LDA via L1-PCA of Whitened Data

Rubén Martín-Clemente, *Member, IEEE*, and Vicente Zarzoso, *Senior Member, IEEE*

***Abstract*—Principal component analysis (PCA) and Fisher's linear discriminant analysis (LDA) are widespread techniques in data analysis and pattern recognition. Recently, the L1-norm has been proposed as an alternative criterion to classical L2-norm in PCA, drawing considerable research interest on account of its increased robustness to outliers. The present work proves that, combined with a whitening preprocessing step, L1-PCA can perform LDA in an unsupervised manner, i.e., sparing the need for labelled data. Rigorous proof is given in the case of data drawn from a mixture of Gaussians. A number of numerical experiments on synthetic as well as real data confirm the theoretical findings.**

***Index Terms*—Fisher's linear discriminant analysis, L1-norm, principal component analysis.**

I. Introduction

FISHER'S linear discriminant analysis (LDA) and principal component analysis (PCA) can be considered as two pillars of data analysis [1]. Given a dataset with two classes, LDA finds a projection of the data points onto a one-dimensional space where the classes are well separated. When the two classes are Gaussian with equal covariances, LDA yields the optimal Bayes classifier. LDA is a supervised technique, whose performance depends heavily on the availability of correctly labelled data [2]. On the other hand, PCA finds the linear projection best fitting the data in the least-squares sense, which is also the projection of maximum variance of the dataset. As a fully data-driven technique, PCA is unsupervised and does not require labelled samples [3].

Spurred by the versatility of PCA, the last decade has witnessed a flurry of research into alternative criteria aimed at enhancing its capabilities or alleviating its limitations in various operating conditions. One such criterion, in particular, is based on replacing the L2-norm of classical PCA by the L1-norm, which is more robust against outliers, thus giving rise to L1-PCA [4]. This technique, compared to other robust versions of PCA [5]–[8], is particularly intuitive and simple, as well as invariant to the rotation of the data, all of which justifies its growing interest and use in a wide range of applications, e.g., sensor-array processing, image fusion, video surveillance, robust face recognition, *et cetera* [9]–[13]. In addition, a number of efficient L1-PCA algorithms have been proposed in the recent literature (see e.g. [14]–[18]).

Even though L1-PCA and LDA are apparently disparate techniques with very different purposes, the present contribution shows that a link actually exists between both approaches. We prove that, for whitened (or sphered) data, L1-PCA can perform LDA in an *unsupervised* fashion, that is, sparing the need for training data. This result is of theoretical interest and opens interesting research perspectives for performing LDA using L1-PCA algorithms. A number of numerical experiments validate the theoretical findings in a variety of simulation scenarios.

The paper is organized as follows. Section II reviews the basics of Fisher's LDA and PCA and other related approaches such as kurtosis optimization. Section III analyzes the L1-LDA criterion using a Gaussian assumption, and shows it to perform LDA in an unsupervised manner under certain prescribed conditions. Links with other techniques are also established and working algorithms reviewed. Generalization to more than two classes is addressed in Section IV. Computer experiments validating the theoretical findings on synthetic and real data are reported in Section V. The concluding remarks of Section VI bring the paper to an end.

The notation used in the paper is as follows. Lightface letters (e.g., a, A) represent scalar quantities, which may be functions, univariate random variables, deterministic constants or indices. The distinction will be clear from the context. Boldface lowercase ($\boldsymbol{x}$) and uppercase ($\boldsymbol{X}$) letters, respectively, stand for vectors and matrices, either deterministic or random depending on the context.

II. Background

This section reviews the basic concepts on LDA and PCA, recalling some connections with other techniques. This material will be useful for the analysis in Section III.

A. Assumptions

The following conditions are assumed throughout the paper. Consider p-dimensional observations $\boldsymbol{x} \in \mathbb{R}^p$, where each observation $\boldsymbol{x}$ belongs to one of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Suppose that the two classes can be described by probability density functions f_1 and f_2, respectively. The prior probability of class $\mathcal{C}_i$ is denoted by π_i, $i = 1, 2$, with $\pi_1 + \pi_2 = 1$. Distribution f_i has a mean $\boldsymbol{\mu}_i = \mathrm{E}\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_i\}$ and an invertible positive definite covariance matrix $\boldsymbol{V}_i = \mathrm{Cov}(\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_i) = \mathrm{E}\{(\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\intercal \mid \boldsymbol{x} \in \mathcal{C}_i\}$, $i = 1, 2$, where $\mathrm{E}\{\cdot\}$ is the expectation operator and $(\cdot)^\intercal$ stands for transpose. The following typical assumption can be made without loss of generality:

Assumption A1: The data have zero mean, i.e., $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{x}\} = \sum_{i=1}^{2} \pi_i \boldsymbol{\mu}_i = \boldsymbol{0}$.

Manuscript received October 13, 2018; revised July 14, 2019; accepted November 13, 2019. Date of publication November 25, 2019; date of current version December 27, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yuantao Gu. This work was supported by the Spanish Ministry of Economy and Competitiveness under project TEC2017-82807-P. The work of R. Martín-Clemente was supported by professorship at the I3S Laboratory in July 2018 funded by an IFA scholarship from Université Côte d'Azur. *(Corresponding author: Ruben Martin-Clemente.)*

R. Martín-Clemente is with the Departamento de Teoría de la Señal y Comunicaciones, Escuela Superior de Ingeniería, Universidad de Sevilla, Seville 41092, Spain (e-mail: ruben@us.es).

V. Zarzoso is with the Université Côte d'Azur, CNRS, I3S Laboratory, Sophia Antipolis Cedex 06903, France (e-mail: vicente.zarzoso@univ-cotedazur.fr).

Digital Object Identifier 10.1109/TSP.2019.2955860

Under this assumption, one can easily show that:

$$\boldsymbol{\mu}_1 = -\pi_2 \Delta\boldsymbol{\mu} \tag{1a}$$

$$\boldsymbol{\mu}_2 = \pi_1 \Delta\boldsymbol{\mu} \tag{1b}$$

where $\Delta\boldsymbol{\mu} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

B. LDA Formulation

LDA [1] aims at projecting the data along a direction $\boldsymbol{a} \in \mathbb{R}^p$, resulting in the projection

$$y = \boldsymbol{a}^\intercal \boldsymbol{x}. \tag{2}$$

A reasonable criterion is to select a vector $\boldsymbol{a}$ such that the corresponding projections y obtained from $\mathcal{C}_1$ and $\mathcal{C}_2$ are separated as much as possible. To this end, the derivation of LDA involves the so-called *within-class* scatter matrix, defined as

$$\boldsymbol{S}_W = \sum_{i=1}^{2} \pi_i \boldsymbol{V}_i \tag{3}$$

which quantifies the average variation of the data in the classes. Similarly, the *between-class* scatter matrix

$$\boldsymbol{S}_B = \sum_{i=1}^{2} \pi_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal = \pi_1 \pi_2 \Delta\boldsymbol{\mu} \Delta\boldsymbol{\mu}^\intercal \tag{4}$$

contains the distance between the class centers. Equation (1) has been invoked to reach the last term of Eqn. (4). To find direction $\boldsymbol{a}$, LDA maximizes the Rayleigh quotient:

$$J_{\mathrm{LDA}}(\boldsymbol{a}) = \frac{\boldsymbol{a}^\intercal \boldsymbol{S}_B \boldsymbol{a}}{\boldsymbol{a}^\intercal \boldsymbol{S}_W \boldsymbol{a}} \tag{5}$$

which is simply obtained at [1]

$$\boldsymbol{S}_W \boldsymbol{a}^* = \Delta\boldsymbol{\mu}. \tag{6}$$

The rationale behind LDA cost (5) can be explained by the following: observe that the mean and variance of the projected data of class $\mathcal{C}_i$ can be expressed, respectively, as

$$m_i = \boldsymbol{a}^\intercal \boldsymbol{\mu}_i \quad \text{and} \quad \sigma_i^2 = \boldsymbol{a}^\intercal \boldsymbol{V}_i \boldsymbol{a} \qquad i = 1, 2. \tag{7}$$

Then simple algebra shows that

$$J_{\mathrm{LDA}}(\boldsymbol{a}) = \frac{\pi_1 \pi_2 (\Delta m)^2}{\pi_1\, \sigma_1^2 + \pi_2\, \sigma_2^2}$$

with

$$\Delta m \stackrel{\text{def}}{=} m_2 - m_1 = \boldsymbol{a}^\intercal \boldsymbol{\mu}_2 - \boldsymbol{a}^\intercal \boldsymbol{\mu}_1 = \boldsymbol{a}^\intercal \Delta\boldsymbol{\mu}. \tag{8}$$

From here it is apparent that LDA searches for a vector $\boldsymbol{a}$ such that: (i) m_1 and m_2 are far apart from each other, and (ii) the spread of the projections around m_1 and m_2 is small. Both conditions try to prevent the overlapping and maximize the separation of the projected classes, thereby easing their discrimination.

To compute class statistics $\{\boldsymbol{\mu}_i, \boldsymbol{V}_i\}_{i=1}^2$, LDA needs labelled data — also known as training data — where the class to which each observed vector $\boldsymbol{x}$ belongs is known. This is why LDA is a supervised classification technique.

1) Whitened Data: LDA admits an interesting interpretation when the classes are whitened or sphered, as explained next. Let us now consider the case in which the classes are homoscedastic, i.e., they have the same covariance matrix, $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{V}$, which is a commonly accepted assumption in LDA. Then, Eqn. (3) reduces to $\boldsymbol{S}_W = \boldsymbol{V}$ and Eqn. (6) simplifies into

$$\boldsymbol{V} \boldsymbol{a}^* = \Delta\boldsymbol{\mu}. \tag{9}$$

The projection onto vector (9) gives

$$y = \boldsymbol{x}^\intercal \boldsymbol{a}^* = \Delta\boldsymbol{\mu}^\intercal \boldsymbol{V}^{-1} \boldsymbol{x}.$$

It is now useful to calculate the eigendecomposition of the within-class covariance matrix, given by $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\intercal$. Then we have $\boldsymbol{V}^{-1} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\intercal$, and

$$y = \Delta\tilde{\boldsymbol{\mu}}^\intercal \tilde{\boldsymbol{x}}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{D}^{-1/2}\boldsymbol{U}^\intercal \boldsymbol{x}$ and $\Delta\tilde{\boldsymbol{\mu}} = \boldsymbol{D}^{-1/2}\boldsymbol{U}^\intercal \Delta\boldsymbol{\mu}$. This transformation basically *whitens* (or *spheres*) the data because the covariance of $\tilde{\boldsymbol{x}}$ is the identity matrix for both classes, $\mathrm{Cov}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{C}_i) = \boldsymbol{D}^{-1/2}\boldsymbol{U}^\intercal \boldsymbol{V} \boldsymbol{U} \boldsymbol{D}^{-1/2} = \boldsymbol{I}$, $i = 1, 2$. Therefore, LDA is equivalent to whitening each class and then projecting the data along the line joining the whitened class centroids, defined by vector $\Delta\tilde{\boldsymbol{\mu}}$.

2) Kurtosis-Based Unsupervised LDA: The kurtosis, defined as the fourth-order central moment divided by the squared variance, is a measure of the gaussianity, peakedness and bimodality of a distribution [19]. We recall it here because the following theorem, reproduced from [20] using our notation, shows that LDA is also closely related to this statistic.

Theorem 1: Let $\boldsymbol{x}$ be a p-dimensional random variable distributed as $\pi_1\, f_1(\boldsymbol{x}) + \pi_2\, f_2(\boldsymbol{x})$, where $\pi_1 + \pi_2 = 1$ and f_i, $i = 1, 2$, is a Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{V}_i = \boldsymbol{V}$, the same for both distributions. Let $\boldsymbol{a}$ be a unit-norm vector on $\mathbb{R}^p$ and $y = \boldsymbol{a}^\intercal \boldsymbol{x}$. If $\boldsymbol{a}$ satisfies

$$\boldsymbol{V}\, \boldsymbol{a} = \Delta\boldsymbol{\mu} \tag{10}$$

then it maximizes the absolute kurtosis of y. Furthermore, these directions minimize the kurtosis if $|\pi_1 - 1/2| < 1/\sqrt{12}$, and maximize it otherwise.

A related result for the case of different covariance matrices can be also found in [20]. It is interesting to observe that (10) is equivalent to (9) but can be obtained without prior knowledge of the data allocation, i.e., without the need for training data. This result opens the way for kurtosis-based unsupervised classification techniques [21]. The main drawback of this approach is that kurtosis is very sensitive to outliers since, by raising the projections to the fourth power, the effects of the data points far from the nominal distribution can easily overshoot, leading to poor results in the presence of faulty data.

3) Generalized LDA: Finally, the LDA solution (6) can be also generalized to weighted within-class scatter matrices of the form:

$$\bar{\boldsymbol{S}}_W = \sum_{i=1}^{2} \pi_i \beta_i \boldsymbol{V}_i \tag{11}$$

where β_i, $i = 1, 2$, are constants such that $\bar{\boldsymbol{S}}_W$ is positive or negative definite. Such weighted scatter matrices lead to admissible LDA-type classifiers for multivariate normal distributions

with different covariance matrices [22]. Notice that the L1-norm based criterion studied in Section III will be shown to be a particular generalized LDA classifier under certain conditions.

