Non-linear ICA by Using Isometric Dimensionality Reduction

John A. Lee[1], Christian Jutten[2], and Michel Verleysen[1]

[1] Université catholique de Louvain (DICE)
Place de Levant, 3, B-1348 Louvain-la-Neuve, Belgium
{lee,verleysen}@dice.ucl.ac.be
[2] Institut National Polytechnique de Grenoble (LIS)
Avenue Félix Viallet, 46, 38031 Grenoble Cedex, France
Christian.Jutten@inpg.fr

Abstract. In usual ICA methods, sources are typically estimated by maximizing a measure of their statistical independence. This paper explains how to perform non-linear ICA by preprocessing the mixtures with recent non-linear dimensionality reduction techniques. These techniques are intended to produce a low-dimensional representation of the data (the mixtures), which is isometric to their initial high-dimensional distribution. A detailed study of the mixture model that makes the separation possible precedes a practical example.

1   Introduction

Independent Component Analysis [2, 3] (ICA) aims at recovering a vector of unknown latent variables $\mathbf{x}$ starting from a vector of observed variables $\mathbf{y}$. Usually, the variables in $\mathbf{y}$ are assumed to be (noiseless) linear mixtures of $\mathbf{x}$, according to the generative model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \ , \tag{1}$$

where $\mathbf{A}$ is a full-rank $D \times P$ 'mixing' matrix, with $D \geq P$. In order to retrieve $\mathbf{x}$, the ICA model also assumes that all components of $\mathbf{x}$ have zero mean and are statistically independent from each other. Therefore, the goal of ICA is to identify $\mathbf{A}$ by determining the 'separating' matrix $\mathbf{B}$ in the reversed model

$$\mathbf{x} \approx \hat{\mathbf{x}} = \mathbf{B}\mathbf{y} \ . \tag{2}$$

Practically, ICA proceeds by defining a measure of independence E_{ICA} on $\hat{\mathbf{x}}$ and by maximizing it:

$$\hat{\mathbf{x}} = \arg\max E_{\text{ICA}}(\mathbf{B}\mathbf{y}) \ . \tag{3}$$

Thanks to the independence hypothesis on $\mathbf{x}$ and because the model is linear, it can be proved that E_{ICA} reaches its maximum when $\mathbf{B}\mathbf{A} = \mathbf{\Delta}\mathbf{\Pi}$ where $\mathbf{\Delta}$ and $\mathbf{\Pi}$ are respectively a diagonal and a permutation matrices [2, 3].

When providing ICA with a non-linear model, the same statement should be true too. Unfortunately, it is not difficult to show that non-linear transformations

of any set of variables allows building an infinity of variables that are independent from each other. Hence the maximization of E_{ICA} does not lead to the desired solution anymore. As a matter of fact, this has considerably slown down the investigation of non-linear ICA.

However, although ICA proves incompatible with a non-linear model in its full generality [6], it has been shown by several authors that non-linear ICA is still feasible in some specific cases. In particular, some work [8,6] has been devoted to so-called post-non-linear mixtures (PNL), defined as follows:

$$\mathbf{y} = \mathbf{f}(\mathbf{Ax}) \ , \tag{4}$$

where $\mathbf{f}$ is a vector of D invertible and differentiable functions from $\mathbb{R}$ to $\mathbb{R}$. Under mild conditions on $\mathbf{A}$, the latent variables $\mathbf{x}$ can be retrieved using the same principle as in linear ICA. Indeed, by maximizing the independence of $\hat{\mathbf{x}} = \mathbf{g}(\mathbf{By})$, it is possible to identify $\mathbf{f}$ (as the inverse of $\mathbf{g}$) and $\mathbf{A}$ ($\mathbf{AB} = \mathbf{\Delta\Pi}$).

This paper explores another way to perform non-linear ICA. In PNL mixtures, the inversion of the non-linear functions is achieved by maximizing the independence. A slightly different and more complex model is proposed, consisting of two parts — a linear one and a non-linear one —, as in PNL mixtures. The main difference holds in the fact that the non-linear part of the model is identified by optimizing a criterion that does not relate to statistical independence. Actually, the non-linear part is inverted by computing an isometric transformation of the available data. Section 2 explains how isometric transformations can be integrated in an ICA model in a very natural way. In particular, Subsection 2.1 describes the particular metric that is used in the isometry. Section 3 gives some experimental results and Section 4 comments them. Finally, Section 5 draws the conclusions.

