LETTER Communicated by Erkki Oja

Linear Geometric ICA: Fundamentals and Algorithms

Fabian J. Theis
fabian.theis@mathematik.uni-regensburg.de
Institute of Biophysics, University of Regensburg, Germany

Andreas Jung
andreas.jung@physik.uni-regensburg.de
Institute for Theoretical Physics, University of Regensburg, Germany

Carlos G. Puntonet
carlos@atc.ugr.es
Elmar W. Lang
elmar.lang@biologie,uni-regensburg.de
Department of Architecture and Computer Technology, University of Granada, Spain

Geometric algorithms for linear independent component analysis (ICA) have recently received some attention due to their pictorial description and their relative ease of implementation. The geometric approach to ICA was proposed first by Puntonet and Prieto (1995). We will reconsider geometric ICA in a theoretic framework showing that fixed points of geometric ICA fulfill a geometric convergence condition (GCC), which the mixed images of the unit vectors satisfy too. This leads to a conjecture claiming that in the nongaussian unimodal symmetric case, there is only one stable fixed point, implying the uniqueness of geometric ICA after convergence. Guided by the principles of ordinary geometric ICA, we then present a new approach to linear geometric ICA based on histograms observing a considerable improvement in separation quality of different distributions and a sizable reduction in computational cost, by a factor of 100, compared to the ordinary geometric approach. Furthermore, we explore the accuracy of the algorithm depending on the number of samples and the choice of the mixing matrix, and compare geometric algorithms with classical ICA algorithms, namely, Extended Infomax and FastICA. Finally, we discuss the problem of high-dimensional data sets within the realm of geometrical ICA algorithms.

1 Introduction

Given a random vector, independent component analysis (ICA) tries to find its statistically independent components. This idea can also be used to solve the blind source separation (BSS) problem, which is, given only

the mixtures of some underlying independent source signals, to separate the mixed signals—henceforth called sensor signals—thus recovering the original sources. In contrast to correlation-based transformations such as principal component analysis (PCA), ICA renders the output signals as statistically independent as possible by evaluating higher-order statistics. The idea of ICA was first expressed by Jutten, Hérault, Comon, and Sorouchiary (1991), while the term *ICA* was later coined by Comon (1994). However, the field became popular only with the seminal paper by Bell and Sejnowski (1995), who elaborated on the Infomax principle first advocated by Linsker (1989, 1992). Many other ICA algorithms have been proposed, with the FastICA algorithm (Hyvärinen, 1999) being the most efficient among them.

Recently, geometric ICA algorithms have received further attention due to their relative ease of implementation (Puntonet & Prieto, 1995). They have been applied successfully to the analysis of real-world biomedical data (Bauer, Puntonet, Rodriguez-Alvarez, & Lang, 2000) and have been extended as well to nonlinear ICA problems (Puntonet, Alvarez, Prieto, & Prieto, 1999).

2 Basics

In linear BSS, a random vector $X\colon \Omega \to \mathbb{R}^n$ called sensor signal is given; it originates from an unknown independent random vector $S\colon \Omega \to \mathbb{R}^n$, which will be denoted as source signal, via a mixing process with an unknown mixing matrix $A \in \mathrm{Gl}(n)$, that is, $X = A \circ S$. Note that we assume as many sensors as sources. Here, Ω denotes a fixed probability space and $\mathrm{Gl}(n) := \{W \in \mathrm{Mat}(n \times n; \mathbb{R}) \mid \det(W) \neq 0\}$ the general linear group in $\mathbb{R}^n$. Only the sensor signal is observable, and the task is to recover A and thereby $S = A^{-1} \circ X$.

Uniqueness of the solutions can be acertained if we allow at most one of the source variables $S_i := \pi_i \circ S$, where $\pi_i\colon \mathbb{R}^n \to \mathbb{R}$ denotes the projection on the ith coordinate, to be gaussian. Then *any* solution to the BSS problem, that is, any $B \in \mathrm{Gl}(n)$ such that $B \circ X$ is independent, is equivalent to A^{-1}, where equivalent means that B can be written as $B = LPA^{-1}$ with an invertible diagonal matrix (scaling matrix) $L \in \mathrm{Gl}(n)$ and an invertible matrix with unit vectors in each row (permutation matrix) $P \in \mathrm{Gl}(n)$ (Comon, 1994). Vice versa, any matrix B that is equivalent to A^{-1} solves the BSS problem, since the transformed mutual information is invariant under scaling and permutation of coordinates.

3 Geometric Considerations

The basic idea of the geometric separation method lies in the fact that in the source space $\{s_1, \ldots, s_\lambda\} \subset \mathbb{R}^n$, where s_i represent a fixed number of samples of the source vector S with zero mean, the data clusters along the axes of the coordinate system are transformed by A into data clusters along

