On the Link Between L1-PCA and ICA

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

Abstract—Principal component analysis (PCA) based on L1-norm maximization is an emerging technique that has drawn growing interest in the signal processing and machine learning research communities, especially due to its robustness to outliers. The present work proves that L1-norm PCA can perform independent component analysis (ICA) under the whitening assumption. However, when the source probability distributions fulfil certain conditions, the L1-norm criterion needs to be minimized rather than maximized, which can be accomplished by simple modifications on existing optimal algorithms for L1-PCA. If the sources have symmetric distributions, we show in addition that L1-PCA is linked to kurtosis optimization. A number of numerical experiments illustrate the theoretical results and analyze the comparative performance of different algorithms for ICA via L1-PCA. Although our analysis is asymptotic in the sample size, this equivalence opens interesting new perspectives for performing ICA using optimal algorithms for L1-PCA with guaranteed global convergence while inheriting the increased robustness to outliers of the L1-norm criterion.

Index Terms—Feature extraction or construction, interactive data exploration and discovery, independent component analysis, principal component analysis, L1-norm, multivariate statistics, feature representation, feature evaluation and selection

1 Introduction

Finding interesting projections of observed multidimensional data is a fundamental problem in signal processing and machine learning, with a wide range of applications in data representation and visualization, classification, clustering, and source separation, among others. Carefully selected projections often reveal much of the data's structure, where the concept of 'structure' may be defined in several ways depending on the application at hand [11]. Arguably one of the most popular projection pursuit techniques is the classical *principal component analysis (PCA)* [14]. PCA aims at retaining as much as possible the variance present in the dataset, and can be seen as the maximization of the L2-norm of the vector of projected data samples. The last two decades have witnessed the development of *independent component analysis (ICA)*, whose aim is to search for statistically independent projections [5], [6], [13]. Typically based on higher-order statistics, ICA is the standard tool for blind separation of independent source signals. The interest in L1-norm PCA, another data projection technique, has gradually been growing in recent years [3], [17], [18], [20], [21], [22], [23], [25], [28], [29], [30]. This interest can mainly be attributed to the L1-norm improved robustness against outliers as compared to the L2-norm and the higher-order statistical criteria used, respectively, in L2-PCA and ICA. Another attractive feature of L1-PCA is that it can be performed by optimal algorithms that, though computationally intensive, offer guaranteed global convergence, as recently shown in [22].

Although the above data projection techniques are apparently disparate, they are actually connected. It is well-known that ICA can be simplified by *whitening* or *sphering* the data via L2-PCA, though this transformation is generally not sufficient for estimating the independent components [6]. The present contribution takes a step further in establishing this link, and proves that ICA can actually be performed by L1-PCA after L2-norm whitening. We derive the conditions under which the solutions associated with independent components are stationary points of the L1-norm criterion, and provide some numerical examples to illustrate and support our findings. Hence, we connect an attractive emerging technique (L1-PCA) to a mature multivariate analysis method with well-established theoretical foundations. These results open interesting new perspectives for performing ICA using optimal algorithms for L1-PCA while inheriting the improved robustness to outliers of the L1-norm criterion.

The paper is structured as follows. Section 2 reviews the mathematical formulations and some results of the different data projection techniques considered in this work. Section 3, the core of our contribution, establishes the link between ICA and L1-PCA. Illustrative examples supporting the theoretical derivations are presented and discussed in Section 4, including a comparative performance analysis of the different techniques for ICA via L1-PCA studied here. The concluding remarks of Section 5 bring the paper to an end. For the sake of clarity, proofs of the theoretical results have been deferred to the Appendix.

Throughout the paper, we adopt standard mathematical notations: lightface lowercase (x), boldface lowercase ($\mathbf{x}$) and boldface uppercase characters ($\mathbf{X}$) represent, respectively, scalars, vectors and matrices. Symbol $(\cdot)^\dagger$ denotes the matrix transpose operator, and $\mathrm{E}\{\cdot\}$ is the mathematical expectation, while $\mathbf{I}_N$ and $\mathbf{0}_N$ stand for the $(N \times N)$ identity matrix and the vector of zero entries with dimension N,

- *R. Martín-Clemente is with the Departamento de Teoría de la Señal y Comunicaciones, Escuela Superior de Ingeniería, Universidad de Sevilla, Avda. Descubrimientos s/n, Seville 41092, Spain. E-mail: ruben@us.es.*
- *V. Zarzoso is with the I3S Laboratory, UMR 7271, University of Nice Sophia Antipolis, CNRS, CS 40121, Sophia Antipolis Cedex 06903, France. E-mail: zarzoso@i3s.unice.fr.*

Manuscript received 4 Aug. 2014; revised 9 Mar. 2016; accepted 5 Apr. 2016. Date of publication 21 June 2016; date of current version 13 Feb. 2017. Recommended for acceptance by M. A. Carreira-Perpinan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2557797

respectively. Finally, $\mathbf{e}_i$ represents the ith canonical basis vector, with a one in the ith entry and zeros elsewhere; its dimensions will be clear from the context.

2 Problem Formulation and Assumptions

This section reviews some existing results on L1-PCA and ICA, highlighting the gaps that our work aims to bridge.

2.1 L1-PCA

We observe T realizations $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T\}$ of a zero-mean N-dimensional random variable $\mathbf{z}$, which are stored in the $(N \times T)$ data matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T]$. Projection of the data in the direction of the unit-length vector $\mathbf{w} \in \mathbb{R}^N$ yields the T-dimensional vector $\mathbf{y}$ given by

$$\mathbf{y}^\dagger = \mathbf{w}^\dagger \mathbf{Z}. \tag{1}$$

As is well known, the classical PCA of the data seeks to maximize the L2-norm criterion

$$\max_{\|\mathbf{w}\|_2=1} \|\mathbf{y}\|_2^2 = \max_{\|\mathbf{w}\|_2=1} \sum_{t=1}^{T} \left(\mathbf{w}^\dagger \mathbf{z}_t\right)^2 \tag{2}$$

and can readily be solved in terms of the singular value decomposition (SVD) of matrix $\mathbf{Z}$. As presented in [22], a natural extension of PCA using the L1 norm is given by

$$\max_{\|\mathbf{w}\|_2=1} \|\mathbf{y}\|_1 = \max_{\|\mathbf{w}\|_2=1} \sum_{t=1}^{T} \left|\mathbf{w}^\dagger \mathbf{z}_t\right|. \tag{3}$$

The interest in L1-norm based PCA has grown in recent years due to the fact that it is more robust against outliers than classical PCA, as large outliers are favoured over small data values by the square in eqs. (2) [3], [9], [17], [18], [20], [21], [22], [23], [25], [28], [29], [30].

2.1.1 Algorithms for L1-PCA

Interestingly, Markopoulos and co-workers [22] have recently identified the equivalence between L1-norm maximization and an L2-norm maximization problem in the binary field widely known as 'binary quadratic programming'. The following lemma, reproduced here from [22] for the sake of clarity, formally states this equivalence.

Lemma 1. *We have that*

$$\max_{\|\mathbf{w}\|_2=1} \|\mathbf{w}^\dagger \mathbf{Z}\|_1 = \max_{\mathbf{c} \in \{\pm 1\}^T} \|\mathbf{Z}\mathbf{c}\|_2, \tag{4}$$

where notation $\mathbf{c} \in \{\pm 1\}^T$ means that each of the T entries of column vector $\mathbf{c}$ takes one possible value from the set with elements $\{-1, 1\}$, and at the optimum

$$\mathbf{w}_{\text{opt}} = \mathbf{Z}\mathbf{c}_{\text{opt}} / \|\mathbf{Z}\mathbf{c}_{\text{opt}}\|_2. \tag{5}$$

Binary quadratic programming has been extensively studied in the literature. The optimal solution has the following property [22]:

$$\mathbf{c}_{\text{opt}} \in \mathcal{S}_1 \overset{\text{def}}{=} \bigcup_{\|\mathbf{w}\|_2=1} \text{sign}(\mathbf{Z}^\dagger \mathbf{w}). \tag{6}$$

In [15], an auxiliary-angle approach is developed to calculate efficiently the $|\mathcal{S}_1| = \sum_{n=0}^{\text{rank}(\mathbf{Z})-1} \binom{T-1}{n}$ vectors of the candidate solution set $\mathcal{S}_1$. Among them, the optimal solution is computed as the maximizer of criterion (4). The algorithm is also scalable in the sense that $\mathcal{S}_1$ can be estimated from a low-rank approximation of the data and be refined in further stages at a reduced cost [15]. Alternatively, one could use the polynomial-time approaches in [1], [2], [10], [19]. Based on [15], Markopoulos and co-workers [22] have proposed an optimal algorithm for computing the L1 principal component with global convergence and complexity $\mathcal{O}(T^{\text{rank}(\mathbf{Z})})$. A simplified and faster, yet suboptimal, version of [22] consists in iteratively producing a sequence of binary vectors $\mathbf{c}$, where each one differs from the immediately previous and following vectors in a single position only. At each iteration, the algorithms chooses to flip the component yielding the highest increase of $\|\mathbf{Z}\mathbf{c}\|_2$ [16].

A less expensive algorithm for the computation of $\mathbf{c}_{\text{opt}}$ and $\mathbf{w}_{\text{opt}}$ was proposed earlier by Kwak in [17], where the following update is repeated iteratively until convergence

$$\mathbf{c} = \text{sign}(\mathbf{Z}^\dagger \mathbf{w}) \tag{7a}$$

$$\mathbf{w}^+ = \mathbf{Z}\mathbf{c} / \|\mathbf{Z}\mathbf{c}\|_2. \tag{7b}$$

Symbol $\mathbf{w}^+$ represents the estimated extracting vector after the current iteration. Although simple, this iterative algorithm is not guaranteed to converge to a global maximum and has indeed been shown to get trapped in spurious local extrema [22].

2.2 ICA

ICA is a multivariate analysis technique for estimating the N-dimensional latent variable vector $\mathbf{s}$ in the data model

$$\mathbf{z} = \mathbf{Q}\mathbf{s}, \tag{8}$$

where $\mathbf{Q}$ is some invertible $(N \times N)$ matrix. The hypotheses commonly used in ICA, and that we adopt in the sequel, are the following:

- A1) Components s_i, also called sources, are mutually statistically independent, and have zero mean and unit variance.
- A2) The observed data $\mathbf{z}$ have been whitened (or sphered) by a suitable transformation.

Assumptions A1-A2 mean in particular that sources and observations have an identity covariance matrix, $\text{E}\{\mathbf{s}\mathbf{s}^\dagger\} = \text{E}\{\mathbf{z}\mathbf{z}^\dagger\} = \mathbf{I}_N$, and imply that the mixing matrix $\mathbf{Q}$ is unitary, since $\text{E}\{\mathbf{z}\mathbf{z}^\dagger\} = \mathbf{Q}\text{E}\{\mathbf{s}\mathbf{s}^\dagger\}\mathbf{Q}^\dagger = \mathbf{Q}\mathbf{Q}^\dagger$. The whitening transformation is always possible and can essentially be performed using L2-PCA [6].

