{C}}_N(\rho)^{-1}\tilde{\mathbf{y}}_j^{(i)}\right)\tilde{\mathbf{y}}_j^{(i)}(\tilde{\mathbf{y}}_j^{(i)})^{\mathrm{T}} + \rho\mathbf{I}_N, \tag{10}$$

where $\rho \in (0,1]$ is a fixed regularization parameter, and u satisfies the following properties:

- u is a nonnegative, nonincreasing, bounded, and continuous function on $\mathbb{R}^+$,
- $\phi : x \mapsto xu(x)$ is non-decreasing and bounded, with $\phi_\infty \triangleq \lim_{x\to\infty}\phi(x) \leq \frac{1}{c}$.[3]

Examples of common u functions include

$$u_{\text{Tyler}}(x) \triangleq \frac{1}{c}\frac{1}{x} \tag{11}$$

$$u_{\text{Huber}}(x) \triangleq \frac{1}{c}\min\left\{1, \frac{1}{x}\right\}. \tag{12}$$

These "Tyler's" and "Huber's" estimators verify the conditions above, with $\phi_\infty = 1/c$.

The solution $\hat{\mathbf{C}}_N(\rho)$ to (10) is a hybrid estimator reminiscent of the original Maronna's M-estimator of scatter [12] and Ledoit-Wolf's shrinkage estimator [13]. Here, like β in the RSCM (4), ρ is a regularization parameter that determines the tradeoff between bias (the shrinkage target, $\mathbf{I}_N$) and variance (the pooled SCM). Using $\hat{\mathbf{C}}_N(\rho)^{-1}$ as a plug-in estimator of $\mathbf{C}_N^{-1}$ in (2), we should obtain a robust version of RLDA, which we coin M-RLDA. An important question is whether M-RLDA performs as well as the standard (RSCM-based) RLDA approach when the data are clean. As we will show, the answer is yes, at least under the large-dimensional regime (Assumption 1).

A. Asymptotic performance of M-RLDA

In this section, we will make an additional (rather loose) technical assumption needed to prove our main result:

Assumption 3.
$\nu_n \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i(\mathbf{C}_N)}$ satisfies $\nu_n \to \nu$ weakly with $\nu \neq \delta_0$ almost everywhere.

We also introduce the following function that will be important in obtaining an asymptotic equivalent of $\hat{\mathbf{C}}_N(\rho)$:
Definition. Define $v : [0,\infty) \to (0, u(0)]$ as $v(x) = u(g^{-1}(x))$ where g^{-1} denotes the inverse function of $g(x) = \frac{x}{1-(1-\rho)c\phi(x)}$, which maps $[0,\infty)$ onto $[0,\infty)$.

The function v is continuous, non-increasing and onto. The fact that $\phi : x \mapsto xu(x)$ is non-decreasing and bounded and that $\phi_\infty \triangleq \lim_{x\to\infty}\phi(x) \leq \frac{1}{c}$ guarantees that g (and thus v) is properly defined for all $\rho \in (0,1]$. Notice that

[3] We note that the fact that $\phi_\infty \leq 1/c$ (where, recall, $N/n \to c$ as $N, n \to \infty$) is not necessary to ensure the existence of a solution to (10). Nevertheless, it is shown in [15] that choosing such an upper bound is helpful to manage some technical derivations when dealing with estimator asymptotic equivalents, while remaining non-binding in practice.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2021.3075150, IEEE Transactions on Signal Processing

v has essentially the same properties as the u function. For the example u functions of Tyler's and Huber's estimators in (11) and (12), the corresponding v functions take the form:

$$v_{\text{Tyler}}(x) = \frac{1}{\rho c}\frac{1}{x}$$
$$v_{\text{Huber}}(x) = \frac{1}{c}\min\left\{1, \frac{1}{\rho x}\right\}.$$

With these definitions, we can now present the following result, with the proof relegated to Section VI-A.

Lemma 1. *[Asymptotic behavior of $\hat{\mathbf{C}}_N(\rho)$] Define $\mathcal{I}$ a compact set included in $(0,1]$. Let $\hat{\mathbf{C}}_N(\rho)$ be the unique solution to (10). Then, as $N, n \to \infty$, under Assumption 1 and Assumption 3,*

$$\sup_{\rho\in\mathcal{I}} \left\|\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)\right\| \xrightarrow{\text{a.s.}} 0, \tag{13}$$

where

$$\hat{\mathbf{S}}_N(\rho) \triangleq (1-\rho)v(\gamma(\rho))\tfrac{1}{n-2}\textstyle\sum_{i=1}^{n}\tilde{\mathbf{y}}_i\tilde{\mathbf{y}}_i^{\mathrm{T}} + \rho\mathbf{I}_N,$$

with $\gamma(\rho)$ the unique positive solution to

$$\gamma(\rho) = \tfrac{1}{N}\mathrm{Tr}\left[\mathbf{C}_N\left(\tfrac{(1-\rho)v(\gamma(\rho))}{1+c(1-\rho)v(\gamma(\rho))\gamma(\rho)}\mathbf{C}_N + \rho\mathbf{I}_N\right)^{-1}\right]. \tag{14}$$

Furthermore, the function $\rho \mapsto \gamma(\rho)$ is bounded, continuous on $(0,\infty]$ and greater than zero.

Lemma 1 shows that, asymptotically, Maronna's M-estimators (solutions to (10)) behave like a classical regularized estimator. In fact, up to dividing $\hat{\mathbf{S}}_N(\rho)$ by ρ, we retrieve the form of the RSCM: $\hat{\mathbf{S}}_N(\rho)/\rho = \hat{\mathbf{R}}(\beta_\rho)$, with β_ρ set to $\frac{(1-\rho)}{\rho}v(\gamma(\rho))$.

Lemma 1 is similar to [15, Theorem 2]; the main difference being that the u function can be non-decreasing, and that we work with re-centered data samples, inducing technical difficulties in the proof (see Section VI-A).

The asymptotic equivalence can be exploited to prove that the bilinear forms $\rho^k\mathbf{a}^{\mathrm{T}}\hat{\mathbf{C}}_N(\rho)^{-k}\mathbf{b}$ are asymptotically close to their RSCM counterparts $\mathbf{a}^{\mathrm{T}}\hat{\mathbf{R}}(\beta_\rho)^{-k}\mathbf{b}$, a result that will be important for proving that the classification error of M-RLDA and that of RLDA are asymptotically the same.

Lemma 2. *For a given $\rho \in (0,1]$, we define $\beta_\rho = \frac{(1-\rho)}{\rho}v(\gamma(\rho))$, with γ the solution to (14). We then have $\hat{\mathbf{S}}_N(\rho)/\rho = \hat{\mathbf{R}}(\beta_\rho)$, and for $\mathbf{a},\mathbf{b} \in \mathbb{R}^N$ such that $\limsup_N ||\mathbf{a}|| < \infty$ a.s., $\limsup_N ||\mathbf{b}|| < \infty$ a.s., and $k = 1, 2$, we have for all $\rho \in (0,1]$,*

$$\left|\rho^k\mathbf{a}^{\mathrm{T}}\hat{\mathbf{C}}_N(\rho)^{-k}\mathbf{b} - \mathbf{a}^{\mathrm{T}}\hat{\mathbf{R}}(\beta_\rho)^{-k}\mathbf{b}\right| \xrightarrow{\text{a.s.}} 0, \tag{15}$$

as $N, n \to \infty$.

Proof. For $k = 1$, first note that

$$\left|\mathbf{a}^{\mathrm{T}}(\rho\hat{\mathbf{C}}_N(\rho)^{-1} - \hat{\mathbf{R}}(\beta_\rho)^{-1})\mathbf{b}\right| \leq K||\rho\hat{\mathbf{C}}_N(\rho)^{-1} - \hat{\mathbf{R}}(\beta_\rho)^{-1}||$$
$$\overset{(a)}{\leq} K||\rho\hat{\mathbf{C}}_N(\rho)^{-1}|| \cdot ||\hat{\mathbf{R}}(\beta_\rho)^{-1}|| \cdot ||\hat{\mathbf{C}}_N(\rho)/\rho - \hat{\mathbf{R}}(\beta_\rho)||,$$

where (a) is due to the resolvent identity[4], and where $K = ||\mathbf{a}|| \cdot ||\mathbf{b}||$. The fact that $\rho, \beta_\rho > 0$ ensures that $||\hat{\mathbf{C}}_N(\rho)^{-1}|| < \infty$ and $||\hat{\mathbf{R}}(\beta_\rho)^{-1}|| < \infty$. From Lemma 1, we also have $\left\|\hat{\mathbf{C}}_N(\rho)/\rho - \hat{\mathbf{R}}(\beta_\rho)\right\| = \left\|\hat{\mathbf{C}}_N(\rho)/\rho - \hat{\mathbf{S}}_N(\rho)/\rho\right\| \xrightarrow{\text{a.s.}} 0$. Along with $\limsup_N ||\mathbf{a}|| < \infty$ a.s., $\limsup_N ||\mathbf{b}|| < \infty$ a.s., (15) is proved for $k = 1$.