2   Mixture Model

The following generative model is considered:

$$\mathbf{y} = \mathbf{f}(\mathbf{z}) = \mathbf{f}(\mathbf{Ax}) \ , \tag{5}$$

where $\mathbf{f}$ is a smooth ($\mathcal{C}^\infty$) function from $\mathbb{R}^P$ to $\mathbb{R}^D$, $\mathbf{A}$ is a square $P \times P$ matrix and the vector $\mathbf{y}$ is assumed to be isometric to the vector $\mathbf{z}$. By 'isometric' it is meant that the distance measured between two realizations of $\mathbf{y}$ equals the distance measured between the corresponding realizations of $\mathbf{z}$.

Of course, if the Euclidean distance is used for both $\mathbf{y}$ and $\mathbf{z}$, then the mixing function $\mathbf{f}$ can only be a rotation matrix. More precisely, $\mathbf{f}(\mathbf{z}) = \mathbf{Qz}$, where $\mathbf{Q}$ is a $D \times P$ matrix, resulting from the concatenation of P unit-norm vectors. This obviously raises little interest. Fortunately enough, the use of the Euclidean distance proves not mandatory at all. A couple of recent works [7,4,5] suggest using different metrics to measures distances on $\mathbf{y}$ and $\mathbf{z}$. In particular, the use of the so-called geodesic distance for $\mathbf{y}$ is advised, while keeping the Euclidean distance for $\mathbf{z}$.

2.1   Geodesic Distances

Geodesic distances are used in the fields of manifold learning and non-linear dimensionality reduction (NLDR) by distance preservation [7, 4, 5]. Given a P-dimensional smooth ($\mathcal{C}^\infty$) manifold $\mathcal{M}$ in a D-dimensional space, the geodesic distance between two points $\mathbf{y}_i$ and $\mathbf{y}_j$ of the manifold is measured along the manifold, unlike the Euclidean distance, which is measured along the line segment connecting the two points. Actually, the geodesic distance $\delta(\mathbf{y}_i, \mathbf{y}_j)$ is computed as the minimum arc length between the two points, an arc $\boldsymbol{\gamma}$ being a smooth one-dimensional submanifold. Hence,

$$\delta(\mathbf{y}_i, \mathbf{y}_j) = \min_{\gamma(\zeta)} \int_{\zeta_i}^{\zeta_j} \|\mathbf{J}_\zeta \mathbf{f}(\boldsymbol{\gamma}(\zeta))\|_2 d\zeta \ , \qquad (6)$$

where $\mathbf{y}_i = \boldsymbol{\gamma}(\zeta_i)$ and $\mathbf{y}_j = \boldsymbol{\gamma}(\zeta_i)$ are in $\mathcal{M}$, the function $\mathbf{f}(\mathbf{z})$ designates the parametric equations of $\mathcal{M}$ and $\mathbf{J}_\zeta$ is the Jacobian matrix with respect to ζ. It is easy to see that geodesic distances are equivalent to Euclidean ones if the manifold is linear (planar).

Figure 1 illustrates the purpose of geodesic distances in NLDR: such a metric allows measuring distances that are (almost) independent of the manifold embedding. Contrarily to Euclidean distances, geodesic ones do not change if the 'C'-shaped manifold in Fig. 1 is unrolled or unfolded. This shows how NDLR by distance preservation works: a low-dimensional embedding of the manifold is computed as the result of a (nearly) isometric transformation from $\mathbb{R}^D$ to $\mathbb{R}^P$.
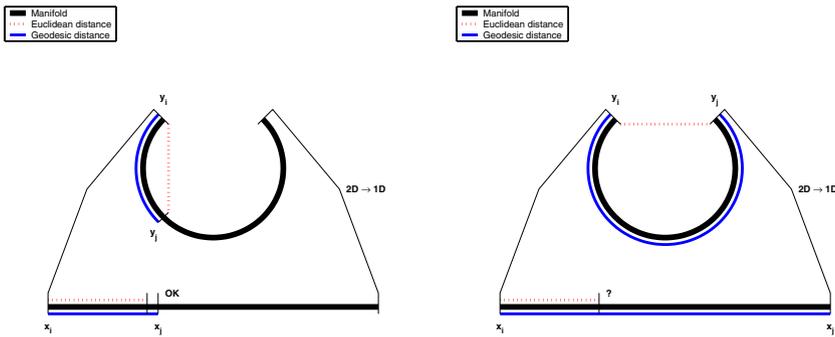
Fig. 1. Geodesic distances for dimensionality reduction in the case of a 'c'-shaped curve: for short (left) as well as long (right) distances, the geodesic distances makes possible the isometry between the manifold and its low-dimensional embedding.