C. L2-PCA and L1-PCA

In its original formulation [3], classical PCA seeks directions maximizing the variance of the projection (2) according to the quadratic criterion:

$$\max_{\boldsymbol{a}} \mathrm{E}\{y^2\} \text{ subject to } \|\boldsymbol{a}\|_2 = 1. \tag{12}$$

This technique is also known as L2-PCA because, when operating on finite sample observations, the expectation in (12) is usually replaced by the L2-norm of the vector of projected samples. PCA was originally developed for dimensionality reduction of multivariate data, in general. The solution to problem (12) can simply be computed as the dominant eigenvector of the overall data covariance matrix, which under the assumptions of Section II-A is given by

$$\boldsymbol{V}_0 = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^\intercal\} = \boldsymbol{S}_W + \boldsymbol{S}_B. \tag{13}$$

If the classes are well separated, the between-class scatter typically dominates the within-class scatter, i.e.,

$$\|\boldsymbol{S}_W\|_{\mathrm{Fro}} \ll \|\boldsymbol{S}_B\|_{\mathrm{Fro}} \tag{14}$$

where $\|\cdot\|_{\mathrm{Fro}}$ represents the Frobenius norm, and then:

$$\boldsymbol{V}_0 \approx \boldsymbol{S}_B = \pi_1\pi_2\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}^\intercal. \tag{15}$$

As a result, L2-PCA defines a projection in the direction joining the class centroids, $\Delta\boldsymbol{\mu}$. This result reminds us of the LDA solution for whitened classes as presented in the previous section but, in contrast to LDA, only in certain situations it provides a good direction of projection for separating the two classes. For the interested reader, a comparison between LDA and PCA in image recognition tasks is given for example in [23].

Though mathematically appealing, the L2-norm is rather sensitive to impulsive noise or outliers since squaring the projections overemphasizes the effects of far-off data points. As presented in [4], a natural extension of L2-PCA to deal with this drawback is given by

$$\max_{\boldsymbol{a}} \mathrm{E}\{|y|\} \text{ subject to } \|\boldsymbol{a}\|_2 = 1. \tag{16}$$

The absolute value in Eqn. (16) is replaced in practice by the L1-norm of the vector of projected samples and thus the above criterion is referred to as L1-PCA. Note that the above two PCA criteria are not equivalent in general, though they can provide similar results when the data contain no outliers [4], [16]. In addition, note also that none of them require labeled training data.

III. Link between L1-PCA and LDA

Inspired by the results reviewed in the previous section, we consider the following unsupervised criterion for linear discriminant analysis:

$$\max_{\boldsymbol{a}} \mathrm{E}\{|y|\} \text{ subject to } \mathrm{E}\{y^2\} = 1 \tag{17}$$

with $y = \boldsymbol{a}^\intercal\boldsymbol{x}$ and hence $\mathrm{E}\{y^2\} = \mathrm{E}\{(\boldsymbol{a}^\intercal\boldsymbol{x})^2\} = \boldsymbol{a}^\intercal\boldsymbol{V}_0\boldsymbol{a}$. Observe that the constraint differs from that used in L1-PCA ($\|\boldsymbol{a}\|_2 = 1$) but still prevents $\mathrm{E}\{|y|\}$ from increasing by a mere increase of the magnitude of $\boldsymbol{a}$. In the rest of the paper, the above criterion is referred to as *L1-norm based unsupervised LDA (L1-uLDA)*. The L1-uLDA criterion is analyzed in the remainder of this section. In particular, the unit-variance constraint is shown to be the ingredient confering discriminative capabilities. The L1-uLDA criterion is quite general in that it does not require the data to be prewhitened. Without limiting the foregoing, it is well-known that the unit-variance constraint equates to a unit-norm constraint for whitened data, so that L1-uLDA becomes L1-PCA in this case. This observation enables us to make the link between L1-PCA and LDA explicit in Section III-D. It is concluded that, for whitened data, L1-PCA can perform LDA in an unsupervised fashion, without information about class labels. For unwhitened data, which is the more general case, this result still holds since a simple sphering or whitening preprocessing step allows us to carry out L1-uLDA by L1-PCA.

A. Justification of the Criterion

An intuitive explanation of how the criterion (17) works is that the absolute value can be considered as a measure of distance from the origin; as a result, it seeks a direction maximizing the average distance from zero of the projected data points, which in the end, thanks to the zero-mean assumption, favours the emergence of two distinct clusters on either side of the origin. The second power used in L2-PCA (12) can also be considered as another measure of distance from the origin but it leads to the direction of maximum variability of the data, which, as it is well-known, does not necessarily produce linear discriminant results. The main point is that L1-PCA is endowed with discriminative power thanks to the unit-variance constraint. We will show that, after whitening the data, L1-PCA is able to find the direction of the line joining the class centroids.

Furthermore, as previously shown in Section II-B2, there exists a close connection between LDA and the optimization of the kurtosis of the linearly projected data. An additional motivation for criterion (17) lies in the fact that, as recently shown in [24], the kurtosis is related to the mean of the absolute value of the random variable. To see this, let us standardize the random variable y to have zero mean and unit variance and, assuming that moments exist up to order four, consider the fourth-order Gram-Charlier expansion of its probability density function [25]:

$$f(y) \approx \frac{\exp(-y^2/2)}{\sqrt{2\pi}}\left[1 + \frac{\lambda_3}{6}(y^3 - 3y) + \frac{\kappa_4}{24}(y^4 - 6y^2 + 3)\right]$$

where $\lambda_3 = \mathrm{E}\{y^3\}$ and $\kappa_4 = (\mathrm{E}\{y^4\} - 3)$ denotes the excess kurtosis (i.e., the kurtosis minus three) of y. Then, using that $\int_0^\infty y^u \exp(-y^2/2)dy = 0.5^{-((u-1)/2)}\Gamma((u+1)/2)$, $\Gamma(\cdot)$ being the Gamma function, some algebra shows that

$$\mathrm{E}\{|y|\} = \int_{-\infty}^{\infty} |y| f(y)\mathrm{d}y = \sqrt{\frac{2}{\pi}}\left(1 - \frac{\kappa_4}{24}\right). \tag{18}$$

As a result, maximizing (resp. minimizing) $\mathrm{E}\{|y|\}$ is equivalent to minimizing (resp. maximizing) the kurtosis. Although one may question whether the Gram-Charlier expansion can represent adequately the distribution of the data, this equivalence

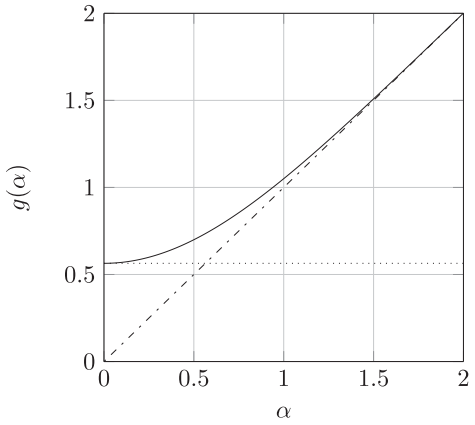


Fig. 1. Plot of function $g(\alpha)$ defined by eqn. (23). The constant $1/\sqrt{\pi}$ and the straight line $y = \alpha$ are represented by the dotted and dashed-dotted lines, respectively.

together with Theorem 1 (Section II-B2) suggest exploring the link between the averaged absolute value and LDA.

B. Theoretical Analysis in the Gaussian Case

As it is usual in the study of LDA, we assume that the clusters are normally distributed, i.e., $\boldsymbol{x} \mid \mathcal{C}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{V}_i), i = 1, 2$. This is a probabilistic model that has been extensively applied in signal processing. It follows that the observations are distributed as a mixture of Gaussians, $\boldsymbol{x} \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{V}_i)$. Hence, any linear projection as in Eqn. (2) will also be distributed according to a Gaussian mixture model $y \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(m_i, \sigma_i^2)$, with probability density function

$$f(y) = \sum_{i=1}^{2} \frac{\pi_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y-m_i)^2}{2\sigma_i^2}\right) \tag{19}$$

where m_i and σ_i^2 are defined in (7). As a consequence of Assumption A1, the mean value of y is also zero.

The objective function and the constraint can be expressed as follows. The second-order moment of y is readily given by

$$\mathrm{E}\{y^2\} = \boldsymbol{a}^\intercal \boldsymbol{V}_0 \boldsymbol{a} \tag{20}$$

where $\boldsymbol{V}_0$ is the covariance matrix of $\boldsymbol{x}$ defined in (13). Next, using the change of variable $x = (y - m_i)/\sigma_i$ and letting

$$\alpha_i = \frac{m_i}{\sqrt{2}\sigma_i} \tag{21}$$

some algebraic manipulations show that

$$\mathrm{E}\{|y|\} = \int_{-\infty}^{\infty} |y| f(y) \mathrm{d}y = \sqrt{2} \sum_{i=1}^{2} \pi_i \sigma_i g(\alpha_i) \tag{22}$$

with

$$g(\alpha_i) = \frac{1}{\sqrt{\pi}} \exp(-\alpha_i^2) + \alpha_i \,\mathrm{erf}\,(\alpha_i) \tag{23}$$

where the error function is defined as $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t$. It holds that (23) is an even function and that $\lim_{|\alpha|\to\infty} g(\alpha) = |\alpha|$. Fig. 1 plots this function.

Appendix A shows that the stationary point of the constrained optimization problem (17) is given by

$$\bar{\boldsymbol{S}} \boldsymbol{a} = \delta \Delta\boldsymbol{\mu} \tag{24}$$

where

$$\bar{\boldsymbol{S}} = \sum_{i=1}^{2} \pi_i \, \beta_i \boldsymbol{V}_i \tag{25a}$$

$$\beta_i = \frac{\sqrt{2/\pi}}{\sigma_i} \exp(-\alpha_i^2) - \mathrm{E}\{|y|\} \qquad i = 1, 2 \tag{25b}$$

$$\delta = \pi_1 \, \pi_2 \, (\gamma_2 - \gamma_1) \tag{25c}$$

$$\gamma_i = m_i \mathrm{E}\{|y|\} - \mathrm{erf}(\alpha_i) \qquad i = 1, 2 \tag{25d}$$

and α_i, $i = 1, 2$, were defined in Eqn. (21).

Furthermore, it is important to point out that solution (24) locally *maximizes* criterion (17) when $|\alpha_i|$, $i = 1$, 2, are sufficiently large (see Appendix A). To shed some light on this result, some comments are in order. Firstly, observe that when $|\alpha_1|$ and $|\alpha_2|$ are large, the magnitude of the projected means $|m_i|$ are large relative to the standard deviations σ_i. Therefore, the projected data form clusters and, because of the zero-mean assumption, one cluster is to the left and the other to the right of the origin, which is the desirable outcome in linear discrimination [22].

Secondly, when $|\alpha_i|$ is sufficiently large, $\exp(-\alpha_i^2) \approx 0$ and $\mathrm{erf}(\alpha_i) \approx \mathrm{sign}(\alpha_i)$. In that case, and recalling that $\lim_{|\alpha|\to\infty} g(\alpha) = |\alpha|$, expression (22) simplifies to

$$\begin{aligned} \mathrm{E}\{|y|\} &\approx \pi_2 |m_2| + \pi_1 |m_1| = \pi_2 m_2 - \pi_1 m_1 \\ &= 2\pi_1 \pi_2 \boldsymbol{a}^\intercal \Delta\boldsymbol{\mu} = 2\pi_1 \pi_2 \Delta m \end{aligned} \tag{26}$$

where we have assumed, without loss of generality, that $m_1 < 0 < m_2$. Observe that maximizing (26) with a unit-norm constraint, as in the original L1-PCA criterion (16), would yield a projection on the direction of the line joining the class centroids, $\Delta\boldsymbol{\mu}$, which is the direction of maximum variance: Section II-C recalled that, under condition (14), L2-PCA also finds the direction $\Delta\boldsymbol{\mu}$. It is the unit-variance constraint in L1-uLDA variant (17) that endows L1-PCA with LDA's discriminative power, as shown throughout this paper.

C. Link With LDA

Now we are ready to establish the formal connection with LDA. In general, if β_1 and β_2 in Eqns. (24)–(25) are both positive (resp. negative) then $\bar{\boldsymbol{S}}$ is positive (resp. negative) definite because $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are both positive definite. In this case, the L1-uLDA solution (24)–(25) is a particular instance of LDA classifier (6) with generalized covariance (11). The only additional requirement for (24) to define an LDA solution is that $\delta \neq 0$, which, according to Eqns. (25c)–(25d), is equivalent to the constraint:

$$\Delta m \neq \frac{\mathrm{erf}(\alpha_2) - \mathrm{erf}(\alpha_1)}{\mathrm{E}\{|y|\}}. \tag{27}$$

Next, suppose that the projected data form clusters where the mean is large relative to the standard deviation, in such a way that $|\alpha_i|$, $i = 1, 2$, are sufficiently large (asymptotic regime). Hence, the exponentials $\exp(-\alpha^2)$ vanish in (25b), yielding $\beta_1 = \beta_2 \approx$

$-\mathrm{E}\{|y|\}$. Then, the generalized covariance matrix (25a) can readily be expressed as

$$\bar{\boldsymbol{S}} \approx -\mathrm{E}\{|y|\} \sum_{i=1}^{2} \pi_i \boldsymbol{V}_i \propto \boldsymbol{S}_W$$

which leads to Fisher's LDA solution (6). Taking into account Eqn. (26), the condition on δ now becomes:

$$\Delta m \neq \frac{1}{\sqrt{\pi_1 \pi_2}} \tag{28}$$

where we have supposed that $m_1 < 0 < m_2$ without loss of generality. Appendix B proves that condition (28) is always satisfied.