The case $k = 2$ is handled by noting that

$$\left|\mathbf{a}^{\mathrm{T}}(\rho^2\hat{\mathbf{C}}_N(\rho)^{-2} - \hat{\mathbf{R}}(\beta_\rho)^{-2})\mathbf{b}\right| \leq K||\rho^2\hat{\mathbf{C}}_N(\rho)^{-2} - \hat{\mathbf{R}}(\beta_\rho)^{-2}||$$
$$\overset{(a)}{\leq} K||\rho^2\hat{\mathbf{C}}_N(\rho)^{-2}||||\hat{\mathbf{R}}(\beta_\rho)^{-2}|| \cdot ||\hat{\mathbf{C}}_N(\rho)^2/\rho^2 - \hat{\mathbf{R}}(\beta_\rho)^2||$$

where (a) is again due to the resolvent identity. Using the same arguments as for $k = 1$, we have $||\rho^2\hat{\mathbf{C}}_N(\rho)^{-2}|| < \infty$ a.s. and $||\hat{\mathbf{R}}(\beta_\rho)^{-2}|| < \infty$ a.s. The convergence $||\hat{\mathbf{C}}_N(\rho)^2/\rho^2 - \hat{\mathbf{R}}(\beta_\rho)^2|| \xrightarrow{\text{a.s.}} 0$ follows as a consequence of Lemma 1 and Weyl's theorem [25], which ensures convergence of individual eigenvalues. This proves the result for $k = 2$. □

Remark 1. *Note that Lemma 1 and Lemma 2 hold even if the data are not Gaussian, as long as the entries of the random vectors $\tilde{\mathbf{y}}_j^{(i)}$, $j = 1, \cdots, n_i$, $i = 0, 1$, have finite $(8+\sigma)$-th moment ($\sigma > 0$). However, Proposition 1 and subsequent results in the paper depend on the Gaussianity of the data, one of the assumptions of LDA.*

Remark 2. *In the proof of Lemma 2, we only use the a.s. point-wise convergence of $\left\|\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)\right\|$ to 0, while Lemma 1 states that this convergence holds uniformly on $\rho \in \mathcal{I}$. In Section VI-A, we proved this more general result as it could be important for different applications (e.g., [19]).*

Denote the class-conditional classification error of M-RLDA by

$$\epsilon_i^{\text{M-RLDA}}(\rho) \triangleq \epsilon_i^{\text{LDA}}(\rho\hat{\mathbf{C}}_N(\rho)^{-1}).$$

Our main technical result is the following:

Theorem 1 (Deterministic equivalent of the classification error of M-RLDA)**.** *Let Assumptions 1-3 hold. For each $\rho \in (0,1]$, as $N, n \to \infty$,*

$$|\epsilon_i^{\text{M-RLDA}}(\rho) - \bar{\epsilon}_i^{\text{M-RLDA}}(\rho)| \xrightarrow{\text{a.s.}} 0,$$

where

$$\bar{\epsilon}_i^{\text{M-RLDA}}(\rho) \triangleq \bar{\epsilon}_i^{\text{RLDA}}(\rho, \beta_\rho)$$

with $\bar{\epsilon}_i^{\text{RLDA}}(\rho, \beta_\rho)$ given in Proposition 1, and where we recall that $\beta_\rho = \frac{(1-\rho)}{\rho}v(\gamma(\rho))$ for $\rho \in (0,1]$. Furthermore, $\rho \mapsto \beta_\rho$ is onto on $[0,\infty)$.

Proof. We will first show that $\left|\epsilon_i^{\text{M-RLDA}}(\rho) - \epsilon_i^{\text{RLDA}}(\rho, \beta_\rho)\right| \xrightarrow{\text{a.s.}} 0$ as $n, N \to \infty$.

[4] For invertible $\mathbf{U}, \mathbf{V}$, it is true that $\mathbf{U}^{-1} - \mathbf{V}^{-1} = \mathbf{V}^{-1}(\mathbf{V}-\mathbf{U})\mathbf{U}^{-1}$.

To do so, we will prove that

$$\left|G_i(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \rho\hat{\mathbf{C}}_N(\rho)^{-1}) - G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta_\rho)^{-1})\right| \xrightarrow{\text{a.s.}} 0 \tag{16}$$

$$\left|D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \rho\hat{\mathbf{C}}_N(\rho)^{-1}) - D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta_\rho)^{-1})\right| \xrightarrow{\text{a.s.}} 0, \tag{17}$$

where we recall that $G_i(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \rho\hat{\mathbf{C}}_N(\rho)^{-1})$ and $D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \rho\hat{\mathbf{C}}_N(\rho)^{-1})$, given in (7) and (8), are used to compute the class-conditional classification error $\epsilon_i^{\mathrm{M-RLDA}}(\rho) = \epsilon_i^{\mathrm{LDA}}(\rho\hat{\mathbf{C}}_N(\rho)^{-1})$ in (6).

To prove the result, we will apply Lemma 2 with $\mathbf{a} = \boldsymbol{\mu}_i - (\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1)/2$ and $\mathbf{b} = \hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1$. Assumption 2 implies $\limsup_N ||\mathbf{a}|| < \infty$ a.s., $\limsup_N ||\mathbf{b}|| < \infty$ a.s., by the law of large numbers. With this, (16) follows from (15) by taking $k = 1$. To prove (17), using (8), we note that

$$\begin{aligned}
&\left|D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \rho\hat{\mathbf{C}}_N(\rho)^{-1}) - D(\mathbf{C}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta_\rho)^{-1})\right| \\
&= \left|\operatorname{Tr} \mathbf{C}_N\left(\rho^2\hat{\mathbf{C}}_N(\rho)^{-1}\mathbf{b}\mathbf{b}^{\mathrm{T}}\hat{\mathbf{C}}_N(\rho)^{-1} - \hat{\mathbf{R}}(\beta_\rho)^{-1}\mathbf{b}\mathbf{b}^{\mathrm{T}}\hat{\mathbf{R}}(\beta_\rho)^{-1}\right)\right| \\
&\overset{(a)}{\leq} ||\mathbf{C}_N|| \cdot \left|\mathbf{b}^{\mathrm{T}}\left(\rho^2\hat{\mathbf{C}}_N(\rho)^{-2} - \hat{\mathbf{R}}(\beta_\rho)^{-2}\right)\mathbf{b}\right|,
\end{aligned}$$

where (a) uses the fact that $\operatorname{Tr}[\mathbf{AB}] \leq ||\mathbf{A}|| \operatorname{Tr}[\mathbf{B}]$ for a symmetric matrix $\mathbf{A}$ and non-negative definite matrix $\mathbf{B}$. Using (15) with $k = 2$ leads to (17).

From (6), using (16), (17), and the fact that $\sqrt{\cdot}$ and $\Phi(\cdot)$ are continuous functions, we have proved

$$\left|\epsilon_i^{\mathrm{M-RLDA}}(\rho) - \epsilon_i^{\mathrm{RLDA}}(\rho, \beta_\rho)\right| \xrightarrow{\text{a.s.}} 0, \quad n, N \to \infty.$$

Combining with Proposition 1 gives the convergence result.

In addition, it is shown in the proof of Lemma 1 that $\rho \mapsto \gamma(\rho)$ is continuous and bounded. This, combined with the fact that v is continuous and non-increasing, shows that the mapping $\rho \mapsto \beta_\rho = \frac{(1-\rho)}{\rho} v(\gamma(\rho))$ is onto on $[0, \infty)$. □

Theorem 1 shows that in the absence of outliers both the standard RLDA method and M-RLDA are equivalent over the range of the regularization parameter, up to a simple transformation. That is, the mapping $\rho \mapsto \beta_\rho$ suffices to retrieve the equivalent RSCM estimator from the robust estimator (and vice-versa using the inverse mapping).

B. Choice of estimator in outlier-free data

We have seen that in an outlier-free context, the asymptotic classification error associated with an M-estimator (with a given u function) is the same as that of *any* other M-estimator (i.e., with any other u function) or that of the RSCM, up to a change in the regularization parameter. This is in line with recent results obtained in robust estimation theory under double asymptotics (see e.g., [15]), but it differs from the known fact that, in finite dimensions and in the absence of outliers, robust estimators are often sub-optimal compared to their non-robust counterparts (see, e.g., [16]). In classical scenarios where $n \gg N$, there exists a trade-off between accuracy and robustness (see, e.g., [10]). However, in the large-dimensional regime where n and N grow large together, there is no need to 'choose' between accuracy and robustness, as we have shown that in a clean data scenario, all estimators, robust or non-robust, are indeed equivalent.

1) Calibration of the regularization parameter: Among the possible $\rho \in (0, 1]$, one should choose the parameter that minimizes the classification error $\epsilon^{\mathrm{M-RLDA}}(\rho)$. This quantity is inaccessible in practice, however one may build a data-based estimate of it. To this end, we can exploit such an estimate $\hat{\epsilon}^{\mathrm{RLDA}}(\kappa, \beta)$ of the true classification error $\epsilon^{\mathrm{RLDA}}(\kappa, \beta)$ for RLDA which verifies, for all $\kappa, \beta > 0$,

$$\left|\hat{\epsilon}^{\mathrm{RLDA}}(\kappa, \beta) - \epsilon^{\mathrm{RLDA}}(\kappa, \beta)\right| \xrightarrow{\text{a.s.}} 0$$

as $N, n \to \infty$ [5, 6]. This result is specifically recalled in Lemma 3 (see Section VI-C). In this section, and thereafter in the simulations, we will assume that we have equal priors, i.e., $\pi_0 = \pi_1$. In this case, the choice of κ in the RSCM is no longer relevant, and we shall use the notation $\hat{\epsilon}^{\mathrm{RLDA}}(\beta)$ in lieu of $\hat{\epsilon}^{\mathrm{RLDA}}(\kappa, \beta)$. Using this estimate, it is easy to identify a regularization parameter β^o (e.g., via a line-search) that minimizes the classification error. Knowing β^o, we can then exploit Theorem 1, and use the mapping $\rho \mapsto \beta_\rho$ to identify a suitable regularization parameter minimizing the classification error or M-RLDA[5]. This procedure is summarized in Algorithm 1.