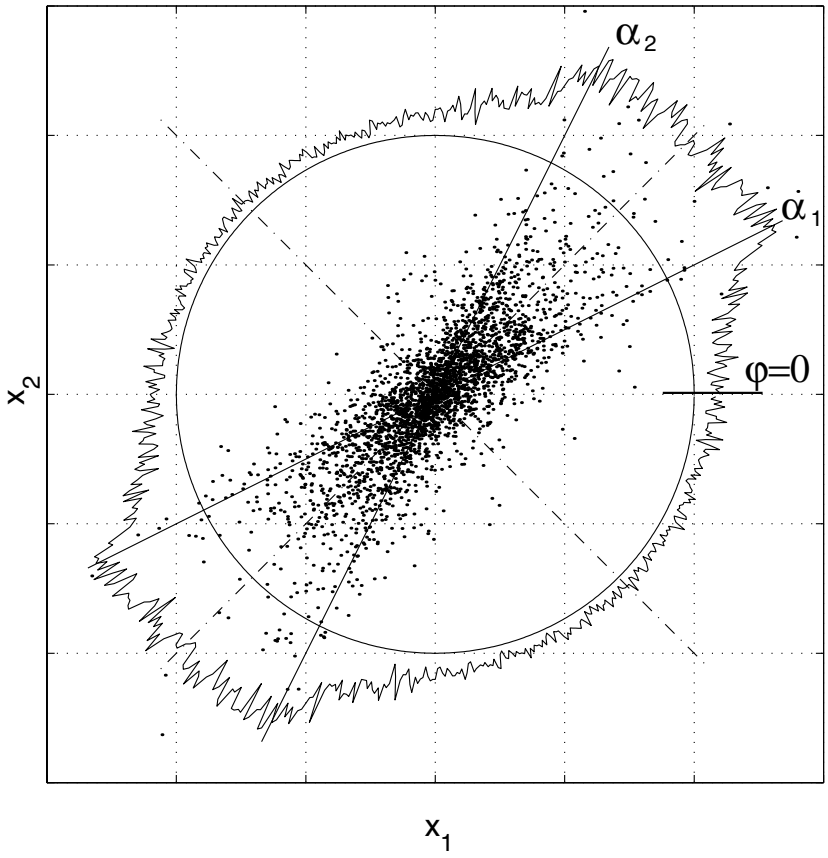

Figure 1: A two-dimensional scatter plot of a mixture of two Laplacian signals with identical variance. The signals have been mixed by a matrix A mapping the unit vectors onto vectors inclined with angle α_i to the x_1-axis. The ragged line along the circle is the histogram of the observations after projection onto the circle plotted in polar coordinates. Dashed lines show borders of the receptive fields—the borders of the circle sections that lie closest to the angles α_i or $-\alpha_i$, respectively.

transformed coordinate axes through the origin. The detection of these n new axes allows determining a demixing matrix B that is equivalent to A^{-1} (see Figure 1).

We now consider the learning process to be terminated and describe precisely how to recover the matrix A then—after the axes, which span the observation space, have been extracted from the data successfully. Let

$$\Lambda := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \exists i \, x_i > 0, x_j = 0 \text{ for all } j \neq i\}$$

be the set of positive coordinate axes, and denote with $\Lambda' := A\Lambda$ the image of this set under the transformation A. Note that due to A being bijective, Λ' intersects the unit $(n-1)$-sphere,

$$S^{n-1} := \{x \in \mathbb{R}^n \mid |x| = 1\},$$

in exactly n distinct points $\{p_1, \dots, p_n\}$ and that those p_i's form a basis of $\mathbb{R}^n$.

Define the matrix $M_{p_1,\dots,p_n} \in \mathrm{Gl}(n)$ to be the linear mapping of e_i onto p_i for $i = 1, \dots, n$, that is,

$$M_{p_1,\dots,p_n} = (p_1|\cdots|p_n).$$

This matrix thus effects the linear coordinate change from the standard coordinates (e_i) to the new basis (p_i). Note that for this coordinate transformation, the following lemma holds:

Lemma 1. *For any permutation $\sigma \in \mathcal{S}_n$, the two matrices $M_{p_1,\dots,p_n}$ and $M_{p_{\sigma(1)},\dots,p_{\sigma(n)}}$ are equivalent.*

Now we can state the following theorem:

Theorem 1 (Uniqueness of the Geometric Method). *The matrix $M_{p_1,\dots,p_n}$ is equivalent to A.*

Proof. By construction of $M_{p_1,\dots,p_n}$, we have $M_{p_1,\dots,p_n}(e_i) = p_i = f(e_i) = \frac{Ae_i}{|Ae_i|}$, so there exists a $\lambda_i \in \mathbb{R}\backslash\{0\}$ such that $M_{p_1,\dots,p_n}(e_i) = \lambda_i A e_i$. Defining L such that $L(e_i) := \lambda_i e_i$ yields an invertible diagonal matrix $L \in \mathrm{Gl}(n)$ such that $M_{p_1,\dots,p_n} = LA$. This shows the claim.

Corollary 1. *The matrix $M^{-1}_{p_1,\dots,p_n}$ solves the BSS problem.*

4 The Ordinary Geometric Algorithm

For now, we will restrict ourselves to the two-dimensional case for simplicity; extensions to higher dimensions are discussed in section 11. Let $S\colon \Omega \longrightarrow \mathbb{R}^2$ be an independent two-dimensional Lebesgue-continuous random vector describing the source pattern distribution; its density function is denoted by $\rho\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$. As S is independent, ρ factorizes $\rho(x, y) = \rho_1(x)\rho_2(y)$, with $\rho_i\colon \mathbb{R} \longrightarrow \mathbb{R}$ denoting the corresponding marginal source density functions. As a further simplification, we will assume the source variables S_i to have zero mean $E(S) = 0$ and to be distributed symmetrically, that is, $\rho_i(x) = \rho_i(-x)$ for $x \in \mathbb{R}$ and $i = 1, \dots, n$. To ensure stability of the geometric algorithm, we have to assume that the signal distributions are nongaussian and unimodal. In practice, these restrictions are often met at least approximately.

As above, let X denote the sensor signal vector and A the invertible mixing matrix such that $X = A \circ S$. Without loss of generality, assume that A is of the form

$$A = \begin{pmatrix} \cos\alpha_1 & \cos\alpha_2 \\ \sin\alpha_1 & \sin\alpha_2 \end{pmatrix},$$

where $\alpha_i \in [0, \pi)$ denote two angles. The ordinary geometric learning algorithm (Puntonet & Prieto, 1995) for symmetric distributions in its simplest form then is as follows:

Pick four starting vectors w_1, w_1', w_2, and w_2' on S^1 such that w_i and w_i' are opposite each other (i.e., $w_i = -w_i'$ for $i = 1, 2$) and w_1 and w_2 are linearly independent vectors in R^2. Usually one takes the unit vectors $w_1 = e_1$ and $w_2 = e_2$. Furthermore, fix a learning rate η: $\mathbb{N} \longrightarrow \mathbb{R}$ with (Cottrell, Fort, & Pagès, 1994) $\eta(t) > 0$, $\sum_{n\in\mathbb{N}} \eta(n) = \infty$ and $\sum_{n\in\mathbb{N}} \eta(n)^2 < \infty$. Then iterate the following step until an appropriate abort condition has been met:

Choose a sample $x(t) \in \mathbb{R}^2$ according to the distribution of X. If $x(t) = 0$, pick a new one. Note that this case happens with probability zero since the probability density function (pdf) ρ_X of X is assumed to be continuous. Project $x(t)$ onto the unit sphere to get $y(t) := \frac{x(t)}{|x(t)|}$. Let i be in $\{1, 2\}$ such that w_i or w_i' is the vector closest to y with respect to a Euclidean metric. Then update $w_i(t)$ according to the following update rule,

$$w_i(t+1) := \mathrm{pr}\left(w_i(t) + \eta(t)\frac{y(t) - w_i(t)}{|y(t) - w_i(t)|}\right),$$

where pr: $\mathbb{R}^2\backslash\{0\} \longrightarrow S^1$ represents the projection onto the unit sphere, and

$$w_i'(t+1) := -w_i(t+1).$$

The other two w's are not moved in this iteration.