We adopt the deflationary approach to ICA whereby the independent components are found sequentially, one after another [6], [13]. Each source is extracted by linearly projecting the data on a suitable N-dimensional direction $\mathbf{w}$

$$y = \mathbf{w}^\dagger \mathbf{z}, \tag{9}$$

where $\|\mathbf{w}\|_2 = 1$ to preserve the unit-variance constraint on y as in the whitening conditions.

According to eq. (8), if we define the change of variable $\mathbf{g} = \mathbf{Q}^\dagger \mathbf{w}$ then the projection can be expressed as

$$y = \mathbf{g}^\dagger \mathbf{s}, \tag{10}$$

where, again, $\|\mathbf{g}\|_2 = 1$. Hence, to find independent component s_i, vector $\mathbf{g}$ must be equal to $\pm\mathbf{e}_i$. Equivalently, the source recovery condition is achieved when $\mathbf{w}$ is, up to an irrelevant sign, the ith column $\mathbf{q}_i$ of the unitary mixing matrix $\mathbf{Q}$, so that $\mathbf{w}^\dagger\mathbf{z} = \pm\,\mathbf{q}_i^\dagger\mathbf{Q}\mathbf{s} = \pm\, s_i$, as can be seen by combining eqs. (8)-(9).

Remark that, because ICA is a statistical approach by nature, the use of the mathematical expectation becomes necessary in its theoretical formulations, as in the covariance matrices defined above. In practice, finite-length data are observed and expectations are replaced by sample averages in order to develop working algorithms.

2.2.1 Quasi-Newton Algorithms for ICA Computation

It is interesting to realize that some existing algorithms originally proposed to solve the ICA problem present certain similarities with algorithms for L1-norm optimization. The Central Limit Theorem states that the distribution of the sum of independent random variables is approximately Gaussian. ICA may be seen as the reverse process of this theorem: as we want to undo the mixing of independent variables, we have to search for directions that maximize the nonGaussianity, or negentropy, of the projection (9). To this end, Hyvärinen proposed to maximize the following approximation of negentropy [13]:

$$J(\mathbf{w}) = \left[\mathrm{E}\{G(y)\} - \mathrm{E}\{G(\nu)\}\right]^2 \tag{11}$$

under constraint $\|\mathbf{w}\|_2 = 1$, where ν is a normalized Gaussian variable and G is any sufficiently smooth nonquadratic even function. Because J is the squared difference of the expectation $\mathrm{E}\{G(y)\}$ from what it would be if y were Gaussian, it can be considered as a natural measure of how far the distribution of y is from a Gaussian distribution.

A result of [12] establishes the link between the stationary points of $\mathrm{E}\{G(y)\}$ and the independent components. Expressing this theorem in our notation, we have

Theorem 1. *Assume that the input data follow model (8) under conditions A1-A2, and that G is a sufficiently smooth even function. Then the local maxima (resp. minima) of $\mathrm{E}\{G(\mathbf{w}^\dagger\mathbf{z})\}$ under the constraint $\|\mathbf{w}\|_2 = 1$ include the columns $\mathbf{q}_i$ of the unitary mixing matrix $\mathbf{Q}$ such that the corresponding independent components s_i satisfy*

$$\mathrm{E}\{s_i\, G'(s_i) - G''(s_i)\} \;>\; 0 \quad (resp.\; <\; 0),$$

where G' is the derivative and G'' the second derivative of G.

An approximate Newton algorithm to optimize this criterion leads to the fixed-point iteration [13]

$$\tilde{\mathbf{w}} = \mathrm{E}\{\mathbf{z}G'(\mathbf{w}^\dagger\mathbf{z})\} - \mathrm{E}\{G''(\mathbf{w}^\dagger\mathbf{z})\}\mathbf{w} \tag{12a}$$

$$\mathbf{w}^+ = \tilde{\mathbf{w}}/\|\tilde{\mathbf{w}}\|_2, \tag{12b}$$

which is widely known as *FastICA*. Under the ICA model (8), the FastICA algorithm shows at least quadratic local convergence, which becomes cubic in the case of sources with symmetric distributions. For the kurtosis-based nonlinearity $G(y) = y^4$, convergence is global and cubic. These convergence results require in particular the computation of the fourth derivative of $G(y)$ [13, Appendix A].

2.3 Gaps in ICA Theory Based on the Absolute Value Function

Assuming ergodicity conditions, the L1-norm $\|\mathbf{y}\|_1$ becomes (up to an irrelevant scale factor) proportional to $\mathrm{E}\{|y|\}$ for large enough sample size, $\|\mathbf{y}\|_1 \underset{T\to+\infty}{\longrightarrow} T\mathrm{E}\{|y|\}$. Because $\mathrm{E}\{G(\nu)\}$ is constant for fixed nonlinearity G, it follows that the L1-norm criterion (3) and the ICA criterion (11) with $G(y) = |y|$ are related. But this connection has to be handled with care for several reasons that are explained in this section.

First, the result of Theorem 1 assumes the differentiability of G, whereas that of $G(y) = |y|$ fails at the origin. Let us suppose that $G'(y) = \mathrm{sign}(y)$ (admitting the differentiability of G) and $G''(y) = 2\delta(y)$ (admitting again the differentiability of G'), where $\delta(y)$ denotes Dirac's delta function. Defining $f_i(\cdot)$ as the probability density function (pdf) of s_i, Theorem 1 can be expressed as

- If $\mathrm{E}\{|s_i|\} > 2f_i(0)$, then source s_i defines a local maximum of criterion (11).
- If $\mathrm{E}\{|s_i|\} < 2f_i(0)$, then source s_i defines a local minimum of criterion (11).

Although this is indeed the result that will be found in the analysis of Section 3, strictly speaking Theorem 1 does not apply here because $G(y) = |y|$ is not a differentiable function.

Secondly, for $G(y) = |y|$ the first step of FastICA iteration (12) would read

$$\tilde{\mathbf{w}} = \mathrm{E}\{\mathbf{z}\,\mathrm{sign}(\mathbf{w}^\dagger\mathbf{z})\} - 2f_y(0)\mathbf{w}, \tag{13}$$

where $f_y(0)$ stands for the pdf of $y = \mathbf{w}^\dagger\mathbf{z}$ evaluated at the origin. To reach this expression, we have applied the result

$$\int_{-\infty}^{\infty} h(x)\delta^{(k)}(x)\,\mathrm{d}x = (-1)^k h^{(k)}(0) \tag{14}$$

for any differentiable function $h(\cdot)$, where symbol $(\cdot)^{(k)}$ denotes the kth derivative. Update (13) requires the pdf of the extractor output to be estimated at each iteration, which may be impractical. Nevertheless, if the second term on the right-hand side is dropped, the sample counterpart of iteration (13), including the normalization step, is equivalent to Kwak's algorithm in eq. (7). Interestingly, the same algorithm is obtained in [32] for the extraction of an independent component under partial knowledge of its positive support. These algorithms, however, show convergence problems in practical settings.

Thirdly, and related to the previous point, the proof of local convergence of the FastICA algorithm (12) given in [13, Appendix A] relies on the expectations $\mathrm{E}\{G^{(3)}(s_i)\}$ and $\mathrm{E}\{G^{(4)}(s_i)\}$, where s_i is the source being targeted. When $G(y) = |y|$ the convergence characteristics come to depend on the derivatives of the source pdf at the origin, since $\mathrm{E}\{G^{(3)}(s_i)\} = -2f_i'(0)$ and $\mathrm{E}\{G^{(4)}(s_i)\} = 2f_i''(0)$, as can be seen using the delta function property (14). These derivatives, however, are undefined in important cases such as the generalized Gaussian distribution $f_i(u) \propto \exp(-|u|^\alpha)$ with parameter $\alpha \leq 1$, including, e.g., the double exponential (Laplacian) density.

In a bid to surmount these shortcomings, one may use differentiable approximations of the absolute value

function, which are actually very popular in the ICA literature, such as

$$G(u) = \frac{1}{a}\log\,\cosh(a\,u) \tag{15}$$

with $1 \leq a \leq 2$ [13]. Such approximations would in addition spare the need, noticed in eq. (13), for estimating the extractor output pdf at each iteration. However, as will be seen in the numerical experiments presented in Section 4, these approximations do not offer the same convergence characteristics and robustness to outliers as the absolute value.

We conclude from this preliminary analysis that, contrary to what may be intuitively expected from previous results, the standard theory of nonlinear functions for ICA needs to be taken cautiously when the smoothness condition on G required by Theorem 1 is violated. The present contribution bridges this gap by establishing in a more rigorous manner the exact link between the local extrema of $\mathrm{E}\{|y|\}$ and the separating solutions, depending on certain characteristics of the source distributions. Our development does not rely on any differentiable approximations, but on the absolute value function itself. By virtue of the globally convergent algorithms reviewed in Section 2.1.1, the equivalence shown in this work opens the possibility of performing ICA based on L1-PCA with guaranteed convergence to the right solution.

3 ICA via L1-Norm PCA

This section establishes the precise link between the stationary points of the absolute value criterion $\mathrm{E}\{|y|\}$ and the independent components under the ICA model (8). To this end, the criterion is first cast into convenient closed-form expressions (Section 3.1) that will prove useful in characterizing its local extrema (Sections 3.2 and 3.3). It will be shown in particular that the criterion is actually minimized for sources fulfilling certain conditions involving their probability distributions. These conditions can also be linked to the source kurtosis values (Section 3.4). The theoretical developments conclude by showing how algorithms for L1-norm maximization can be modified to perform minimization (Section 3.5). Such modified algorithms are necessary to extract independent components minimizing the absolute value criterion.

3.1 Closed-Form Expressions for the Expectation of the Absolute Value

Let $y = \mathbf{g}^\dagger\mathbf{s} = \sum_{n=1}^{N} g_n\, s_n$ be a linear combination of the sources s_n with weights given by g_n. To shorten notation, we write $\Upsilon(y) := \frac{1}{2}\mathrm{E}\{|y|\}$ in the sequel, where the scale factor is included for mathematical convenience.

Lemma 2. *We have that*

$$\Upsilon(y) = \sum_{n=1}^{N} g_n\, \mathcal{G}_n, \tag{16}$$

where $\mathcal{G}_n := P(y > 0)\mathrm{E}\{s_n \,|\, y > 0\}$, $P(y > 0)$ is the probability of $y > 0$ and $\mathrm{E}\{s_n \,|\, y > 0\}$ denotes the conditional expectation of the nth source when the event $y > 0$ has occurred.

