proceedings

MDPI

Proceedings

L1 Norm Based PCA for Unsupervised Classification †

J.L. Camargo-Olivares [1], R. Martín-Clemente [1,*], S. Hornillo-Mellado[1] and V. Zarzoso [2]

[1] Signal Processing and Communications Department, University of Seville, 41004 Sevilla, Spain; jlcamargo@yahoo.es (J.L.C.-O.); ruben@us.es, susanah@us.es (R.M.-C.)

[2] I3S Laboratory, University of Cotê d'Azur, 06410 Biot, France; zarzoso@i3s.unice.fr

* Correspondence: ruben@us.es

† Presented at the 1st International Electronic Conference on Applied Sciences, 10–30 November 2020; Available online: https://sciforum.net/conference/ASEC2020

Published: 10 November 2020

check for updates

Abstract: Principal component analysis (PCA) is a widespread technique for the analysis of multivariate data, which finds applications in the fields of machine learning and artificial intelligence. Standard PCA seeks to calculate the subspace that minimizes the Euclidean distance (L2-norm) of the data points to it. Unfortunately, PCA is extremely sensitive to the presence of large outliers in the data. 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 proposed contribution shows that, when combined with a whitening preprocessing step, L1-norm based PCA is endowed with discriminative power and can perform data classification in an *unsupervised manner*, i.e., sparing the need for labelled data. By minimizing the L1-norm in the feature space, the technique mimics the action of *common spatial patterns* (CSP), a supervised feature extraction method used in brain computer interfaces. This result is of theoretical interest and opens new interesting research perspectives for L1-PCA. Furthermore, it enables us to perform classification using algorithms for optimizing the L1-norm, which inherit the improved robustness to outliers of the L1-norm criterion. Several numerical experiments will confirm the theoretical findings.

Keywords: principal component analysis; binary classification; machine learning

1. Introduction

L1-norm based criteria are becoming increasingly popular in the fields of machine learning and signal processing. In particular, there is growing interest for the development of L1-norm Principal Component Analysis (L1-PCA) [1,2]. L1-PCA is a variant of traditional PCA which offers enhanced robustness against large outliers. This is an interesting feature because outliers, which are erroneous measurements that lie far apart from the main bulk of the data, are very common in experimental datasets, due to the imperfections in the measuring instruments or the environmental conditions. Specifically, L1-PCA has proven to be highly effective for the restoration of faulty data, in the reconstruction of occluded images or in dimensionality reduction problems [1–4]. However, as negative points, L1-PCA algorithms are either computationally intensive and time consuming [3], despite efforts to simplify their operation [4], or prone to fall into local optima [1]. Furthermore, L1-PCA is a difficult subject to analyze mathematically because, implicitly, it involves the higher-order statistics of the data. For one reason or another, only a few attempts have been made to explain the behavior of L1-PCA in practical situations. Among them, to cite an example, [5] showed that L1-PCA is able to perform Independent Component Analysis (ICA) if the data follows the ICA model. The present contribution continues to investigate the properties of L1-PCA. Here, we report that L1-PCA, after a minor modification, replicates the operation of the technique known as Common Spatial Patterns (CSP), a *supervised* feature extraction method used in brain computer interfaces [6]. As a result, L1-PCA can be

used to separate overlapping populations that are normally distributed and perform data classification in an *unsupervised manner*, i.e., sparing the need for labelled data, which is a remarkable feature. This finding opens new interesting research perspectives for L1-PCA in the field of machine learning. Furthermore, it enables us to develop classification algorithms based on the L1-norm, which inherit the improved robustness to outliers of the L1-norm criterion.

The paper is organized as follows: Section 2 introduces the L1-norm from standard PCA. Section 3 shows that the L1-norm is endowed with discriminative properties in binary classification scenarios. Section 4 illustrates the performance of the approach through computer simulations. Finally, Section 5 brings the paper to an end.