In Figures 2 and 3, the learning algorithm has been visualized on the sphere and after the projection onto $[0, \pi)$.

This weight update rule resembles unsupervised competitive learning rules used in many clustering algorithms like k-means, vector quantization, or Kohonen's self-organizing maps, but with the modifications that the step size along the direction of a sample does not depend on distance and the learning process takes place on S^1, not in $\mathbb{R}$.

5 Formal Model of the Geometric Algorithm

Now we present a formal theoretical framework for geometric ICA that will be used in the next section to formulate a proper convergence condition.

First, we show using the symmetry of S that it is in fact not necessary to have two vectors w_i and w_i' moving around on the same axis. Indeed, we should not speak of vectors but of lines in $\mathbb{R}^2$, so the w_i's would be in

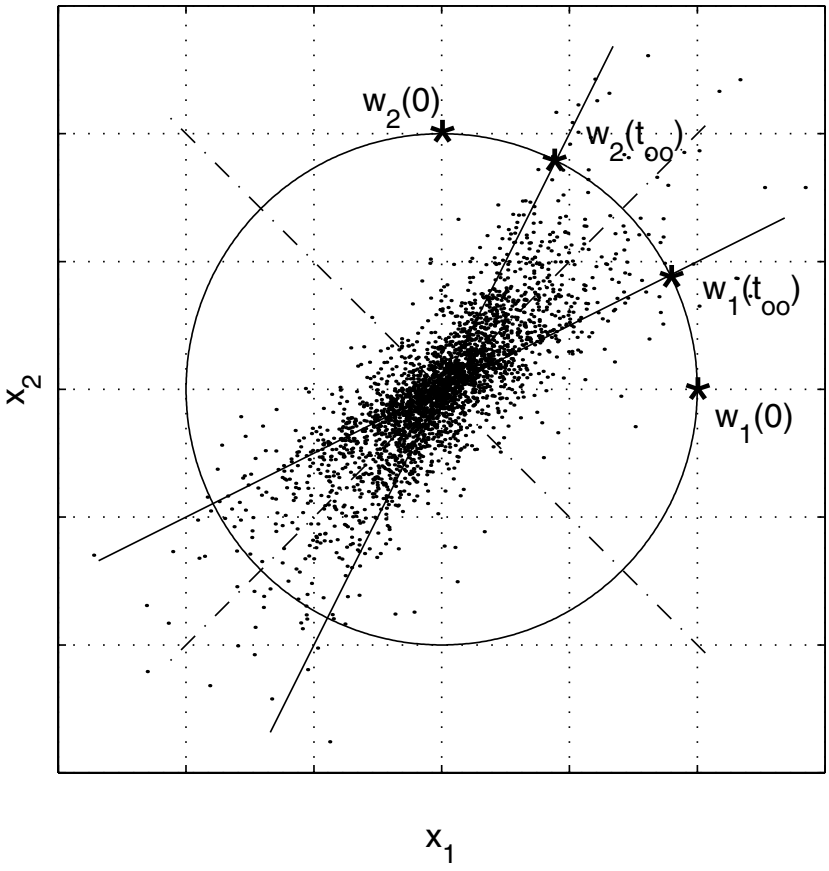

Figure 2: Visualization of the geometric algorithm with starting points $w_1(0)$ and $w_2(0)$ and end points $w_1(\infty)$ and $w_2(\infty)$. Dash-dotted lines mark receptive field borders.

the real projective space $\mathbb{RP}^1 = S^1/\sim$, where $\sim$ identifies antipodal points. This is the manifold of all one-dimensional subvector spaces of $\mathbb{R}^2$. A metric is defined by setting $d([x],[y]) := \min\{|x-y|, |x+y|\}$ for $[x],[y] \in \mathbb{RP}^1$. Alternatively, one can picture the w's in $S^1_+ := S^1 \cap \{(x_1,x_2) \in \mathbb{R}^2 \mid x_2 \geq 0\}/\sim$, where $\sim$ identifies the two points $(1,0)$ and $(-1,0)$. Let $\zeta\colon S^1 \longrightarrow S^1_+$ represent the canonical projection. Furthermore, it is useful to introduce polar coordinates $\varphi\colon S^1_+ \longrightarrow [0,\pi)$ on S^1_+ with the stratification $\varphi'\colon \mathbb{R} \longrightarrow S^1_+$ such that $\varphi' \circ \varphi = \mathrm{id}$, where id denotes the identity. Let $\chi := \varphi \circ \varphi'\colon \mathbb{R} \longrightarrow [0,\pi)$ be the modulo π map. We are interested in the projected random sensor signal vector $\mathrm{pr} \circ X\colon \Omega \longrightarrow S^1$, so after cutting open the circle S^1 and identification of opposite points, we want to approximate the transformed