Algorithm 1 Regularization parameter optimization

```
1) Compute the optimized regularization parameter of the
   RSCM via a numerical search
       $\beta^o = \arg\min_{\beta>0} \left\{\hat{\epsilon}^{\mathrm{RLDA}}(\beta)\right\}.$
2) For a given $u$ function and its associated M-estimator
   $\hat{\mathbf{C}}_N$, find a solution $\rho^o$ to the equation in $\rho$
       $\frac{1}{\rho}\frac{1}{N}\operatorname{Tr}\left[\hat{\mathbf{C}}_N(\rho)\right] = 1 + \beta^o.$
3) Construct the discriminant rule with $\hat{\mathbf{C}}_N(\rho^o)^{-1}$.
```

We note that step 2 of Algorithm 1 requires the computation of $\hat{\mathbf{C}}_N(\rho)$ for ρ taken from a (sufficiently dense) discrete set in $(0, 1]$. In practice, $\hat{\mathbf{C}}_N(\rho)$ can be easily computed via a simple iterative algorithm (see e.g., [26]), which usually takes only a few iterations to converge. Nevertheless, each iteration requires the inversion of a (possibly large-dimensional) matrix, which is computationally expensive. Therefore, there is a robustness/complexity trade-off when it comes to choosing the appropriate estimator for a given classification task. Nevertheless, we note that unlike popular cross-validation schemes which require dividing the labeled data into a training set and a validation set repeatedly to identify the best regularization parameters, an advantage of Algorithm 1 is that it provides a robust, online estimation of the optimal ρ, with no need for resampling.

[5] The mapping $\rho \mapsto \beta_\rho$ is only onto on $(0, \infty)$, and thus the uniqueness of the optimal regularization parameter for M-RLDA is not guaranteed.

2) Example with synthetic data: In Fig. 2, we plot the empirical classification error associated with the RSCM, Huber's estimator, and Tyler's estimator as a function of the regularization parameters β and ρ (top and bottom x-axes, respectively), in an outlier-free scenario. The deterministic classification error, computed using Theorem 1 and Proposition 1, and $\epsilon^{\mathrm{LDA}}(\hat{\mathbf{H}} = \mathbf{C}_N^{-1})$ ("oracle" estimator) are also shown. Simulations show a very good match between empirical and theoretical values, validating Theorem 1. The regularization parameters ρ_H^o, ρ_T^o, corresponding to Huber's and Tyler's estimator respectively, obtained using Algorithm 1, are identified with arrows. We remark that careful calibration of the regularization parameter is important: if not carefully chosen, the classification error can reach 47%, a substantial increase compared with the minimal 25% error.

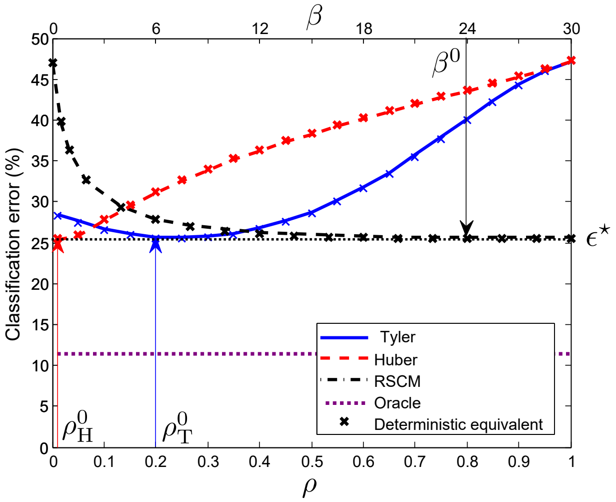

Fig. 2: Classification error when using the RSCM, Huber's, and Tyler's estimator as β and ρ vary (top and bottom x-axes, respectively), for $N = 150$, $n_0 = n_1 = 50$, averaged over 1,000 realizations, with a test sample size of 5,000 samples per class. $[\mathbf{C}_N]_{ij} = 0.9^{|i-j|}$, with eigendecomposition $\mathbf{C}_N = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}$, and $\boldsymbol{\mu} \propto \mathbf{V}\mathbf{1}_N$. The oracle estimator's classification error ($\mathbf{C}_N$ known) is also shown.

IV. Application to Real Data

A. MNIST data set

Here we compare the performance of M-RLDA and RLDA using the MNIST database [27], a set of handwritten digits composed of 60,000 training images and 10,000 testing images from 10 classes, corresponding to the digits, $0, 1, \cdots, 9$. The images are constituted of $N = 28 \times 28 = 784$ pixels.

For all 45 pairs of classes C_0/C_1 corresponding to digits 0/1, 0/2, etc., we computed the RSCM, Huber's, and Tyler's covariance estimators from all the available training data set ($\sim$6,000 samples per class) and tested the corresponding classifiers on the testing data set ($\sim$1,000 samples per class). For each pair, we determined the minimal testing classification error for each classifier (i.e., based on each covariance estimator), obtained after a sweep over all regularization parameters ρ and β. This provides a bound on the lowest classification error achievable on the given testing data set. Huber's and Tyler's estimators are computed using a simple iterative algorithm found in [26]. For the RSCM, we sweep β over the range $[0.001, 200]$, with increments of 0.01 in $[0.001, 1]$, 0.1 in $[1, 15]$, and 5 in $[15, 200]$. For M-RLDA, we take a range of $[0.01, 1]$ for ρ, with increments of 0.01. In Fig. 3 we plot, for each class pair, the classification error of M-RLDA for Tyler's and Huber's estimator against that of RLDA. In general, all estimators perform fairly well, with an error of at most 4.1% for RLDA (for the class pairs $(7, 9)$ and $(5, 8)$, indicated on the graph), and as little as 0.1% (for the class pair $(0, 1)$). In all but one case, Huber's estimator leads to a performance equal to or better than that of the RSCM. In some cases, the gain is fairly substantial. Most notable is the case of the class pair $(4, 9)$, for which Huber's estimator leads to a 16% relative improvement in classification error relative to RLDA. In Fig. 4, we give examples of testing images for the class pair $(4, 9)$ that were correctly classified with both RLDA and the Huber-based classifier, and examples of testing images that were correctly classified when using Huber's estimator but misclassified by RLDA. In contrast to the nearly uniform gains of Huber's estimator, Tyler's estimator shows mixed results in Fig. 3, returning a classification error that is sometimes smaller than and sometimes higher than the RSCM, depending on the considered class pairs. Overall, when using all available training images, these results suggest that Huber's estimator prevails over Tyler's for this specific classification task. A reason for this behavior might be the shape of the underlying u function: in Tyler's case some data points (with Mahalanobis distance smaller than average) may be overweighted compared to others, while the same data points would have their weight capped in Huber's case.

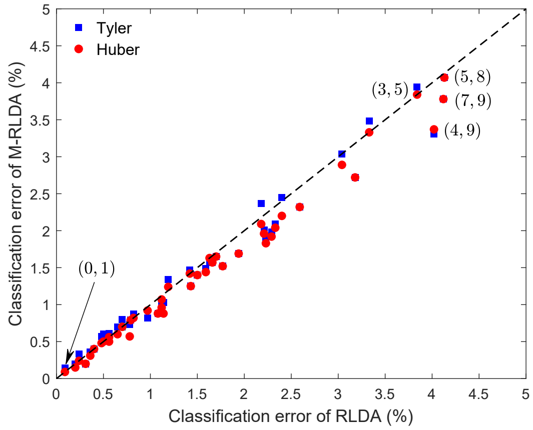

Fig. 3: Classification error of M-RLDA, relative to that of RLDA, for all class pairs of MNIST.

In Fig. 5, for the class pairs $(4, 9)$ and $(1, 7)$ we show the classification error on all the testing data when the number of training data points $n = n_0 + n_1$ is smaller. For this,

Fig. 4: Examples of test images ($(4, 9)$ class pair) that were correctly classified by M-RLDA with Huber's estimator and RLDA (top row), and of test images that were correctly classified by M-RLDA but misclassified by RLDA (bottom row).

we randomly split all the training sample and use only part of it for training. We repeat this 100 times and report the average results for each classifier and for an increasing number of training data points. We observe that in cases where the number of training images is smaller, M-RLDA still performs slightly better than the RLDA.

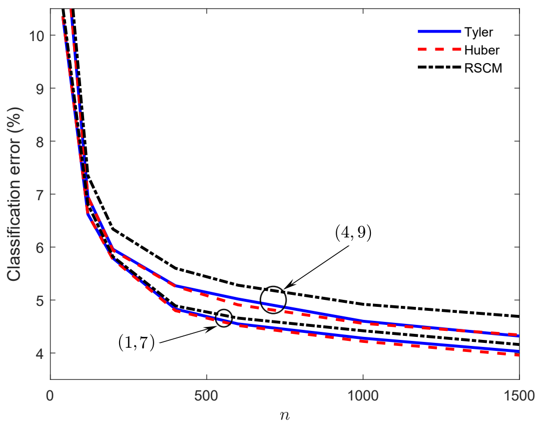

Fig. 5: Classification error of M-RLDA, relative to that of RLDA, as the number of training data points increases.

