ROBUST LINEAR DISCRIMINANT ANALYSIS USING TYLER'S ESTIMATOR: ASYMPTOTIC PERFORMANCE CHARACTERIZATION

Nicolas Auguin[⋆]*, David Morales-Jimenez*[†]*, Matthew R. McKay*[⋆]

[⋆]Hong Kong University of Science and Technology, ECE Department, Hong Kong
[†]Queen's University Belfast, ECIT Institute, Belfast, United Kingdom

ABSTRACT

We consider a robust version of regularized discriminant analysis (RDA) classifiers to account for potential spurious or mislabeled observations in the training data set. To build a robust discriminant rule, a robust estimation of the covariance matrix is essential. In this work, we propose to use a regularized version of Tyler's covariance estimator, in the regime where both the number of variables and the number of training samples are large and of similar order. Building upon fundamental results from random matrix theory, we show that the robust classifier is asymptotically equivalent to traditional, non-robust classifiers when the training data is free from outliers. Simulations on synthetic and real datasets confirm our theoretical observations and further attest to the benefits brought by the robust classifier when the data is corrupted by outliers.

Index Terms— Robust estimation, covariance matrices, linear discriminant analysis

1. INTRODUCTION

Discriminant analysis is a common classification method used in statistics, machine learning, and pattern recognition to identify the combination of features that best separate a number of classes of events or objects [1]. Discriminant analysis belongs to the wide class of parametric classification methods [2], which assume that the data follows a certain distribution (for example, a Gaussian mixture model). Based on labeled data, from which estimates of the class means and covariance matrices are obtained, a discrimination rule is learned, which is then used to determine the class which an unseen data sample most likely pertains to.

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, classical estimators of covariance matrices like the sample covariance matrix (SCM) typically fail. To solve this issue, regularized versions of discriminant analysis have been proposed [3], based on regularized versions of the SCM. Regularized discriminant analysis has since then established itself as the go-to choice in practice. In a series of recent works [4, 5], RDA has been studied from a random matrix theory perspective. Specifically, when the number of variables and the number of 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 RDA.

A common problem arising in discriminant analysis is the fact that the data, although assumed to arise from a Gaussian mixture model, is often not Gaussian: data samples may instead follow a heavy-tailed distribution, or some of the training samples may be outliers, the origin of which can be diverse (for a review, see [6]). This is a critical issue in practice; as the discriminant rule learned from the training data necessitates the estimation of the covariance matrix of the data, when outliers are present and/or if the data is not Gaussian, using a simple estimate like the sample covariance matrix can lead to underwhelming results. It is then natural to aim for a robust estimation of the covariance matrix. In this work, we consider a regularized version of Tyler's estimator of covariance, proposed in [7, 8], and we study the asymptotic performance of the associated discriminant rule in the context of linear discriminant analysis (LDA). To do so, we build on a series of recent works concerned with the performance analysis of RDA [4, 5] and the asymptotic behavior of Tyler's estimator [9, 10]. We first demonstrate that, when no outliers are present in the data, there is no performance loss when using regularized Tyler's estimator rather than the regularized SCM (RSCM). We then validate this observation with simulations on synthetic data and on the MNIST dataset. It is also shown empirically that, when the data is corrupted by a certain type of outliers, there is a clear benefit in using this robust estimator rather than the RSCM.

Notation: The superscript $^{\mathrm{T}}$ means transpose. $\mathrm{tr}\mathbf{A}$ represents the trace of the matrix $\mathbf{A}$. $||.||$ denotes the Euclidean norm for vectors and the spectral norm for matrices. $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution. The arrow $\xrightarrow{\text{a.s.}}$ designates the almost sure convergence of a random variable.

2. DISCRIMINANT ANALYSIS

2.1. Model

In discriminant analysis, a discriminant rule is determined so as to decide to which class a given (unseen) data vector most likely belongs. Such rule is built based on an available training data set composed of n samples pertaining to, say, 2 classes, C_0, C_1. Assume that the $n_i > 0$ observations from

The work of N. Auguin and M. R. McKay was supported by the Hong Kong Research Grants Council under grant number 16203315.

class C_i are independent and sampled from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i \in \mathbb{R}^{N\times 1}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$, with $\boldsymbol{\Sigma} \succeq \mathbf{0}$.