2. Background

Let $\boldsymbol{x} \in \mathbb{R}^p$ be a multivariate random variable measured or observed during an experiment. For simplicity, we assume that $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{0}$, where $\mathrm{E}[\cdot]$ is the expectation operator. The aim of standard PCA is to find the best-fit *low*-dimensional subspace for the data points. This is the subspace that minimizes the average squared distance of the data points to it. It can be also shown that this problem is equivalent to finding linear projections of the variables that have maximal variance [7]. A projection onto the direction of a unit vector $\boldsymbol{a}$ is given by

$$y = \boldsymbol{a}^\top \boldsymbol{x}.$$

The variance of the projected data equals

$$\sigma^2(\boldsymbol{a}) = \mathrm{E}\left[y^2\right] \tag{1}$$

The first principal component is the vector that solves the problem

$$\arg\max_{\|\boldsymbol{a}\|_2=1} \sigma^2(\boldsymbol{a}) \tag{2}$$

The $n-$th principal component is the vector that solves the optimization problem (2) subject to the additional constraint of being orthogonal to the previous $n-1$ principal components. The desired best-fitting subspace, finally, is the span of the first few principal components. They are, in other words, the most significant directions characterizing the point cloud of $\boldsymbol{x}$.

However, it is well-known that standard PCA overreacts to large outliers because it takes the square of the projected data in (1). In order to palliate this weakness, [1] proposed the replacement of the square function by the absolute value, yielding the following alternative criterion:

$$\arg\max_{\|\boldsymbol{a}\|_2=1} \mathrm{E}[|y|] \tag{3}$$

In practice, given a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from the random variable $\boldsymbol{x}$, (3) is approximated by its sample based estimate

$$\max_{\|\boldsymbol{a}\|_2=1} \quad \frac{1}{N} \sum_{i=1}^{N} |\boldsymbol{a}^\top \boldsymbol{x}_i| \tag{4}$$

Because $\sum_{i=1}^{N} |\boldsymbol{a}^\top \boldsymbol{x}_i|$ represents the L1-norm of the vector $\boldsymbol{y}$ whose kth entry is given by $y_k = \boldsymbol{a}^\top \boldsymbol{x}_i$, PCA based on criterion (3) is usually referred to as 'L1-norm based PCA' or, simply, 'L1-PCA'. Working algorithms for solving (3) have been proposed in [1,3,4].

2.1. L1-PCA in the Case of Gaussian Data

To gain some insight into the performance of L1-PCA, let us make the usual assumption that the probability density function of the data is a p-variate normal density function of the form

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} \det(\boldsymbol{C})^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}^{-1}\boldsymbol{x}} \tag{5}$$

where $\boldsymbol{C} = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ is the data covariance matrix. Let $y = \boldsymbol{a}^\top \boldsymbol{x}$ be the projection of $\boldsymbol{x}$ into the direction defined by $\boldsymbol{a} \in \mathbb{R}^p$. The probability density function of y is given by

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \tag{6}$$

where $\sigma^2 = \boldsymbol{a}^\top \boldsymbol{C}\boldsymbol{a}$ is the variance of the projected data. Now, some calculus shows that

$$\mathrm{E}[|y|] = \int_{-\infty}^{\infty} |y|\, p(y)\, \mathrm{d}y = \sqrt{\frac{2}{\pi}}\, \sigma$$

Then, as maximizing the standard deviation σ is equivalent to maximizing the variance σ^2, one sees that L1-PCA behaves in this case like traditional PCA, while offering robustness against the presence of large outliers in the data [1–3].