The fact that for this classification task RLDA generally does fairly well is likely because many of the MNIST images are relatively easy to separate. To further test discriminative power under 'noisier' conditions, we considered scenarios where the training images were subjected to salt-and-pepper (impulse) noise [28]. An example is shown in Fig. 6. Considering the class pairs $(3, 5)$ and $(1, 7)$, we compute the minimal testing classification error of RLDA and M-RLDA (with Huber's estimator) as a function of the noise density. All the training images are used, and the performance is measured on the available test set ($\sim$ 1000 samples per class). Results are averaged over 100 realizations and shown in Fig. 7. As the noise increases, M-RLDA achieves considerable performance improvement over RLDA, demonstrating the value of robust estimation.

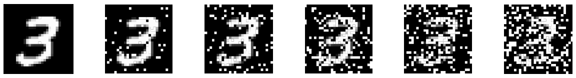

Fig. 6: A training image subjected to salt-and-pepper noise with different levels of noise density: from left to right, roughly 0%, 10%, 20%, 30%, 40%, and 50% of the pixels of the image are affected.

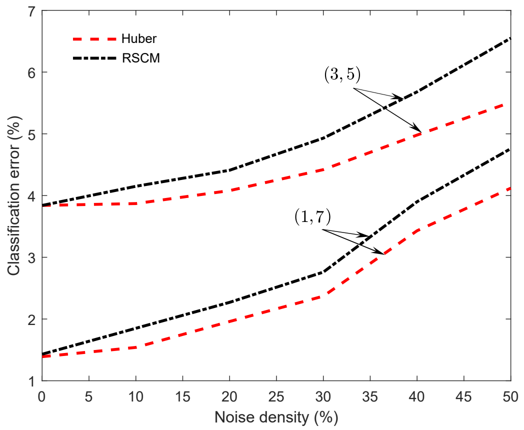

Fig. 7: Classification error of RLDA and M-RLDA (Huber) with training images subjected to salt-and-pepper-noise for two pairs of classes.

B. Phoneme data set

Next we perform simulations on the phoneme data set from the TIMIT Acoustic-Phonetic Continuous Speech Corpus. This data set was also used previously in [29]. The data sample is composed of log-periodograms of 32-millisecond speech frames sampled at a rate of 16 kHz. Each sample represents one of five phonemes: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". Here we focus on the classification between the phonemes "aa" and "ao". For a random split of the available sample between training and testing data, we compute the classification error of the RSCM, Tyler's estimator, and Huber's estimator, as we increase the number of training data points. The data not used for training is used for testing purposes. We repeat this 100 times. The average performance is reported in Fig. 8. Here, we observe that the classification performance is worse than for the MNIST example, suggesting that this is a more challenging classification problem. As the number of samples decreases, the performance degrades (as expected), while the benefits of robust estimation, particularly Huber, become apparent under data-limited settings.

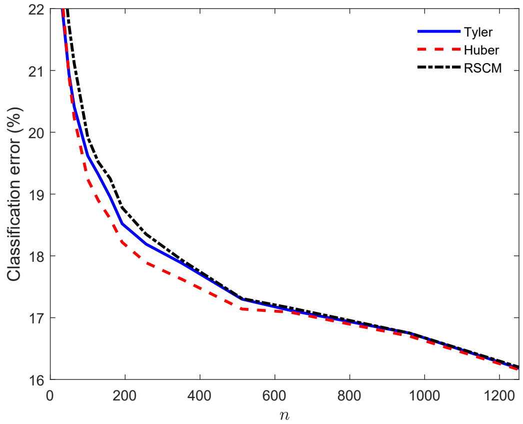

Fig. 8: Classification error of RLDA and M-RLDA on the phoneme data set as the number of training sample points increases.

V. Conclusions

We have studied the asymptotic performance of a robust version of regularized LDA classifiers, based on robust M-estimators of covariance within the Maronna's class of estimators. Considering the large-dimensional regime, we have shown that in the absence of outliers the robust LDA classifier performs asymptotically equivalently to standard regularized LDA (RLDA). This equivalence is subject to a transformation of the estimator's regularization parameter. We also put forward a practical algorithm to estimate the optimal regularization parameter for the robust RLDA classifier. Through simulations using the MNIST data set, we have demonstrated that the proposed robust classifiers can return superior performance when the data are corrupted by impulsive noise. Similar conclusions have been inferred in data-limited setting when using a phoneme data set. Such superiority depends on the choice of covariance estimator: when using Huber's estimator, the robust classifier performs consistently better than standard RLDA, but this was not the case for Tyler's estimator. This observation is consistent with previous studies which established differences in the behavior of these robust estimators under different outlying data models [15]. Similar systematic studies in the context of LDA remain an interesting avenue for future research. Another possible extension would be to consider the problem of multiple classes, for example in the context of Fisher's discriminant analysis [30].

VI. Appendix

A. Proof of Lemma 1

The proof follows along similar lines to that of [15, Theorem 2], but here we need to deal with re-centered data; i.e., for each class, the sample mean is subtracted from the data, introducing dependencies. To address this technical difficulty, we will adopt arguments similar to [19], with adaptations to account for the facts that two sample means (one for each class) are involved in the calculation, and we consider the family of u functions within Maronna's M-estimator class, while [19] focused on Tyler's estimator.

We first note that the existence and uniqueness of $\hat{\mathbf{S}}_N(\rho)$, as well as the continuity of the mapping $\rho \mapsto \gamma(\rho)$, follow from the proof of [15, Theorem 2]. The corresponding steps in that proof (see [15, Appendix A]) can be mirrored exactly, noticing that everything holds as well when the function $\phi : x \mapsto xu(x)$ is non-decreasing, rather than strictly increasing as it was assumed in [15].

Turn now to the main object of the proof, the a.s. convergence of $||\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)||$. Letting $\mathbf{Y}_i = [\mathbf{y}_1^{(i)}, \cdots, \mathbf{y}_{n_i}^{(i)}]$ be the data matrix of the n_i observations from class C_i, there exists $\mathbf{X}_i = \mathbf{C}_N^{1/2}\mathbf{Z}_i$ such that $\mathbf{Y}_i = \boldsymbol{\mu}_i \mathbf{1}_{n_i}^{\mathrm{T}} + \mathbf{X}_i$, with $\mathbf{Z}_i = [\mathbf{z}_1^{(i)}, \cdots, \mathbf{z}_{n_i}^{(i)}]$ a collection of independent random vectors with standard multivariate Gaussian distribution. In the following, similarly to $\tilde{\mathbf{y}}_j^{(i)}$ we will denote by $\tilde{\mathbf{x}}_j^{(i)}$ the re-centered version of sample $\mathbf{x}_j^{(i)} \in C_i$, $i = 0, 1$, $j \in \{0, \cdots, n_i\}$, that is:

$$\tilde{\mathbf{x}}_j^{(i)} = \mathbf{x}_j^{(i)} - \frac{1}{n_i}\mathbf{X}_i \mathbf{1}_{n_i}.$$

With this, (10) can be rewritten as

$$\hat{\mathbf{C}}_N(\rho) = \frac{(1-\rho)}{n-2}\sum_{i=0}^{1}\sum_{j=1}^{n_i} u\left((\tilde{\mathbf{x}}_j^{(i)})^{\mathrm{T}}\hat{\mathbf{C}}_N(\rho)^{-1}\tilde{\mathbf{x}}_j^{(i)}\right)\tilde{\mathbf{x}}_j^{(i)}(\tilde{\mathbf{x}}_j^{(i)})^{\mathrm{T}} + \rho\mathbf{I}_N. \tag{18}$$

Without loss of generality, we reorganize the samples $\mathbf{x}_j^{(i)}$ and the re-centered samples $\tilde{\mathbf{x}}_j^{(i)}$ for $j \in \{1, \cdots, n_i\}, i = 1, 2$ by writing $\mathbf{x}_q = \mathbf{x}_q^{(0)}$, $\tilde{\mathbf{x}}_q = \tilde{\mathbf{x}}_q^{(0)}$ for $q \in \{1, \cdots, n_0\}$ and $\mathbf{x}_{n_0+q} = \mathbf{x}_q^{(1)}$, $\tilde{\mathbf{x}}_{n_0+q} = \tilde{\mathbf{x}}_q^{(1)}$ for $q \in \{1, \cdots, n_1\}$.

With these notations, (18) can be rewritten in a more convenient form. Specifically, for a fixed $\rho \in I$,

$$\hat{\mathbf{C}}_N(\rho) = (1-\rho)\frac{1}{n-2}\sum_{q=1}^{n} v(\tilde{d}_q(\rho))\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}} + \rho\mathbf{I}_N, \tag{19}$$

where $\tilde{d}_q(\rho) \triangleq \frac{1}{N}\tilde{\mathbf{x}}_q^{\mathrm{T}}\hat{\mathbf{C}}_{(q)}(\rho)^{-1}\tilde{\mathbf{x}}_q$ and $\hat{\mathbf{C}}_{(q)}(\rho) = \hat{\mathbf{C}}_N(\rho) - (1-\rho)\frac{1}{n}v\left(\frac{1}{N}\tilde{\mathbf{x}}_q^{\mathrm{T}}\hat{\mathbf{C}}_N^{-1}(\rho)\tilde{\mathbf{x}}_q\right)\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}$.