The linear discriminant rule consists in assigning a new (test) measurement $\mathbf{x}$ to class C_k if

$$k = \underset{i\in\{0,1\}}{\operatorname{argmin}} \left\{ (\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \log\pi_i^{(n)} \right\}, \quad (1)$$

with $\pi_i^{(n)} \triangleq \frac{n_i}{n}$ the *a priori* probability of class C_i.The LDA rule therefore assigns the label 0 to observation $\mathbf{x}$ if $\mathbb{P}(\mathbf{x}|\mathbf{x}\in C_0) > \mathbb{P}(\mathbf{x}|\mathbf{x}\in C_1)$, and the label 1 otherwise.

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

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

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

where

$$\hat{\mathbf{S}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i-1}(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)^{\mathrm{T}}, \quad i\in\{0,1\}.$$

The estimator $\hat{\mathbf{S}}$ is usually referred to as the "pooled SCM" in the literature. The main issue with these sample estimates is that they are known to perform poorly when the number of samples is of the same order as the number of variables (or possibly smaller). In practice, to alleviate the potential ill-conditioning of the sample covariance matrix, a regularized estimator, referred to as RSCM, is typically used [3]

$$\overline{\boldsymbol{\Sigma}}(\rho) = \mathbf{I}_N + \rho\hat{\mathbf{S}}. \quad (2)$$

2.2. Classification error of linear discriminant analysis

Let $\hat{\mathbf{H}}$ be an estimator of $\boldsymbol{\Sigma}^{-1}$. Then, conditioned on the training data $\mathbf{x}_1,\cdots,\mathbf{x}_n$, the probability of misclassification is given by [4]

$$\epsilon^{\mathrm{LDA}}(\hat{\mathbf{H}}) = \pi_0^{(n)}\epsilon_0^{\mathrm{LDA}}(\hat{\mathbf{H}}) + \pi_1^{(n)}\epsilon_1^{\mathrm{LDA}}(\hat{\mathbf{H}}) \quad (3)$$

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

$$\epsilon_i^{\mathrm{LDA}}(\hat{\mathbf{H}}) = \Phi\left(\frac{(-1)^{i+1}G(\hat{\mathbf{H}}) + (-1)^i\log\frac{\pi_1^{(n)}}{\pi_0^{(n)}}}{\sqrt{D(\hat{\mathbf{H}})}}\right), \quad (4)$$

where

$$G(\hat{\mathbf{H}}) = \left(\boldsymbol{\mu}_i - \frac{\hat{\mathbf{x}}_0+\hat{\mathbf{x}}_1}{2}\right)^{\mathrm{T}}\hat{\mathbf{H}}(\hat{\mathbf{x}}_0-\hat{\mathbf{x}}_1) \quad (5)$$

$$D(\hat{\mathbf{H}}) = (\hat{\mathbf{x}}_0-\hat{\mathbf{x}}_1)^{\mathrm{T}}\hat{\mathbf{H}}\boldsymbol{\Sigma}\hat{\mathbf{H}}(\hat{\mathbf{x}}_0-\hat{\mathbf{x}}_1). \quad (6)$$

An issue with standard RDA is that it is not robust if outlying samples are present in the training data. In such cases, of much practical relevance, we can resort to a robust covariance estimator, as discussed next.