In practice, computing geodesic distances from a finite-size sample $\mathbf{Y} = [\ldots, \mathbf{y}_i, \ldots, \mathbf{y}_j, \ldots]_{1 \leq i,j \leq N}$ is difficult. Fortunately, geodesic distances can be approximated by so-called graph distances, as illustrated in Fig. 2. The quality of that approximation is assessed in [1] (theoretical point of view) and [4, 5] (practical issues).
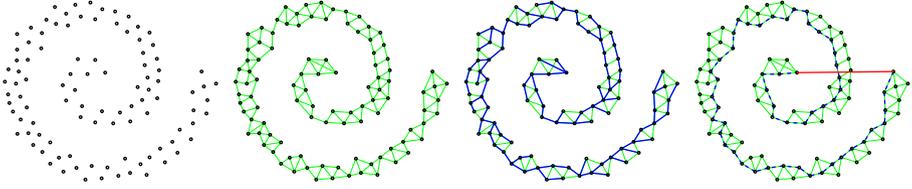
Fig. 2. Procedure to compute graph distances: (1st plot) a few manifold points are available, (2nd plot) each point becomes a graph vertex and is connected with its closest neighbors in order to obtain a graph, (3rd plot) after labeling the graph edges with their length, Dijkstra's algorithm is run on the graph, with the central point of the spiral as source vertex, (4th plot) the Euclidean and graph distances between the same two points.

2.2 Isometry

Back to the generative model in Eq. 5, it may be assumed as in the previous section that the vector $\mathbf{z}$ and the function $\mathbf{f}$ respectively contain the parameters of a manifold and its parametric equations. The hypothesis of isometry amounts to state that $\delta(\mathbf{y}_i, \mathbf{y}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$ for any corresponding pairs of realizations of $\mathbf{y}$ and $\mathbf{z}$. A manifold which satisfies that hypothesis is said to be Euclidean and has nice properties. For example, the minimization involved in the computation of the geodesic distance $\delta(\mathbf{y}_i, \mathbf{y}_j)$ may be dropped in Eq. 6. Indeed, by virtue of the isometry, it comes that

$$
\begin{aligned}
\delta(\mathbf{y}_i, \mathbf{y}_j) &= \|\mathbf{z}_i - \mathbf{z}_j\|_2 \\
&= \|\mathbf{z}_i - (\mathbf{z}_i + \alpha(\mathbf{z}_j - \mathbf{z}_i))\|_2 + \|(\mathbf{z}_i + \alpha(\mathbf{z}_j - \mathbf{z}_i)) - \mathbf{z}_j\|_2 \\
&= \delta(\mathbf{f}(\mathbf{z}_i), \mathbf{f}(\mathbf{z}_i + \alpha(\mathbf{z}_j - \mathbf{z}_i))) + \delta(\mathbf{f}(\mathbf{z}_i + \alpha(\mathbf{z}_j - \mathbf{z}_i)), \mathbf{f}(\mathbf{z}_j)) \ , \quad (7)
\end{aligned}
$$

where α is a real number between 0 and 1. These equalities simply demonstrate that the shortest geodesic arc between $\mathbf{y}_i$ and $\mathbf{y}_j$ is the image by $\mathbf{f}$ of the line segment going from $\mathbf{z}_i$ to $\mathbf{z}_j$: all points $\mathbf{f}(\mathbf{z}_i + \alpha(\mathbf{z}_j - \mathbf{z}_i))$ on that segment must also lie on the shortest path. Therefore, in the case of a Euclidean manifold, the arc $\boldsymbol{\gamma}(\zeta)$ in Eq. 6 can be written as

$$
\boldsymbol{\zeta} : [0,1] \subset \mathbb{R} \to \mathbb{R}^P, \zeta \mapsto \mathbf{z} = \boldsymbol{\gamma}(\zeta) = \mathbf{z}_i + \zeta(\mathbf{z}_j - \mathbf{z}_i) \tag{8}
$$

and the minimization in Eq. 6 becomes useless. Using the last result and knowing that geodesic distances equal Euclidean ones in a vector space, it comes that

$$
\|\mathbf{z}_i - \mathbf{z}_j\|_2 = \delta(\mathbf{z}_i, \mathbf{z}_j) = \int_0^1 \|\mathbf{J}_\zeta \boldsymbol{\gamma}(\zeta)\|_2 d\zeta \quad \text{and} \tag{9}
$$

$$
\delta(\mathbf{y}_i, \mathbf{y}_j) = \int_0^1 \|\mathbf{J}_\zeta \mathbf{f}(\boldsymbol{\gamma}(\zeta))\|_2 d\zeta = \int_0^1 \|\mathbf{J}_{\boldsymbol{\gamma}(z)} \mathbf{f}(\boldsymbol{\gamma}(\zeta)) \, \mathbf{J}_\zeta \boldsymbol{\gamma}(\zeta)\|_2 d\zeta \ . \tag{10}
$$