Without loss of generality, we can further assume that $\tilde{d}_1(\rho) \leq \cdots \leq \tilde{d}_n(\rho)$. Then, using the fact that v is non-increasing, and that $\mathbf{A} \succeq \mathbf{B} \Rightarrow \mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$ for positive Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\begin{aligned}\tilde{d}_n(\rho) &= \frac{1}{N}\tilde{\mathbf{x}}_n^{\mathrm{T}}\left((1-\rho)\frac{1}{n-2}\sum_{q=1}^{n-1} v(\tilde{d}_q(\rho))\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}} + \rho\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_n \\ &\leq \frac{1}{N}\tilde{\mathbf{x}}_n^{\mathrm{T}}\left((1-\rho)\frac{1}{n-2}\sum_{q=1}^{n-1} v(\tilde{d}_n(\rho))\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}} + \rho\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_n,\end{aligned}$$

and since $\tilde{\mathbf{x}}_n \neq 0$ with probability 1,

$$\tilde{\mathbf{x}}_n^{\mathrm{T}}\left(\frac{1-\rho}{n-2}\sum_{q=1}^{n-1}\tilde{d}_n(\rho)v(\tilde{d}_n(\rho))\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}} + \rho\tilde{d}_n(\rho)\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_n \geq N. \tag{20}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2021.3075150, IEEE Transactions on Signal Processing

Similarly,

$$\tilde{\mathbf{x}}_1^{\mathrm{T}}\left(\frac{1-\rho}{n-2}\sum_{q=2}^{n}\tilde{d}_1(\rho)v(\tilde{d}_1(\rho))\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}+\rho\tilde{d}_1(\rho)\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_1\leq N.$$

We want to show that $\gamma(\rho)$ is a good deterministic approximation of $\tilde{d}_q(\rho)$ (for all $q=1,\cdots,n$), a result that we will later leverage to show that $\hat{\mathbf{C}}_N(\rho)$ in (19) is asymptotically similar to $\hat{\mathbf{S}}_N(\rho)$. Specifically, we will show that

$$\sup_{\rho\in\mathcal{I}}\max_{1\leq q\leq n}\left|\tilde{d}_q(\rho)-\gamma(\rho)\right|\xrightarrow{\text{a.s.}}0. \tag{21}$$

This will be proven by a contradiction argument: assume there exists a sequence $\{\rho_n\}_{n=1}^{\infty}$ over which $\tilde{d}_n(\rho_n)>\gamma(\rho_n)+l$ infinitely often, for some $l>0$ fixed. Consider a subsequence of $\{\rho_n\}_{n=1}^{\infty}$ such that $\rho_n\to\rho_1$; since $\{\rho_n\}_{n=1}^{\infty}$ is bounded, such subsequence exists by the Bolzano-Weierstrass theorem. On this subsequence and for all large n, (20) yields

$$1\leq\frac{\tilde{\mathbf{x}}_n^{\mathrm{T}}}{N}\left(\frac{1-\rho_n}{n-2}\sum_{q=1}^{n-1}\psi(\gamma(\rho_n)+l)\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}+\rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_n, \tag{22}$$

where $\psi(x)\triangleq xv(x)$ is a non-decreasing function. Denote

$$\tilde{e}_q\triangleq\frac{\tilde{\mathbf{x}}_q^{\mathrm{T}}}{N}\left(\frac{1-\rho_n}{n-2}\sum_{j\neq q}\psi(\gamma(\rho_n)+l)\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^{\mathrm{T}}+\rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\tilde{\mathbf{x}}_q,$$

so that, with (22), we have $\tilde{e}_n\geq 1$ on the chosen subsequence. It turns out that $\tilde{e}_n$ is asymptotically equivalent to

$$e_n\triangleq\frac{\mathbf{x}_n^{\mathrm{T}}}{N}\left(\frac{1-\rho_n}{n-2}\sum_{j\neq n}\psi(\gamma(\rho_n)+l)\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}+\rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\mathbf{x}_n$$

which involves the zero-mean data samples $\mathbf{x}_j$. Specifically:

Proposition 2. *As $N,n\to\infty$,*

$$\max_{1\leq q\leq n}|\tilde{e}_q-e_q|\xrightarrow{\text{a.s.}}0. \tag{23}$$

Proof. See Section VI-B. □

We can now leverage a fact, proven in [15, Theorem 2]: on the chosen subsequence and for any given $\rho_1\in(0,1]$, $e_n\xrightarrow{\text{a.s.}}e<1$. Due to Proposition 2, this implies that $\tilde{e}_n\xrightarrow{\text{a.s.}}e<1$, in contradiction with $\tilde{e}_n\geq 1$ from (22).

It follows that there is no sequence of ρ_n such that $\tilde{d}_n(\rho_n)>\gamma(\rho_n)+l$ infinitely often. Consequently, $\tilde{d}_n(\rho)\leq\gamma(\rho)+l$ for all large n a.s., uniformly on $\rho\in I$. Following the same strategy, we prove that $\tilde{d}_1(\rho)\geq\gamma(\rho)-l$ for all large n a.s. uniformly on $\rho\in I$. As this is true for arbitrarily small $l>0$, we then have that $\sup_{\rho\in I}\max_{1\leq q\leq n}|\tilde{d}_q(\rho)-\gamma(\rho)|\xrightarrow{\text{a.s.}}0$. By continuity of v, we also have

$$\sup_{\rho\in I}\max_{1\leq q\leq n}|v(\tilde{d}_q(\rho))-v(\gamma(\rho))|\xrightarrow{\text{a.s.}}0. \tag{24}$$

Now, note that

$$\sup_{\rho\in I}\left\|\hat{\mathbf{C}}_N(\rho)-\hat{\mathbf{S}}_N(\rho)\right\|\leq\left\|\frac{1}{n-2}\sum_{q=1}^{n}\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}\right\|$$
$$\times\sup_{\rho\in I}\max_{1\leq q\leq n}(1-\rho)\left|v(\tilde{d}_q(\rho))-v(\gamma(\rho))\right|. \tag{25}$$

We will show that the right-hand side of (25) goes to 0 a.s. With (24), this follows by showing that

$$\limsup_n\left\|\frac{1}{n-2}\sum_{q=1}^{n}\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}\right\|=<\infty\text{ a.s.}$$

Recall that $\tilde{\mathbf{x}}_j^{(i)}=\mathbf{x}_j^{(i)}-\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}$, so that

$$\frac{1}{n-2}\sum_{q=1}^{n}\tilde{\mathbf{x}}_q\tilde{\mathbf{x}}_q^{\mathrm{T}}=\frac{1}{n-2}\sum_{i=1}^{1}\sum_{j=1}^{n_i}\tilde{\mathbf{x}}_j^{(i)}(\tilde{\mathbf{x}}_j^{(i)})^{\mathrm{T}}=M_0+M_1,$$

with

$$M_i=\frac{1}{n-2}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)}\left(\mathbf{x}_j^{(i)}\right)^{\mathrm{T}}+\frac{1}{n-2}\sum_{j=1}^{n_i}\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\left(\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\right)^{\mathrm{T}}$$
$$-\frac{1}{n-2}\sum_{j=1}^{n_i}\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\left(\mathbf{x}_j^{(i)}\right)^{\mathrm{T}}-\frac{1}{n-2}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)}\left(\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\right)^{\mathrm{T}}. \tag{26}$$

We will show that the spectral norm of each term on the RHS of (26) is bounded for all large n a.s. For the first term, we have $\limsup_n\left\|\frac{1}{n-2}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)}\left(\mathbf{x}_j^{(i)}\right)^{\mathrm{T}}\right\|<\infty$ a.s. as a direct consequence of Assumption 3 and [31]. For the second term in (26), we have

$$\limsup_n\left\|\frac{1}{n-2}\sum_{j=1}^{n_i}\left(\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\right)\left(\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\right)^{\mathrm{T}}\right\|$$
$$\leq\limsup_n\left\|\frac{1}{n-2}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right\|<\infty\text{ a.s.} \tag{27}$$

For the third term (and similarly for the fourth term),

$$\left\|\frac{1}{n-2}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)}\left(\frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}\right)^{\mathrm{T}}\right\|$$
$$\overset{(a)}{=}\frac{n_i}{n-2}\left\|\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right)^{\mathrm{T}}\mathbf{C}_N\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right)\right\|$$
$$\leq\frac{n_i}{n-2}\|\mathbf{C}_N\|\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right)^{\mathrm{T}}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right),$$

where (a) uses the fact that $\mathbf{x}_j^{(i)}=\mathbf{C}_N^{1/2}\mathbf{z}_j^{(i)}$ for $j=1,\cdots,n_i$. By the law of large numbers, as $N,n\to\infty$,

$$\left|\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right)^{\mathrm{T}}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_j^{(i)}\right)-c_i\right|\xrightarrow{\text{a.s.}}0,$$

where we introduced $c_i \triangleq c/\pi_i$, the limiting ratio of the number of variables to the number of samples from class C_i. It then follows from Assumption 1 that

$$\limsup_n \left\| \frac{1}{n-2} \sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} \left(\frac{1}{n_i} \mathbf{X}_i \mathbf{1}_{n_i} \right)^{\mathrm{T}} \right\| \leq c_i \left\| \mathbf{C}_N \right\| < \infty \text{ a.s.}$$