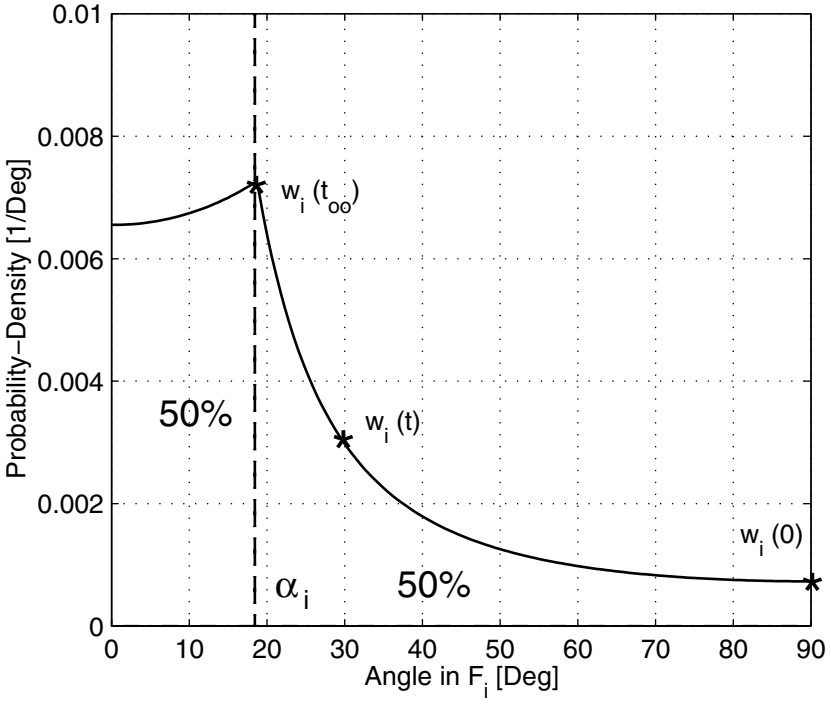

Figure 3: Plot of the density ρ_Y of a mixture of two Laplacian signals with identical variance. The weight adaptation by the geometric algorithm is also visualized. See also Figure 2.

random variable $Y := \varphi \circ \zeta \circ \mathrm{pr} \circ X\colon\ \Omega \longrightarrow [0, \pi)$ in a suitable manner. Note that using the symmetry of ρ, the density function ρ_Y of the transformed sensor signal Y can be calculated from the density ρ_X of the original sensor signal X by

$$\rho_Y(\varphi) = 2|\det A|^{-1} \int_0^\infty \rho(A^{-1}(r\cos\varphi, r\sin\varphi)^\top) r\, dr.$$

Then the geometric learning algorithm induces the following discrete Markov process $W(t)\colon\ \Omega \longrightarrow \mathbb{R}^2$, defined recursively by $W(0) = (w_1, w_2)$ and $W(t+1) = \chi^2(W(t) + \eta(t)\vartheta((Y(t), Y(t)) - W(t)))$, where

$$\vartheta(x, y) := \begin{cases} (\mathrm{sgn}(x), 0) & |y| \geq |x| \\ (0, \mathrm{sgn}(y)) & |x| > |y| \end{cases}$$

and $Y(0), Y(1), \ldots$ is a sequence of independent and identically distributed random variables $\Omega \longrightarrow \mathbb{R}$ with the same distribution as Y. These random

variables will be needed to represent the independence of the successive sampling experiments. Note that the modulo π map χ guarantees that $W(t+1) \in [0, \pi)$. Indeed, this is just winner-takes-all learning with a signum function in $\mathbb{R}$, but taking into account the fact that we have to stay in $[0, \pi)$. Note that the metric used here is the planar metric, which obviously is equivalent to the metric on S^1_+ induced by the Euclidean metric on $S^1 \subset \mathbb{R}^2$.

We furthermore can assume that after enough iterations, there is one point $a \in S^1$ that will not be transversed anymore, and without loss of generality, we assume a to be 0 (otherwise, cut S^1 open at a and project along this resulting arc), so that the above algorithm simplifies to the planar case with the recursion rule $W(t+1) = W(t) + \eta(t)\vartheta((Y(t), Y(t)) - W(t))$.

Without the sign function and the additional fact that the probability distribution of Y is log concave, it has been shown (Cottrell & Fort, 1987; Ritter & Schulten, 1988; Benaim, Fort, & Pagès, 1998) that the process $W(t)$ converges to a unique constant fixed-point process $W \equiv w \in \mathbb{R}^2$ such that

$$\int_{\beta_1(F_i)}^{w_i} \rho_Y(\varphi)\, d\varphi = \int_{w_i}^{\beta_2(F_i)} \rho_Y(\varphi)\, d\varphi$$

for $i = 1, 2$, where $F_i := F(w_i) := \{\varphi \in [0, \pi) \mid \chi(|\varphi - w_i|) \leq \chi(|\varphi - w_j|)$ for all $j \neq i\}$ with $F_i = \chi([\beta_1(F_i), \beta_2(F_i)])$ denotes the receptive field of w_i and $\beta_j(F_i)$ designate the receptive field borders. However, it is not clear how to generalize the proof to the geometric case, especially because we do not have (and do not want) log concavity of Y, as this would lead to a unique fixed point. Therefore, we will assume convergence in a sense stated in the following section.

6 Limit Points of the Geometric Algorithm

In this section, we study the end points of geometric ICA, so we will assume that the algorithm has already converged. The idea is to formulate a condition that the end points will have to satisfy and to show that the solutions are among them.

Definition 1 (Geometric Convergence Condition). *Two angles $l_1, l_2 \in [0, \pi)$ satisfy the geometric convergence condition (GCC) if they are the medians of Y restricted to their receptive fields, that is, if l_i is the median of $\rho_Y \mid F(l_i)$.*

Definition 2. *A constant random vector $\hat{W} \equiv (\hat{w}_1, \hat{w}_2) \in \mathbb{R}^2$ is called fixed point of geometric ICA in the expectation if $E(\vartheta(Y - \hat{W}(t))) = 0$.*

Hence, the expectation of a Markov process $W(t)$ starting at a fixed point of geometric ICA will not be changed by the geometric update rule because

$$E(W(t+1)) = E(W(t)) + \eta(t)E(\vartheta(Y(t) - W(t))) = E(W(t)).$$

Theorem 2. *Given that the geometric algorithm converges to a constant random vector $W(\infty) \equiv (w_1(\infty), w_2(\infty))$, then $W(\infty)$ is a fixed point of geometric ICA in the expectation, if and only if the $w_i(\infty)$ satisfy the GCC.*