As $\|\mathbf{z}_i - \mathbf{z}_j\|_2 = \delta(\mathbf{y}_i, \mathbf{y}_j)$, the equality $\|\mathbf{J}_\zeta \boldsymbol{\gamma}(\zeta)\|_2 = \|\mathbf{J}_{\boldsymbol{\gamma}(\zeta)} \mathbf{f}(\boldsymbol{\gamma}(\zeta)) \, \mathbf{J}_\zeta \boldsymbol{\gamma}(\zeta)\|_2$ must hold. This means that the Jacobian of a Euclidean manifold must be a D-by-P matrix whose columns are orthogonal vectors with unit norm. This leaves

the norm of $\mathbf{J}_\zeta\boldsymbol{\gamma}(\zeta)$ unchanged after left multiplication by $\mathbf{J}_{\boldsymbol{\gamma}(z)}\mathbf{f}(\boldsymbol{\gamma}(\zeta))$. More precisely, the Jacobian matrix can be written in a generic way as $\mathbf{J}_\mathbf{z}\mathbf{f}(\mathbf{z}) = \mathbf{Q}\mathbf{V}(\mathbf{z})$, where $\mathbf{Q}$ is a constant orthonormal matrix (a rotation matrix in the D-dimensional space) and $\mathbf{V}(\mathbf{z})$ a D-by-P matrix with unit-norm columns and only one non-zero entry per row. The last requirement ensures that the columns of $\mathbf{V}(\mathbf{z})$ are always orthogonal, independently from the value of $\mathbf{z}$.

Because of the particular form of its Jacobian matrix, a Euclidean P-manifold embedded in a D-dimensional space can always be written with the following 'canonical' parametric equations:

$$\mathbf{y} = \mathbf{Q}\,\mathbf{f}(\mathbf{z}) = \mathbf{Q}\,[f_1(z_{1\le p\le P}),\ldots,f_D(z_{1\le p\le P})]^T \ , \tag{11}$$

where $\mathbf{Q}$ is the same as above, $\mathbf{J}_\mathbf{z}\mathbf{f}(\mathbf{z}) = \mathbf{V}(\mathbf{z})$ and $f_1,\ldots,f_D$ are constant, linear or non-linear continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Hence, if $\mathbf{Q}$ is omitted, the parametric equation of each coordinate in the D-dimensional space of a Euclidean manifold depends on at most a single latent variable z_p.

Visually, in a three-dimensional space, a manifold is Euclidean if it looks like a curved sheet of paper.

2.3 Isometric Dimensionality Reduction

Using the above-mentioned ideas, it may be stated that if a manifold is Euclidean, its latent variables can be retrieved. More formally, knowing a sufficiently large set $\mathbf{Y} = [\ldots,\mathbf{y}_i,\ldots,\mathbf{y}_j,\ldots]_{1\le i,j\le N}$ of points drawn from a Euclidean P-dimensional manifold, it is possible to determine the corresponding values of the latent variables, up to a translation and a rotation, by finding an isometric P-dimensional representation $\mathbf{Z}$ of $\mathbf{Y}$.

From a practical point of view, an estimation $\hat{\mathbf{Z}}$ of $\mathbf{Z}$ can be computed using NLDR methods [7, 4, 5] that work by distance preservation. These methods precisely attempt to find a low-dimensional representation of high-dimensional points that is 'as isometric as possible'. If these methods use geodesic distances in the D-dimensional space and Euclidean distances in the P-dimensional space and if the manifold is Euclidean, then a prefect isometry is possible. This means that starting from an observation $\mathbf{y}_i$ the corresponding value $\mathbf{z}_i$ of the latent variables can be recovered or, in other words, that the function $\mathbf{y} = \mathbf{f}(\mathbf{z})$ can be perfectly inverted, up to the above-mentioned undeterminacies (translation and rotation). Within the framework of ICA, this also means that the non-linear part of the generative model in Eq. 5 can be inverted to find $\hat{\mathbf{z}} \approx \mathbf{z} = \mathbf{A}\mathbf{x}$; next, $\hat{\mathbf{x}}$ can be recovered by using a classical linear ICA method.