3. ROBUST LDA IN CLEAN DATA

3.1. Robust estimation of the covariance matrix

We propose to use a robust estimator of $\boldsymbol{\Sigma}$ known as regularized Tyler's estimator, defined as the unique solution $\hat{\mathbf{C}}(\beta)$ to the following fixed-point equation [7, 8]:

$$\hat{\mathbf{C}}(\beta) = \frac{(1-\beta)}{n-2}\sum_{i=0}^{1}\sum_{j=1}^{n_i}\frac{(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)^{\mathrm{T}}}{\frac{1}{N}(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)^{\mathrm{T}}\hat{\mathbf{C}}^{-1}(\beta)(\mathbf{x}_j^{(i)}-\hat{\mathbf{x}}_i)} + \beta\mathbf{I}_N, \quad (7)$$

where $\beta \in (\max\{0, 1-n/N\}, 1]$. This covariance matrix estimator is a hybrid estimator reminiscent of the original Tyler's estimator of scale [11] and Ledoit-Wolf shrinkage estimator [12]. Here, like ρ in the case of the RSCM (2), β is a regularization parameter that determines the tradeoff between bias (the shrinkage target, $\mathbf{I}_N$) and variance (the pooled SCM). Using the inverse of $\hat{\mathbf{C}}(\beta)$ as a plug-in estimator of $\boldsymbol{\Sigma}^{-1}$ in (1), we should effectively end up with a robust version of LDA, which we coin Tyler-LDA. We note that, in [13], the authors also proposed to estimate the population matrix $\boldsymbol{\Sigma}$ using regularized M-estimators (although excluding Tyler's estimator) in the context of discriminant analysis, but did not provide a systematic analysis of the performance of their method.

An important question that naturally arises is whether Tyler-LDA performs (at least) as well as standard LDA approaches when the data is clean. This will be answered in the next section.

3.2. Asymptotic performance of Tyler-LDA

We operate under the following assumptions [5]:

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

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

Assumption 1 characterizes the growth regime under consideration, while Assumption 2, which is concerned with the covariance and mean scaling of the training data, ensures that a non-trivial (i.e., neither 0 or 1) asymptotic classification accuracy can be achieved [5].

In [4, 5], the authors studied the asymptotic performance (in terms of the total classification error) of the RSCM (2). In particular, they showed that, under Assumptions 1-2, the class-conditional classification error of the RSCM, $\epsilon_i^{\mathrm{RSCM}}(\rho) \triangleq \epsilon_i^{\mathrm{LDA}}(\overline{\boldsymbol{\Sigma}}^{-1}(\rho))$ converges to the deterministic quantity $\bar{\epsilon}_i^{\mathrm{RSCM}}(\rho)$, in the sense

$$|\epsilon_i^{\mathrm{RSCM}}(\rho) - \bar{\epsilon}_i^{\mathrm{RSCM}}(\rho)| \xrightarrow{\text{a.s.}} 0, \quad n,N\to\infty,$$

with $\bar{\epsilon}^{\mathrm{RSCM}}(\rho)$ depending only on the true means of each class and the underlying covariance matrix. The specifics of this result are recalled in Lemma 1, in Appendix.

As we will show, the asymptotic misclassification probability of Tyler-LDA is exactly the same as that of the RSCM when the training data is free of outliers. To proceed, let us define the asymptotic class-conditional classification error of

Tyler-LDA as $\epsilon_i^{\text{Tyler}}(\beta) \triangleq \epsilon_i^{\text{LDA}}(\beta\hat{\mathbf{C}}^{-1}(\beta))$. Equipped with these assumptions and notations, we can now state our main result (proved in Section 5):

Theorem 1. *[Deterministic equivalent] Let Assumptions 1-2 hold. For a given* $\beta \in \mathcal{R}_\zeta \triangleq [\zeta + \max\{0, 1-c^{-1}\}, 1]$, *where* $\zeta \in (0, \min\{1, c^{-1}\})$, *define* $\rho_\beta = \frac{1}{\beta\gamma(\beta)} \frac{1-\beta}{1-(1-\beta)c}$, *where* $\gamma(\beta)$ *is the unique positive solution to the equation*

$$\frac{1}{N} \operatorname{Tr} \boldsymbol{\Sigma}(\gamma(\beta)\beta\mathbf{I}_N + (1-\beta)\boldsymbol{\Sigma})^{-1} = 1.$$