The discussion simplifies when the classes are homoscedastic, i.e., $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{V}$. Then, the L1-PCA scatter matrix (25a) reads:

$$\bar{\boldsymbol{S}} = \left(\sum_{k=1}^{2} \pi_k \, \beta_k \right) \boldsymbol{V} \propto \boldsymbol{V}$$

so that Eqn. (24) always turns out to be equivalent to Eqn. (9) defining LDA in the homoscedastic scenario as recalled in Section II-B. This result also coincides with Theorem 1 in Section II-B2 for the kurtosis-based unsupervised classifier, and provides additional evidence of the relation between the proposed unsupervised criterion and LDA.

The above analysis can be summarized as follows:

i) The critical points of L1-uLDA criterion (17) present the general form of solutions (6) with general covariance given by Eqn. (25).
ii) The equivalence between L1-uLDA and LDA is exact if the projection onto the directions they define produces two sufficiently separated clusters on both sides of the origin so that $|\alpha_i|$, $i = 1, 2$, are sufficiently large.
iii) Under this condition, LDA is obtained in an unsupervised fashion by maximizing the proposed L1-uLDA criterion.

Note that the condition guaranteeing the equivalence between LDA and L1-uLDA is mild, as it coincides with the condition for LDA to yield proper discrimination results. A complementary discussion on the conditions under which $|\alpha_i|$, $i = 1, 2$, can be considered sufficiently large is given in Appendix C. Despite the theoretical equivalence in ideal conditions, L1-uLDA operates in an unsupervised fashion. This theoretical analysis will be put to the test in the experimental analysis of Section V.

D. Algorithm

1) Derivation: We have seen that the unit-variance constraint in (17) is crucial for L1-uLDA to yield discriminative solutions. However, classical algorithms for L1-PCA work under a unit-norm constraint, as in Eqn. (16). A sphering or whitening pre-processing step transforms the variance constraint into the norm constraint, thus enabling the use of classical algorithms for L1-PCA. Given $\boldsymbol{V}_0$, the covariance matrix of $\boldsymbol{x}$, and its eigenvalue decomposition (EVD) $\boldsymbol{V}_0 = \boldsymbol{Q}\boldsymbol{D}_0\boldsymbol{Q}^\intercal$, consider the change of variable $\boldsymbol{z} = \boldsymbol{D}_0^{-1/2}\boldsymbol{Q}^\intercal\boldsymbol{x}$. This transformation spheres or whitens the data, as the covariance matrix of $\boldsymbol{z}$ becomes the identity. Let $\boldsymbol{w} = \boldsymbol{D}_0^{1/2}\boldsymbol{Q}^\intercal\boldsymbol{a}$ and observe that $y = \boldsymbol{w}^\intercal\boldsymbol{z} = \boldsymbol{a}^\intercal\boldsymbol{x}$, implying that

$$\mathrm{E}\{y^2\} = \boldsymbol{w}^\intercal \mathrm{E}\{\boldsymbol{z}\boldsymbol{z}^\intercal\}\boldsymbol{w} = \boldsymbol{w}^\intercal\boldsymbol{w} = \|\boldsymbol{w}\|_2^2.$$

As a result, for whitened data the proposed criterion (17) can be rewritten as

$$\max_{\boldsymbol{w}} \mathrm{E}\{|y|\} \text{ subject to } \|\boldsymbol{w}\|_2 = 1 \tag{29}$$

where $y = \boldsymbol{w}^\intercal\boldsymbol{z}$. Hence, the unit-power constraint on the projected data is transformed into a unit-norm constraint on the projection vector. The constrained optimization problem (29) is the basis of L1-PCA and has been studied in a number of recent works [4], [16]. In particular, it can be shown that the fixed point iteration

1) $y = \boldsymbol{w}_n^\intercal\boldsymbol{z}$
2) $\boldsymbol{w}_{n+1} = \dfrac{\mathrm{E}\{\boldsymbol{z}\,\mathrm{sign}(y)\}}{\|\mathrm{E}\{\boldsymbol{z}\,\mathrm{sign}(y)\}\|_2}$

monotonically increases the absolute value criterion $\mathrm{E}\{|y|\}$ after each iteration [4], and so the algorithm converges at least to a local maximum — similarly, given a set of unlabeled data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ sampled from the random variable $\boldsymbol{x}$, step 2 transforms into

$$\boldsymbol{w}_{n+1} = \frac{\sum_{t=1}^{T} \boldsymbol{z}_t \mathrm{sign}(\boldsymbol{w}_n^\intercal\boldsymbol{z}_t)}{\| \sum_{t=1}^{T} \boldsymbol{z}_t \mathrm{sign}(\boldsymbol{w}_n^\intercal\boldsymbol{z}_t))\|_2}.$$

The most notable feature of this simple algorithm is that no parameters need to be tuned. Alternatively, one can use the polynomial time approaches in [26]–[28]. L1-PCA algorithms with guaranteed convergence to a global maximum have also been proposed, although they come at the expense of increased computational complexity [16], [29]. In addition, a simplified and faster, yet suboptimal, version of [16] can be found in [17].

In summary, the algorithm (assuming unwhitened data) would involve the following steps:

1) Compute and subtract the mean from the data.
2) Whitening: $\boldsymbol{z} \leftarrow \boldsymbol{D}_0^{-1/2}\boldsymbol{Q}^\intercal\boldsymbol{x}$, where $\boldsymbol{V}_0 = \boldsymbol{Q}\boldsymbol{D}_0\boldsymbol{Q}^\intercal$ is the EVD of the covariance matrix of $\boldsymbol{x}$.
3) Solve (29) to obtain $\boldsymbol{w}^*$. This step is the standard L1-PCA of the sphered data and any of the above mentioned algorithms (e.g., [4], [16], [17]) can be used for its computation.
4) Undo whitening: return $\boldsymbol{a}^* = \boldsymbol{Q}\mathbf{D}_0^{-1/2}\boldsymbol{w}^*$.

Interestingly, this algorithm has recently been shown to also carry out independent component analysis (ICA) with increased protection against outliers [24]. In other words, if the data follow the ICA model, this algorithm can allow its estimation. Otherwise, if the data have two classes, this same algorithm can perform LDA.

2) Further Comments: After sphering, the data covariance matrix equals the identity, i.e., $\boldsymbol{V}_0 = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^\intercal\} = \boldsymbol{I}$. Since $\boldsymbol{V}_0 = \boldsymbol{S}_W + \boldsymbol{S}_B$, it follows that

$$\boldsymbol{S}_W = \boldsymbol{V}_0 - \boldsymbol{S}_B = \boldsymbol{I} - \pi_1\pi_2\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}^\intercal$$

where we have invoked Eqns. (4) and (13). This matrix can be easily inverted using the well-known Sherman-Morrison formula, yielding:

$$\boldsymbol{S}_W^{-1} = \boldsymbol{I} - \frac{\pi_1\pi_2\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}^\intercal}{1 + \pi_1\pi_2\|\Delta\boldsymbol{\mu}\|^2}.$$

It follows that the optimal LDA projection is in the direction of the line joining the class means:

$$\boldsymbol{a}^* = \boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu} = \frac{\Delta\boldsymbol{\mu}}{1+\pi_1\pi_2\|\Delta\boldsymbol{\mu}\|^2}.$$

If the classes are well-separated, $\Delta\boldsymbol{\mu}$ also coincides with the direction of maximum variance (as recalled in Section II-C). Consequently, we may think that $\Delta\boldsymbol{\mu}$ (and hence $\boldsymbol{a}$) can be easily estimated in an unsupervised manner by L2-PCA. However, after data sphering all directions yield the same variance and hence $\Delta\boldsymbol{\mu}$ cannot be identified from a variance criterion. In other words, L2-PCA becomes inoperative. To fix this problem, the proposed L1-uLDA algorithm replaces L2-PCA with L1-PCA. Having replaced the square by the absolute value in the objective function, L1-PCA is able to identify $\Delta\boldsymbol{\mu}$ as shown throughout this section.

IV. The multiclass case

In the preceding section of the paper, the link between LDA and the constrained L1-PCA giving rise to L1-uLDA has been established in the binary (two-class) case. This section extends these results to the general scenario where data may be composed of more than two classes. We begin the derivations by recalling LDA in the multiclass case.

1) Multiclass LDA: In the multiclass case, where $\boldsymbol{x}$ is drawn from one of c classes $\mathcal{C}_1, \ldots, \mathcal{C}_c$, the *between-class* and *within-class* scatter matrices are defined respectively as [30]:

$$\boldsymbol{S}_B^c = \sum_{i=1}^{c} \pi_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\intercal = \sum_{i=1}^{c} \pi_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal$$

$$\boldsymbol{S}_W^c = \sum_{i=1}^{c} \pi_i \boldsymbol{V}_i.$$

where we have assumed again that the total mean vector of $\boldsymbol{x}$ is zero, i.e., $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{x}\} = \sum_{i=1}^{c} \pi_i \boldsymbol{\mu}_i = \boldsymbol{0}$, and maintained the previous notation: $\boldsymbol{\mu}_i = \mathrm{E}\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_i\}$ is the mean of $\mathcal{C}_i$, $\boldsymbol{V}_i = \mathrm{Cov}(\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}_i) = \mathrm{E}\{(\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\intercal \mid \boldsymbol{x} \in \mathcal{C}_i\}$ is the class covariance matrix, and $\pi_i = P(\boldsymbol{x} \in \mathcal{C}_i)$ is the corresponding a priori probability. Regardless of the value of c, it always holds that the data covariance matrix $\boldsymbol{V}_0 = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^\intercal\}$ can be decomposed as

$$\boldsymbol{V}_0 = \boldsymbol{S}_B^c + \boldsymbol{S}_W^c. \tag{30}$$

Multiclass LDA addresses the more general problem of projecting the data onto a q-dimensional subspace, with $q < p$, while also retaining as much of the class discriminating information as possible for the subsequent classification stage. The LDA projection is given by $\boldsymbol{y} = \boldsymbol{A}^\intercal\boldsymbol{x}$, where $\boldsymbol{A}$ is the $(p \times q)$ matrix whose columns are the dominant eigenvectors of the matrix pencil $(\boldsymbol{S}_W^c, \boldsymbol{S}_B^c)$, given by the solutions of the generalized eigenvalue (GEVD) problem [30]:

$$\boldsymbol{S}_W^c\boldsymbol{a} = \lambda\boldsymbol{S}_B^c\boldsymbol{a}. \tag{31}$$

This can be seen as the solution of two coupled optimization problems involving quadratic forms subject to the same unit variance constraint. In the first place, because of relation (30), the GEVD problem (31) can readily be expressed as:

$$\boldsymbol{S}_B^c\boldsymbol{a} = \lambda'\boldsymbol{V}_0\boldsymbol{a} \tag{32}$$

where $\lambda' \stackrel{\mathrm{def}}{=} 1/(1+\lambda)$. Some algebraic manipulations prove that Eqn. (32) is the solution of the constrained problem:

$$\max\ \boldsymbol{a}^\intercal\boldsymbol{S}_B^c\boldsymbol{a} \quad \text{subject to} \quad \mathrm{E}\{y^2\} = 1 \tag{33}$$

since $\mathrm{E}\{y^2\} = \mathrm{E}\{(\boldsymbol{a}^\intercal\boldsymbol{x})^2\} = \boldsymbol{a}^\intercal\boldsymbol{V}_0\boldsymbol{a}$. Likewise, problem (31) can also be expressed as:

$$\boldsymbol{S}_W^c\boldsymbol{a} = \lambda''\boldsymbol{V}_0\boldsymbol{a} \tag{34}$$

with $\lambda'' \stackrel{\mathrm{def}}{=} \lambda/(1+\lambda)$, which is the solution of

$$\min\ \boldsymbol{a}^\intercal\boldsymbol{S}_W^c\boldsymbol{a} \quad \text{subject to} \quad \mathrm{E}\{y^2\} = 1. \tag{35}$$

Problems (32)–(33) and (34)–(35) are indeed inextricably intertwined because, under the unit-variance constraint:

$$\begin{aligned}\arg\max_{\boldsymbol{a}} \boldsymbol{a}^\intercal\boldsymbol{S}_B^c\boldsymbol{a} &= \arg\max_{\boldsymbol{a}} \boldsymbol{a}^\intercal(\boldsymbol{V}_0 - \boldsymbol{S}_W^c)\boldsymbol{a} \\ &= \arg\max_{\boldsymbol{a}} (1 - \boldsymbol{a}^\intercal\boldsymbol{S}_W^c\boldsymbol{a}) \\ &= \arg\min_{\boldsymbol{a}} \boldsymbol{a}^\intercal\boldsymbol{S}_W^c\boldsymbol{a}\end{aligned}$$

and $\lambda' + \lambda'' = 1$. Therefore, the dominant eigenpairs of (32) and the maxima of (33) are directly linked to the minor (least significant) eigenpairs of (34) and the minima of (35), respectively. Finally, Eqn. (32) can be also expressed as

$$\lambda'\boldsymbol{V}_0\boldsymbol{a} = \boldsymbol{S}_B^c\,\boldsymbol{a} = \left(\sum_{i=1}^{c} \pi_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\intercal\right)\boldsymbol{a} = \sum_{i=1}^{c} (\pi_i m_i)\,\boldsymbol{\mu}_i, \tag{36}$$

where $m_i = \boldsymbol{\mu}_i^\intercal\boldsymbol{a}$. Hence, we see that the columns of the optimal matrix $\boldsymbol{A}$ lie in the space spanned by $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c$. In addition, pre-multiplying Eqn. (36) by $\boldsymbol{a}^\intercal$ and imposing the unit-variance constraint shows that the dominant generalized eigenvector of (32) is given by

$$\lambda'_{\max} = \max\ \boldsymbol{a}^\intercal\boldsymbol{S}_B^c\,\boldsymbol{a} = \max \sum_{i=1}^{c} \pi_i(m_i)^2. \tag{37}$$

In other words, the LDA solution lies in the direction maximizing the average squared means of the projected classes under the unit-variance constraint. In a totally similar fashion, it is simple to observe that the minor generalized eigenvector of (34) under the unit-variance constraint is given by

$$\lambda''_{\min} = \min\ \boldsymbol{a}^\intercal\boldsymbol{S}_W^c\,\boldsymbol{a} = \min \sum_{i=1}^{c} \pi_i\sigma_i^2$$

which minimizes the average variance of the projected classes.