3 Experimental Results

In order to illustrate how isometric ICA works, the two following sources are proposed:

$$\mathbf{x} = \begin{bmatrix} \arccos(\cos(0.034\pi t)) \\ \sin(0.006\pi t) \end{bmatrix} \ . \tag{12}$$

Thousand observations of $\mathbf{x}$ are available in the interval $[0.000 \le t \le 0.999]$ and are shown on the left of Fig. 3. Sources are artificially mixed as follows:

$$\mathbf{y} = 0.5 \begin{bmatrix} +\sqrt{2} & +1 & +1 \\ -\sqrt{2} & +1 & +1 \\ 0 & -\sqrt{2} & +\sqrt{2} \end{bmatrix} \begin{bmatrix} \cos(\pi z_1) \\ \sin(\pi z_1) \\ \pi z_2 \end{bmatrix} \quad , \text{ where } \quad \mathbf{z} = \begin{bmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{bmatrix} \mathbf{x} \ . \quad (13)$$

It is easy to see that the above-stated conditions to apply isometric ICA are fulfilled. The mixtures are shown on the right of Fig. 3.
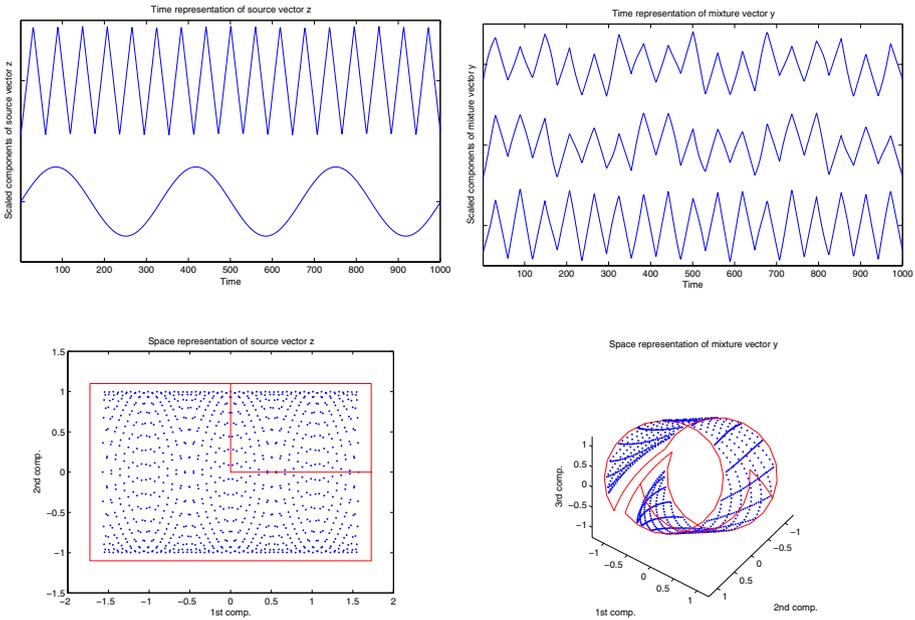
Fig. 3. Sources (left) and mixtures (right); time (top) and space (bottom) representations are given.

Starting from the mixtures, a first attempt to separate the sources consists in running a classical (linear) ICA method. For example, FastICA (deflation, tanh non-linear function) yields the result shown on the left of Fig. 4. During the whitening step, PCA indicates that *three* components are needed to explain 95% of the variance. Linear ICA succeeds rather well in recovering the serrated source (already clearly visible in the mixtures) but fails in the case of the sine.