Proof is in Appendix A. Assuming that the source of interest is s_i, we may express y as $y = g_i s_i + \sum_{n\neq i} g_n s_n$. Let σ_i be the standard deviation of $\sum_{n\neq i} g_n s_n$ (i.e., $\sigma_i^2 = \sum_{n\neq i} g_n^2$). The contribution of the interfering sources can be represented in terms of the standardized variable $b_i = (\sum_{n\neq i} g_n s_n)/\sigma_i$ as follows:

$$y = g_i s_i + \sigma_i b_i. \tag{17}$$

We are now in a position to give an expression for $\mathcal{G}_i$:

Theorem 2. *We have that*

$$\begin{aligned}\mathcal{G}_i &:= P(y > 0)\mathrm{E}\{s_i \,|\, y > 0\} \\ &= \int_{-\infty}^{\infty} s f_i(s) C_{b_i}\left(-\frac{g_i}{\sigma_i}s\right)\mathrm{d}s,\end{aligned} \tag{18}$$

where $f_i(\cdot)$ is the probability density function (pdf) of s_i and $C_{b_i}(\cdot)$ is the complementary cumulative distribution function (ccdf) of b_i, i.e., $C_{b_i}(x) = P(b_i > x) = 1 - F_{b_i}(x)$, with $F_{b_i}(\cdot)$ being the cumulative distribution function (cdf) of b_i.

Proof is given in Appendix B. For example, let us evaluate this expression for the case $g_i = 0$. It follows that $y = \sigma_i b_i$ is a function of b_i only and, consequently, y is independent of s_i. Noting that $\sigma_i > 0$, it follows that $C_{b_i}(0) = P(b_i > 0) = P(\sigma_i b_i > 0) = P(y > 0)$ and substituting these values above we have that $\mathrm{E}\{s_i \,|\, y > 0\} = \frac{C_{b_i}(0)}{P(y>0)}\int_{-\infty}^{\infty} s f_i(s)\,\mathrm{d}s = \mathrm{E}\{s_i\} = 0$ as might be expected. Note also that we can select any index j different from i and write

$$b_i = \left(g_j s_j + \sum_{n\neq i,j} g_n s_n\right)\Big/\sigma_i.$$

This expression will facilitate the calculation of $C_{b_i}(x)$. Our third theorem is as follows

Theorem 3. *We have that*

$$C_{b_i}(\alpha) = 1 - \int_{-\infty}^{\infty} f_j(x) F_{ij}(\sigma_i\alpha - g_j x)\,\mathrm{d}x,$$

where $f_j(\cdot)$ is the pdf of s_j and $F_{ij}(\cdot)$ is the cdf of $\sum_{n\neq i,j} g_n\, s_n$, with $i \neq j$.

Proof is in Appendix C. Substituting this expression into eq. (18), we get

$$\begin{aligned}\mathcal{G}_i &= P(y > 0)\mathrm{E}\{s_i \,|\, y > 0\} \\ &= \int_{-\infty}^{\infty} s f_i(s) C_{b_i}\left(-\frac{g_i}{\sigma_i}s\right)\mathrm{d}s \\ &= \int_{-\infty}^{\infty} s f_i(s)\left[1 - \int_{-\infty}^{\infty} f_j(x) F_{ij}(-g_i s - g_j x)\,\mathrm{d}x\right]\mathrm{d}s \\ &= -\iint_{\mathbb{R}^2} s f_i(s) f_j(x) F_{ij}(-g_i s - g_j x)\,\mathrm{d}x\,\mathrm{d}s.\end{aligned} \tag{19}$$

Note that $\int_{-\infty}^{\infty} s f_i(s)\,\mathrm{d}s$ cancels out under the assumption of zero-mean sources. This formula has the virtue of making explicit the dependence of the left hand side of eq. (19) with respect to any coefficient g_j, and will be very useful for calculating derivatives in the next section.

3.2 Optimization of the L1 Cost

As mentioned earlier, under the assumption that the data follows the ICA data model, the L1-PCA problem can be

cast as: maximize the expectation of $|y|$ with respect to $\mathbf{g}$ under the constraint $\|\mathbf{g}\|_2 = 1$. The link between L1-PCA and ICA will be established if the extremal points are the canonical base vectors $\mathbf{g} = \pm\mathbf{e}_i$, which define the separating solutions (Section 2.2). We therefore consider the problem

$$\max_{\|\mathbf{g}\|_2=1} \Upsilon(y). \tag{20}$$

Because $\|\mathbf{g}\|_2^2 = 1$ is an equivalent constraint, using eq. (16) the Lagrangian $\mathcal{L}$ becomes

$$\mathcal{L}(\mathbf{g}, \lambda) = \sum_{n=1}^{N} g_n \mathcal{G}_n + \lambda(\|\mathbf{g}\|_2^2 - 1).$$

Then, we have

$$\frac{\partial \mathcal{L}}{\partial g_i} = \mathcal{G}_i + \sum_{n=1}^{N} g_n \frac{\partial \mathcal{G}_n}{\partial g_i} + 2\lambda g_i,$$

where λ is the Lagrange multiplier. Differentiating (19) we get after some calculations

$$\frac{\partial \mathcal{G}_i}{\partial g_i} = \iint_{\mathbb{R}^2} s^2 f_i(s) f_j(x) f_{ij}(-g_i s - g_j x)\, \mathrm{d}s\, \mathrm{d}x \tag{21}$$

$$\frac{\partial \mathcal{G}_n}{\partial g_i} = \iint_{\mathbb{R}^2} xs f_n(s) f_i(x) f_{ni}(-g_n s - g_i x)\, \mathrm{d}s\, \mathrm{d}x,\ n \neq i, \tag{22}$$

where $f_{ij}(\cdot)$ is the first derivative of $F_{ij}(\cdot)$ or, in other words, the pdf of $\sum_{n\neq i,j} g_n\, s_n$.

Let us study the problem under the assumption that the projection y is equal to one of the sources, e.g., $y = s_1$. In this case

$$g_1 = 1 \tag{23}$$

$$g_2 = g_3 = \ldots = g_N = 0 \tag{24}$$

$$\mathcal{G}_1 \propto E\{s_1 | y > 0\} = E\{s_1 | s_1 > 0\} \tag{25}$$

$$\mathcal{G}_n \propto E\{s_n | y > 0\} = E\{s_n | s_1 > 0\} = E\{s_n\} = 0, n \neq 1. \tag{26}$$

From eq. (24), the random variable $\sum_{n\neq 1,j} g_n\, s_n$ becomes zero so that its pdf can be modeled as a Dirac's delta function, $f_{1j}(\cdot) = \delta(\cdot)$.

Substituting all these values into (21) and (22), we get

$$\frac{\partial \mathcal{G}_1}{\partial g_1} = \iint_{\mathbb{R}^2} s^2 f_1(s) f_j(x) \delta(-s)\, \mathrm{d}s\, \mathrm{d}x = 0 \tag{27}$$

$$\frac{\partial \mathcal{G}_1}{\partial g_i} = \iint_{\mathbb{R}^2} xs f_1(s) f_i(x) \delta(-s)\, \mathrm{d}s\, \mathrm{d}x = 0, \qquad i \neq 1. \tag{28}$$

Replacing now (23)-(28) into the above expression for $\partial\mathcal{L}/\partial g_i$, we get

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial g_1} &= \mathcal{G}_1 + 2\lambda \\ \frac{\partial \mathcal{L}}{\partial g_i} &= 0, \qquad i \neq 1. \end{aligned}$$

Obviously we are at a critical point. The value of the Lagrange multiplier is obtained simply by setting to zero the first of the above equations

$$\begin{aligned} \lambda &= -\frac{\mathcal{G}_1}{2} = -\frac{1}{2} \mathrm{E}\{s_1 \,|\, s_1 > 0\} P(y > 0) \\ &= -\frac{1}{2} \int_0^\infty s f_1(s)\, \mathrm{d}s. \end{aligned} \tag{29}$$

Thus we can conclude that the ICA solutions are extrema candidates of the L1-norm PCA criterion because they satisfy the Lagrange conditions. Next section employs second-derivative tests to discern the nature (maxima or minima) of these critical points.

3.3 Stationary Point Analysis

Let $\mathbf{L}(\mathbf{g}, \lambda)$ be the Hessian matrix of the Lagrangian with respect to $\mathbf{g}$, that is, $\mathbf{L}(\mathbf{g}, \lambda) = \mathbf{F}(\mathbf{g}) + \lambda \mathbf{H}(\mathbf{g})$, where $\mathbf{F}(\mathbf{g})$ is the Hessian of $\Upsilon(y)$ and

$$\mathbf{H}(\mathbf{g}) \overset{\text{def}}{=} \left[\frac{\partial^2}{\partial g_i \partial g_j} \sum_{n=1}^{N} g_n^2 \right]_{i,j} \qquad i, j = 1, \ldots, N.$$

Matrix $\mathbf{L}(\mathbf{g}, \lambda)$ plays a role similar to that of $\mathbf{F}(\mathbf{g})$ in unconstrained maximization problems. The next theorem recalls the conditions for local maxima and minima in terms of $\mathbf{L}$. Before proceeding, however, a definition should be recalled: the *tangent space* at a point $\mathbf{g}$ on the unit sphere $\|\mathbf{g}\|_2^2 = 1$ is the set $\mathcal{T}(\mathbf{g}) = \{\mathbf{v} \in \mathbb{R}^N : \mathbf{v}^\dagger \mathbf{g} = 0\}$, i.e., $\mathcal{T}$ comprises the vectors that are orthogonal to the gradient of the sphere at point $\mathbf{g}$.

Theorem 4. *Assume that $\mathbf{g}^*$ and λ^* satisfy the Lagrange conditions (i.e., $\mathbf{g}^*$ is a critical point of the given objective function subject to equality constraint). We use the notation $\mathbf{L}^* = \mathbf{L}(\mathbf{g}^*, \lambda^*)$ and $\mathcal{T}^* = \mathcal{T}(\mathbf{g}^*)$. Then*

- *If $\mathbf{v}^\dagger \mathbf{L}^* \mathbf{v} < 0$ for all $\mathbf{v} \in \mathcal{T}^* \setminus \{\mathbf{0}_N\}$ then $\mathbf{g}^*$ is a strict local maximum.*
- *If $\mathbf{v}^\dagger \mathbf{L}^* \mathbf{v} > 0$ for all $\mathbf{v} \in \mathcal{T}^* \setminus \{\mathbf{0}_N\}$ then $\mathbf{g}^*$ is a strict local minimum.*
- *If $\mathbf{v}^\dagger \mathbf{L}^* \mathbf{v}$ is positive for some vector $\mathbf{v} \in \mathcal{T}^*$ and negative for another such vector, then $\mathbf{g}^*$ is neither a local maximum nor a local minimun.*

The corresponding proof can be found, for example, in [4, chapter 19]. In the previous section, we found that the pair

$$(\mathbf{g}^*, \lambda^*) = \left(\mathbf{e}_1, -\frac{1}{2} \int_0^\infty s f_1(s)\, \mathrm{d}s \right)$$

satisfies the Lagrange conditions. Now we have to determine the Hessian matrix of the Lagrangian, although, for the sake of readability, we have deferred such calculation to Appendix D. The Hessian becomes surprisingly simple

$$\mathbf{L}^* = \mathrm{diag}[0, f_1(0), \ldots, f_1(0)] + 2\lambda^* \mathbf{I}_N.$$

The tangent space is

$$\mathcal{T}^* = \{\mathbf{v} \in \mathbb{R}^N : \mathbf{e}_1^\dagger \mathbf{v} = 0\} = \{\mathbf{v} \in \mathbb{R}^N : \mathbf{v} = [0, v_2, \ldots, v_N]^\dagger\}.$$

Hence, for each $\mathbf{v} \in \mathcal{T}^* \setminus \{\mathbf{0}_N\}$

$$\mathbf{v}^\dagger \mathbf{L}^* \mathbf{v} = \left(\sum_{n=2}^{N} v_n^2\right)\left[f_1(0) - \int_0^\infty s f_1(s)\,\mathrm{d}s\right].$$

It follows that $\mathbf{L}^*$ is negative definite for $f_1(0) < \int_0^\infty s f_1(s)\,\mathrm{d}s$, and thus we are at a maximizer. Otherwise, if the quantity in brackets is strictly positive, the point $\mathbf{g}^*$ is a minimizer. Repeating the above derivations for other source extracting solutions, we can generalize the previous results with the following theorem.