In summary, multiclass LDA tries to find projections maximizing the average squared mean while minimizing the average variance of the projected classes under a unit-norm constraint on the projected data. Inspired by this result, we show next that L1-uLDA follows a related approach in the multiclass scenario.

2) Multiclass L1-Ulda in Low Dispersion Regime: A slight variant of LDA is obtained by replacing the squares by absolute values in Eqn. (37), yielding

$$\max \sum_{i=1}^{c} \pi_i|m_i| \quad \text{subject to} \quad \mathrm{E}\{y^2\} = 1. \tag{38}$$

Some algebra shows that the solution fulfils

$$\boldsymbol{V}_0\boldsymbol{a} \propto \sum_{i=1}^{c} \pi_i \text{sign}(m_i)\boldsymbol{\mu}_i \tag{39}$$

which, as in classical LDA, lies in the subspace spanned by the class means [cf. Eqn. (36)].

This alternative criterion remains supervised, since computing m_i requires prior knowledge of the class labels. We now prove that the L1-uLDA criterion optimizes (38) in an unsupervised fashion when the data classes are tightly concentrated around their means. To complete the link with LDA, the next section will analyze the other extreme case where the within-class spread is large.

Recall that the L1-uLDA criterion is given by

$$\max \text{E}\{|y|\} \quad \text{subject to} \quad \text{E}\{y^2\} = 1. \tag{40}$$

We first consider the case where the classes are tightly concentrated around their means, i.e., $\text{trace}(\boldsymbol{V}_i) \approx 0$ and $\sigma_i \to 0$, $\forall i = 1, 2, \ldots, c$. Hence, the probability density function (pdf) of the projected data accepts the approximation:

$$f_y(u) \underset{\sigma_i \to 0}{\longrightarrow} \sum_{i=1}^{c} \pi_i \delta(u - m_i).$$

Therefore:

$$\text{E}\{|y|\} = \int_{-\infty}^{+\infty} |u| f_y(u) du$$

$$\underset{\sigma_i \to 0}{\longrightarrow} \sum_{i=1}^{c} \pi_i \int_{-\infty}^{+\infty} |u| \delta(u - m_i) du = \sum_{i=1}^{c} \pi_i |m_i|.$$

The L1-uLDA criterion (40) indeed asymptotically optimizes (38) unsupervisedly, i.e., without requiring knowledge of the class labels. Remark that this result is independent of the underlying data distribution. The only requirement is that each class be closely concentrated around its mean.

3) Multiclass L1-uLDA in High Dispersion Regime: We now turn to the case where the within-class dispersion is large as compared with the between-class dispersion. To address this more involved scenario, the assumption is now made that the data $\boldsymbol{x}$ from class $\mathcal{C}_i$ are Gaussian distributed. The pdf of $\boldsymbol{x}$ is therefore $\boldsymbol{x} \sim \sum_{i=1}^{c} \pi_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{V}_i)$. Under this model it can be shown that

$$\text{E}\{|y|\} = \int_{-\infty}^{\infty} |y| f(y) \text{d}y = \sqrt{2} \sum_{i=1}^{c} \pi_i \sigma_i g(\alpha_i) \tag{41}$$

where α_i and $g(\alpha_i)$ were respectively defined in Eqns. (21) and (23), and σ_i^2 is the variance of the projected ith class. A series of algebraic manipulations, similar to those for the two-class case, yield that the stationary points of the L1-uLDA criterion verify the equation

$$\boldsymbol{V}_0\boldsymbol{a} = \frac{1}{\text{E}\{|y|\}} \sum_{i=1}^{c} \pi_i \left[\frac{\sqrt{2/\pi}}{\sigma_i} e^{-\alpha_i^2} \boldsymbol{V}_i \boldsymbol{a} + \text{erf}(\alpha_i) \boldsymbol{\mu}_i \right]. \tag{42}$$

If the spread of the clusters is large in relation to their respective means, α_i are close to zero, implying $\exp(-\alpha_i^2) \approx 1$ and $\text{erf}(\alpha_i) \approx 0$. Then (42) can be rewritten as the GEVD problem

$$\bar{\boldsymbol{S}}_W^c \, \boldsymbol{a} = \lambda''' \boldsymbol{V}_0 \boldsymbol{a} \tag{43}$$

where

$$\bar{\boldsymbol{S}}_W^c = \sum_{i=1}^{c} \pi_i \, \beta_i \boldsymbol{V}_i, \text{ with } \beta_i = \frac{\sqrt{2/\pi}}{\sigma_i}.$$

Equation (43) is the multiclass LDA solution with a generalized weighted within-class scatter matrix. Furthermore, it becomes equivalent to the traditional LDA solution (34) when all classes have equal covariance matrices, $\boldsymbol{V}_i = \boldsymbol{V}$, $\forall i = 1, 2, \ldots, c$. On the other hand, remark that in the low dispersion regime α_i become large and then $\exp(-\alpha_i^2) \approx 0$ and $\text{erf}(\alpha_i) \approx \text{sign}(\alpha_i)$, leading to the same conclusions as in the previous section.

We have established the link between LDA and L1-uLDA in extreme cases of within-class dispersion. In the general case, the actual solutions of the L1-uLDA criterion (40) will lie somewhere between these two extremes, preserving in any case the ability of multiclass LDA to distinguish between the different classes, as demonstrated by the numerical experiments reported next.

V. Numerical Experiments

A number of computer simulations in a variety of experimental conditions are performed to validate the theoretical study developed in this paper and to test the equivalence to LDA while sparing the need for labelled data (unsupervised operation). Other techniques for unsupervised and semi-supervised LDA have been proposed in the literature (see e.g. [31]–[33]). It should be kept in mind, however, that to limit the scope of the paper our focus is only on the connection between L1-PCA and LDA, without claiming its potential superiority over existing alternative techniques for classification or clustering. A comparative performance analysis should be the topic of future research. Note also that other complementary analyses of L1-PCA (robustness against outliers, linear dimension reduction capabilities, comparison with other robust PCA approaches, etc.) can already be found in the cited literature (see e.g. [4, 17]).

Our experiments include synthetically generated as well as real data. For simplicity, we apply our tests to L1-uLDA, which is carried out by applying L1-PCA to prewhitened data. To perform L1-PCA in step 3 of the L1-uLDA optimization algorithm in Section III-D, and unless otherwise stated, we employ the bit flipping algorithm presented in [17]. A free MATLAB implementation of this algorithm is provided in [34].

A. Synthetic Data

Bivariate data can be easily visualized as points in a two-dimensional scatter plot. For this reason, we use them to illustrate the performance of L1-uLDA first before moving on to the general case of more than two dimensions.

1) Finite Sample Size: The theoretical analysis of the previous section relies on an ensemble or distributional characterization with the implicit assumption of infinite sample size. The first experiment evaluates the ancillary L1-uLDA criterion in short data records, and shows the existence of spurious

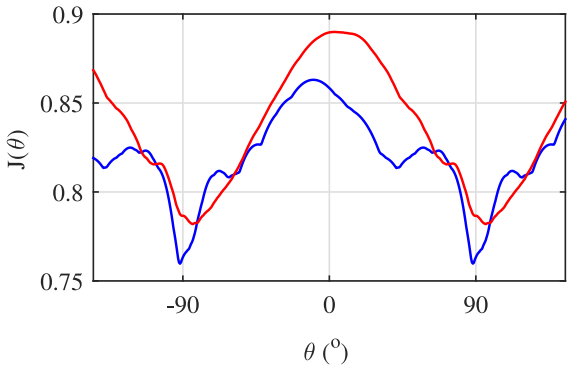


Fig. 2. Equivalent L1-uLDA criterion (44) computed from a realization of $T = 200$ points (red line) and $T = 50$ points (blue line) as a function of the angle formed by the projection direction with the horizontal axis. Spurious local maxima are clearly visible in the latter case.

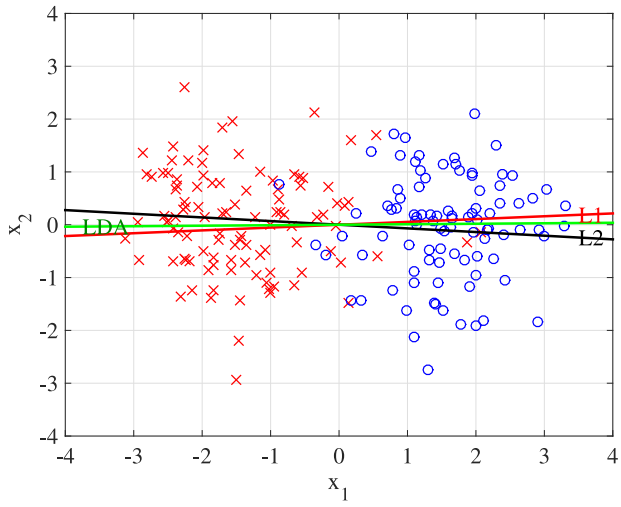


Fig. 3. Scatter plot of $T = 200$ samples from a mixture of Gaussian clusters. Red line: direction of the optimal projection according to the proposed criterion (L1-uLDA). Green line: optimal Fisher's discriminant direction (LDA). Black line: direction of maximum variance (L2-PCA).

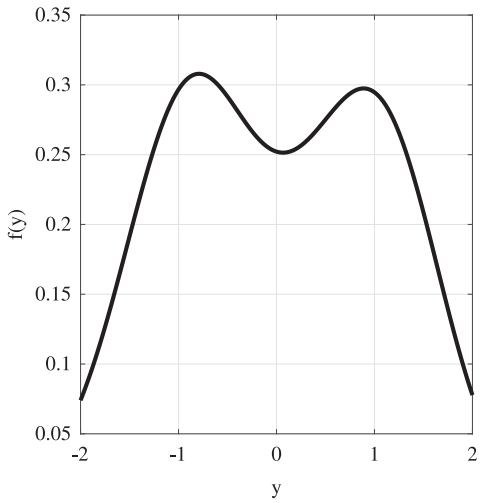


Fig. 4. Histogram of the data projected onto the direction determined by L1-uLDA in Fig. 3. The two modes point out the existence of two clusters.

local maxima for insufficient sample size. We generate two-dimensional random samples from a mixture of two equiprobable ($\pi_1 = \pi_2 = 0.5$) Gaussian distributions with means $\boldsymbol{\mu}_1 = -[1.5, 0]^\intercal$ and $\boldsymbol{\mu}_2 = [1.5, 0]^\intercal$, and identical covariance matrix $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Let $y = \boldsymbol{a}(\theta)^\intercal \boldsymbol{x}$ be the one-dimensional projection of the data vector $\boldsymbol{x}$ in the direction $\boldsymbol{a}(\theta) = [\cos(\theta), \sin(\theta)]^\intercal$. It can be readily shown that the L1-uLDA criterion (17) is also equivalent to maximizing the unconstrained function

$$J(\theta) = \frac{\mathrm{E}\{|y|\}}{\sigma_y} \tag{44}$$

where σ_y is the standard deviation of y. Indeed, the optimization of a Rayleigh quotient as that of (44) is equivalent to optimizing the numerator while constraining the denominator. The maxima and the minima of (17) can be easily studied by plotting $J(\theta)$. The red line in Fig. 2 plots function $J(\theta)$ estimated from a realization of $T = 200$ data samples. The curve passes through a global maximum at $\theta = 0$ degrees, which is consistent with the fact that the optimal direction resulting from Fisher's discriminant is parallel to the x_1-axis as is easy to check (line joining the means; see Eqn. (9)).

Next, the experiment is repeated, though this second time the objective function is estimated from only the first 50 samples of the previous 200-samples dataset. The corresponding objective function is shown in the blue line of Fig. 2. Observe that several spurious local maxima appear when reducing the sample size. We conclude that it is convenient to carry out L1-PCA using globally convergent algorithms such as that of [16] if the sample size is small. The bit flipping algorithm of [17], though suboptimal, achieves good convergence in all our experiments.