Proof. Assume $W(\infty)$ to be a fixed point of geometric ICA in the expectation. Without loss of generality, let $[\beta_1, \beta_2]$ be the receptive field of $w_1(\infty)$ such that $\beta_i \in [0, \pi)$. Since $W(\infty)$ is a fixed point of geometric ICA in the expectation, we have $E(\chi_{[\beta_1,\beta_2]}(Y(t)) \operatorname{sgn}(Y(t) - w_1(\infty))) = 0$, where $\chi_{[\beta_1,\beta_2]}$ denotes the characteristic function of that interval. But this means $\int_{\beta_1}^{w_1(\infty)} (-1)\rho_Y(\varphi)\, d\varphi + \int_{w_1(\infty)}^{\beta_2} 1\rho_Y(\varphi)\, d\varphi = 0$ and therefore $\int_{\beta_1}^{w_1(\infty)} \rho_Y(\varphi)\, d\varphi = \int_{w_1(\infty)}^{\beta_2} \rho_Y(\varphi)\, d\varphi$, so $w_1(\infty)$ satisfies GCC. The same calculation for $w_2(\infty)$ shows one direction of the claim. The other direction follows by simply reading the above proof backward, which completes the proof.

As before, let $p_i := Ae_i$ be the transformed unit vectors, and let $q_i := \varphi \circ \zeta \circ \operatorname{pr}(p_i) \in [0, \pi)$ be the corresponding angles for $i = 1, 2$.

Theorem 3. *The transformed angles q_i satisfy GCC.*

Proof. Because of the symmetry of the claim, it is enough to show that q_1 satisfies GCC. Without loss of generality, let $0 < \alpha_1 < \alpha_2 < \pi$ using the symmetry of ρ. Then, due to construction, $q_i = \alpha_i$. Let $\beta_1 := \frac{\alpha_1+\alpha_2}{2} - \frac{\pi}{2}$ and $\beta_2 := \beta_1 + \frac{\pi}{2}$. Then the receptive field of q_1 can be written (modulo π) as $F(q_1) = [\beta_1, \beta_2]$. Therefore, we have to show that $q_1 = \alpha_1$ is the median of ρ_Y restricted to $F(q_1)$, which means $\int_{\beta_1}^{\alpha_1} \rho_Y(\varphi)\, d\varphi = \int_{\alpha_1}^{\beta_2} \rho_Y(\varphi)\, d\varphi$.

We will reduce this to the orthogonal standard case $A = \mathrm{id}$ by transforming the integral as follows:

$$\begin{aligned}\int_{\beta_1}^{\alpha_1} \rho_Y(\varphi)\, d\varphi &= 2|\det A|^{-1} \int_{\beta_1}^{\alpha_1} d\varphi \int_0^{\infty} r\, dr\, \rho(A^{-1}(r\cos\varphi, r\sin\varphi)^{\top}) \\ &= 2|\det A|^{-1} \int_K dx\, dy\, \rho(A^{-1}(x, y)^{\top}),\end{aligned}$$

where $K := \{(x, y) \in \mathbb{R}^2 \mid \beta_1 \le \arctan(y/x) \le \alpha_1\}$ denotes the cone of opening angle $\alpha_1 - \beta_1$ starting from angle β_1. Using the transformation formula, we continue $\int_{\beta_1}^{\alpha_1} \rho_Y(\varphi)\, d\varphi = 2\int_{A^{-1}(K)} dx\, dy\, \rho(x, y)$. Now note that the transformed cone $A^{-1}(K)$ is a cone ending at the x-axis of opening angle $\pi/4$, because A is linear; therefore, we are left with the following integral:

$$\begin{aligned}\int_{\beta_1}^{\alpha_1} \rho_Y(\varphi)\, d\varphi &= 2\int_0^{\infty} dx \int_{-x}^{0} dy\, \rho(x, y) = 2\int_0^{\infty} dx \int_0^{x} dy\, \rho(x, -y) \\ &= 2\int_0^{\infty} dx \int_0^{x} dy\, \rho(x, y) = \int_{\alpha_1}^{\beta_2} \rho_Y(\varphi)\, d\varphi,\end{aligned}$$

where we have used the same calculation for $[\alpha_1, \beta_2]$ as for $[\beta_1, \alpha_1]$ at the last step. This completes the proof of the theorem.

Combining both theorems, we have therefore shown:

Theorem 4. *Let Φ be the set of fixed points of geometric ICA in the expectation. Then there exists $(\hat{w}_1, \hat{w}_2) \in \Phi$ such that $M^{-1}_{\hat{w}_1,\hat{w}_2}$ solves the BSS problem. The stable fixed points in Φ can be found by the geometric ICA algorithm.*

Furthermore, we believe that in the special case of unimodal, symmetric, and nongaussian signals, the set Φ consists of only two elements: a stable and an unstable fixed point, where the stable fixed point will be found by the algorithm:

Conjecture 1. *Assume that the sources S_i are unimodal, symmetric, and nongaussian. Then there are only two fixed points of geometric ICA in the expectation.*

We can prove this conjecture for the special case of two sources with identical distributions that are nicely super- or subgaussian in the sense that ρ_Y has only four extremal points.

Theorem 5. *Assume that the sources S_i are unimodal and symmetric and that $\rho_1 = \rho_2$. Assume that $\rho_Y \mid [0, \pi)$ with $A = \mathrm{id}$ has exactly two local maxima and two local minima Then there exist only two fixed points of geometric ICA in the expectation.*

Proof. The same calculation as in the proof of theorem 3 shows that we can assume $A = \mathrm{id}$ without loss of generality. Then by theorem 2 and by $\rho_1 = \rho_2$, the two pairs $\{0, \frac{\pi}{2}\}$ and $\{\frac{\pi}{4}, \frac{3\pi}{4}\}$ satisfy GCC. We have to show that no other pair fulfills this condition.

First, note that the symmetry of ρ_1 and ρ_2 shows that $\rho_Y(\frac{n\pi}{2} - \varphi) = \rho_Y(\frac{n\pi}{2} + \varphi)$ for $n \in \mathbb{Z}$ and $\varphi \in \mathbb{R}$, and $\rho_1 = \rho_2$ induces even $\rho_Y(\frac{n\pi}{4} - \varphi) = \rho_Y(\frac{n\pi}{4} + \varphi)$. Due to assumption, ρ_Y has only two maxima and two minima in $[0, \pi)$; the above equations then show that those have to be at $\{0, \frac{\pi}{4}, \frac{\pi}{2}\, \frac{3\pi}{4}\}$.