Theorem 5. *Consider the optimization problem (20) under ICA model (8)-(10). Then*

- *C1) If $f_i(0) < \int_0^\infty s f_i(s)\,\mathrm{d}s$, vectors $\pm\mathbf{e}_i$ are strict local maxima of the constrained optimization problem.*
- *C2) If $f_i(0) > \int_0^\infty s f_i(s)\,\mathrm{d}s$, vectors $\pm\mathbf{e}_i$ are strict local minima of the constrained optimization problem.*

Because the sources have zero mean, $\mathrm{E}\{|s_i|\} = 2\int_0^\infty s f_i(s)\,\mathrm{d}s$, and conditions C1-C2 are equivalent to those found in Section 2.3 for the approximate negentropy contrast (11) with $G(y) = |y|$. However, the conditions found in Section 2.3 might not have held because Theorem 1 assumes the differentiability of $G(y)$, which is not the case here. Our development, which works directly on the absolute value function without simplifications, proves the validity of these conditions.

Finally, remark that for Gaussian distributions $f_i(0) = \int_0^\infty s f_i(s)\,\mathrm{d}s$, and so neither C1 nor C2 are fulfilled. This result is closely related to the fact that ICA is unable to deal with Gaussian sources, as reflected by the negentropy approximation in eq. (11).

3.4 Link with Kurtosis Optimization

If the sources have symmetric distributions, the above conditions can be cast in terms of kurtosis, the normalized fourth-order cumulant. Consider the Gram-Charlier expansion of the ith source pdf $f_i(s)$ truncated at order 4 [7]:

$$f_i(s) \simeq \frac{1}{\sqrt{2\pi}}\exp(-s^2/2)\left[1 + \frac{\kappa_i}{4!}(s^4 - 6s^2 + 3)\right], \qquad (30)$$

where $\kappa_i = \mathrm{E}\{s_i^4\} - 3$ denotes the fourth-order cumulant of s_i. Under normalization assumption A1, the fourth-order cumulant becomes the kurtosis. It follows that:

$$f_i(0) - \int_0^\infty s f_i(s)\,\mathrm{d}s = \frac{\kappa_i}{\sqrt{72\pi}}.$$

We conclude that symmetric sources with negative (resp. positive) kurtosis are maximizers (resp. minimizers) of the L1-norm criterion $\mathrm{E}\{|y|\}$. Positive- and negative-kurtosis pdfs are called, respectively, superGaussian and subGaussian distributions in the ICA literature [6].

To further deepen this characterization, we note that, being y a zero-mean variable, the L1-norm criterion can be easily rewritten as $\Upsilon(y) = \int_0^\infty x f_y(x)\,\mathrm{d}x$. Let us assume that $f_y(x)$ can be accurately approximated by a Gram-Charlier expansion like that in eq. (30), but replacing κ_i by $\kappa_y = \mathrm{E}\{y^4\} - 3$, the fourth-order marginal cumulant of y. Substituting above, we easily get

$$\Upsilon(y) = \frac{1}{\sqrt{2\pi}}\left(1 - \frac{\kappa_y}{4!}\right).$$

As a result, maximizing (resp. minimizing) Υ is equivalent to minimizing (resp. maximizing) κ_y, and L1-PCA becomes a kurtosis optimization criterion, which has extensively been studied in the literature; see, e.g., [6], [8], [26], [31] and references therein. The following theorem—reproduced here from [8] for the sake of clarity and completeness—summarizes the main results of kurtosis optimization

Theorem 6. *Suppose that $\kappa_i > 0$ for $i = 1, \ldots, r$ and that $\kappa_i < 0$ for $i = r+1, \ldots, N$ (we can set $r = 0$ or $r = N$ if necessary).*

- *If $r \neq 0$ and $r \neq N$, the arguments of the local maxima of κ_y subject to $\|\mathbf{g}\|_2 = 1$ are the vectors $\pm\mathbf{e}_i$ for $i = 1, \ldots, r$, and the arguments of the local minima are the vectors $\pm\mathbf{e}_i$ for $i = r+1, \ldots, N$.*
- *If $r = 0$, the arguments of the local minima of κ_y subject to $\|\mathbf{g}\|_2 = 1$ are the vectors $\pm\mathbf{e}_i$ for $i = 1, \ldots, N$, and the arguments of the local maxima are reduced to the vectors $\mathbf{g}$ for which $g_i^2 = \delta/\kappa_i$ for each i, where $\delta = (\sum_{n=1}^N 1/\kappa_n)^{-1}$.*
- *If $r = N$, the arguments of the local maxima of κ_y subject to $\|\mathbf{g}\|_2 = 1$ are the vectors $\pm\mathbf{e}_i$ for $i = 1, \ldots, N$, and the arguments of the local minima are reduced to the vectors $\mathbf{g}$ for which $g_i^2 = \delta/\kappa_i$ for each i, where $\delta = (\sum_{n=1}^N 1/\kappa_n)^{-1}$.*

By virtue of the connection between L1-PCA and kurtosis optimization established above, maximizing (resp. minimizing) Υ recovers the sources with negative (resp. positive) kurtosis. Now, applying the multilinearity property of higher-order cumulants on model (10), we have, in particular, $\kappa_y = \sum_{n=1}^N g_n^4 \kappa_n$ [24]. It follows that we may apply the following simple rule for deciding whether to maximize or minimize the L1-PCA criterion to perform ICA

- If $\kappa_y < 0$ then at least one source has negative kurtosis: maximize $\mathrm{E}\{|y|\}$.
- If $\kappa_y > 0$ then at least one source has positive kurtosis: minimize $\mathrm{E}\{|y|\}$.

After the successful extraction of the first independent component, we can apply a deflation step to remove the contribution of the extracted source from the observations. This procedure, rather standard in ICA, is then recursively applied to extract sequentially the rest of the estimated independent components [6].

3.5 Minimizing the L1-Norm Criterion

According to Theorem 5, independent components fulfilling condition C2 can be extracted by minimizing the absolute value criterion. Finding the minima can be carried out by existing algorithms for L1-norm maximization with suitable rather minor modifications. For this purpose, we note the following result, which holds asymptotically in the sample size T (Appendix E).

Lemma 3. *For any $\mathbf{c} \in \mathcal{S}_1$, the candidate solution set defined in eq. (6), we have as $T \to +\infty$*

$$\|\mathbf{Z}\mathbf{c}_{\text{opt}}\|_2 \leq \|\mathbf{Z}\mathbf{c}\|_2,$$

where $\mathbf{c}_{\text{opt}} = \text{sign}(\mathbf{Z}^\dagger \mathbf{q}_i)$ *and* $\mathbf{q}_i$ *is the column of the unitary mixing matrix associated with an independent component fulfilling condition C2.*

As recalled in Section 2.1.1, the algorithm for L1-norm maximization of [22] efficiently computes all candidate solutions that form set $\mathcal{S}_1$ and then selects the maximizer of $\|\mathbf{Z}\mathbf{c}\|_2$. The direction for data projection is finally computed as in eq. (5). As a result of Lemma 3, among all candidate vectors $\mathbf{c} \in \mathcal{S}_1$, one just needs to select the solution yielding the minimum value of $\|\mathbf{Z}\mathbf{c}\|_2$ in order to minimize the L1-norm criterion. Hence, it suffices to replace '$\arg\max$' by '$\arg\min$' in Step 4 of [22, Fig. 2]. This simple modification leads to an algorithm for L1-norm minimization with guaranteed global convergence and the same computational cost as the original algorithm for L1-norm maximization.

Remark that this result is asymptotic under the ICA model, for which, up to scale, $\mathbf{Z}\mathbf{c}_{\text{opt}} \underset{T\to+\infty}{\longrightarrow} \mathbf{w}_{\text{opt}}$. In general, the L1-norm minimizer may not be found in the direction of $\mathbf{Z}\mathbf{c}$ for some $\mathbf{c} \in \mathcal{S}_1$ because the L1-norm and the L2-norm minimization problems are not related as in Lemma 1.

4 Experimental Analysis

This section aims at illustrating the theoretical results developed in this work with the aid of some numerical experiments[1]. Throughout these experiments, zero-mean unit-variance independent sources have either uniform or Laplacian distributions. The uniform distribution is given by $f_u(s) = \frac{1}{2\sqrt{3}}$ for $s \in [-\sqrt{3}, \sqrt{3}]$, and zero otherwise. It follows that $f_u(0) = \frac{1}{2\sqrt{3}} < \int_0^\infty s f_u(s)\,\mathrm{d}s = \frac{9}{2}\frac{1}{2\sqrt{3}}$, thus fulfilling condition C1. Therefore, Theorem 5 shows that independent components of this type are extracted in the ICA model by maximizing the L1 cost. On the other hand, the distribution of a standardized Laplacian reads $f_\ell(s) = \frac{1}{\sqrt{2}}\exp\left(-\sqrt{2}|s|\right)$. Since $f_\ell(0) = \frac{1}{\sqrt{2}} > \int_0^\infty s f_\ell(s)\,\mathrm{d}s = \frac{1}{2\sqrt{2}}$, condition C2 is fulfilled and we have to minimize $\mathrm{E}\{|y|\}$ to extract such independent components according to Theorem 5.

Because $\|\mathbf{w}\|_2 = \|\mathbf{g}\|_2 = 1$, i.e., both vectors undergo the same constraint to optimize the contrasts, working on eq. (9) or eq. (10) is equivalent. Hence, the latter is used in all experiments below to generate data projections directly from the source signals.

4.1 Illustrative Examples

We first consider $N = 2$ sources with uniform densities. Under the whitening assumption, we have $\|\mathbf{g}\|_2^2 = g_1^2 + g_2^2 = 1$, and then both coefficients can be parameterized by a single angle, yielding $y = g_1 s_1 + g_2 s_2 = \cos(\theta) s_1 + \sin(\theta) s_2$. We generate $T = 10^4$ independent samples of the sources, and then compute criterion $\mathrm{E}\{|y|\}$ as a function of θ. The result is shown in Fig. 1a. The values of θ that maximize the criterion are located at $\theta = 0$ and $\theta = \pi/2$, points where y equals one of the independent components. The minimizer is located at $\theta = \pi/4$, and corresponds to the 'maximally-mixed' solution $y = (s_1 + s_2)/\sqrt{2}$. Fig. 1b shows the same criterion when the sources have a Laplacian distribution: the separating solutions are actually minimizers of the criterion, as predicted by Theorem 5.