Then, $|\epsilon_i^{\text{Tyler}}(\beta) - \bar{\epsilon}_i^{\text{Tyler}}(\beta)| \xrightarrow{\text{a.s.}} 0$ *for each* $\beta \in \mathcal{R}_\zeta$, *as* $N, n \to \infty$, *with*

$$\bar{\epsilon}_i^{\text{Tyler}}(\beta) \triangleq \bar{\epsilon}_i^{\text{RSCM}}(\rho_\beta), \tag{8}$$

with the expression of $\bar{\epsilon}_i^{\text{RSCM}}$ *recalled in Lemma 1.*

Theorem 1 shows that, when data is outlier-free, both the standard regularized LDA method and Tyler-LDA share the same asymptotic performance, up to a transformation of the regularization parameters. This is an important message, as it shows that there is nothing to lose by using Tyler's estimator over other conventional methods when there are no outlying observations in the training data.

Regularization parameter optimization. Among possible choices of $\beta \in \mathcal{R}_\zeta$, in practice one shall choose the regularization parameter that minimizes the misclassification error $\epsilon^{\text{Tyler}}(\beta)$. To do so, one can first use the optimization procedure proposed in [5] to find the optimal $\rho^\star$ minimizing $\bar{\epsilon}^{\text{RSCM}}(\rho)$, and then use the inverse of the mapping $\beta \mapsto \rho$ given in Theorem 1 to identify a parameter $\beta^\star$ that minimizes $\bar{\epsilon}_i^{\text{Tyler}}(\beta)$.[1] It then remains to estimate this (unknown) parameter based on the training sample. We will give more precise details in an extended version of the paper.

4. SIMULATIONS

4.1. Synthetic data

In all simulations, we used the sample mean as the mean estimator for $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. In Fig. 1, we plot the empirical classification error associated with the RSCM and Tyler's estimator as a function of the regularization parameters ρ and β (top and bottom x-axes, respectively), in an outlier-free scenario. The deterministic classification error, computed using Theorem 1 and Lemma 1, and $\epsilon^{\text{LDA}}(\hat{\mathbf{H}} = \boldsymbol{\Sigma})$ ("oracle" estimator) are also shown. Simulations show a very good match between empirical and theoretical values, which validates Theorem 1. We note that the optimal regularization parameters $\beta^\star$ (for Tyler's estimator) and $\rho^\star$ (for the RSCM), identified on the figure, verify $\rho^\star = \rho_{\beta^\star}$ per the mapping given in the theorem statement.

Now consider a toy-example scenario where outliers (distributed as $\mathcal{N}(5\boldsymbol{\mu}, \mathbf{I}_N)$) are introduced in the training sample. To simplify the discussion, we fix $\rho = \rho^\star$ and $\beta = \beta^\star$, i.e., the optimal regularization parameters in the outlier-free case (setting of Fig. 1). Fig. 2 shows the classification error of both estimators as the proportion of outliers

[1] We remark that the mapping $\beta \mapsto \rho$ is only onto on $(0, \infty)$, and thus the uniqueness of $\beta^\star$ is not guaranteed.

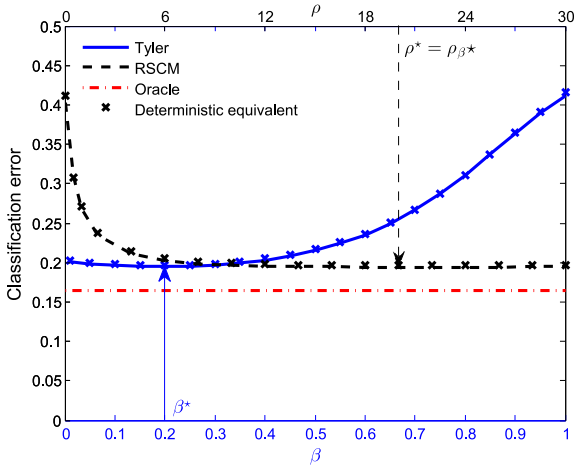


Fig. 1: Classification error of the RSCM and Tyler's estimator as ρ and β vary (top and bottom x-axes, respectively), for $N = 100$, $n_0 = n_1 = 200$, averaged over 1000 realizations, with a test sample size of 5000 samples per class. $\boldsymbol{\Sigma}$ is such that $[\boldsymbol{\Sigma}]_{ij} = 0.8^{|i-j|}$, and has eigenvalue decomposition $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^{\text{T}}$. $\boldsymbol{\mu}$ is such that $\boldsymbol{\mu} \propto \mathbf{V}\mathbf{1}_N$. The oracle estimator's classification error ($\boldsymbol{\Sigma}$ assumed to be known) is also shown.