Now we claim that for $\beta_1 \neq \frac{n\pi}{4}$, $n \in \mathbb{Z}$, the median of $\rho_Y \mid [\beta_1, \beta_1 + \frac{\pi}{2}]$ does not equal $\beta_1 + \frac{\pi}{4}$. Note that this shows theorem 5.

For the proof of the claim, consider the smallest integer $p \in \mathbb{Z}$ such that $\gamma_1 := \frac{p\pi}{2} > \beta_1$, and set $\gamma_2 := \gamma_1 + \frac{\pi}{4}$. Let $\beta_2 := \beta_1 + \frac{\pi}{2}$ and $\alpha := \beta_1 + \frac{\pi}{4}$. We have to show that the median of $\rho_Y \mid [\beta_1, \beta_2]$ does not equal α. A visualization of these relations and the following definitions is given in Figure 4.

Using the symmetry noted above, we have for $\delta_1 := \gamma_1 + (\gamma_1 - \beta_1) = 2\gamma_1 - \beta_1$: $C_1 := \int_{\beta_1}^{\gamma_1} \rho_Y = \int_{\gamma_1}^{\delta_1} \rho_Y$. Note that $\delta_1 \neq \alpha$, or else $p\pi = 2\gamma_1 =$

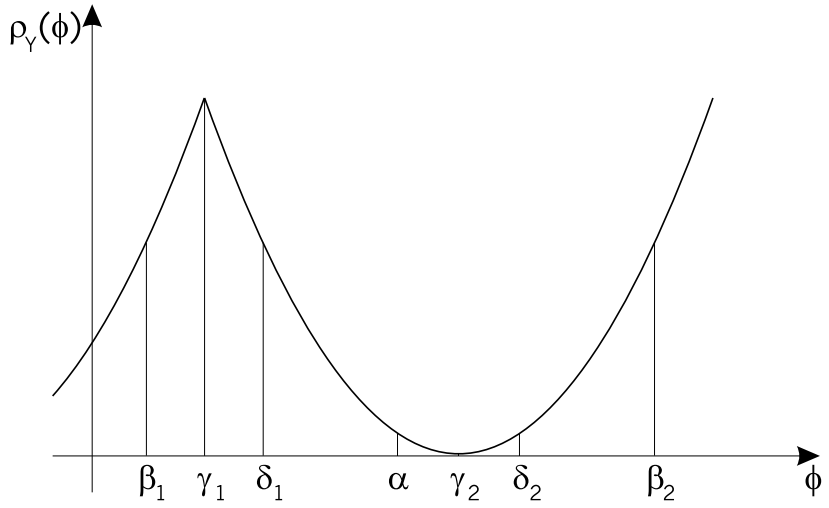

Figure 4: Visualization of the proof of theorem 5. α, β_i, γ_i, and δ_i are defined in the text.

$\alpha + \beta_1 = 2\beta_1 + \frac{\pi}{2}$, which contradicts the assumption to β_1. Without loss of generality, let $\alpha > \delta_1$; the other case can be proven similarly.

Setting $\delta_2 := \gamma_2 + (\gamma_2 - \alpha) = 2\gamma_1 - \alpha$, again symmetry shows that $C_2 := \int_\alpha^{\gamma_2} \rho_Y = \int_{\gamma_2}^{\delta_2} \rho_Y$ and $C_3 := \int_{\delta_1}^{\alpha} \rho_Y = \int_{\delta_2}^{\beta_2} \rho_Y$; the second equation follows because $\gamma_2 - \delta_1 = \beta_2 - \gamma_2$ and $\gamma_2 - \alpha = \delta_2 - \gamma_2$.

Now assume in contradiction to our claim that α is the median of $\rho_Y \mid [\beta_1, \beta_2]$. Then we have $2C_1 + C_3 = \int_{\beta_1}^{\alpha} \rho_Y = \int_\alpha^{\beta_2} \rho_Y = 2C_2 + C_3$ and therefore $C_1 = C_3$. As shown above, $\rho_Y(\gamma_1)$ and $\rho_Y(\gamma_2)$ are the only extremal points of $\rho_Y \mid [\beta_1, \beta_2]$. Without loss of generality, let $\rho_Y(\gamma_1)$ be a maximum; then $\rho_Y(\gamma_2)$ has to be a minimum. But this means that $C_1 = \int_{\beta_1}^{\gamma_1} \rho_Y > (\gamma_1 - \beta_1)\rho_Y(\alpha) \geq \int_\alpha^{\gamma_2} \rho_Y = C_2$ which contradicts the above. This completes the proof of the theorem.

Conjecture 1 states that there are only two fixed points of geometric ICA. In fact, we claim that of those two, only one fixed point is stable in the sense that slight perturbations of the initial conditions preserve the convergence. Then, depending on the kurtosis of the sources, either the stable (supergaussian case) or the unstable (subgaussian case) fixed point represents the image of the unit vectors. This is stated in the following conjecture.

Conjecture 2. *Assume that the sources S_i are unimodal, symmetric, and nongaussian. Then by conjecture 1, there are only two fixed points $(\hat{w}_1, \hat{w}_2)$ and*

$(\tilde{w}_1, \tilde{w}_2)$ of geometric ICA in the expectation. We claim:

i. There is only one stable fixed point $(\hat{w}_1, \hat{w}_2)$.

ii. If the sources are supergaussian, $M^{-1}_{\hat{w}_1,\hat{w}_2}$ solves the BSS problem.

iii. If the sources are subgaussian, $M^{-1}_{\tilde{w}_1,\tilde{w}_2}$ solves the BSS problem.

7 Update Rules Without Sign Functions

We have shown that the geometric update step requires the signum function as follows: $w_i(t+1) = w_i(t) + \eta(t)\,\mathrm{sgn}(y(t) - w_i(t))$. Then (normally), the w_i converge to the medians in their receptive field. Note that the medians do not have to coincide with any maxima of the sensor signal density distribution on the sphere, as shown in Figure 5. Therefore, in general, any algorithm searching for the maxima of the distribution (Prieto, Prieto, Puntonet, Canas, & Martin-Smith, 1999) will not end at the medians, which are the correct images of the unit vectors under the given mixing transformation. Only given special restrictions of the sources (same supergaussian distribution of each component, as, for example, speech signals), the medians correspond to the maxima, and a maximum searching algorithm will converge to the correct fixed points of geometric ICA.