3. L1-norm Based Classification

Binary classification problems are ubiquitous in many real-life applications. Consider that we observe random samples drawn from two different populations ω_1 and ω_2 with the same population mean, assumed to be zero. It is supposed that the distribution of the random samples can be modeled as a mixture of Gaussians, i.e.,

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} \sum_{i=1}^{2} \pi_i \det(\boldsymbol{C}_i)^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}_i^{-1}\boldsymbol{x}}$$

where π_1 and π_2 are the *a priori* probabilities of occurrence of ω_1 and ω_2, with $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ the corresponding class covariance matrices ($\boldsymbol{C}_1 \neq \boldsymbol{C}_2$). Consider again the L1-norm criterion

$$J(\boldsymbol{a}) = \mathrm{E}[|y|] = \mathrm{E}[|\boldsymbol{a}^\top \boldsymbol{x}|] \tag{7}$$

Similar calculus as above shows that

$$J(\boldsymbol{a}) = \sqrt{\frac{2}{\pi}}\, (\pi_1\, \sigma_1(\boldsymbol{a}) + \pi_2\, \sigma_2(\boldsymbol{a})) \tag{8}$$

where $\sigma_i^2(\boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{C}_i \boldsymbol{a}$ is the variance of the ith class in the direction of the unit vector $\boldsymbol{a} \in \mathbb{R}^p$.

Let us assume hereafter, without any loss of generality, that the data are *whitened*. A random variable $\boldsymbol{x}$ is whitened by multiplying it by a matrix $\boldsymbol{Q}$ so that the result $\boldsymbol{Q}\boldsymbol{x}$ has covariance $\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^\top = \boldsymbol{I}$, where $\boldsymbol{C} = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ and $\boldsymbol{I}$ is the identity matrix. This goal can be achieved in practice by setting $\boldsymbol{Q} = \boldsymbol{C}^{-1/2}$. To keep the notation simple, the whitened data are also denoted, with some abuse, by $\boldsymbol{x}$. Likewise, the whitened class covariance matrices are still denoted by $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. Whitening implies that

$$\boldsymbol{C} = E[\boldsymbol{x}\boldsymbol{x}^\top] = \pi_1 \boldsymbol{C}_1 + \pi_2 \boldsymbol{C}_2 = \boldsymbol{I} \tag{9}$$

$$E[y^2] = \boldsymbol{a}^\top E[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{a} = \pi_1 \sigma_1^2 + \pi_2 \sigma_2^2 = 1 \tag{10}$$

Furthermore, Equation (8) still holds true. The real utility of whitening is that it introduces a constraint, namely, Equation (10), on the class variances: when one of them increases the other

decreases, and *vice versa*. As a consequence, a thorough analysis leads to the following result (proof is omitted):

Theorem 1. *Under the whitening assumption, the minimizers of* (8) *with the constraint* $\|\boldsymbol{a}\| = 1$ *maximize or minimize the power ratio*

$$R(\boldsymbol{a}) = \frac{\sigma_1^2}{\sigma_2^2} \tag{11}$$

This Theorem can be put in relation to the useful technique known as common spatial patterns (CSP), which is widely used in brain-computer interfaces (BCIs). Typically, electroencephalogram (EEG) samples are acquired under two different experimental conditions (e.g., imagining left and right hand movements). CSP linearly projects the data onto directions where the ratio (11) is maximal or minimal or, in simple words, where the variance of the projected data points is significantly higher for one class than for the other. The projected data variances are then be used as features for classification [6]. It follows that the L1 criterion possesses the discriminative capabilities of CSP. Quite interestingly, CSP is a supervised technique, whose performance relies heavily on the availability of correctly labeled data. On the contrary, minimizing the L1 criterion (8) can be performed in a completely unsupervisedly fashion.