2) Comparison with LDA and L2-PCA: Since L1-PCA is related with L2-PCA (Section II-C), one may naturally wonder about their comparative behavior. To this end, Fig. 3 shows the scatter plot of the $T = 200$ data points used in the previous experiment, where data from class 1 and class 2 appear as red crosses and blue circles, respectively. The red line marked 'L1' is the direction of the optimal L1-uLDA projection. Projecting the whole dataset onto this line gives the histogram plotted in Fig. 4, clearly showing the clustered structure of the data. Similarly, the black line labelled as 'PCA' in Fig. 3 is the direction of maximum variance of the data, constructed by classical L2-PCA, whereas the green line marked 'LDA' represents Fisher's discriminant direction for this dataset.

Even though the three approaches present a similar behavior in this example, strictly speaking classical L2-norm based PCA is not a linear discrimination technique and, therefore, it may yield different results in certain cases. Supporting this claim, Fig. 5 shows the projection directions obtained after repeating the experiment under the same conditions except for the covariance matrices, which are set to $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \mathrm{diag}(1, 3)$, so that the direction of maximum variance lies now along the x_2-axis.

3) More Than Two Dimensions: Let us now extend the previous analysis to more than two dimensions. Consider data from two classes with prior probabilities $\pi_1 = \pi_2 = 1/2$ in a p-dimensional space, with $p > 2$. Data from class i, $i = 1, 2$, are drawn from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{V}_i)$, with $\boldsymbol{\mu}_1 =$

TABLE I
ANGLES IN DEGREES BETWEEN LDA AND L1-uLDA DIRECTIONS ($\Delta\theta_{\mathrm{L1}}$), AND LDA AND L2-PCA DIRECTIONS ($\Delta\theta_{\mathrm{L2}}$). VALUES SHOWN REPRESENT MEAN $\pm$ STANDARD DEVIATION OVER 10 INDEPENDENT REALIZATIONS OF $T = 200p$ DATA SAMPLES. p: DATA DIMENSIONALITY; μ: DISTANCE BETWEEN CLASS MEANS

	(a) $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \mathbf{I}$				(b) $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \mathrm{diag}(1, p, \ldots, p)$			
	$p = 5$		$p = 10$		$p = 5$		$p = 10$	
μ	$\Delta\theta_{\mathrm{L1}}$	$\Delta\theta_{\mathrm{L2}}$	$\Delta\theta_{\mathrm{L1}}$	$\Delta\theta_{\mathrm{L2}}$	$\Delta\theta_{\mathrm{L1}}$	$\Delta\theta_{\mathrm{L2}}$	$\Delta\theta_{\mathrm{L1}}$	$\Delta\theta_{\mathrm{L2}}$
3	5.1 ± 2.7	4.5 ± 1.9	7.4 ± 1.7	6.1 ± 1.2	2.9 ± 1.5	87.5 ± 2.7	2.6 ± 0.5	89.4 ± 0.4
5	1.5 ± 0.4	4.6 ± 1.7	1.6 ± 0.6	4.1 ± 0.9	1.0 ± 0.5	9.4 ± 4.5	0.5 ± 0.2	88.1 ± 0.9
7	0.3 ± 0.6	3.7 ± 1.1	0.2 ± 0.4	4.0 ± 0.8	0.2 ± 0.2	4.3 ± 1.7	0.1 ± 0.1	13.5 ± 2.8
10	0.0 ± 0.0	3.5 ± 1.0	0.0 ± 0.0	3.9 ± 1.0	0.0 ± 0.0	2.7 ± 1.0	0.0 ± 0.0	3.4 ± 0.5

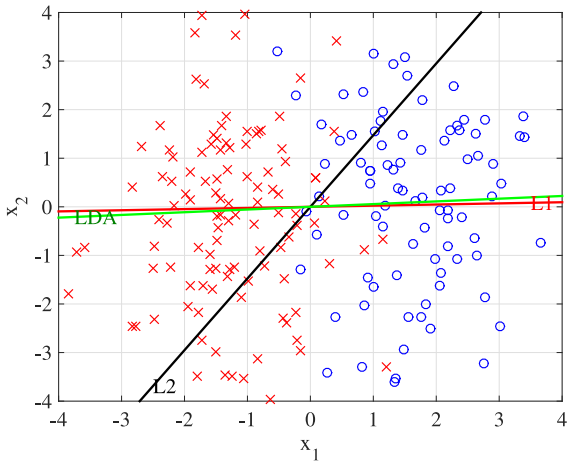


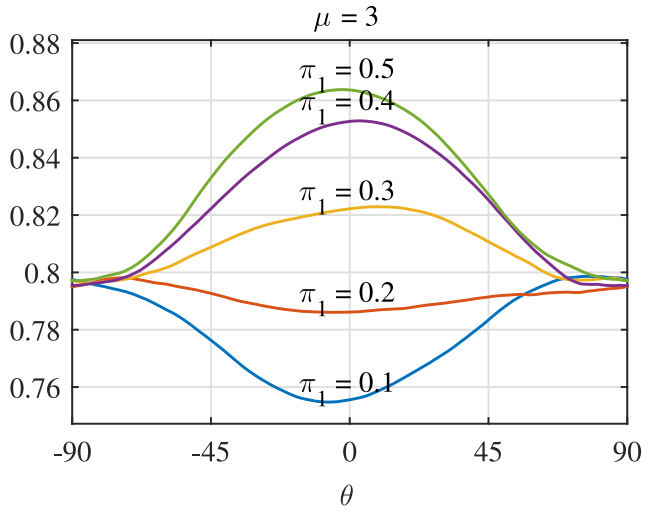
Fig. 5. Linear discrimination performance for classes with non-isotropic covariance matrices, under the same general conditions of the experiment of Fig. 3. L2-PCA fails to find a discriminant solution, whereas L1-uLDA projector still lies close to LDA's.

$-\frac{\mu}{2}\boldsymbol{e}_1, \boldsymbol{\mu}_2 = \frac{\mu}{2}\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the vector with 1 in the first entry and 0's elsewhere. Accordingly, parameter μ represents the distance between the class means. The covariance matrices are chosen either as (a): $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{I}$, or (b): $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \mathrm{diag}(1, p, \ldots, p)$. In both cases, the direction resulting from Fisher's discriminant is parallel to the x_1-axis. In (b), the x_1-axis is also the direction of minimum variance of the classes. The LDA projection vector ($\boldsymbol{a}_{\mathrm{LDA}}$), the L2-PCA principal direction of the data ($\boldsymbol{a}_{\mathrm{L2}}$) and the proposed L1-uLDA projection vector ($\boldsymbol{a}_{\mathrm{L1}}$) are estimated from the same $T = 200\,p$ data points and normalized to unit length. To compare Fisher's discriminant with the two other methods, the angles formed by the LDA vector $\boldsymbol{a}_{\mathrm{LDA}}$ with $\boldsymbol{a}_{\mathrm{L1}}$ and $\boldsymbol{a}_{\mathrm{L2}}$ are calculated by the formula $\Delta\theta_x = \arccos(|\boldsymbol{a}_{\mathrm{LDA}}^T \boldsymbol{a}_x|)$, where $x \in \{\mathrm{L1}, \mathrm{L2}\}$, for different parameter pairs (p, μ). The absolute value simply forces the angle to lie in the first quadrant. Table I shows the mean angle values $\pm$ standard deviations (in degrees) for 10 independent data realizations. Observe that θ_{L1} tends to 0° as μ increases, which means that $\boldsymbol{a}_{\mathrm{L1}}$ gradually aligns itself with $\boldsymbol{a}_{\mathrm{LDA}}$, achieving a perfect fit when the clusters are well separated, as predicted by our theoretical analysis. While the different approaches indeed become equivalent when the classes are clearly distinct, L1-uLDA keeps closer to LDA than L2-PCA when the classes tend to overlap. This observation is particularly evident in case (b), where $\boldsymbol{a}_{\mathrm{L2}}$ lies almost orthogonal to $\boldsymbol{a}_{\mathrm{LDA}}$ for small μ.

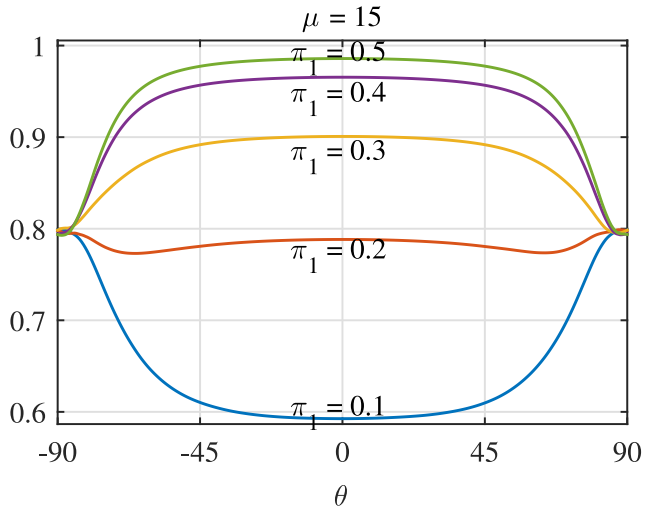
4) Data Corrupted With Outliers: Let us return to the case $p = 10$ and $\mu = 5$ with $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{I}$ of the previous experiment. We replace the 20% of the data from each class by outliers generated by raising normalized Gaussian random samples to the third power and centering them on the corresponding class means. Some outliers considerably far from the clusters are obtained in this manner. For the corrupted data, the mean angle between the optimal L1-uLDA projection vector and the horizontal axis, averaged over 10 independent experiments, is equal to 8.8° with an standard deviation of 4.9°. This mean angle is equal to 9.4° for LDA applied on the same corrupted data (standard deviation equal to 3.3°). Both means are compared with the aid of Student's *t*-test and do not differ significantly at the 5% level. For comparison, the angle is 75.9° (standard deviation 1.4°) when calculated by L2-PCA. This experiment suggests that the proposed L1-uLDA method is as robust to outliers as the original LDA method while operating in an unsupervised manner.

5) Unbalanced Clusters: The L1-uLDA criterion may fail when one of the classes is much more frequent than the other, so that the cluster masses become too unbalanced. Consider two-dimensional random samples from a mixture of two Gaussian distributions with means $\boldsymbol{\mu}_1 = -\frac{\mu}{2}\boldsymbol{e}_1, \boldsymbol{\mu}_2 = \frac{\mu}{2}\boldsymbol{e}_1$, where $\boldsymbol{e}_1 = [1, 0]^\intercal$, and common covariance matrix $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Unlike the above experiments for classes with the same prior probabilities $\pi_1 = \pi_2 = 0.5$, now the clusters are not assumed to have a similar number of points. Fig. 6a plots the objective function $J(\theta)$, defined in Eqn. (44), for $\mu = 3$ and different values of the class prior probabilities π_1 and π_2, with $\pi_1 + \pi_2 = 1$. The plots show that, for unevenly sized clusters (e.g., $\pi_1 = 0.1$ or $\pi_1 = 0.2$), the maxima of $J(\theta)$ occur at $\theta = \pm 90^\circ$, and hence L1-uLDA would produce directions orthogonal to Fisher's discriminant. This outcome can be interpreted as follows. When the priors π_i are highly unbalanced, the zero-mean constraint forces the distribution with larger mass to be concentrated around zero. Therefore, the mean of the projected class is always close to zero as well, no matter on which direction we project the data. This prevents the projected data from forming a cluster where the mean is large relative to the standard deviation, which is required for L1-uLDA to work. Fig. 6b plots the new function $J(\theta)$ obtained after setting $\mu = 14$, resulting in a clearer cluster separation. Observe that the curve $\pi_1 = 0.2$ now passes through a maximum at $\theta = 0^\circ$, although it is not the global maximum. Therefore, L1-uLDA may fail again in this case. See Appendix C for a further explanation.

6) The Multiclass Case: Consider a $p = 30$–dimensional data set, with $c = 30$ classes, and 500 samples per class (implying that the prior probabilities are the same for all classes). The class means $\boldsymbol{\mu}_i$ are generated from a normal distribution with



(a) Close clusters.



(b) Well separated clusters.