increases. It appears that outliers affect the performance of the RSCM more than that of Tyler's estimator, with up to a 6% difference in misclassification probablity when the data has only 5% outliers. This shows that, when data is corrupted by such outliers, Tyler's estimator prevails over the RSCM, and it suggests that using Tyler's estimator would be preferable over standard, non-robust methods when little is known about the quality of the training data.

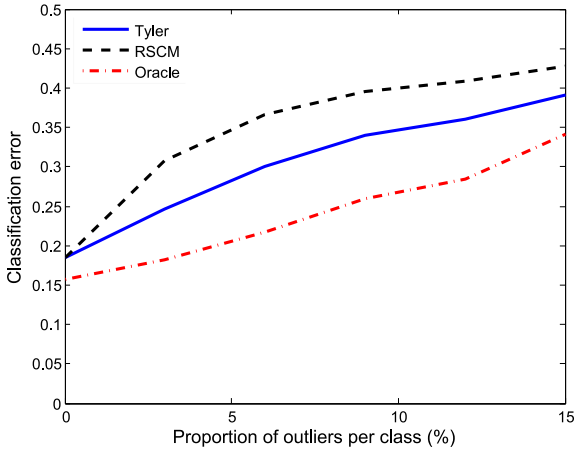
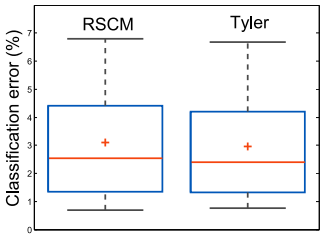


Fig. 2: Classification error of the RSCM and Tyler's estimator, for $N = 100$, $n_0 = n_1 = 200$ ($c_N = 1/4$), averaged over 1000 realizations. Outliers follow a multivariate Gaussian distribution $\mathcal{N}(5\boldsymbol{\mu}, \mathbf{I}_N)$.

4.2. MNIST data

We also performed simulations on the MNIST data set [14]. For all 45 possible pairs of classes C_0/C_1 corresponding to digits 0/1, 0/2, ... etc., we computed the RSCM and Tyler's estimator on all the available training data set (∼5800 samples per class) and tested it on the testing data set (∼1000

samples per class). Fig. 3 shows the statistics (as box plots) of the minimal testing classification error results for both estimators and all class pairs, obtained after a sweep over possible regularization parameters ρ and β. Note that this provides a lower bound of the lowest classification error achievable on the given testing set, ignoring any effects of estimating the regularization parameter. The performance on some representative class pairs is also reported.



Classes	RSCM	Tyler
0/7	**0.9 %**	1.2 %
4/8	2.0 %	2.0 %
2/6	4.4 %	4.3 %
1/5	1.8 %	1.7 %
3/9	4.5 %	**3.5 %**

Fig. 3: Box plot of the classification error of the RSCM and Tyler-LDA for all class pairs of the MNIST dataset (left), and comparison of the classification error for some specific class pair examples (right).

We observe that the RSCM and Tyler's estimator perform similarly (with a slight advantage, although non-statistically significant, for Tyler's estimator), which appears consistent with our analysis, if one were to assume that the MNIST data is free of outliers. In practice, one should however estimate the optimal regularization parameter based on training data. To do so, the procedure briefly described in Subsection 3.2 can be used. As indicated, precise details on this shall be given in an extended version of this paper. It also remains to study the theoretical performance of the RSCM and of Tyler's estimator under less clean scenarios; for example, with samples from different classes having different variances, or with outlying samples coming from a different distribution. This problem is currently under investigation.