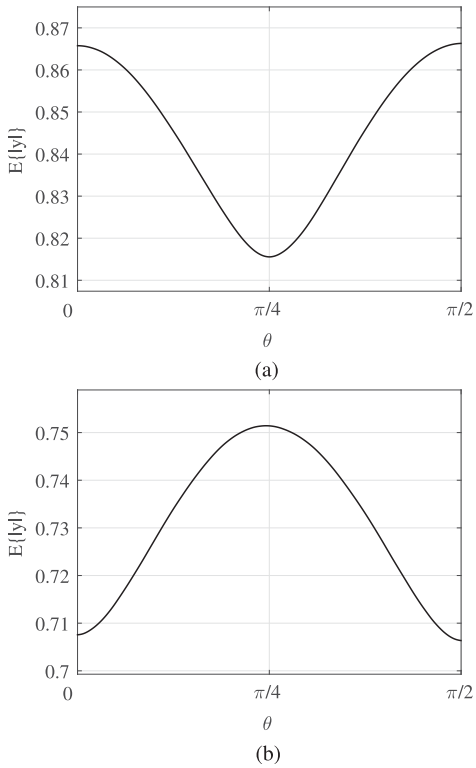

Fig. 1. Criterion $\mathrm{E}\{|y|\}$ as a function of angle θ for an orthogonal mixture of two sources with (a) uniform distribution and (b) Laplacian distribution. In the first case, the independent sources maximize the criterion, whereas they minimize it in the second case. The criterion values have been estimated from random unitary mixtures of independent source realizations composed of 10^4 samples.

The purpose of the second example is to show that we can use algorithms originally devised for L1-PCA to perform ICA. In particular, we consider Kwak's suboptimal iterative algorithm [17] (denoted 'iterative L1-PCA' hereafter) [summarized by eq. (7)] and Markopoulos' globally convergent method [22] (that we refer to as 'optimal L1-PCA'). These algorithms were reviewed in Section 2.1.1. Consider a mixture of $N = 3$ uniform sources: $y = g_1 s_1 + g_2 s_2 + (\sqrt{1 - g_1^2 - g_2^2}) s_3$. The contour lines of the criterion for a realization of $T = 10^4$ samples are depicted in Fig. 2a. Also displayed is the search path followed by the iterative L1-PCA algorithm, which converges near the maximum at $g_1 = g_2 = 0$, thus finding the independent component s_3. The optimal L1-PCA algorithm yields the same solution.

We now consider a mixture of three Laplacian sources. Fig. 3 shows that criterion $\mathrm{E}\{|y|\}$ presents a minimum at $g_1 = g_2 = 0$. The contour lines of $\mathrm{E}\{|y|\}$ and the search path followed by the iterative L1-PCA algorithm are depicted in Fig. 2b. As expected, the algorithm maximizes $\mathrm{E}\{|y|\}$ again, yielding the 'maximally-mixed' solution

$$y = \frac{1}{\sqrt{3}}(s_1 + s_2 + s_3)$$

1. MATLAB code for some of these experiments is available as supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/tpami.2016.2557797.

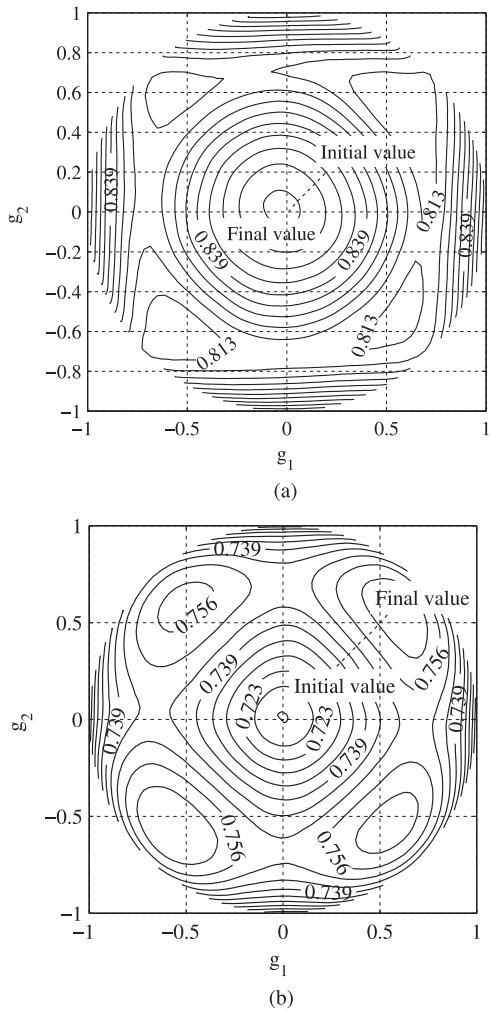

Fig. 2. Contour plots of criterion $\mathrm{E}\{|y|\}$ as a function of g_1 and g_2 (solid lines) and evolution of the iterative L1-PCA algorithm (7) (dashed lines) in a three-source scenario. (a) Sources fulfilling condition C1 (uniform distributions): The algorithm converges to a valid separating solution. (b) Sources fulfilling conditions C2 (Laplacian distributions): the algorithm fails to converge to a valid separating solution.

and failing to extract one of the independent components. Indeed, Laplacian sources fulfil condition C2, and so we actually need to minimize the cost function as established by Theorem 5. This minimization can be done using the optimal L1-PCA method modified as in Section 3.5, thus recovering source s_1 from the same data realization

$$y = 0.9945s_1 + 0.0936s_2 + 0.0742s_3.$$

To quantify source extraction performance more concisely, one can use the the interference-to-signal ratio (ISR) [6], defined as

$$\mathrm{ISR} = \frac{1 - \max_n |g_n|^2}{\max_n |g_n|^2} = \frac{1}{\max_n |g_n|^2} - 1. \tag{31}$$

This index cancels out when one of the sources has been perfectly extracted, and takes values greater than zero otherwise. The above source recovery result by the optimal L1-PCA yields an ISR of -19.6 dB.

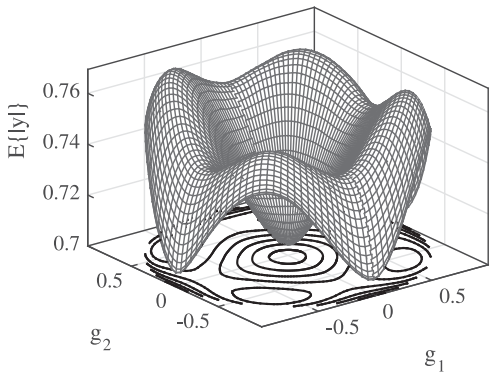

Fig. 3. Criterion $\mathrm{E}\{|y|\}$ as a function of g_1 and g_2 for a mixture of three sources satisfying condition C2 (Laplacian distributions).

4.2 Finite-Sample Performance

To quantitatively assess the comparative performance of the different algorithms considered in this work, a series of Monte Carlo experiments are conducted. Apart from the iterative and the optimal L1-PCA algorithms (Section 2.1.1), we also consider the FastICA algorithm (12) with absolute value nonlinearity ('FastICA-abs'), the differentiable approximation (15) ('FastICA-logcosh') and the fourth-power nonlinearity giving rise to the kurtosis-based FastICA ('FastICA-4power'). Although the fourth-power nonlinearity is obviously not related to the L1-norm criterion and lacks robustness to outliers, it features very fast (cubic) global convergence in ideal conditions and is particularly well suited to subGaussian distributions in terms of asymptotic variance [13], so it is used here as a benchmark. In FastICA-abs update (13), $f_y(0)$ is computed through the kernel density estimate with Gaussian kernel

$$\hat{f}_y(u) = \frac{1}{T\sqrt{2\pi}h} \sum_{t=1}^{T} \exp\left(-\frac{(u - y_t)^2}{2h^2}\right) \tag{32}$$

and Silverman's optimal bandwidth $h_{\mathrm{opt}} = \left(4\sigma_y^5/(3T)\right)^{1/5}$, where σ_y stands for the standard deviation of the extractor output samples $\{y_n\}_{n=1}^{T}$ at the current iteration. For FastICA-logcosh, we use $a = 1$ in eq. (15), a value in the range recommended in [13]. The iterative algorithms are stopped when consecutive updates lie sufficiently parallel to each other according to the criterion $1 - |\mathbf{w}^{\dagger}\mathbf{w}^{+}| < 10^{-3}N/T$ or when a maximum number of 1,000 iterations is reached.

The first experiment evaluates the methods ability to target one of the independent components as a function of the sample size T. To this end, the ISR (31) is averaged over ν independent Monte Carlo runs, with $\nu = \lceil 10^5/T \rceil$, where $\lceil \cdot \rceil$ represents the ceiling operator (smallest integer not less than its argument). At each run different source realizations and extracting vector initializations are randomly generated, but all methods operate on the same data and all iterative algorithms are initialized at the same point. The experiments are carried out on MATLAB R2015a software running on a 1.3 GHz Intel Core i5 processor.

The results in the two-source scenario ($N = 2$) are plotted in Figs. 4a and 4b. The fourth-power nonlinearity provides the best performance, as expected for subGaussian sources in the absence of outliers, followed by the L1-PCA