8 Histogram-Based Algorithm: FastGeo

So far in geometric ICA, mostly on-line algorithms with competitive learning rules as in section 4 have been used (Puntonet & Prieto, 1995). As shown above, they search for points satisfying the GCC. In the following, we will establish a new geometric algorithm based on the GCC alone. For this, let $n = 2$ and

$$A := \begin{pmatrix} \cos(\alpha_1) & \cos(\alpha_2) \\ \sin(\alpha_1) & \sin(\alpha_2) \end{pmatrix}. \tag{8.1}$$

Theorem 3 shows that the vectors $(\cos(\alpha_i), \sin(\alpha_i))^\top$ satisfy the GCC. Therefore, the vectors w_i will converge to the medians in their receptive fields. This enables us to compute these positions directly using a search on the histogram of Y (see Figure 6), which reduces the computation time by a factor of about 100 or more. In the FastGeo algorithm, we scan through the different receptive fields and test GCC. In practice, this means discretizing the distribution f_Y of Y using a given bin size $\beta > 0$ and then testing the π/β different receptive fields. The algorithm will be formulated more precisely in the following.

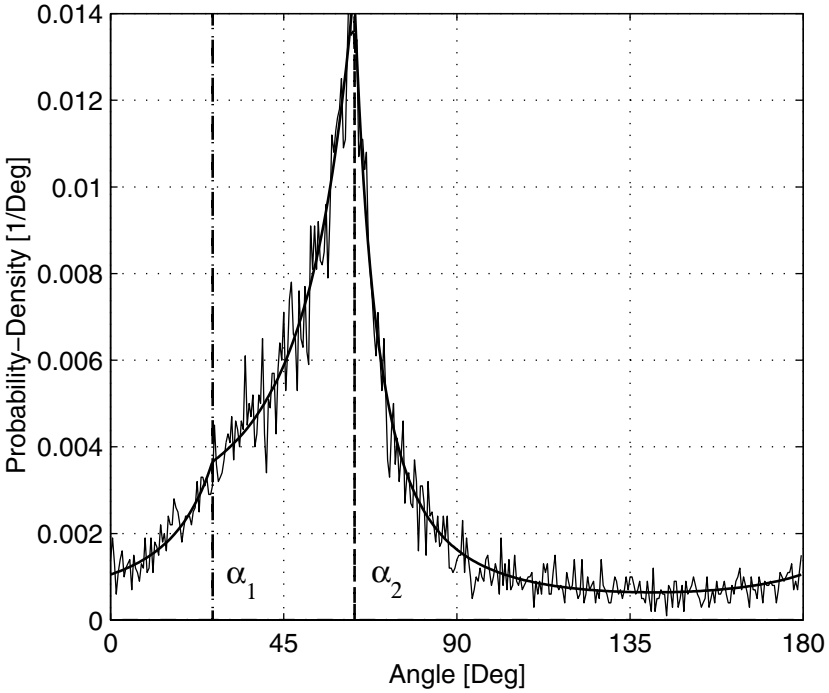

Figure 5: Projected density distribution ρ_Y of a mixture of two Laplacian signals with different variances, with the mixture matrix mapping the unit vectors e_i to $(\cos\alpha_i, \sin\alpha_i)$ for $i = 1, 2$. Dark line = theoretical density function; gray line = histogram of a mixture of 10,000 samples.

For simplicity, let us assume that the cumulative distribution F_Y of Y is invertible; this means that F_Y is nowhere constant. Define a function

$$\begin{aligned} \mu\colon [0,\pi) &\longrightarrow \mathbb{R} \\ \varphi &\longmapsto \frac{l_1(\varphi)+l_2(\varphi)}{2} - \left(\varphi + \frac{\pi}{2}\right), \end{aligned} \tag{8.2}$$

where

$$l_i(\varphi) := F_Y^{-1}\left(\frac{F_Y(\varphi + i\frac{\pi}{2}) + F_Y(\varphi + (i-1)\frac{\pi}{2})}{2}\right) \tag{8.3}$$

is the median of $Y \mid [\varphi + (i-1)\frac{\pi}{2}, \varphi + i\frac{\pi}{2}]$ in $[\varphi + (i-1)\frac{\pi}{2}, \varphi + i\frac{\pi}{2}]$ for $i = 1, 2$.

Lemma 2. *Let φ be a zero of μ in $[0, \pi)$. Then the $l_i(\varphi)$ satisfy the GCC.*

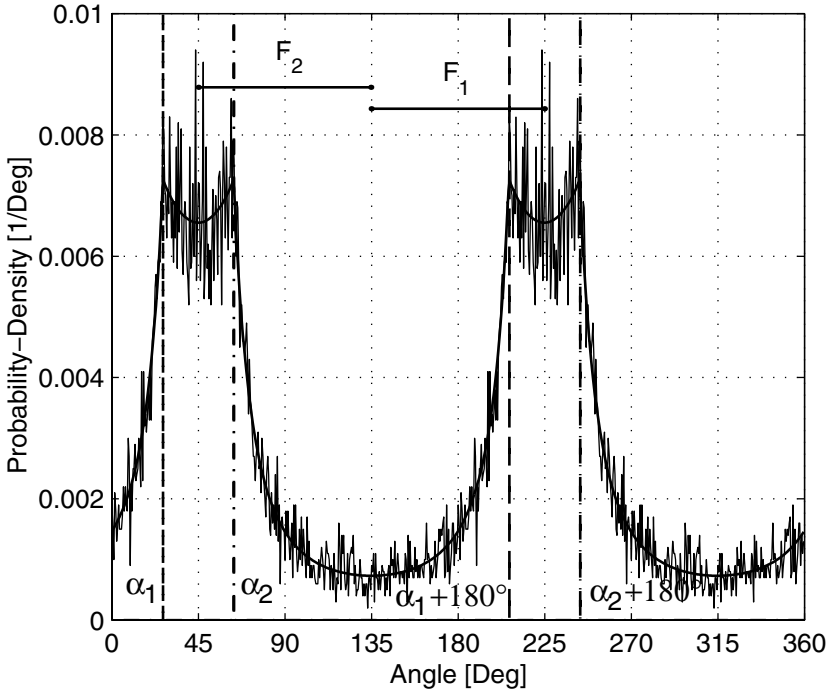

Figure 6: Probability density function f_Y of Y from Figure 1 with the mixing angles α_i and their receptive fields F_i for $i = 1, 2$.