4. Computer Experiments

Some experiments are now conducted to illustrate the potential of the L1-approach

4.1. Experiment 1

To illustrate Theorem 1, let us consider a mixture in a bidimensional space of two equiprobable Gaussian classes, i.e., $\pi_1 = \pi_2 = 1/2$, with zero-means and respective covariances

$$\boldsymbol{C}_1 = \begin{pmatrix} 1 & 0.68 \\ 0.68 & 1 \end{pmatrix} \text{ and } \boldsymbol{C}_2 = \begin{pmatrix} 1 & -0.68 \\ -0.68 & 1 \end{pmatrix}. \tag{12}$$

Observe that matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ fulfill the whitening condition (9). Figure 1 represents the theoretically exact value of the cost function $J(\theta) = \mathrm{E}[|\boldsymbol{a}(\theta)^\top \boldsymbol{x}|]$, with $\boldsymbol{a}(\theta) = [\cos(\theta), \sin(\theta)]^\top$, calculated from Equation (8). For reference, we also plot the power ratio $R(\theta)$, defined as in (11), in the same Figure. We see that the minima of the L1-cost $J(\theta)$ correspond with either the maximum or the minimum of $R(\theta)$, as predicted by the Theorem. We also see that the maxima of $J(\theta)$ are at 0 and $\pm\pi/2$ rad. At these points, the standard deviations of the projected populations are the same, i.e., $\sigma_1 = \sigma_2$, with $\sigma_i^2 = \boldsymbol{a}^\top \boldsymbol{C}_i \boldsymbol{a}$. It follows that the projected populations are totally mixed, because the different classes cannot be distinguished from each other.

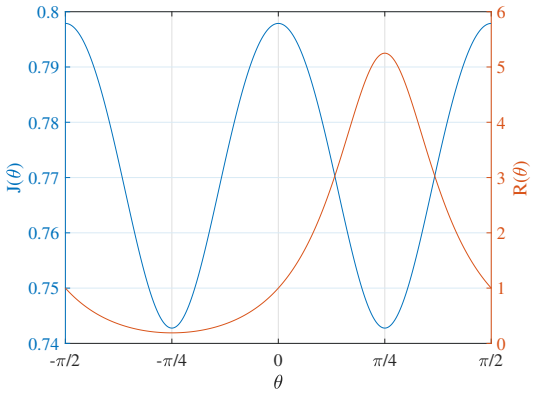

Figure 1. L1-norm function $J(\theta) = \mathrm{E}[|\boldsymbol{a}(\theta)^\top \boldsymbol{x}|]$ and power ratio $R(\theta) = \frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}$, illustrating the matching between their extrema.

4.2. Experiment 2

To test the L1-norm approach in a multidimensional setting, we perform several experiments with $p \in \{2, 5, 10, 15, 20, 25, 30\}$. In each one, we draw $N = 50\,p$ samples per each of the two Gaussian classes, and the covariance matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are randomly generated. After applying a *whitening* pre-processing to the data, the gradient descent algorithm in [8] is applied to find the orthogonal directions that (globally or locally) minimize the L1-norm criterion (7). The closeness to the subspace spanned by the line in the direction of the global minimum is used as unsupervised criterion to classify the random samples into one cluster or the other. Figure 2 shows the accuracy of the classification, averaged over 100 independent experiments. Furthermore, L1-norm criteria are also expected to exhibit robustness against large outliers. To test this property, we repeat the experiment with the difference that the whitened data points are now corrupted by replacing 10 per cent of the data samples, at randomly chosen time instants, by Gaussian noise realizations with identity covariance matrix and mean $\boldsymbol{\mu}_{\text{outliers}} = [10, 10, \ldots, 10]^\top$. The new results are also represented in Figure 2, proving the reliability of the L1-norm. In both cases, we see that the performance increases with the dimensionality of the input representation. This finding reflects the well-known fact that it is usually easier to perform classification in high-dimensional spaces.