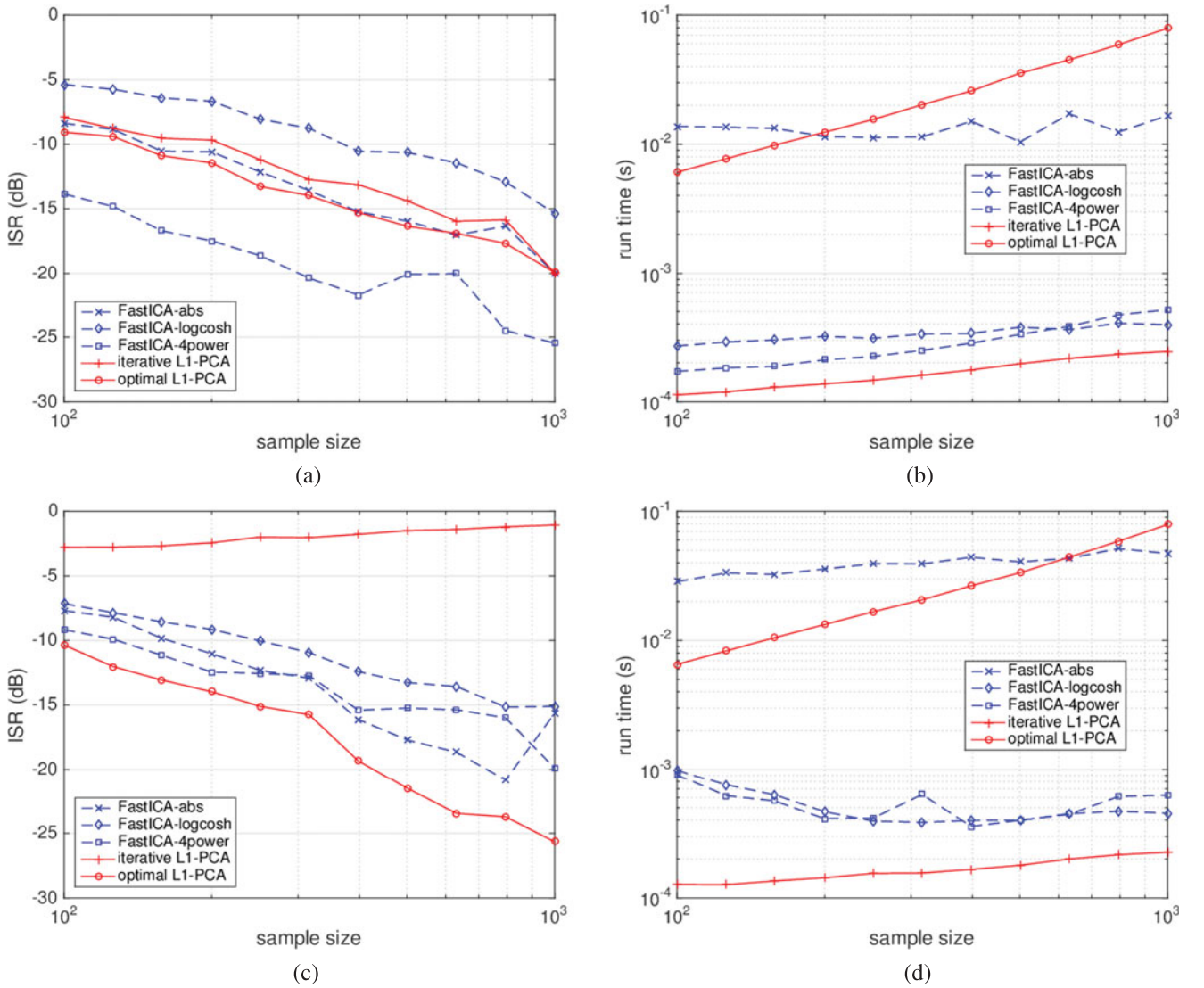

Fig. 4. Independent source extraction performance as a function of the sample size T, for $N = 2$ sources with uniform distributions [plots (a)-(b)] and Laplacian distributions [plots (c)-(d)].

algorithms and FastICA-abs. Clearly, FastICA-logcosh yields the worst results. All iterative methods provide a result in a fraction of a millisecond, except for FastICA-abs, which reaches the maximum number of iterations in about 10 percent of runs over the whole sample size range. The iterative L1-PCA shows the fastest performance in this experiment, whereas the optimal L1-PCA is on average over two orders of magnitude more time consuming.

The experiment is repeated but using sources with Laplacian distribution, whose extraction requires the minimization of the L1-norm criterion according to condition C2. Hence, the optimal L1-PCA method is modified to minimize the L1 norm as described in Section 3.5. Figs. 4c and 4d show that the iterative L1-PCA keeps its superior cost-effectiveness, but fails to extract an independent component as it wrongly tries to maximize the criterion. The modified optimal L1-PCA provides the best performance, up to 10 dB below the FastICA-abs algorithm. These results confirm, once more, that the L1-norm is to be minimized to target independent components fulfilling condition C2 and, in addition, that our simple modification can effectively accomplish this task. Again, the FastICA-abs algorithm shows the highest iteration count, reaching the maximum number of iterations in 30-35 percent of runs over the sample size interval considered in this experiment. FastICA-logcosh and FastICA-4power only reach the iteration limit around 2 percent of runs for small observation windows (< 300 samples).

4.3 Robustness Against Outliers

Keeping an observation length of $T = 100$ samples, we now consider mixtures of $N = 3$ sources and corrupt the observations at randomly chosen time instants by replacing the original samples by Gaussian noise realizations with mean $[10, 10, 10]^\dagger$ and identity covariance matrix. For uniformly distributed sources with varying corruption rate, the results are displayed in Figs. 5a and 5b. As expected, FastICA-4power is severely affected by the presence outliers, whereas the other methods show an improved robustness. It can also be remarked that the differentiable nonlinearity used by FastICA-logcosh incurs a loss in protection against outliers over the techniques directly relying on the absolute value function (FastICA-abs, iterative L1-PCA and optimal L1-PCA), which present very similar ISR trends. The iterative L1-PCA

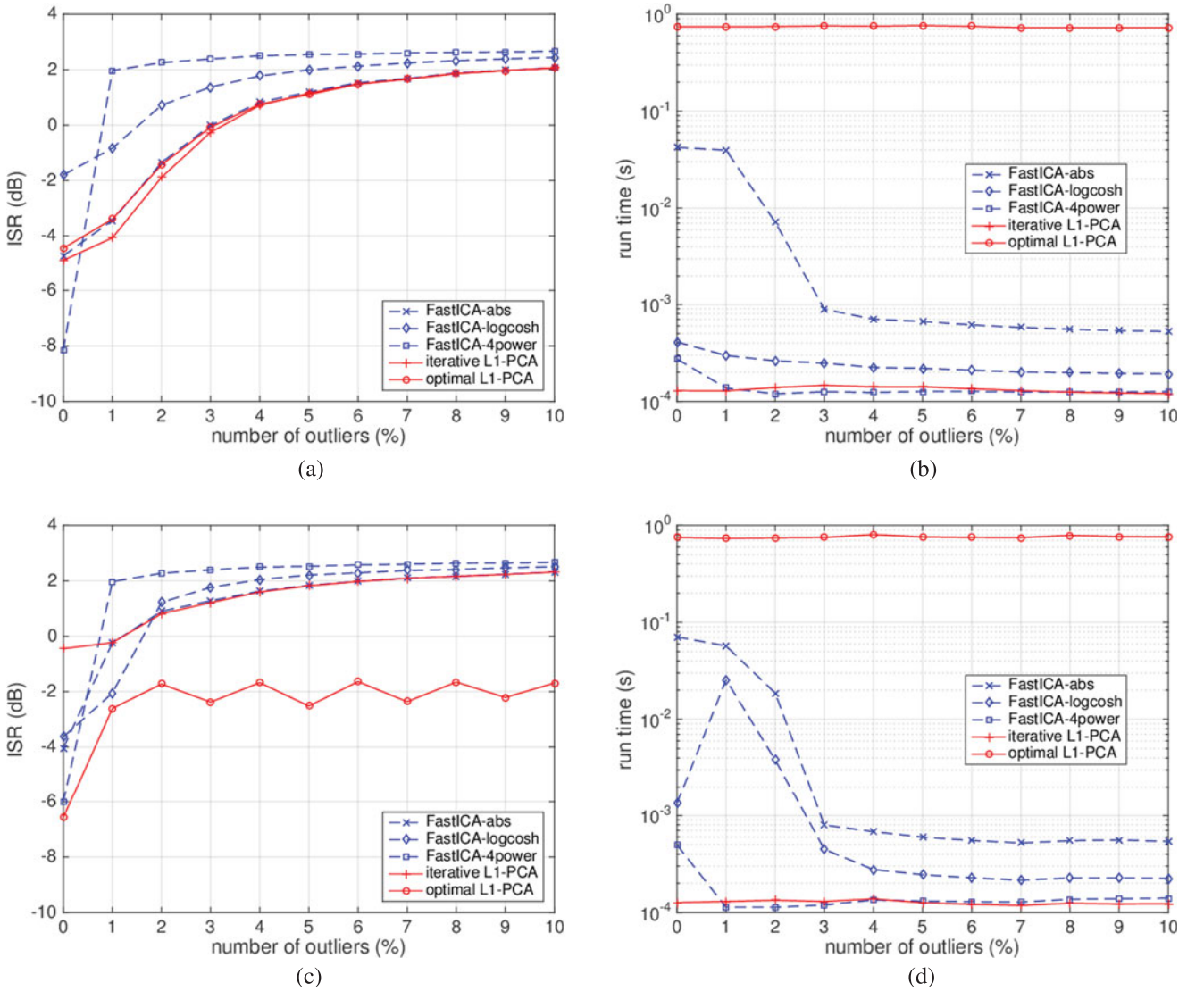

Fig. 5. Independent source extraction performance as a function of the outlier contamination rate, for $N = 3$ sources with uniform distributions [plots (a)-(b)] and Laplacian distributions [plots (c)-(d)]. The observation window length is fixed at $T = 100$ samples.

provides the same performance as the optimal L1-PCA at a fraction of the computational cost. The computation time of the optimal L1-PCA becomes constant, as it only depends on the sample size and the observation dimension (Section 2.1.1), which are fixed in this experiment. Again, FastICA-abs often requires a high number of iterations for uncorrupted data, reaching the iteration limit in 30-35 percent of runs.

The experiment is repeated with Laplacian distributed sources, and the results are represented in Figs. 5c and 5d. Now, the optimal L1-PCA algorithm with the modification proposed in Section 3.5 clearly yields the best ISR performance, although still at the expense of heavier computational demands as compared with the other methods. FastICA-logcosh shows a slight improvement over FastICA-abs, but presents similar convergence difficulties when dealing with few outliers. Indeed, both methods reach the iteration limit in around 80 percent of runs. The iterative L1-PCA presents again the best cost-effectiveness but, as in the experiment of the previous section, cannot deal with the

TABLE 1
Percentage of Monte Carlo Runs Where the Maximum Number of Iterations Is Reached in the Experiments of Figs. 4 and 5 (without Outliers): 'Uni': Uniform Sources. 'Lap': Laplacian Sources

	Case 1 $(N,T) = (2,100)$		Case 2 $(N,T) = (2,1000)$		Case 3 $(N,T) = (3,100)$	
	Uni	Lap	Uni	Lap	Uni	Lap
FastICA-abs	11.6	21.4	11	34	31.9	75.8
FastICA-logcosh	0	1.4	0	0	0.2	2.8

In case 3, FastICA-logcosh attains the iteration bound in 80.8 percent of runs for 1 percent of outliers.

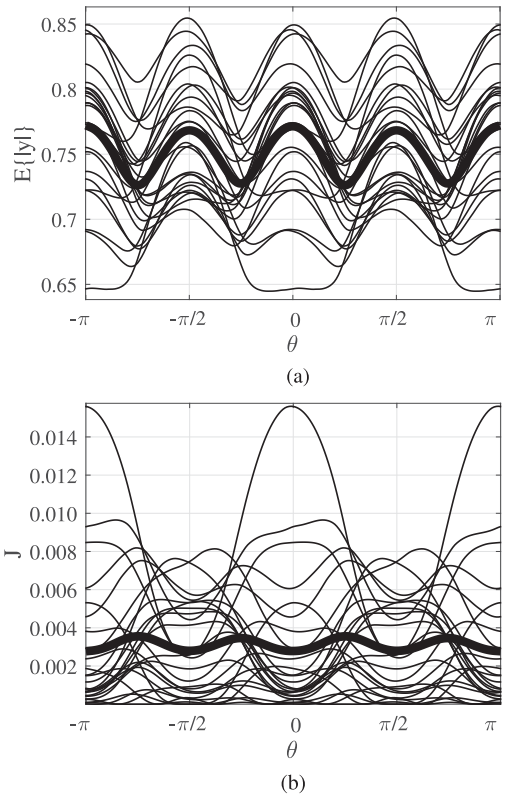

Fig. 6. Plots of statistical criteria obtained from random unitary mixtures of two sources with uniform distributions corrupted by outliers. Each line represents one of 25 independent data realizations, whereas the thick line plots the average of the 25 independent runs. (a) L1-norm criterion $\mathrm{E}\{|y|\}$. (b) ICA criterion $J(\mathbf{w})$ with $G(y) = \log\cosh(y)$. In the second case, the local maxima deviate significantly from valid separating solutions and, as a result, maximization of the criterion fails to extract an independent component in most cases.

source distribution of the nominal data and consequently it also offers limited ability to fight against outliers in this scenario.