Putting everything together, the spectral norm of M_i, $i = 0, 1$ is bounded for all large n a.s, which implies $\limsup_n \left\| \frac{1}{n-2} \sum_{q=1}^{n} \tilde{\mathbf{x}}_q \tilde{\mathbf{x}}_q^{\mathrm{T}} \right\| < \infty$ a.s. Together with (24) and (25), we eventually have

$$\sup_{\rho \in \mathcal{I}} \left\| \hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho) \right\| \xrightarrow{\text{a.s.}} 0.$$

B. Proof of Proposition 2

For $q \in \{1, \cdots, n\}$, denote

$$\tilde{\mathbf{E}}_q = \left(\frac{1-\rho_n}{n-2} \sum_{j \neq q} \psi(\gamma(\rho_n) + l) \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^{\mathrm{T}} + \rho_n (\gamma(\rho_n) + l) \mathbf{I}_N \right)^{-1}$$

and

$$\mathbf{E}_q = \left(\frac{1-\rho_n}{n-2} \sum_{j \neq q} \psi(\gamma(\rho_n) + l) \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} + \rho_n (\gamma(\rho_n) + l) \mathbf{I}_N \right)^{-1},$$

such that

$$\tilde{e}_q = \frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \tilde{\mathbf{E}}_q \tilde{\mathbf{x}}_q, \qquad e_q = \frac{1}{N} \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q .$$

We have, for all $\zeta > 0$,

$$\begin{aligned} P\left(\max_{1 \leq q \leq n} |\tilde{e}_q - e_q| > \zeta \right) &\overset{(a)}{\leq} \sum_{q=1}^{n} P\left(|\tilde{e}_q - e_q| > \zeta \right) \\ &\overset{(b)}{\leq} n \frac{E[|\tilde{e}_q - e_q|^p]}{\zeta^p}, \end{aligned} \tag{28}$$

where (a) and (b) are respectively due to Boole's and Markov's inequalities. In the following we will prove that

$$E[|\tilde{e}_q - e_q|^p] \leq \frac{K_p}{N^p} \tag{29}$$

for some $p \geq 1$, where K_p depends on p but not on N. With this result, the proof of Proposition 2 then follows by taking $p > 2$ and applying the Borel-Cantelli lemma.

In proving (29), we start by fixing $q \in \{1, \cdots, n\}$ such that the associated $\mathbf{x}_q$ belongs to class C_0. The case where $\mathbf{x}_q$ belongs to class C_1 can be handled similarly. Write

$$\begin{aligned} \tilde{e}_q - e_q &= \frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \tilde{\mathbf{E}}_q \tilde{\mathbf{x}}_q - \frac{1}{N} \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \\ &= \overbrace{\frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} (\tilde{\mathbf{E}}_q - \mathbf{E}_q) \tilde{\mathbf{x}}_q}^{-D} + \overbrace{\frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \mathbf{E}_q \tilde{\mathbf{x}}_q - \frac{1}{N} \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q}^{-A-B+C}, \end{aligned}$$

with

$$\begin{aligned} A &\triangleq \frac{1}{N} \frac{1}{n_0} \mathbf{1}_{n_0}^{\mathrm{T}} \mathbf{X}_0^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \\ B &\triangleq \frac{1}{N} \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \frac{1}{n_0} \mathbf{X}_0 \mathbf{1}_{n_0} \\ C &\triangleq \frac{1}{N} \frac{1}{n_0} \mathbf{1}_{n_0}^{\mathrm{T}} \mathbf{X}_0^{\mathrm{T}} \mathbf{E}_q \frac{1}{n_0} \mathbf{X}_0 \mathbf{1}_{n_0} \\ D &\triangleq D_1 + D_2 + D_3 , \end{aligned}$$

with

$$\begin{aligned} D_1 =& \frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \tilde{\mathbf{E}}_q (1-\rho_n) \psi(\gamma(\rho_n) + l) \frac{1}{n-2} \\ & \times \frac{1}{n_0^2} \sum_{\substack{1 \leq j \leq n_0 \\ j \neq q}} \mathbf{X}_0 \mathbf{1}_{n_0} \mathbf{1}_{n_0}^{\mathrm{T}} \mathbf{X}_0^{\mathrm{T}} \mathbf{E}_q \tilde{\mathbf{x}}_q \\ D_2 =& - \frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \tilde{\mathbf{E}}_q (1-\rho_n) \psi(\gamma(\rho_n) + l) \frac{1}{n-2} \\ & \times \frac{1}{n_0} \sum_{\substack{1 \leq j \leq n_0 \\ j \neq q}} (\mathbf{x}_j \mathbf{1}_{n_0}^{\mathrm{T}} \mathbf{X}_0^{\mathrm{T}} - \mathbf{X}_0 \mathbf{1}_{n_0} \mathbf{x}_j^{\mathrm{T}}) \mathbf{E}_q \tilde{\mathbf{x}}_q \\ D_3 =& \frac{1}{N} \tilde{\mathbf{x}}_q^{\mathrm{T}} \tilde{\mathbf{E}}_q (1-\rho_n) \psi(\gamma(\rho_n) + l) \frac{1}{n-2} \times \\ & \left(\frac{1}{n_1^2} \sum_{j=n_0+1}^{n} \mathbf{X}_1 \mathbf{1}_{n_1} \mathbf{1}_{n_1}^{\mathrm{T}} \mathbf{X}_1^{\mathrm{T}} - \frac{1}{n_1} \sum_{j=n_0+1}^{n} (\mathbf{x}_j \mathbf{1}_{n_1}^{\mathrm{T}} \mathbf{X}_1^{\mathrm{T}} - \mathbf{X}_1 \mathbf{1}_{n_1} \mathbf{x}_j^{\mathrm{T}}) \right) \\ & \times \mathbf{E}_q \tilde{\mathbf{x}}_q , \end{aligned}$$

obtained using the resolvent identity and the fact that $\tilde{\mathbf{x}}_q = \mathbf{x}_q - \frac{1}{n_0} \mathbf{X}_0 \mathbf{1}_{n_0}$. From Minkowski's inequality,

$$\begin{aligned} & E\left[|\tilde{e}_q - e_q|^p \right] \leq \\ & \left(E^{1/p}[|A|^p] + E^{1/p}[|B|^p] + E^{1/p}[|C|^p] + E^{1/p}[|D|^p] \right)^p . \end{aligned} \tag{30}$$

Thus, it is enough to show that $E[|A|^p] \leq \frac{K_{pA}}{N^p}$, $E[|B|^p] \leq \frac{K_{pB}}{N^p}$, $E[|C|^p] \leq \frac{K_{pC}}{N^p}$ and $E[|D|^p] \leq \frac{K_{pD}}{N^p}$.

Denote $\mathbf{X}^{(q)}$ the data matrix $\mathbf{X}$ from which the data sample $\mathbf{x}_q$ was removed. We start by writing

$$\begin{aligned} E[|A|^p] &= \frac{1}{N^p} \frac{1}{n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{X}_0 \mathbf{1}_{n_0} \mathbf{1}_{n_0}^{\mathrm{T}} \mathbf{X}_0^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] \\ &= \frac{1}{N^p} \frac{1}{n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \left(\mathbf{x}_q + \mathbf{X}_0^{(q)} \mathbf{1}_{n_0 - 1} \right) \right. \right. \\ & \qquad \left. \left. \times \left(\mathbf{x}_q + \mathbf{X}_0^{(q)} \mathbf{1}_{n_0 - 1} \right)^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] . \end{aligned} \tag{31}$$

Developing (31) and using Jensen's inequality, we can write $E[|A|^p] \overset{(a)}{\leq} A_1 + A_2 + A_3 + A_4$ with

$$\begin{aligned} A_1 &= \frac{4^{p/2-1}}{N^p n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] \\ A_2 &= \frac{4^{p/2-1}}{N^p n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{X}_0^{(q)} \mathbf{1}_{n_0 - 1} \mathbf{1}_{n_0 - 1}^{\mathrm{T}} \mathbf{X}_0^{(q)\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] \\ A_3 &= \frac{4^{p/2-1}}{N^p n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{X}_0^{(q)} \mathbf{1}_{n_0 - 1} \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] \\ A_4 &= \frac{4^{p/2-1}}{N^p n_0^p} E\left[\left| \mathbf{x}_q^{\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \mathbf{1}_{n_0 - 1}^{\mathrm{T}} \mathbf{X}_0^{(q)\mathrm{T}} \mathbf{E}_q \mathbf{x}_q \right|^{p/2} \right] . \end{aligned}$$

For term A_1,

$$A_1 = \frac{1}{N^p}\frac{1}{n_0^p}\cdot 4^{p/2-1} E\left[\left|\mathbf{x}_q^{\mathrm{T}}\mathbf{E}_q\mathbf{x}_q\right|^p\right]$$
$$\leq \frac{1}{N^p}\frac{1}{n_0^p}\cdot 4^{p/2-1} E\left[\|\mathbf{z}_q\|^{2p}\|\mathbf{E}_q\|^p\|\|\mathbf{C}_N\|^p\right] .$$

We have

$$\|\mathbf{E}_q\|^p \leq \frac{1}{(\gamma(\rho_n)+l)^p\rho_n^p} ,$$

and by Minkowski's inequality,

$$E[\|\mathbf{z}_q\|^{2p}] = E(\sum_{j=1}^{N} z_{j,q}^2)^p \leq N^p E|z_{1,q}|^{2p} \leq K_p N^p .$$