5. PROOF OF THEOREM 1

The proof relies on understanding the asymptotic behavior of bilinear forms of the type $\mathbf{a}^{\mathrm{T}}\hat{\mathbf{C}}^{-k}(\beta)\mathbf{b}$ ($k = 1, 2$, $\beta \in \mathcal{R}_\zeta$), which appear in the expressions of $G(\beta\hat{\mathbf{C}}(\beta)^{-1})$ and $D(\beta\hat{\mathbf{C}}(\beta)^{-1})$ in (5) and (6), used to compute the class-conditional classification error $\epsilon_i^{\mathrm{Tyler}}(\beta) = \epsilon_i^{\mathrm{LDA}}(\beta\hat{\mathbf{C}}(\beta)^{-1})$ in (4). This type of functional of $\hat{\mathbf{C}}^{-k}(\beta)$ is intricate, because $\hat{\mathbf{C}}(\beta)$ is only implicitly defined. However, along similar lines to the proof of [10, Lemma 3], it can be proved that[2] for $\beta \in \mathcal{R}_\zeta = [\zeta + \max\{0, 1 - c^{-1}\}, 1]$, where $\zeta \in (0, \min\{1, c^{-1}\})$,

$$\left\| \hat{\mathbf{C}}(\beta)/\beta - \overline{\mathbf{\Sigma}}(\rho_\beta) \right\| \xrightarrow{\text{a.s.}} 0, \quad n, N \to \infty, \tag{9}$$

with ρ_β defined in Theorem 1. From this, it can be shown that the bilinear forms $\beta^k \mathbf{a}^{\mathrm{T}}\hat{\mathbf{C}}^{-k}(\beta)\mathbf{b}$ are asymptotically close to their RSCM counterparts $\mathbf{a}^{\mathrm{T}}\overline{\mathbf{\Sigma}}(\rho_\beta)^{-k}\mathbf{b}$; the behavior of which is well-understood. Specifically, for a given $\beta \in \mathcal{R}_\zeta$, 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 $k = 1, 2$,

$$\left| \beta^k \mathbf{a}^{\mathrm{T}}\hat{\mathbf{C}}^{-k}(\beta)\mathbf{b} - \mathbf{a}^{\mathrm{T}}\overline{\mathbf{\Sigma}}^{-k}(\rho_\beta)\mathbf{b} \right| \xrightarrow{\text{a.s.}} 0, \tag{10}$$

as $N, n \to \infty$. Taking for example $k = 1$, it is proved by first noting

$$\left| \mathbf{a}^{\mathrm{T}}(\beta\hat{\mathbf{C}}^{-1}(\beta) - \overline{\mathbf{\Sigma}}^{-1}(\rho_\beta))\mathbf{b} \right| \leq K ||\beta\hat{\mathbf{C}}^{-1}(\beta) - \overline{\mathbf{\Sigma}}^{-1}(\rho_\beta)||$$
$$\leq K \cdot ||\beta\hat{\mathbf{C}}^{-1}(\beta)|| \cdot ||\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta)|| \cdot ||\hat{\mathbf{C}}(\beta)/\beta - \overline{\mathbf{\Sigma}}(\rho_\beta)||,$$

where the last inequality is due to the resolvent identity [15], and where $K = ||\mathbf{a}|| \cdot ||\mathbf{b}||$. The fact that $\beta, \rho_\beta > 0$ ensure that $||\hat{\mathbf{C}}^{-1}(\beta)||, ||\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta)|| < \infty$, which, along with (9) and $\limsup_N ||\mathbf{a}|| < \infty$ a.s., $\limsup_N ||\mathbf{b}|| < \infty$ a.s., leads to (10), for $k = 1$. The case $k = 2$ is handled similarly.