Table 1 summarizes the percentage of Monte Carlo runs that FastICA-abs and FastICA-logcosh reach the maximum number of iterations for three cases considered in these experiments. FastICA-abs convergence problems are particularly remarkable for Laplacian sources, even when increasing the sample size, demonstrating that the convergence results of [13, Appendix A] do not hold for the absolute value nonlinearity under these distributions, as anticipated in Section 2.3. For comparison, the iterative L1-PCA never reaches the iteration limit in the cases considered, even though it operates in the same working conditions (observed data, initialization and stopping criterion) as the other iterative methods. To a lesser extent, FastICA-abs convergence problems persist for uniform sources, and may be due to the Gaussian kernel in the pdf estimate (32), which is probably suboptimal for such distributions. The differentiable approximation used in FastICA-logcosh alleviates these problems, although, as remarked above, at the expense of reduced robustness to outliers.

To gain some insights into the robustness of L1-PCA and related ICA algorithms against outliers, a final experiment is conducted to compare the absolute value criterion and the ICA objective function defined in (11) using the log-cosh nonlinearity (15) with $a = 1$. We generate a mixture of $N = 2$ uniform sources with $T = 10^4$ samples in which an outlier is added every $1,000$ samples. Outliers are randomly drawn from a zero-mean uniform distribution with variance equal to 25. Fig. 6a shows the dependence of $\mathrm{E}\{|y|\}$ on the angle θ (defined as in Section 4.1) for 25 independent random realizations of the sources; the thick line represents the average of the 25 realizations. In most cases, $\mathrm{E}\{|y|\}$ is maximized near integer multiples of $\pi/2$, ensuring that the sources can still be accurately estimated by maximizing the absolute value cost in spite of the outliers. Fig. 6b shows the differentiable approximation of objective function (11) evaluated on the same data. In this case, the ICA solution will often fail to recover the independent components, as the maxima of the criterion deviate considerably from the separating solutions.

5 Conclusion

We have shown that, under the ICA data model after whitening, L1-PCA leads to independent component extraction when the source distribution fulfils certain conditions defined in this work. In this case, since whitening can be carried out by L2-PCA, we may conclude that ICA can be accomplished by L2-PCA followed by L1-PCA. Under alternative conditions, we have proven that the L1 cost has actually to be minimized to extract independent sources. Algorithms originally devised for L1-norm maximization can easily be modified to perform L1-norm minimization for independent source extraction, as illustrated in this paper.

Our research basically complements the theory of ICA about the non-differentiable L1 (absolute value) cost function, and demonstrates that optimal L1-norm algorithms may give better accuracy and robustness than those of conventional ICA methods. The experimental analysis has also shown that iterative optimization techniques, whether based on direct use of the absolute value criterion or on differentiable approximations thereof, present difficulties converging to the right solution, especially when dealing with outliers. Globally convergent algorithm for L1-PCA overcomes this limitation at the expense of increased computational complexity that can become prohibitive for high sample size and/or observation dimensions. However, this issue can be alleviated since, in the L1-norm optimal algorithm, the elements of $\mathcal{S}_1$ (i.e., the candidate solutions) can be computed independently of each other, implying that the algorithm is fully parallelizable besides scalable [15], [22]. Future developments for reducing the computational burden, such as [16], will make these optimal L1-PCA algorithms a more attractive option for ICA in general practical settings.

These results open interesting new perspectives for performing ICA using optimal algorithms for L1-PCA with guaranteed global convergence, while enjoying the attractive robustness to outliers of the L1-norm criterion. Research into L1-PCA, a relatively recent topic, may benefit from the ICA body of knowledge, a well-established discipline. A fruitful cross fertilization of ideas is hence expected between these research areas. For example, the ICA algorithm [27] can be thought of as an extension of the L1-PCA algorithm in [17] for the case of complex-valued data.

As perspectives for the continuation of this work, other candidate solutions satisfying the Lagrange conditions may exist apart from those derived here, and their

characterization should be the subject of further research. Because ICA is a statistical technique by nature, our theoretical results only hold asymptotically in the sample size, and a finite-sample analysis may be worthwhile. In addition, future work should evaluate the source extraction capabilities of L1-PCA in real-world applications. Fresh interpretations of L1-PCA and ICA in the light of their relationship would be interesting, as well as their application in classification and data clustering.

Appendix A
Proof of Lemma 2

Consider determining the cumulative distribution function of y knowing that $y > 0$. Recalling the definition of conditional probability, we have that, if $x \geq 0$,

$$P(y \leq x \,|\, y > 0) = \frac{P(y \leq x, y > 0)}{P(y > 0)} = \frac{F_y(x) - F_y(0)}{P(y > 0)},$$

where $F_y(x) = P(y < x)$ is the cdf of y. Differentiating with respect to x we obtain

$$f(x \,|\, y > 0) = \begin{cases} \frac{f_y(x)}{P(y>0)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

where $f_y(x)$ is the pdf of y. The conditional mean is then given by

$$E\{y \,|\, y > 0\} = \int_0^\infty x f(x \,|\, y > 0)\, \mathrm{d}x = \frac{\int_0^\infty x f_y(x)\, \mathrm{d}x}{P(y > 0)}. \quad (33)$$

Because y has zero mean (by assumption A1), $\mathrm{E}\{|y|\} = 2\int_0^\infty x f_y(x)\, \mathrm{d}x$. Hence, $E\{y \,|\, y > 0\} = \frac{1}{2} E\{|y|\}/P(y > 0)$. Proof is completed by noting that $y = \sum_n g_n s_n$ and therefore $E\{y \,|\, y > 0\} = \sum_n g_n E\{s_n \,|\, y > 0\}$.

Appendix B
Proof of Theorem 2

Consider determining the cdf of s_i knowing that $y > 0$. Recalling the definition of conditional probability, we have that $P(s_i \leq s \,|\, y > 0) = \frac{P(y>0, s_i \leq s)}{P(y>0)}$. The joint probability can be obtained as $P(y > 0, s_i \leq s) = \int_{-\infty}^{s} P(y > 0 \,|\, s_i = x) f_i(x)\, \mathrm{d}x$, where $f_i(\cdot)$ denotes the pdf of s_i. Recalling eq. (17), the probability of event $y > 0$ happening is the same as the probability of event $b_i > -g_i s_i/\sigma_i$ and, noting the independence of s_i and b_i (assumption A1), $P(y > 0 \,|\, s_i = x) = P(b_i > -g_i s_i/\sigma_i \,|\, s_i = x) = P(b_i > -g_i x/\sigma_i) = C_{b_i}(-g_i x/\sigma_i)$. Thus we have

$$P(s_i \leq s \,|\, y > 0) = \frac{1}{P(y > 0)} \int_{-\infty}^{s} C_{b_i}(-g_i x/\sigma_i) f_i(x)\, \mathrm{d}x.$$

Differentiating with respect to s we obtain the pdf of s_i knowing that $y > 0$

$$f_i(s \,|\, y > 0) = \frac{1}{P(y > 0)} C_{b_i}(-g_i s/\sigma_i) f_i(s).$$

Substituting this in $\mathrm{E}\{s_i \,|\, y > 0\} := \int_{-\infty}^{\infty} s f_i(s \,|\, y > 0)\, \mathrm{d}s$ completes the proof.

Appendix C
Proof of Theorem 3

As $b_i = (\sum_{n \neq i} g_n s_n)/\sigma_i$, we select any index $j \neq i$ and write $b_i = (g_j s_j + \sum_{n \neq i,j} g_n s_n)/\sigma_i$. Then $C_{b_i}(x) = P(b_i > x) = \int_{-\infty}^{\infty} P(b_i > x \,|\, s_j = s) f_j(s)\, \mathrm{d}s$. From this point onwards, the proof is similar to that of the previous theorem. The probability of event $b_i > x$ happening is the same as the probability of $b_{ij} := \sum_{n \neq i,j} g_n s_n > \sigma_i x - g_j s_j$, and noting again the source independence assumption A1, we have $P(b_i > x \,|\, s_j = s) = P(b_{ij} > \sigma_i x - g_j s_j \,|\, s_j = s) = P(b_{ij} > \sigma_i x - g_j s) = 1 - F_{ij}(\sigma_i x - g_j s)$. Substituting this expression in the equation for $C_{b_i}(x)$ completes the proof.

Appendix D
Calculation of the Hessian of L1-Norm Criterion

From eq. (16), we have that

$$\frac{\partial^2}{\partial g_i^2} \Upsilon(y) = 2 \frac{\partial \mathcal{G}_i}{\partial g_i} + \sum_{n=1}^{N} g_n \frac{\partial^2 \mathcal{G}_n}{\partial g_i^2}$$

$$\frac{\partial^2}{\partial g_i\, \partial g_j} \Upsilon(y) = \frac{\partial \mathcal{G}_i}{\partial g_j} + \frac{\partial \mathcal{G}_j}{\partial g_i} + \sum_{n=1}^{N} g_n \frac{\partial^2 \mathcal{G}_n}{\partial g_i \partial g_j}, \qquad i \neq j.$$

We study the problem under the assumption that $y = s_1$. Since $g_1 = 1$ and $g_n = 0$ for $n \neq 1$, the above derivatives become

$$\frac{\partial^2}{\partial g_i^2} \Upsilon(y) = 2 \frac{\partial \mathcal{G}_i}{\partial g_i} + \frac{\partial^2 \mathcal{G}_1}{\partial g_i^2} \quad (34)$$

$$\frac{\partial^2}{\partial g_i\, \partial g_j} \Upsilon(y) = \frac{\partial \mathcal{G}_i}{\partial g_j} + \frac{\partial \mathcal{G}_j}{\partial g_i} + \frac{\partial^2 \mathcal{G}_1}{\partial g_i \partial g_j}, \qquad i \neq j. \quad (35)$$