Fig. 6. Objective function $J(\theta)$ [eqn. (44)] estimated from $T = 200$ samples in the bivariate Gaussian case with identity covariance matrix for both classes and different class priors π_1 ($\pi_2 = 1 - \pi_1$). Each curve is obtained by averaging over 50 independent data realizations.

covariance matrix $4\boldsymbol{I}_{30}$, where $\boldsymbol{I}_n$ is the identity matrix of order n. We also assume equal class covariance matrices $\boldsymbol{V}_i = \boldsymbol{I}_{30}$ for $i = 1, 2, \ldots, 30$. Under these conditions, we are not in the high or the low dispersion regime, but somewhere in between. In this experiment, we successively project the data onto $\ell = 1$ to 29 dimensions, using the matrices $\boldsymbol{A} \in \mathbb{R}^{30\times\ell}$ for which the multiclass LDA criterion and the L1-uLDA criterion are maximal. To generate several projection directions, we follow the approach in [4], i.e., we run L1-PCA several times on the whitened data with the additional constraint that the solution found in the kth run had to be orthogonal to the previously found $(k-1)$ solutions. Finally, to test the separability of the projected samples, they are classified using the K-means algorithm. Fig. 7 shows the accuracy of the classification averaged over 100 independent experiments for multiclass LDA and for the L1-uLDA criterion. As can be seen, both methods perform

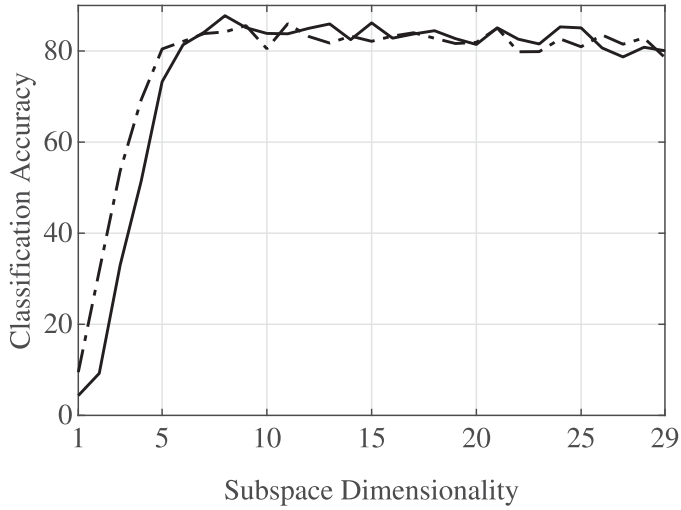


Fig. 7. Classification accuracy as a function of the dimensionality of the projected subspace, for 30 classes and 30 features. Solid line: L1-uLDA. Dashed line: LDA.

similarly excepting for very low (1 to 5) dimensions, in which the standard supervised criterion is superior to the unsupervised one. This can be explained by the observed tendency of multiclass L1-uLDA to form one-dimensional projections with only two well-separated clusters, one on each side of the origin.

B. Experiments With Real Data

The theoretical analysis of Section III assumes that the density is Gaussian for both classes. This section presents three additional experiments to test the behavior of the proposed method when that assumption is violated, as is often the case when analyzing real data. To this end, we use three classical datasets with a deliberately small number of samples so that we can also test the impact of short sample length.

1) Iris Dataset: Originally used by Fisher when introducing LDA [35], this is perhaps the most famous dataset in the pattern recognition literature. It contains the sepal and petal length and width of 150 iris flowers from three different species (*setosa*, *versicolor* and *virginica*). One of them (*setosa*) is linearly separable from the other two; the latter are not linearly separable from each other. The dataset contains 50 samples from each of the three species of iris. It can be downloaded from the UCI Machine Learning Repository [36], [37].

The database can be written as a matrix $\boldsymbol{X} \in \mathbb{R}^{4\times 150}$, corresponding to 4 measurements from 150 flowers. After centering and sphering the data, we calculated the unit-norm L1-uLDA projection vector $\boldsymbol{a}_{\mathrm{L1}} \in \mathbb{R}^4$ of the data matrix. Fig. 8 plots the probability density function, obtained by a kernel density estimator method, of the projection of the data onto the direction of $\boldsymbol{a}_{\mathrm{L1}}$. The distribution is clearly bimodal, which suggests that the sample is not homogeneous but arises from at least two different populations [22]. In fact, the mode on the left exclusively corresponded to the *setosa* individuals, while the other mode represented the mixture of the other two classes.

Next, we separate the projected data into two groups using as threshold the local minimum between the two peaks in the

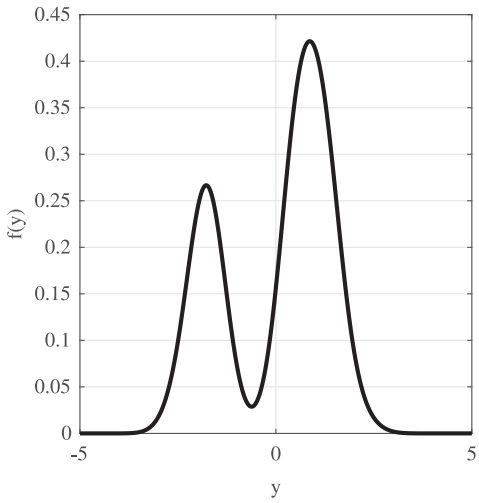


Fig. 8. Experiment on the Iris dataset. Histogram of the projection of the data onto the L1-uLDA optimal direction.

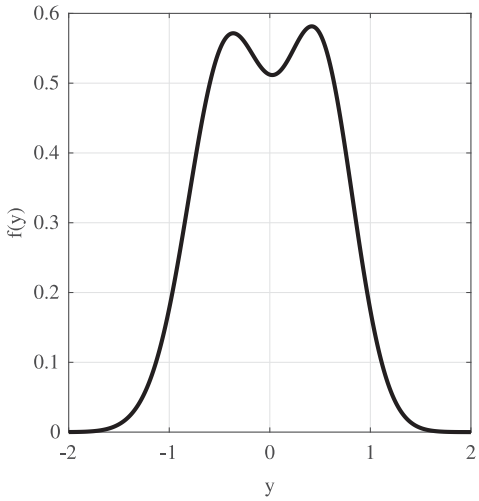


Fig. 9. Experiment on the Iris dataset. Histogram of the projection of the data containing only *versicolor* and *virginica* individuals (corresponding to the right-hand mode of Fig. 8) onto the L1-uLDA optimal direction.

bimodal distribution. It is thus obtained one cluster that only contains iris *setosa* and another cluster that only contains *versicolor* and *virginica*. Then we repeat the procedure (centering, sphering and L1-PCA) on the cluster with the *versicolor* and *virginica* individuals. A bimodal distribution is again apparent as a result (see Fig. 9). Dividing these data once more into two groups, using zero as threshold for deciding to which class each flower is allocated, only three flowers are misclassified: one *virginica* is misclassified as a *versicolor* and two *versicolor* are misclassified as *virginica*. This 98% accuracy is achieved by L1-uLDA in unsupervised operation, without knowledge of the sample labels.

2) Wisconsin Diagnostic Breast Cancer Dataset: This dataset is used for breast cancer diagnosis. It contains 569 instances of 30 real-valued features from images of a fine needle aspirate of a breast mass, that describe the characteristics of the cell nuclei present in the image. Of these 569 instances, 357

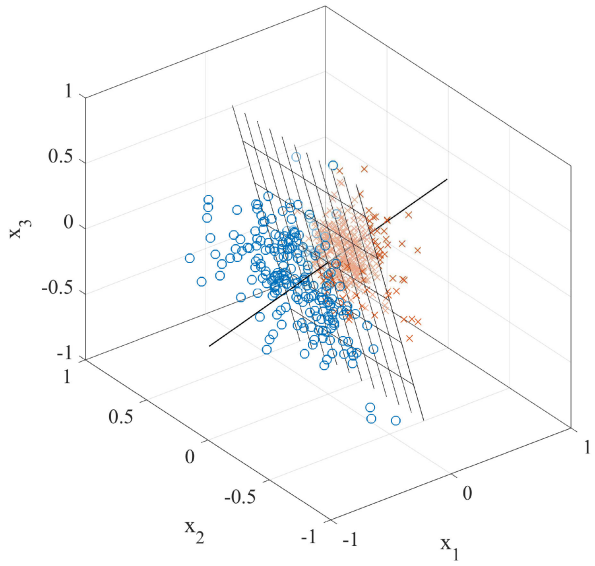


Fig. 10. Scatter plot of the three first principal components of the Wisconsin Diagnostic Breast Cancer Dataset [38]. The plane represents the decision boundary that partitions the space into positive and negative projections onto the L1-uLDA optimal direction, which is plotted as a black line.

correspond to 'benign' (non-cancerous) cases while 212 are 'malignant' (cancerous). The dataset is downloaded from the UCI Machine Learning Repository [36], [38].

First of all, as a preprocessing step, we standardize the 30 features to have zero mean and unit variance. Secondly, we obtain the three major principal components from the standardized features to aid visualization. Fig. 10 shows the scatter plot of these three principal components. The first class, 'benign,' is represented as 'crosses,' and the second class, 'malignant,' as 'circles'. Thirdly, the L1-PCA algorithm of [17] is applied to these principal components (which are already sphered by construction), and not to the original data.

The histogram of data projected onto the direction resulting from L1-PCA again suggests that the sample arises from different populations. The data are classified into two groups using zero as threshold; the plane in Fig. 10 represents the corresponding classification boundary. Interestingly, the sensitivity of this classifier (i.e., the probability of a correct classification given that the instance is 'malignant') is 0.91 while the specificity (i.e., the probability of a correct classification given that the instance is 'benign') equals 0.90. For comparison, we also classify by assigning each instance to the class with the closest mean using the Malahanobis distance as a metric, which is equivalent to Fisher's linear classification rule [2]. In this case, the sensitivity and specificity equal 0.98 and 0.87 respectively. As a further experiment, we consider that training data often contain misclassified items, as a result of the mistakes made by annotators who have performed the labelling. To model this effect, the data are classified again with Fisher's rule after having interchanged the labels of the 56 individuals lying closer to the classification boundary. Since most of them corresponded to 'benign' samples, sensitivity is almost unaltered but specificity decreases to 0.74. The results obtained by L1-uLDA are not affected since this method, being unsupervised, does not make use of the data labels.

Fig. 11. Sample images of the AT&T face database.

3) Face Recognition Via Subspace Projection: 2D images are usually vectorized by stacking their columns into a 1D vector. Suppose that we are given a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ of face images $\boldsymbol{x}_i \in \mathbb{R}^p$ labeled with the person's identity. Given an unlabeled test image, the goal is to identify the featured person. Dimension-reduction based techniques are an effective approach to this problem. These techniques comprise in essence the following steps: 1) transform the data to a space of fewer dimensions, 2) project the training samples and the test image into the low-dimensional space, 3) assign the projected test image to the closest projected training image. The Eigenfaces method described in [39], which is based in classical L2-PCA, and the Fisherfaces method [40], inspired by Fisher's LDA, are among the most successful dimension reduction techniques in this field. The aim of the present experiment is to test the ability of L1-uLDA to solve image classification problems. After sphering the data, the iterative L1-PCA algorithm of [4] is now preferred over the bit flipping algorithm of [17] as the latter becomes too costly in this experimental setting. To generate several projection directions, we run L1-PCA several times with the additional constraint that the solution found in the kth run had to be orthogonal to the previously found $(k-1)$ solutions (see [4] for details). The images are then projected onto the space spanned by the dominant L1-PCA principal components.

This experiment uses the AT&T face dataset, which contains 10 facial images of each of 40 different subjects (400 images in total). For a given subject, the facial images differ in lighting, pose, expression (open / closed eyes, smiling / not smiling) and details (glasses / no glasses). Each image is 92×112 pixels, with 256 grey levels per pixel. Fig. 11 shows some of the images of the database.

Fig. 12 shows the recognition rate with 10-fold cross-validation versus the dimension of the image projection subspace, where we use 90% of the images for training and 10% for validation, with a slightly superior performance of the L2-PCA based method. We repeat the experiment but, to generate data far from the clusters, 10 images of the training dataset are severely distorted by salt-and-pepper noise in this second case. The result, shown in Fig. 13, illustrates again the robustness of L1-principal subspaces against outliers. L1-uLDA reaches a recognition performance up to the mark of LDA's but with no need for training data.

4) Face Recognition With CNNs: Convolutional Neural Networks (CNNs) constitute the state-of-the-art in the field of face recognition. In these last experiments, we test the capabilities of L1-uLDA for dimensionality reduction when applied to the output of a ResNet network [41], which has been implemented in the `dlib` C++ library [42], with 29 convolutional layers. This CNN transforms the image of a human face into a 128 dimensional vector, where images of the same person are mapped close to each other and images from different people become separated far apart after the transformation.

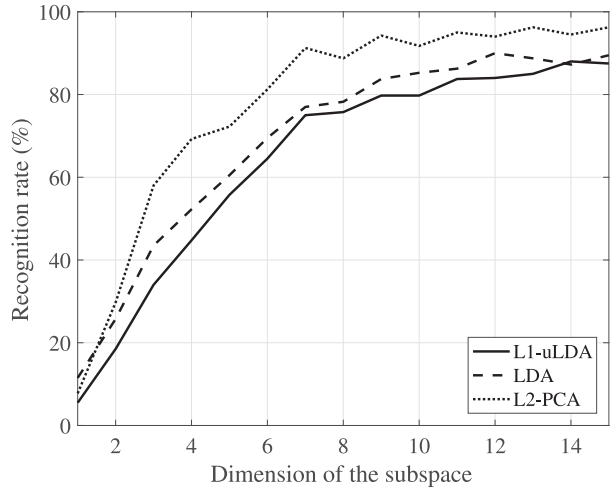


Fig. 12. Classification accuracy versus feature dimension in the AT&T database.