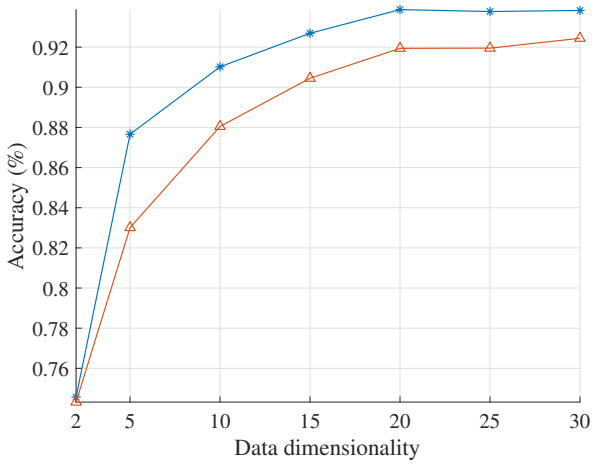

Figure 2. Accuracy in the classification of the data as a function of the data dimensionality. Blue line: accuracy calculated from outlier-free data. Red line: ditto for the outlier-corrupted data.

5. Conclusions

Projecting whitened data onto the few dimensions that minimize the absolute value of the projected data points can perform unsupervised classification in a fully unsupervised fashion, sparing the need for training data and opening new lines of research in the area of L1-PCA. Good performance is shown by numerical experiments.

Author Contributions: All authors participate in conceptualization, investigation and writing—original draft.

Funding: This work is funded by the research project US-1264994 awarded by the Junta de Andalucía (Consejería de Transformación Económica, Industria, Conocimiento y Universidades).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

To enable research reproducibility, the following Matlab code can be used to reproduce Figure 1

```
C1 = [1 0.68; 0.68 1]; C2 = [1 -0.68; -0.68 1]; angles = linspace(-pi/2, pi/2, 100);
for i = 1:numel(angles)
   a = [cos(angles(i)); sin(angles(i))];
```

```
   s1(i) = sqrt(a'*C1*a); s2(i) = sqrt(a'*C2*a);
   J(i) = sqrt(0.5/pi)*(s1(i) + s2(i)); R(i) = [s1(i)/s2(i)]^2;
end
yyaxis left, plot(angles, J), ylabel('J(\theta)'),
yyaxis right, plot(angles, R), grid on, ylabel('R(\theta)')
```

In experiment 2, data have been generated for each class by the Matlab command `mvnrnd`. The basic algorithm for finding a direction minimizing the L1-norm is

```
[p,T] = size(X); % X is the data matrix (num features x num samples)
a = randn(p,1); a = a/norm(a); flag=true;
while(flag)
   a_old = a; a = a - 0.1/T*(X*sign(a'*X)');  a = a/norm(a);
   if norm(a-a_old) < 0.001, flag = false; end
end
```

References

1. Kwak, N. Principal Component Analysis Based on L1-Norm Maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1672–1680.
2. Markopoulos, P.P.; Kundu, S.; Chamadia, S.; Tsagkarakis, N.; Pados, D.A. Outlier-Resistant Data Processing with L1-Norm Principal Component Analysis. In *Advances in Principal Component Analysis*; Springer: Singapore, 2017; pp. 121–135.
3. Markopoulos, P.P.; Karystinos, G.N.; Pados, D.A. Optimal Algorithms for L1-subspace Signal Processing. *IEEE Trans. Signal Process.* **2014**, *62*, 5046–5058.
4. Markopoulos, P.P.; Kundu, S.; Chamadia, S.; Pados, D.A. Efficient L1-Norm Principal-Component Analysis via Bit Flipping. *IEEE Trans. Signal Process.* **2017**, *65*, 4252–4264.
5. Martín-Clemente, R.; Zarzoso, V. On the Link Between L1-PCA and ICA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 515–528.
6. Martín-Clemente, R.; Olias, J.; Thiyam, D.; Cichocki, A.; Cruces, S. Information Theoretic Approaches for Motor-Imagery BCI Systems: Review and Experimental Comparison. *Entropy* **2018**, *20*, 7.
7. Jolliffe, I.T. *Principal Component Analysis*; Springer: New York, NY, USA, 2002.
8. Edelman, A.; Arias, T.A.; Smith, S.T. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. Matrix Anal. Appl.* **1998**, *20*, 303–353.