Thus

$$A_1 \leq \frac{1}{N^p}\frac{K_p\|\mathbf{C}_N\|^p 4^{p/2-1}c_N^p}{(\gamma(\rho_n)+l)^p\rho_n^p} \leq \frac{K_{pA_1}}{N^p} .$$

Next, consider A_2:

$$A_2 \overset{(a)}{\leq} \frac{2^{3p/2-3}}{N^p n_0^p}\left(E\left[\left|\mathbf{x}_q^{\mathrm{T}}\mathbf{Q}_N\mathbf{x}_q - \mathrm{Tr}\left[\mathbf{Q}_N\right]\right|^{p/2}\right]\right.$$
$$\left.+E\left[\left|\mathrm{Tr}\left[\mathbf{Q}_N\right]\right|^{p/2}\right]\right)$$
$$\overset{(b)}{\leq} \frac{K_p}{N^p n_0^p}E\left[\left(E^{p/4}|x_{1,q}|^4\,\mathrm{Tr}[\mathbf{Q}_N\mathbf{Q}_N^{\mathrm{T}}]\right)^{p/4}\right.$$
$$\left.+E|x_{1,q}|^p\,\mathrm{Tr}[(\mathbf{Q}_N\mathbf{Q}_N^{\mathrm{T}})^{p/4}]+E\left|\mathrm{Tr}\left[\mathbf{Q}_N\right]\right|^{p/2}\right]$$
$$= \frac{K_p}{N^p n_0^p}\left(E^{p/4}|x_{1,q}|^4+E|x_{1,q}|^p+1\right)E\|\mathbf{E}_q\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\|^p$$
$$\leq \frac{K_p}{N^p n_0^p}\left(E^{p/4}|x_{1,q}|^4+E|x_{1,q}|^p+1\right)$$
$$\times E\left(\|\mathbf{E}_q\|^p\|\mathbf{C}_N\|^{p/2}\left\|\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\right\|^p\right)$$
$$\leq \frac{1}{N^p}\frac{K_p\|\mathbf{C}_N\|^{p/2}\left(E^{p/4}|x_{1,q}|^4+E|x_{1,q}|^p+1\right)}{(\gamma(\rho_n)+l)^p\rho_n^p}$$
$$\times E\left\|\frac{1}{n_0}\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\right\|^p$$
$$\leq K_{pA_2}/N^p ,$$

where $\mathbf{Q}_N = \mathbf{E}_q\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\mathbf{1}_{n_0-1}^{\mathrm{T}}\mathbf{X}_0^{(q)\mathrm{T}}\mathbf{E}_q$, (a) follows from Jensen's inequality and (b) follows from the trace lemma [32, Lemma B.26].

For A_3,

$$A_3 \leq \frac{4^{p/2-1}}{N^p n_0^p}E^{1/2}\left[\left|\mathbf{x}_q^{\mathrm{T}}\mathbf{E}_q\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\right|^p\right]E^{1/2}\left[\left|\mathbf{x}_q^{\mathrm{T}}\mathbf{E}_q\mathbf{x}_q\right|^p\right].$$

As we have $A_1 \leq K_{pA_1}/N^p$ and $A_2 \leq K_{pA_2}/N^p$, we obtain $A_3 \leq K_{pA_3}/N^p$. Following the same reasoning as for A_3, we also get $A_4 \leq K_{pA_4}/N^p$. Therefore, we obtain

$$E[|A|^p] \leq A_1 + A_2 + A_3 + A_4 \leq K_{pA}/N^p .$$

The same reasoning gives $E[|B|^p] \leq K_{pB}/N^p$.

As for $|C|$, we have

$$E[|C|^p] \leq \frac{1}{N^p}E\left[\left\|\frac{1}{n_0}\mathbf{Z}_0\mathbf{1}_{n_0}\right\|^{2p}\|\mathbf{C}_N\|^p\|\mathbf{E}_q^p\|\right]$$
$$\leq \frac{\|\mathbf{C}_N\|^p}{N^p(\gamma(\rho_n)+l)^p\rho_n^p}E\left[\left\|\frac{1}{n_0}\mathbf{Z}_0\mathbf{1}_{n_0}\right\|^{2p}\right]$$
$$\leq \frac{K_{pC}}{N^p}.$$

Let us now analyze $|D|$. We will show that $E[|D_1|^p] \leq K_{pD_1}/N^p$, $E[|D_2|^p] \leq K_{pD_2}/N^p$, and $E[|D_3|^p] \leq K_{pD_3}/N^p$, which will prove that $E[|D|^p] \leq K_{pD}/N^p$.

We start with D_1, the analysis of D_2 and D_3 following similarly. We have

$$E[|D_1|^p] \overset{(a)}{\leq} \frac{1}{N^p}\left(\psi(\gamma(\rho_n)+l)\right)^p$$
$$\times E^{1/2}[|D_{1a}|^{2p}]E^{1/2}[|D_{1b}|^{2p}]$$
$$\overset{(b)}{\leq} \frac{1}{N^p}\left(\psi(\gamma(\rho_n)+l)\right)^p$$
$$\times(E^{1/2p}[|D_{1c}|^{2p}]+E^{1/2p}[|D_{1d}|^{2p}])^p E^{1/2}[|D_{1b}|^{2p}] \tag{32}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) follows from Minkowski's inequality, and

$$D_{1a} = \left(\mathbf{x}_q - \frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}\right)^{\mathrm{T}}\tilde{\mathbf{E}}_q\frac{n_0-1}{n_0^2}\mathbf{X}_0\mathbf{1}_{n_0}$$
$$D_{1b} = \frac{1}{n-2}\mathbf{1}_{n_0}^{\mathrm{T}}\mathbf{X}_0^{\mathrm{T}}\mathbf{E}_q\left(\mathbf{x}_q - \frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}\right)$$
$$D_{1c} = \mathbf{x}_q^{\mathrm{T}}\tilde{\mathbf{E}}_q\frac{n_0-1}{n_0^2}\mathbf{X}_0\mathbf{1}_{n_0}$$
$$D_{1d} = \left(\frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}\right)^{\mathrm{T}}\tilde{\mathbf{E}}_q\frac{n_0-1}{n_0^2}\mathbf{X}_0\mathbf{1}_{n_0} .$$

We will prove that $E[|D_{1b}|^{2p}] \leq K_{pb}$, $E[|D_{1c}|^{2p}] \leq K_{pc}$ and $E[|D_{1d}|^{2p}] \leq K_{pd}$. Following the analysis of $E[|A|^p]$ and $E[|C|^p]$, we obtain $E[|D_{1b}|^{2p}] \leq K_{pb}$. For $E[|D_{1d}|^{2p}]$,

$$E[|D_{1d}|^{2p}] \leq \frac{1}{n_0^{2p}}E\left[\left\|\tilde{\mathbf{E}}_q\right\|^{2p}\|\mathbf{C}_N\|^{2p}\left\|\frac{1}{n_0}\mathbf{Z}_0\mathbf{1}_{n_0}\right\|^{4p}\right]$$
$$\leq \frac{\|\mathbf{C}_N\|^{2p}}{n_0^{2p}(\gamma(\rho_n)+l)^{2p}\rho_n^{2p}}E\left\|\frac{1}{n_0}\mathbf{Z}_0\mathbf{1}_{n_0}\right\|^{4p}$$
$$\leq K_{pd} .$$

Let us analyze D_{1c}. As $\mathbf{x}_q$ is not independent of $\tilde{\mathbf{E}}_q$, we cannot follow the same procedure as for our analysis of A to determine the order of $E[|D_{1c}|^{2p}]$. To proceed, we decompose $\tilde{\mathbf{E}}_q$ into two parts, one that is independent of $\mathbf{x}_q$, and the remainder. Recalling that $\mathbf{x}_q$ belongs to class C_0, we first write $\sum_{j\neq q}\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^{\mathrm{T}} = \mathbf{E}+\mathbf{F}$, where

$$\mathbf{E}=\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^{\mathrm{T}}+\mathbf{X}_0^{(q)}(\mathbf{X}_0^{(q)})^{\mathrm{T}}-\frac{n_0+1}{n_0^2}\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}(\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1})^{\mathrm{T}}$$
$$\mathbf{F}=-\frac{1}{n_0^2}\mathbf{x}_q(\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1})^{\mathrm{T}}-\frac{1}{n_0^2}\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\mathbf{x}_q^{\mathrm{T}}+\frac{n_0-1}{n_0^2}\mathbf{x}_q\mathbf{x}_q^{\mathrm{T}} ,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2021.3075150, IEEE Transactions on Signal Processing

such that $\mathbf{E}$ is independent of $\mathbf{x}_q$. Then, using the resolvent identity, we rewrite D_{1c} as:

$$D_{1c} = \mathbf{x}_q^{\mathrm{T}}\mathbf{G}\frac{n_0-1}{n_0^2}\mathbf{X}_0\mathbf{1}_{n_0} + \mathbf{x}_q^{\mathrm{T}}\mathbf{H}\frac{n_0-1}{n_0^2}\mathbf{X}_0\mathbf{1}_{n_0}, \quad (33)$$

where

$$\mathbf{G} = \left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{1}{n-2}\mathbf{E} + \rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}$$
$$\mathbf{H} = -\tilde{\mathbf{E}}_q\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{1}{n-2}\mathbf{F}\right)$$
$$\times\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{1}{n-2}\mathbf{E} + \rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}.$$

Using Jensen's inequality,

$$E[|D_{1c}|^{2p}] \leq 2^{2p-1}\left(E[|G|^{2p}] + E[|H|^{2p}]\right),$$

where G and H are the two terms on the RHS of (33).