Equipped with this, we can prove that

$$|G(\beta\hat{\mathbf{C}}(\beta)^{-1}) - G(\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta))| \xrightarrow{\text{a.s.}} 0 \tag{11}$$
$$|D(\beta\hat{\mathbf{C}}(\beta)^{-1}) - D(\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta))| \xrightarrow{\text{a.s.}} 0. \tag{12}$$

Take $\mathbf{a} = \boldsymbol{\mu}_i - (\hat{\mathbf{x}}_0 + \hat{\mathbf{x}}_1)/2$ and $\mathbf{b} = (\hat{\mathbf{x}}_0 - \hat{\mathbf{x}}_1)$. Then, Assumption 2 implies $\limsup_N ||\mathbf{a}|| < \infty$ a.s., $\limsup_N ||\mathbf{b}|| < \infty$ a.s., by the law of large numbers. The convergence in (11) then follows from (10) by taking $k = 1$. For (12), we remark that

$$|D(\beta\hat{\mathbf{C}}(\beta)^{-1}) - D(\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta))|$$
$$= \left| \mathrm{Tr}\,\mathbf{\Sigma}\left(\beta^2\hat{\mathbf{C}}(\beta)^{-1}\mathbf{b}\mathbf{b}^{\mathrm{T}}\hat{\mathbf{C}}(\beta)^{-1} - \overline{\mathbf{\Sigma}}^{-1}(\rho_\beta)\mathbf{b}\mathbf{b}^{\mathrm{T}}\overline{\mathbf{\Sigma}}^{-1}(\rho_\beta)\right)\right|$$
$$\leq ||\mathbf{\Sigma}|| \cdot \left| \mathbf{b}^{\mathrm{T}}\left(\beta^2\hat{\mathbf{C}}(\beta)^{-2} - \overline{\mathbf{\Sigma}}^{-2}(\rho_\beta)\right)\mathbf{b} \right|.$$

Using (10) with $k = 2$ leads to (12).

From (4), using (11), (12), and the fact that $\sqrt{\cdot}$ and $\mathbf{\Phi}(\cdot)$ are continuous functions, we have proved

$$\left| \epsilon_i^{\mathrm{Tyler}}(\beta) - \epsilon_i^{\mathrm{RSCM}}(\rho_\beta) \right| \xrightarrow{\text{a.s.}} 0, \quad n, N \to \infty.$$

Combining this with Lemma 1 concludes the proof.

APPENDIX

Lemma 1. *[5, Corollary 3] Let Assumptions 1-2 hold. As $N, n \to \infty$, we have $|\epsilon_i^{\mathrm{RSCM}}(\rho) - \bar{\epsilon}_i^{\mathrm{RSCM}}(\rho)| \xrightarrow{\text{a.s.}} 0$ for each $\rho > 0$, with*

$$\bar{\epsilon}_i^{\mathrm{RSCM}}(\rho) = \Phi\left(\frac{(-1)^{i+1}\overline{G}_i(\rho) + (-1)^i \log\left(\frac{\pi_0}{\pi_1}\right)}{\sqrt{\overline{D}(\rho)}} \right),$$

with $\overline{G}_i(\rho)$ and $\overline{D}(\rho)$ defined as

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

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

with $\mathbf{A} = \left(\mathbf{I}_N + \frac{\rho}{1+\rho\delta}\mathbf{\Sigma}\right)^{-2}$, and δ the unique solution to

$$\delta = \frac{1}{N}\mathrm{Tr}\,\mathbf{\Sigma}\left(\mathbf{I}_N + \frac{\rho}{1+\rho\delta}\mathbf{\Sigma}\right)^{-1}.$$

[2] Complete details will be given in an extended version.

References

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

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

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

[4] 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, 2015.

[5] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional analysis of regularized discriminant analysis classifiers," *arXiv preprint arXiv:1711.00382*, 2017.

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

[7] Y. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *Proc. IEEE Int. Conf. Acoust. Signal Process.*, 2007, vol. 3, pp. III–1105.

[8] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5640–5651, Sept. 2014.

[9] 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.

[10] 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, 2015.

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

[12] 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.

[13] 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.

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

[15] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*, Cambridge University Press, 2011.