Isometric ICA yields the result shown on the right of Fig. 4. To obtain that result, an isometric representation of the thousand available observations is computed with the method described in [4, 5]. This dimensionality reduction method works by gradient descent, contrarily to other methods which are purely algebraical [7]. The method indicates that an almost isometric two-dimensional rep-
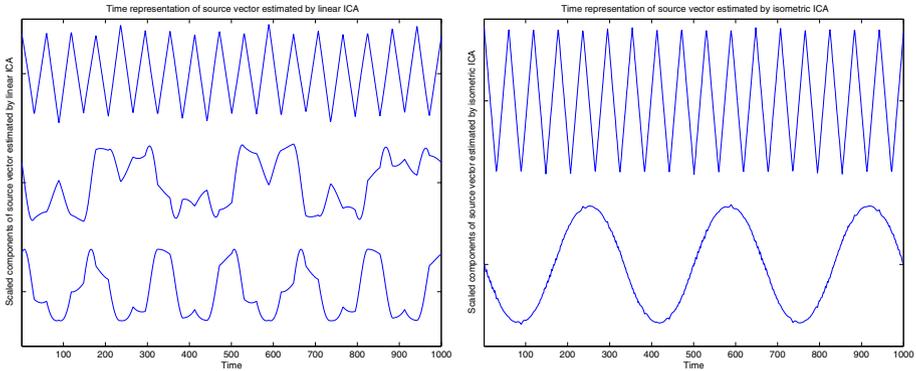
Fig. 4. Results computed by FastICA (left) and isometric ICA (right).

resentation of the observations is possible and computes it. Next, FastICA is run on the two remaining linear mixtures, leading to the result on the right of Fig. 4.

4   Discussion

Because of space constraints, other experiments cannot be included in this paper. However, here are some comments about the advantages and shortcomings of isometric ICA.

The main drawback of isometric ICA holds in the very restrictive conditions that must be satisfied to apply it. Exactly as for post-non-linear mixtures, the non-linear functions involved in the mixture process must be one-to-one. Moreover, all non-linear functions depending on the same component of $\mathbf{z}$ must be 'coupled', otherwise the norm of the corresponding column of the Jacobian matrix cannot be constant. Contrarily to PNL mixtures, isometric mixtures may be further multiplied by any rotation matrix $\mathbf{Q}$.

In practice, it has been shown experimentally that the isometry does not need to be absolutely perfect. Actually, the norms of the Jacobian columns may vary a little and slightly differ from each other. Similarly, the matrix $\mathbf{Q}$ does not need to be perfectly unitary. Even in those cases, the non-linear part of the model can be more or less well inverted, and better results are obtained than when using a simple linear ICA method.

Other practical issues of isometric ICA regard the quality of the isometric representation computed from the observations. Even if all above-mentioned conditions are satisfied, it must be ensured that the methods described in [7, 4, 5] work correctly. It has been shown in [4, 5] that the good approximation of the geodesic distances by the graph distances is very important. If the number of observations is low, if their distribution is very sparse in some regions, or simply if some parameters values are wrong, then the approximation becomes very rough. This may jeopardize the computation of the isometric representation and therefore the inversion of the non-linear function $\mathbf{f}$. As a direct consequence, the subsequent ICA step does not run in an optimal setting.

5   Conclusion

Although true non-linear ICA is impossible, several constrained models have been successfully proposed, especially post-non-linear mixtures. In this paper, a different model is proposed, in which the inversion of the non-linear part is based on geometrical considerations. More precisely, the proposed model assumes that variables at the input and output of its non-linear part are isometric, i.e. distances measured between two corresponding pairs of observations are equal. Because the isometry involves other distances than Euclidean ones, the associated transformation may be non-linear. In the case of the geodesic distance, which has become popular in the field of dimensionality reduction, conditions that makes a perfect isometry possible are studied in details. If those conditions are fulfilled, the non-linear part of the mixture model can be fully inverted and a linear ICA method may be run afterwards. Even if those conditions are rather restrictive, isometric ICA can tackle problems that other linear or non-linear ICA methods cannot solve. A simple example illustrates this fact.

Future work aims at comparing isometric ICA to post-non-linear ICA, from different points of view (computational costs, robustness, etc.). A further study of how isometric ICA behaves when the conditions of its model are not perfectly met is also planned.

References

1. M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, Stanford, December 2000.
2. P. Comon. Independent Component Analysis – A new concept? *Signal Processing*, 36:287–314, 1994.
3. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.
4. J. A. Lee, A. Lendasse, and M. Verleysen. Curvilinear Distances Analysis versus Isomap. In M. Verleysen, editor, *Proceedings of ESANN'2002*, pages 13–20. D-Facto public., Bruges (Belgium), 2002.
5. J. A. Lee and M. Verleysen. Curvilinear distance analysis versus isomap. *Neurocomputing*, 2004. Accepted.
6. A. Taleb and C. Jutten. Source separation in postnonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.
7. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.
8. H. H. Yang, Amari S., and A. Cichocki. Information-theoretic approach to blind separation of sources in non-linear mixtures. *Signal Processing*, 64(3):291–300, 1998.