Using similar reasoning and calculus as before (see e.g., terms A and B above), we can prove that $E[|G|^{2p}] \leq K_{pG}$.

For $E[|H|^{2p}]$, we can write the equation (33) at the top of the next page, which follows from the Cauchy-Schwarz inequality, and since

$$\left\|\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{\mathbf{E}+\mathbf{F}}{n-2} + \rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\right\|$$

and

$$\left\|\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{\mathbf{E}}{n-2} + \rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\right\|$$

are both less than $\frac{1}{\rho_n(\gamma(\rho_n)+l)}$. Let us prove that $E\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}\right\|^{8p} \leq K_{pF}$. Denoting

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3,$$

with

$$\mathbf{F}_1 = -\frac{1}{n_0^2}\mathbf{x}_q(\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1})^{\mathrm{T}}$$
$$\mathbf{F}_2 = -\frac{1}{n_0^2}\mathbf{X}_0^{(q)}\mathbf{1}_{n_0-1}\mathbf{x}_q^{\mathrm{T}}$$
$$\mathbf{F}_3 = \frac{n_0-1}{n_0^2}\mathbf{x}_q\mathbf{x}_q^{\mathrm{T}},$$

we have (34) shown at the top of the next page. Combining (33) with (34), we have proved that $E[|H|^{2p}] \leq K_{pH}$, from which $E[|D_{1c}|^{2p}] \leq K_{pc}$ follows. With $E[|D_{1b}|^{2p}] \leq K_{pb}$, $E[|D_{1c}|^{2p}] \leq K_{pc}$, and $E[|D_{1d}|^{2p}] \leq K_{pd}$, (32) leads to $E[|D_1|^p] \leq K_{pD_1}/N^p$.

Similarly, we can also obtain $E[|D_2|^p] \leq K_{pD_2}/N^p$, $E[|D_3|^p] \leq K_{pD_3}/N^p$, and $E[|D_4|^p] \leq K_{pD_4}/N^p$. As $D = D_1 + D_2 + D_3 + D_4$, the Minkowski inequality gives $E[|D|^p] \leq K_{pD}/N^p$. Using the pth-moment bounds for A, B, C and D in (30), we have shown that (29) holds, and the proof is complete.

C. Estimate of the classification error of RLDA

Lemma 3. *[5, Theorem 2], [6, Theorem 10] Under Assumptions 1 and 2, write*

$$\hat{\epsilon}_i^{\mathrm{RLDA}}(\kappa,\beta) = \Phi\left(\frac{(-1)^{i+1}G_i + \frac{(n_0+n_1-2)\hat{\delta}}{n_i} + (-1)^i\kappa\log\frac{\pi_1}{\pi_0}}{\sqrt{(1+\beta\hat{\delta})^2 D}}\right)$$

with

$$\hat{\delta} = \frac{1}{\beta}\frac{\frac{N}{n_0+n_1-2} - \frac{\mathrm{Tr}[\hat{\mathbf{R}}(\beta)^{-1}]}{n_0+n_1-2}}{1 - \frac{N}{n_0+n_1-2} + \frac{\mathrm{Tr}[\hat{\mathbf{R}}(\beta)^{-1}]}{n_0+n_1-2}},$$

and with $G_i = G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta)^{-1})$ and $D = D(\hat{\mathbf{R}}, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{R}}(\beta)^{-1})$ defined in (7) and (8), respectively. Then for all $\beta > 0$, as $N, n \to \infty$,

$$\hat{\epsilon}_i^{\mathrm{RLDA}}(\kappa,\beta) - \epsilon_i^{\mathrm{RLDA}}(\kappa,\beta) \xrightarrow{\text{a.s.}} 0.$$

References

[1] N. Auguin, D. Morales-Jimenez, and M. R. McKay, "Robust linear discriminant analysis using Tyler's estimator: Asymptotic performance characterization," in *IEEE Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Brighton (UK), May 2019, pp. 5317–5321.

[2] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004, vol. 544.

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

[4] J. H. Friedman, "Regularized discriminant analysis," *J. Am. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[5] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2804–2814, Apr. 2015.

[6] K. Elkhalil, A. Kammoun, R. Couillet, T. Alnaffouri, and M. Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 2464–2479, Apr. 2020.

[7] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *Am. Stat.*, vol. 37, no. 1, pp. 36–48, 1983.

[8] K. Elkhalil, A. Kammoun, R. Calderbank, T. Y. Al-Naffouri, and M.-S. Alouini, "Asymptotic performance of linear discriminant analysis with random projections," in *IEEE Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019, pp. 3472–3476.

[9] H. Sifaou, A. Kammoun, and M.-S. Alouini, "Improved LDA classifier based on spiked models," in *IEEE Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, 2018, pp. 1–5.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data*. Wiley, New York, 1994, vol. 3.

[11] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6059–6070, Nov. 2014.

[12] R. A. Maronna and V. J. Yohai, "Robust estimation of multivariate location and scatter," *Wiley StatsRef: Statistics Reference Online*, 1976.

[13] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, Jul. 2004.

[14] E. Ollila, I. Soloveychik, D. E. Tyler, and A. Wiesel, "Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization," *arXiv preprint arXiv:1608.08126*, 2016.

$$\begin{aligned}
E[|H|^{2p}] &\le ((1-\rho_n)\psi(\gamma(\rho_n)+l))^{2p} \\
&\quad \times E\left[\|\mathbf{C}_N\|^{2p}\left\|\frac{1}{\sqrt{n-2}}\mathbf{z}_q\right\|^{2p}\left\|\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{1}{n-2}\left(\mathbf{E}+\mathbf{F}\right)+\rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\right\|^{2p}\right. \\
&\quad \times \left.\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}\right\|^{2p}\left\|\left((1-\rho_n)\psi(\gamma(\rho_n)+l)\frac{1}{n-2}\mathbf{E}+\rho_n(\gamma(\rho_n)+l)\mathbf{I}_N\right)^{-1}\right\|^{2p}\left\|\frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}\right\|^{2p}\right] \\
&\overset{(a)}{\le} (\psi(\gamma(\rho_n)+l))^{2p}\,\frac{\|\mathbf{C}_N\|^{2p}}{\rho_n^{4p}(\gamma(\rho_n)+l)^{4p}}E^{1/2}\left\|\frac{1}{\sqrt{n-2}}\mathbf{z}_q\right\|^{4p}E^{1/4}\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}\right\|^{8p}E^{1/4}\left\|\frac{1}{n_0}\mathbf{X}_0\mathbf{1}_{n_0}\right\|^{8p}. \qquad (33)
\end{aligned}$$

$$\begin{aligned}
E\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}\right\|^{8p} &\overset{(a)}{\le} 3^{8p-1}\left(E\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}_1\right\|^{8p}+E\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}_2\right\|^{8p}+E\left\|\frac{1}{\sqrt{n-2}}\mathbf{F}_3\right\|^{8p}\right) \\
&\le 3^{8p-1}\left(\frac{2}{n_0^{8p}}\|\mathbf{C}_N\|^{8p}E\left[\left\|\frac{1}{\sqrt{n-2}}\mathbf{z}_q\right\|^{8p}\left\|\frac{1}{n_0}\mathbf{Z}_0^{(q)}\mathbf{1}_{n_0-1}\right\|^{8p}+n^{4p}\left\|\frac{1}{\sqrt{n-2}}\mathbf{z}_q\right\|^{16p}\right]\right)\le K_{pF}. \qquad (34)
\end{aligned}$$

[15] N. Auguin, D. Morales-Jimenez, M. R. McKay, and R. Couillet, "Large-dimensional behavior of regularized Maronna's M-estimators of covariance matrices," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3529–3542, Jul. 2018.

[16] N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *J. of the Royal Statist. Society: Series C (Applied Statist.)*, vol. 29, no. 3, pp. 231–237, 1980.

[17] S. F. Møller, J. von Frese, and R. Bro, "Robust methods for multivariate data analysis," *J. of Chemometrics: A J. of the Chemometrics Society*, vol. 19, no. 10, pp. 549–563, 2005.

[18] R. Couillet and M. R. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Multivar. Anal.*, vol. 131, pp. 99–120, Oct. 2014.

[19] L. Yang, R. Couillet, and M. R. McKay, "A robust statistics approach to minimum variance portfolio optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6684–6697, Dec. 2015.

[20] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Stat.*, pp. 234–251, 1987.

[21] P. J. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 1248–1251.

[22] D. Morales-Jimenez, R. Couillet, and M. R. McKay, "Large dimensional analysis of robust M-estimators of covariance with outliers," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5784–5797, Jul. 2015.

[23] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer Series in Statistics New York, 2001, vol. 1.

[24] R. Couillet, Z. Liao, and X. Mai, "Classification asymptotics in the random matrix regime," in *26th IEEE European Signal Processing Conference (EUSIPCO)*, Rome (Italy), 2018, pp. 1875–1879.

[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[26] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *The Annals of Statistics*, vol. 19, no. 4, pp. 2102–2119, 1991.

[27] Y. LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[28] A. C. Bovik, *Handbook of Image and Video Processing*. Academic Press, 2010.

[29] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, pp. 73–102, 1995.

[30] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenic.*, vol. 7, no. 2, pp. 179–188, 1936.

[31] Z.-D. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices," *Ann. Probab.*, pp. 316–345, 1998.

[32] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010, vol. 20.