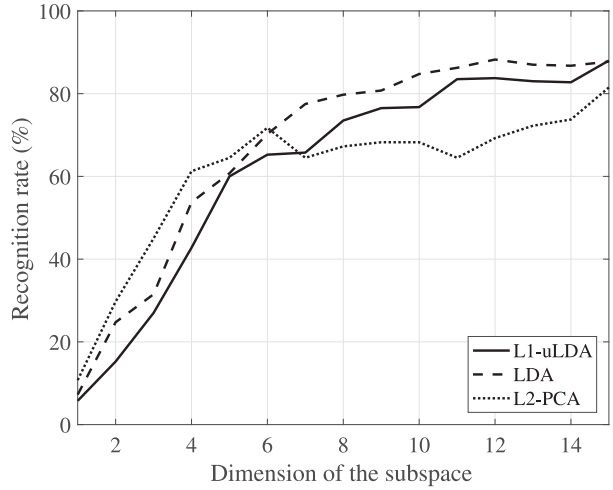


Fig. 13. Classification accuracy versus feature dimension in the AT&T database with outliers.

The experiments are conducted on images from the database "Faces in the wild" [43], which comprises a total of 13 233 labelled images of faces from 5 749 persons, albeit only the 3 023 photographs of the 65 people with more than 20 images in the dataset are used in our experiments. The images vary in height and length, and the faces appear in different angles and scenarios. As a pre-processing, the Histogram of Oriented Gradients (HOG) [44] technique was used to detect the location of the faces, which is necessary before applying the CNN to the pictures. It is empirically observed that about half of the eigenvalues of the autocovariance matrix of the CNN outputs are

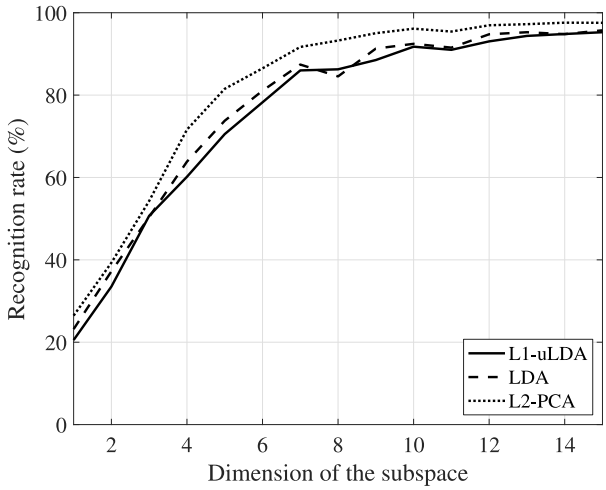


Fig. 14. Classification accuracy versus feature dimension in the 'Faces in the wild" database when the dimensionality reduction is applied to the 128 dimensional outputs of a convolutional neural network.

negligible as compared with the others. This allows to replace the CNN outputs with their low rank SVD reconstruction, which filters the noise while preserving most of the data variance.

Similarly to the previous experiment, we project the 128-dimensional vectors onto a space of fewer dimensions, using training samples to calculate the matrix of the linear transformation by one of the above-referred techniques (i.e. Fisher's LDA, L2-PCA or L1-uLDA). We emphasize that neither PCA nor L1-uLDA use for anything the 'labels,' or 'names,' of the persons in the images. Then, test images are projected into the low-dimensional space and assigned to the label of the closest projected training image. A simple K-nearest neighbors algorithm (K-NN) is used to classify each unlabelled face and determine to which person it belongs. Fig. 14 shows the recognition rate with 10-fold cross-validation (90% of the images for training and 10% for testing) versus the dimension of the image projection subspace. As before, L2-PCA presents a slightly better performance, and unsupervised L1-uLDA is almost equivalent to supervised Fisher's LDA.

VI. Conclusion

The present work has shown that L1-PCA of whitened data can perform LDA in an unsupervised fashion, i.e., without the need for training data. This connection between L1-PCA and LDA had gone previously unnoticed. After whitening, L1-PCA can be carried out with several efficient algorithms recently proposed in the literature. We refer to this L1-PCA variant as L1-uLDA. Compared with L2-PCA and kurtosis, L1-uLDA offers enhanced robustness to outliers, which makes it particularly attractive when processing faulty or unreliable data as confirmed by a variety of computer experiments. Further theoretical research should explore its extension to scenarios with possibly non-Gaussian data. Also, the technique presents limitations in the presence of highly unbalanced clusters, a challenging scenario that should be considered in future works. Finally, our focus was on establishing the connection between L1-PCA and LDA, and space lacked for a comparison with other techniques for unsupervised and semi-supervised LDA, which will be the topic of further investigations.

Appendix

A. Mathematical Derivation of Stationary Point Analysis

This appendix provides details about the stationary point analysis of Section III-B. First, let us review some basic concepts about function optimization with equality constraints, which can be expressed as:

$$\max_{\boldsymbol{a}} f(\boldsymbol{a}) \quad \text{subject to} \quad h(\boldsymbol{a}) = 0 \tag{45}$$

where $f, h : \mathbb{R}^p \to \mathbb{R}$. Let $\mathcal{L}(\boldsymbol{a}, \lambda) = f(\boldsymbol{a}) + \lambda\, h(\boldsymbol{a})$ be the Lagrangian function and let $\boldsymbol{L}(\boldsymbol{a}, \lambda)$ be the Hessian matrix of $\mathcal{L}(\boldsymbol{a}, \lambda)$ with respect to $\boldsymbol{a}$, i.e.,

$$(\boldsymbol{L})_{ij} = \frac{\partial^2 \mathcal{L}}{\partial a_i \partial a_j}(\boldsymbol{a}, \lambda).$$

This matrix can also be decomposed as

$$\boldsymbol{L}(\boldsymbol{a}, \lambda) = \boldsymbol{F}(\boldsymbol{a}) + \lambda \boldsymbol{H}(\boldsymbol{a}) \tag{46}$$

where $\boldsymbol{F}$ and $\boldsymbol{H}$ are the Hessian matrices of $f(\boldsymbol{a})$ and $h(\boldsymbol{a})$, respectively. In addition, the tangent space at a point $\boldsymbol{a}^*$ on the surface $\mathcal{S} = \{\boldsymbol{a} \in \mathbb{R}^p : h(\boldsymbol{a}) = 0\}$ is defined as the set

$$T(\boldsymbol{a}^*) = \{\boldsymbol{v} : \boldsymbol{v}^\intercal \nabla_{\boldsymbol{a}} h(\boldsymbol{a}^*) = 0\}$$

where $\nabla_{\boldsymbol{a}}$ represents the gradient operator. We have the following generic result [45, Chap. 20]:

Theorem 2: Let $\boldsymbol{a}^*$ be a local maximizer of f subject to $h(\boldsymbol{a}) = 0$. Then, there exits $\lambda^* \in \mathbb{R}$ such that

C1) $\nabla_{\boldsymbol{a}} f(\boldsymbol{a}^*) + \lambda^* \nabla_{\boldsymbol{a}} h(\boldsymbol{a}^*) = 0$, and

C2) for all $\boldsymbol{v} \in T(\boldsymbol{a}^*)$, we have $\boldsymbol{v}^\intercal \boldsymbol{L}(\boldsymbol{a}^*, \lambda^*)\boldsymbol{v} < 0$.

Let us particularize this generic result to our problem, where we have $f(\boldsymbol{a}) = \mathrm{E}\{|y|\}$ and $h(\boldsymbol{a}) = \mathrm{E}\{y^2\} - 1 = \boldsymbol{a}^\intercal \boldsymbol{V}_0 \boldsymbol{a} - 1$. From Eqn. (20), we can write:

$$\nabla_{\boldsymbol{a}} h(\boldsymbol{a}) = \nabla_{\boldsymbol{a}} \mathrm{E}\{y^2\} = 2\, \boldsymbol{V}_0 \boldsymbol{a}. \tag{47}$$

Similarly, we can obtain the following formulas:

$$\nabla_{\boldsymbol{a}} \sigma_i = \frac{1}{\sigma_i} \boldsymbol{V}_i \boldsymbol{a}$$

$$\nabla_{\boldsymbol{a}} \alpha_i = \frac{\boldsymbol{\mu}_i}{\sqrt{2}\sigma_i} - \frac{\alpha_i}{\sigma_i^2} \boldsymbol{V}_i \boldsymbol{a}$$

$$g'(\alpha) = \mathrm{erf}(\alpha).$$

The chain rule for differentiating (22) leads to:

$$\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = \sum_{i=1}^{2} \pi_i \left[\frac{\sqrt{2/\pi}}{\sigma_i}\, e^{-\alpha_i^2} \boldsymbol{V}_i \boldsymbol{a} + \mathrm{erf}(\alpha_i) \boldsymbol{\mu}_i \right]. \tag{48}$$

The Lagrangian is, by definition, $\mathcal{L}(\boldsymbol{a}, \lambda) = \mathrm{E}\{|y|\} + \lambda(\mathrm{E}\{y^2\} - 1)$, where λ is the Lagrange multiplier. The stationary points of the problem verify

$$\nabla_{\boldsymbol{a}} \mathrm{E}\{|y|\} = -\lambda\, \nabla_{\boldsymbol{a}} \mathrm{E}\{y^2\}. \tag{49}$$

To find the value of λ, observe that $\boldsymbol{a}^\intercal \nabla_{\boldsymbol{a}} \mathrm{E}\{|y|\} = \mathrm{E}\{|y|\}$ and $\boldsymbol{a}^\intercal \nabla_{\boldsymbol{a}} \mathrm{E}\{y^2\} = 2\mathrm{E}\{y^2\}$. Therefore, premultiplying (49) by $\boldsymbol{a}^\intercal$

we get $\mathrm{E}\{|y|\} = -2\,\lambda\,\mathrm{E}\{y^2\}$ and therefore

$$\lambda^* = -\frac{1}{2}\frac{\mathrm{E}\{|y|\}}{\mathrm{E}\{y^2\}} = -\frac{1}{2}\mathrm{E}\{|y|\}. \tag{50}$$

Combining Eqns. (1a)–(1b), (13), (22) and (47)–(48) proves that expression (49) can indeed be rewritten as in Eqns. (24)–(25).

To ascertain whether this solution is a maximizer of the criterion, we need to consider the second-order condition C2. Differentiating the gradient (48) with respect to the entries of $\boldsymbol{a}$, and after some tedious algebraic manipulations, we can express

$$\boldsymbol{F}(\boldsymbol{a}) = \sqrt{\frac{2}{\pi}}\sum_{i=1}^{2}\pi_i\frac{1}{\sigma_i}\exp(-\alpha_i^2)\boldsymbol{G}_i(\boldsymbol{a}) \tag{51}$$

with

$$\boldsymbol{G}_i(\boldsymbol{a}) = \frac{2\alpha_i^2-1}{\sigma_i^2}(\boldsymbol{V}_i\boldsymbol{a})(\boldsymbol{V}_i\boldsymbol{a})^\intercal + (\boldsymbol{V}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\intercal) - \sqrt{2}\frac{\alpha_i}{\sigma_i}\big((\boldsymbol{V}_i\boldsymbol{a})\boldsymbol{\mu}_i^\intercal + \boldsymbol{\mu}_i(\boldsymbol{V}_i\boldsymbol{a})^\intercal\big).$$

On the other hand, from Eqn. (47) we can easily compute both the Hessian of the constraint

$$\boldsymbol{H}(\boldsymbol{a}) = 2\boldsymbol{V}_0 \tag{52}$$

and the tangent space, $T(\boldsymbol{a}) = \{\boldsymbol{v} : \boldsymbol{v}^\intercal\boldsymbol{V}_0\boldsymbol{a} = 0\}$, where $\boldsymbol{V}_0$ is the covariance matrix of the data. Given the complexity of $\boldsymbol{F}$ in Eqn. (51), checking condition C2 seems a daunting task. But we can at least note that if $|\alpha_i|$ are sufficiently large then $\exp(-\alpha_i^2) \approx 0$, $i = 1, 2$, and therefore Hessian (51) becomes negligible relative to the Hessian of the constraint. As a result, from Eqns. (46), (50) and (52), the Hessian matrix reduces to

$$\boldsymbol{L}(\boldsymbol{a}^*, \lambda^*) \approx \lambda^*\boldsymbol{H}(\boldsymbol{a}^*) = -\mathrm{E}\{|y|\}\boldsymbol{V}_0$$

which is negative definite, thus fulfilling condition C2 and showing that the stationary point is asymptotically a maximizer of the criterion.

B. Proof that Condition (28) is Always Fulfilled

As shown in this paper, when the projected cluster means are large relative to the corresponding standard deviations, the L1-uLDA solution is of the form

$$\boldsymbol{a}^* = \eta\boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu} \tag{53}$$

where η is a normalization constant ensuring the unit variance constraint (20), i.e.,

$$\boldsymbol{a}^\intercal\boldsymbol{V}_0\boldsymbol{a} = 1. \tag{54}$$

Replacing (53) into (54), we readily obtain $\eta^2\Delta\boldsymbol{\mu}^\intercal\boldsymbol{S}_W^{-1}\boldsymbol{V}_0\boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu} = 1$, where $(\boldsymbol{S}_W^{-1})^\intercal = \boldsymbol{S}_W^{-1}$ follows from the symmetry of $\boldsymbol{S}_W$. Therefore:

$$\eta = \frac{1}{\sqrt{\Delta\boldsymbol{\mu}^\intercal\boldsymbol{S}_W^{-1}\boldsymbol{V}_0\boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu}}}.$$

Invoking Eqns. (4) and (13), we readily obtain

$$\eta = \frac{1}{\sqrt{\kappa + \pi_1\pi_2\kappa^2}} \text{ with } \kappa = \Delta\boldsymbol{\mu}^\intercal\boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu}. \tag{55}$$

Now, let us assume that condition (28) is not respected, so that

$$\Delta m = \frac{1}{\sqrt{\pi_1\pi_2}}. \tag{56}$$

According to definition (8) and Eqn. (55):

$$\Delta m = \Delta\boldsymbol{\mu}^\intercal\boldsymbol{a} = \eta\Delta\boldsymbol{\mu}^\intercal\boldsymbol{S}_W^{-1}\Delta\boldsymbol{\mu} = \eta\kappa = \sqrt{\frac{\kappa}{1+\pi_1\pi_2\kappa}}.$$

Substituting this expression in (56) yields

$$\frac{\kappa}{1+\pi_1\pi_2\kappa} = \frac{1}{\pi_1\pi_2}$$

or, equivalently, $1 + \pi_1\pi_2\kappa = \pi_1\pi_2\kappa$, which leads to an absurd solution. It follows that condition (28) is always fulfilled. The case where $m_1 > m_2$ is totally analogous and yields the same conclusion.