We proceed as in eqs. (27)-(28) to find that $\frac{\partial \mathcal{G}_i}{\partial g_i} = f_1(0)$ for $i \neq 1$, and zero for $i = 1$. The proof is simple and can be left to the reader. Similarly, the reader can easily verify that $\frac{\partial \mathcal{G}_i}{\partial g_j} = \frac{\partial \mathcal{G}_j}{\partial g_i} = 0$ for all $i \neq j$. Differentiating (21) we get

$$\frac{\partial^2 \mathcal{G}_1}{\partial g_1^2} = -\iint_{\mathbb{R}^2} s^3 f_1(s) f_i(x) f_{1i}'(-g_1 s - g_i x)\, \mathrm{d}s\, \mathrm{d}x, \quad (36)$$

and, for $i \neq 1$,

$$\frac{\partial^2 \mathcal{G}_1}{\partial g_i^2} = -\iint_{\mathbb{R}^2} x^2 s f_1(s) f_i(x) f_{1i}'(-g_1 s - g_i x)\, \mathrm{d}s\, \mathrm{d}x, \quad (37)$$

where $f_{1i}'(\cdot)$ is the first derivative of $f_{1i}(\cdot)$. From eq. (24), the random variable $\sum_{n \neq 1,i} g_n s_n$ becomes zero so that its pdf tends to Dirac's delta function. Taking into account (14), we obtain $\frac{\partial^2 \mathcal{G}_1}{\partial g_i^2} = f_1(0)$ for $i \neq 1$, and zero for $i = 1$. The calculation of $\partial^2 \mathcal{G}_1 / \partial g_i \partial g_j$ is more involved. For $i \neq 1$ we have that [see eq. (22)]

$$\frac{\partial \mathcal{G}_1}{\partial g_i} = \iint_{\mathbb{R}^2} x s f_1(s) f_i(x) f_{1i}(-g_1 s - g_i x)\,\mathrm{d}s\,\mathrm{d}x.$$

The problem here is that the dependence with g_j, for $j \neq i$, is not explicit. To overcome this drawback, recall that f_{1i} is the pdf of $\sum_{n\neq 1,i} g_n s_n = g_j s_j + \sum_{n\neq 1,i,j} g_n s_n$. Hence, we can express f_{1i} as the convolution of the pdfs of $g_j s_j$ and $\sum_{n\neq 1,i,j} g_n s_n$ and make the differentiation. The interested reader can verify that $\frac{\partial^2 \mathcal{G}_1}{\partial g_i \partial g_j} = 0$ for all $i \neq j$. Substituting above, we find that the Hessian matrix of $\Upsilon(y)$ is diagonal, and is given by $\mathbf{F} = \mathrm{diag}[0, f_1(0), \ldots, f_1(0)]$.

APPENDIX E
PROOF OF LEMMA 3

Let $\mathbf{S}$ define the source matrix composed of T samples (columns). Let us assume without loss of generality that s_1 is the source of interest fulfilling condition C2. According to observation model (8), projecting the data along $\mathbf{w}_{\mathrm{opt}} = \mathbf{q}_1$, where $\mathbf{q}_1$ is the first column of the unitary mixing matrix $\mathbf{Q}$, yields source of interest, i.e., $\mathbf{w}_{\mathrm{opt}}^\dagger \mathbf{Z} = \mathbf{s}_1^\dagger$, where $\mathbf{s}_i^\dagger$ denotes the ith row of matrix $\mathbf{S}$. We define $\mathbf{c}_{\mathrm{opt}} = \mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w}_{\mathrm{opt}}) = \mathrm{sign}(\mathbf{s}_1)$. Now, we can write $\mathbf{Z}\mathbf{c}_{\mathrm{opt}} = \mathbf{Q}\mathbf{S}\,\mathrm{sign}(\mathbf{s}_1)$. The first entry of vector $\mathbf{S}\,\mathrm{sign}(\mathbf{s}_1)$ tends asymptotically (for large sample size) to $\mathrm{E}\{|s_1|\}$ up to an irrelevant non-null scale factor, whereas its jth entry, $j \neq 1$, tends to $\mathrm{E}\{s_j \mathrm{sign}(s_1)\}$, which cancels out by virtue of the source statistical independence assumption. Hence

$$\frac{\mathbf{Z}\mathbf{c}_{\mathrm{opt}}}{\|\mathbf{Z}\mathbf{c}_{\mathrm{opt}}\|_2} \underset{T\to+\infty}{\longrightarrow} \mathbf{w}_{\mathrm{opt}}.$$

According to this relationship, $\|\mathbf{w}_{\mathrm{opt}}^\dagger \mathbf{Z}\|_1 = \mathbf{w}_{\mathrm{opt}}^\dagger \mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w}_{\mathrm{opt}}) = \mathbf{w}_{\mathrm{opt}}^\dagger \mathbf{Z}\mathbf{c}_{\mathrm{opt}} = \|\mathbf{Z}\mathbf{c}_{\mathrm{opt}}\|_2$. Sources fulfilling condition C2 minimize the absolute value criterion, as shown by Theorem 5. The sample equivalent of this result thus establishes that

$$\|\mathbf{w}_{\mathrm{opt}}^\dagger \mathbf{Z}\|_1 \leq \|\mathbf{w}^\dagger \mathbf{Z}\|_1 \qquad \forall \mathbf{w} \;:\; \|\mathbf{w}\|_2 = 1. \tag{38}$$

However, $\|\mathbf{w}^\dagger \mathbf{Z}\|_1 = \mathbf{w}^\dagger \mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w}) = \|\mathbf{w}\|_2 \|\mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w})\|_2 \cos\phi \leq \|\mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w})\|_2$, where ϕ is the angle between vectors $\mathbf{w}$ and $\mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w})$ and we have exploited the fact that $\|\mathbf{w}\|_2 = 1$. Combining these equations, we can write

$$\|\mathbf{Z}\mathbf{c}_{\mathrm{opt}}\|_2 \leq \|\mathbf{Z}\,\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w})\|_2 \qquad \forall \mathbf{w} \;:\; \|\mathbf{w}\|_2 = 1$$

which completes the proof because $\mathrm{sign}(\mathbf{Z}^\dagger \mathbf{w}) \in \mathcal{S}_1$, as defined in eq. (6).

ACKNOWLEDGMENTS

This work is partially funded by the French National Research Agency under contract ANR-2010-JCJC-0303-01 "PERSIST" and the Spanish Ministry of Economy and Competitiveness under project TEC2014-53103-P "CMANS". The authors would like to thank the anonymous reviewers for their interesting comments and useful suggestions, which helped to improve the original version of this manuscript.

REFERENCES

[1] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner, "A polynomial case of unconstrained zero-one quadratic optimization," *Math. Program.*, vol. 91, no. 1, pp. 49–52, Oct. 2001.
[2] 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.
[3] J.P. Brooks, J.H. Dulá, and E.L. Bone, "A pure L_1-norm principal component analysis," *J. Comput. Stat. Data Anal.*, vol. 61, pp. 83–98, May 2013.
[4] E.K.P. Chong and S. H. Zak, *An Introduction to Optimization*. Hoboken, NJ, USA: Wiley, 1996.
[5] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
[6] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Cambridge, MA, USA: Academic Press, 2010.
[7] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1957.
[8] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.
[9] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization," *Proc. Int. Conf. Mach. Learning* 2006, pp. 281–288.
[10] 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.
[11] J.H. Friedman, "Exploratory projection pursuit," *J. Am. Statist. Assoc.*, vol. 82, pp. 249–266, 1987.
[12] A. Hyvärinen and E. Oja, "Independent component analysis by general nonlinear Hebbian-like learning rules," *Signal Process.*, vol. 64, no. 3, pp. 301–313, 1998.
[13] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[14] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.
[15] G. N. Karystinos and A. P. Liavas, "Efficient computation of the binary vector that maximizes a rank-deficient quadratic form," *IEEE Trans. Inform. Theory*, vol. 56, pp. 3581–3593, Jul. 2010.
[16] S. Kundu, P. P. Markopoulos, and D. A. Pados, "Fast computation of the L1-principal component of real-valued data," *Proc. 39th IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 8028–8032.
[17] 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.
[18] N. Kwak and J. Oh, "Feature extraction for one-class classification problems: Enhancements to biased discriminant analysis," *Pattern Recog.*, vol. 42, pp. 17–26, Jan. 2009.
[19] K.-K. Leung, C. W. Sung, M. Khabbazian, and M. A. Safari, "Optimal phase control for equal-gain transmission in MIMO systems with scalar quantization: Complexity and algorithms," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3343–3355, Jul. 2010.
[20] X. Li, Y. Pang, and Y. Yuan, "L1-norm based 2DPCA," *IEEE Trans. Syst., Man Cybern.*, vol. 40, no. 4, pp. 1170–1175, Aug. 2009.
[21] M. McCoy and J. A. Tropp, "Two proposals for robust PCA using semidefinite programming," *Electron. J. Stat.*, vol. 5, pp. 1123–1160, Jun. 2011.
[22] 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.
[23] D. Meng, Q. Zhao, and Z. Xu, "Improved robustness of sparse PCA by L1-norm maximization," *Pattern Recog.*, vol. 45, pp. 487–497, Jan. 2012.
[24] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ, USA: Prentice Hall, 2003.
[25] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with L1-norm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 172–178, Feb. 2010.
[26] J. K. Tugnait, "Identification and deconvolution of multichannel linear non-Gaussian processes using higher order statistics and inverse filter criteria," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 658–672, Mar. 1997.

[27] P.C. Xu, Y.H. Shen, H. Li, J. Wang, and K. Wu, "Independent component analysis of complex valued signals based on first-order statistics," *Radioeng.*, vol. 22, no. 4, pp. 1195–1201, Dec. 2013.
[28] L. Yu, M. Zhang, and C. Ding, "An efficient algorithm for L1-norm principal component analysis,"*Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2012, pp. 1377–1380.
[29] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.
[30] H. Wang, "Block principal component analysis with L1-norm for image analysis," *Pattern Recog. Lett.*, vol. 33, pp. 537–542, Apr. 2012.
[31] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 248–261, Feb. 2010.
[32] V. Zarzoso, R. Martín-Clemente, and S. Hornillo-Mellado, "Independent component analysis based on first-order statistics," *Signal Process.*, vol. 92, pp. 1779–1784, Aug. 2012.

Rubén Martín-Clemente received the M Eng degree in telecommunications engineering and the PhD degree with highest distinction in telecommunications engineering from the University of Sevilla, Sevilla, Spain, in 1996 and 2000, respectively. He is currently an associate professor with the Department of Signal Theory and Communications, University of Sevilla, Spain. He has been a visiting researcher at the University of Regensburg, Regensburg, Germany, in 2001 and 2009. Among other areas, his research interests include multivariate data analysis with emphasis on independent component analysis and its application to biomedical problems. He has authored or co-authored numerous publications on these topics. He has served as 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. He is a member of the IEEE.

Vicente Zarzoso received the graduate degree with highest distinction in telecommunications engineering from the Polytechnic University of Valencia, Valencia, Spain, in 1996. He began his PhD studies at the University of Strachclyde, Glasgow, United Kingdom. In 1999, he received the PhD degree from the University of Liverpool, Liverpool, United Kingdom. He obtained the Habilitation to Lead Researches (HDR) from the University of Nice Sophia Antipolis, France, in 2009. From 2000 to 2005, he held a research fellowship awarded by the Royal Academy of Engineering, United Kingdom. Since 2005, he has been with the Computer Science, Signals and Systems Laboratory of Sophia Antipolis (I3S), University of Nice Sophia Antipolis, France, where he is a full professor and is currently responsible for the "Signal, Image and Systems" (SIS) research team. His research interests include statistical signal and array processing, with emphasis on multidimensional data analysis and signal separation, and its application to biomedical problems and digital communications. He served as an associate editor of 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 a member of the *Institut Universitaire de France*.