C. Lower Bound for the Required Separation Between Classes

This Appendix elaborates on the conditions under which the asymptotic hypothesis made in Section III-C to ensure the equivalence between L1-uLDA and LDA are fulfilled in the binary case. We saw in that section that $|\alpha_i|$, $i = 1, 2$, must be sufficiently large to guarantee the equivalence. Here we derive a lower bound for the separation between classes quantifying more precisely how large these terms must actually be. Although derived under some simplifying assumptions, this bound turns out to be consistent with the experimental results of Section V.

To start our derivation, we recall that the stationary points of the constrained optimization problem satisfy Eqn. (24). This equation is strongly nonlinear and, as such, may have more than one solution. For mathematical tractability, let us assume that the classes have equal covariance matrices, $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{V}$ (homoscedastic case). Under this assumption, Eqn. (24) simplifies into

$$\left(\sum_{i=1}^{2}\pi_i\,\beta_i\right)\boldsymbol{V}\boldsymbol{a} = \delta\Delta\boldsymbol{\mu}. \tag{57}$$

Clearly, the first type of solutions turns out to be equivalent to traditional LDA:

$$\boldsymbol{a}^* = \eta\boldsymbol{V}^{-1}\Delta\boldsymbol{\mu} \tag{58}$$

where [cf. Eqn. (55)]

$$\eta = \frac{1}{\sqrt{\kappa + \pi_1\pi_2\kappa^2}} \text{ with } \kappa = \Delta\boldsymbol{\mu}^\intercal\boldsymbol{V}^{-1}\Delta\boldsymbol{\mu} \tag{59}$$

is a normalization constant ensuring the unit variance constraint $\boldsymbol{a}^\intercal\boldsymbol{V}_0\boldsymbol{a} = 1$. On the other hand, the second type of solutions verifies

$$\sum_{i=1}^{2}\pi_i\,\beta_i = 0 \quad \text{and} \quad \delta = 0.$$

In particular, one can easily check that these conditions are satisfied by $\boldsymbol{a} = \boldsymbol{V}^{-1/2}\mathbf{q}$, where $\mathbf{q}$ is any unit-norm vector orthogonal to $\boldsymbol{V}^{-1/2}\Delta\boldsymbol{\mu}$. Indeed, this solution yields $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\beta_1 = \beta_2 = 0$ and $\delta = 0$, while fulfilling the unit-variance constraint. Clearly, this can be considered as

a worst-case spurious solution, since the projected classes have the same mean and variance, leading to fully overlapped clusters and precluding their discrimination. For classes with Gaussian distribution, we easily obtain from formula (22) that

$$\mathrm{E}\{|y|\} = \sqrt{2/\pi} \tag{60}$$

for these spurious solutions.

Now, intuitively we can consider that the asymptotic regime leading to the equivalence between L1-uLDA and LDA occurs when the cost function associated with the desired solution $\boldsymbol{a}^*$ is larger than that of spurious solutions, that is:

$$\mathrm{E}\{|y|\} > \sqrt{2/\pi}. \tag{61}$$

Because $g(x) \geq |x|$ in (22) (as illustrated in Fig. 1), we can lower-bound the L1-PCA criterion with unit-variance constraint as

$$\mathrm{E}\{|y|\} \geq \sqrt{2} \sum_{i=1}^{2} \pi_i \sigma_i \, |\alpha_i| = \pi_2 |m_2| + \pi_1 |m_1| = 2\pi_1 \pi_2 \Delta\boldsymbol{\mu}^\mathsf{T} \boldsymbol{a}$$

where the last equality holds from Eqn. (26). By replacing the expression of the desired solution (58) and taking into account Eqn. (55), we can write that, for $\boldsymbol{a} = \boldsymbol{a}^*$:

$$\mathrm{E}\{|y|\} \geq 2\pi_1 \pi_2 \Delta\boldsymbol{\mu}^\mathsf{T} \boldsymbol{a}^* = 2\pi_1 \pi_2 \sqrt{\frac{\kappa}{1 + \pi_1 \pi_2 \kappa}}. \tag{62}$$

It follows that a sufficient condition to fulfil inequality (61) is

$$2\pi_1 \pi_2 \sqrt{\frac{\kappa}{1 + \pi_1 \pi_2 \kappa}} > \sqrt{\frac{2}{\pi}} \tag{63}$$

or, equivalently,

$$\kappa > \frac{1}{\zeta} \quad \text{with} \quad \zeta \stackrel{\text{def}}{=} 2\pi(\pi_1 \pi_2)^2 - \pi_1 \pi_2. \tag{64}$$

Since κ represents the squared Mahalanobis distance between the class means [see Eqn. (55)] and is thus a positive quantity, we must have $\min(\pi_1, \pi_2) > 0.1986$ to guarantee $\zeta > 0$ and a meaningful bound in Eqn. (64).

In summary, condition (64) imposes a minimum distance between the classes for the solution of the L1-uLDA criterion not to be spurious, and we argue that this will also be sufficient for the solution to be the equivalent to LDA. Furthermore, as ζ is a concave function attaining its maximum value at $\pi_1 = \pi_2 = 0.5$, the separation increases when the classes are not equiprobable. Although obtained under certain simplifications (homoscedasticity and Gaussianity), these theoretical results are corroborated by experiment 5 of Section V-A, showing in particular that fulfilment of Eqn. (64) guarantees the equivalence between L1-uLDA and LDA.

ACKNOWLEDGMENT

Most of this work was carried out while V. Zarzoso was a member of the Institut Universitaire de France.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer, 2009.

[3] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Berlin, Germany: Springer, 2002.

[4] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[5] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, Jun. 2011, Art. no. 11.

[7] G. Mateos and G. B. Giannakis, "Robust PCA as bilinear decomposition with outlier-sparsity regularization," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5176–5190, Oct. 2012.

[8] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1062–1070.

[9] M. Johnson and A. Savakis, "Fast L1-eigenfaces for robust face recogntion," in *Proc. IEEE Western New York Image Signal Process. Workshop*, 2014, pp. 1–5.

[10] Y. Liu and D. Pados, "Compressed-sensed-domain L1-PCA video surveillance," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 351–363, Mar. 2016.

[11] P. P. Markopoulos, S. Kundu, and D. A. Pados, "L1-fusion: Robust linear-time image recovery from few severely corrupted copies," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 1225–1229.

[12] F. Maritato, Y. Liu, S. Colonnese, and D. Pados, "Face recogn. with L1-norm subspaces," *Proc. SPIE*, vol. 9857, 2016, Art. no. 98570L.

[13] N. Tsagkarakis, P. P. Markopoulos, and D. A. Pados, "Direction finding by complex L1-principal-component analysis," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, 2015, pp. 475–479.

[14] M. McCoy *et al.*, "Two proposals for robust PCA using semidefinite programming," *Electron. J. Statist.*, vol. 5, pp. 1123–1160, 2011.

[15] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy L1-norm maximization," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, no. 1, pp. 1433–1438.

[16] P. P. Markopoulos, G. N. Karystinos, and D. A. Pados, "Optimal algorithms for L1-subspace signal processing," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5046–5058, Jul. 2014.

[17] P. P. Markopoulos, S. Kundu, S. Chamadia, and D. A. Pados, "Efficient L1-norm principal-component analysis via bit flipping," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4252–4264, Aug. 2017.

[18] N. Tsagkarakis, P. P. Markopoulos, G. Sklivanitis, and D. A. Pados, "L1-norm principal-component analysis of complex data," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3256–3267, Jun. 2018.

[19] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychol. Methods*, vol. 2, no. 3, pp. 292–307, Sep. 1997.

[20] D. Peña and F. J. Prieto, "Cluster identification using projections," *J. Amer. Statistical Assoc.*, vol. 96, pp. 1433–1445, 2001.

[21] D. Peña and F. J. Prieto, "The kurtosis coefficient and the linear discriminant function," *Statist. Probability Lett.*, vol. 49, no. 3, pp. 257–261, Sep. 2000.

[22] T. W. Anderson and R. R. Bahadur, "Classification into two multivariate normal distributions with different covariance matrices," *Ann. Math. Statist.*, vol. 33, pp. 420–431, Jun. 1962.

[23] A. Martinez and A. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[24] R. Martín-Clemente and V. Zarzoso, "On the link between L1-PCA and ICA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 515–528, Mar. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7497466/

[25] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[26] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner, "A polynomial case of unconstrained zero-one quadratic optimization," *Math. Program., Ser. B*, vol. A, no. 91, pp. 49–52, Oct. 2001.

[27] W. Ben-Ameur and J. Neto, "A polynomial-time recursive algorithm for some unconstrained quadratic optimization problems," *Discrete Appl. Math.*, vol. 159, pp. 1689–1698, Sep. 2011.

[28] J. A. Ferrez, K. Fukuda, and T. M. Liebling, "Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm," *Eur. J. Oper Res.*, vol. 166, pp. 35–50, 2005.

[29] G. N. Karystinos and A. P. Liavas, "Efficient computation of the binary vector that maximizes a rank-deficient quadratic form," *IEEE Trans. Inf. Theory*, vol. 56, pp. 3581–3593, Jul. 2010.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2001.
[31] J. Chen, Z. Zhao, J. Ye, and H. Liu, "Nonlinear adaptive distance metric learning for clustering," in *Proc. 13th Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 123–132.
[32] C. H. Q. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and k-means clustering," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 521–528.
[33] S. Wang, J. Lu, X. Gu, H. Du, and J. Yang, "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognit.*, vol. 57, pp. 179–189, 2016.
[34] "MATLAB code for the L1-PCA algorithm proposed in [17]," [Online]. Available: https://sites.google.com/view/miloslab/resources. Accessed on: 2018.
[35] R. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–186, 1936.
[36] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.
[37] "Iris flower data set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Iris. Accessed on: 2018.
[38] "Breast Cancer Wisconsin (diagnostic) data set," [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. Accessed on: 2018.
[39] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp. 71–86, 1991.
[40] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778
[42] D. E. King, "Dlib-ML: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009. [Online]. Available: http://dlib.net
[43] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," [Online]. Available: http://vis-www.cs.umass.edu/lfw/index.html. Accessed on: 2019.
[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, pp. 886–893.
[45] E. Chong and S. Zak, *An Introduction to Optimization*, 4th ed. Hoboken, NJ, USA: Wiley, 2013.

Rubén Martín-Clemente received the M. Eng. degree in telecommunications engineering and the Ph.D. degree with highest distinction in telecommunications engineering from the University of Sevilla, Sevilla, Spain, in 1996 and 2000, respectively. He is currently the Head of the Department of Signal Theory and Communications of the University of Sevilla, Spain. He has been a Visiting Researcher with the University of Regensburg, Regensburg, Germany, in 2001 and 2009 and with the University of Nice, France, in 2015, 2016, and 2018, respectively. Among other areas, his research interests include signal processing and machine learning with emphasis on independent component analysis and its application to biomedical problems. He has authored or co-authored numerous publications on these topics. He was as a Program Committee Member for several international conferences and was a Program Committee Chair of the 5th International Conference on Independent Component Analysis and Blind Signal Separation in 2004.

Vicente Zarzoso received the graduate degree with highest distinction in telecommunications engineering from the Polytechnic University of Valencia, Valencia, Spain, in 1996. He began the Ph.D. studies with the University of Strathclyde, Glasgow, U.K. In 1999, he received the Ph.D. degree from the University of Liverpool, Liverpool, U.K. He received the Habilitation to Lead Researches (HDR) from the University of Nice Sophia Antipolis (now member of the Université Côte d'Azur, UCA), France, in 2009. From 2000 to 2005, he held a research fellowship awarded by the Royal Academy of Engineering, U.K. Since 2005, he has been with the Computer Science, Signals and Systems Laboratory of Sophia Antipolis (I3S), UCA, CNRS, France, where he is a Full Professor and since 2016 he has been the Head of the "Signals, Images and Systems" (SIS) research team. His research interests lie in the areas of signal processing and machine learning with emphasis on matrix and tensor factorizations, principal/independent component analysis and related techniques, including theoretical aspects and applications in biomedical problems and digital communications. He was an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2011 to 2015 and has been a Program Committee Member of several international conferences. He was a Program Committee Chair of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA-2010) and a keynote lecturer at the LVA/ICA-2015 Summer School. He is a Senior Member of the IEEE and an Honorary Member of the *Institut Universitaire de France*.