Proof. By definition, $[\frac{l_1(\varphi)+l_2(\varphi)}{2} - \frac{\pi}{2}, \frac{l_1(\varphi)+l_2(\varphi)}{2}]$ is the receptive field of $l_1(\varphi)$. Since $\mu(\varphi) = 0$, the starting point of the above interval is φ, because $\varphi = \frac{l_1(\varphi)+l_2(\varphi)}{2} - \frac{\pi}{2}$. Hence, we have shown that the receptive field of $l_1(\varphi)$ is $[\varphi, \varphi + \frac{\pi}{2}]$, and by construction $l_1(\varphi)$ is the median of Y restricted to the above interval. The claim for $l_2(\varphi)$ then follows.

Algorithm 1 (FastGeo). *Find the zeros of μ.*

μ always has at least two zeros that represent the stable and the unstable fixed point of the ordinary geometric algorithm. In practice, we extract the fixed point, which then gives the proper demixing matrix A^{-1} by picking φ_0 such that $f_Y(l_1(\varphi_0))+f_Y(l_2(\varphi_0))$ is maximal. For unimodal and supergaussian source distributions, conjecture 2 claims that this results in a stable fixed point. For subgaussian sources, choosing φ_0 with $f_Y(l_1(\varphi_0))+f_Y(l_2(\varphi_0))$ being minimal induces the corresponding demixing matrix. Hence, one advantage of this histogram-based algorithm is that without any modifications, we can solve the ICA problem for subgaussian signals too. Furthermore, the

sophisticated parameter choice of the ordinary geometric algorithm is not necessary any more; only one parameter, the bin size, has to be chosen.

In practice, one sometimes notices that due to the discretization of the distribution, the approximated distribution has a rather noisy shape on small scales. This results from μ having zeros being split up into multiple zeros close together. Therefore, a useful improvement of convergence can be established by smoothing this distribution with a kernel function with sufficiently small halfwidth. This smoothing should be performed preferably during the discretization process of the original distribution.

9 Accuracy

In this section, we consider the dependence of the FastGeo ICA algorithm on the number of samples after the bin size β has been fixed. As seen in the previous section, the accuracy of the histogram-based algorithm then depends only on the distribution of the samples X (resp. Y); that is, we can estimate the error made by approximating the mixing matrix A by a finite number of samples. In the following, we present some results of test runs made with this algorithm.

When choosing two arbitrary angles $\alpha_i \in [0, \pi), i = 1, 2$ for the mixing matrix A, we define α as the distance between these two angles modulo $\frac{\pi}{2}$. This will give us an angle in the range between 0 and $\frac{\pi}{2}$. In Figure 7, the relative accuracy of the recovered angles $\frac{\Delta\alpha}{\alpha} = \frac{|\alpha_i - \alpha_i^{\text{recovered}}|}{\alpha}$ is given as a function of the angle α for a fixed number of samples. Obviously, the resulting graph is reasonably constant over a wide range of α, demonstrating that the estimate of the α_i with respect to α is robust over a wide range of α ($\alpha > 10^\circ$) and gets only slightly worse for small values of α. Note that the distortions around the origin are due to the finite bin size; increasing the number of bins increases the accuracy for small α's, but also the computational effort.

To investigate the relation between the error $\Delta\alpha$ and the number of samples, we examine for different α the dependence of the standard deviation of $\Delta\alpha$ on the number of samples (see Table 1), where we chose the following mixing matrix A:

$$A := \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{9.1}$$

The table entry gives the standard deviation of the nondiagonal terms after normalizing each column of the mixing matrix, so that the diagonal elements are unity. For comparison, we also calculated the performance index E_1 (Amari, Cichocki, & Yang, 1996).

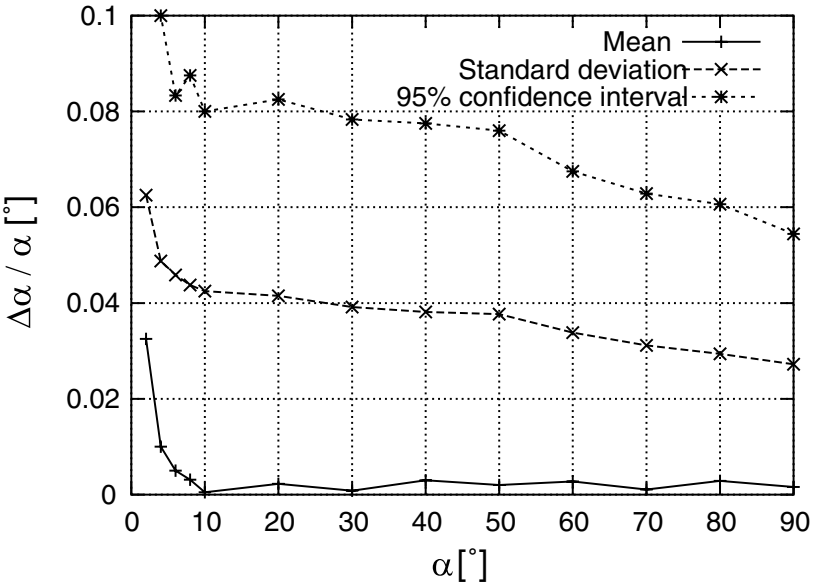

Figure 7: Mixture of 1,000 samples of two Laplacian source signals with identical variances. Plotted are the mean, standard deviation, and 95% confidence interval of $\Delta\alpha/\alpha$ versus α calculated from 100 runs for each angle α.

Table 1: Standard Deviations of the Nondiagonal Terms and the Performance Index E_1 with Different Number of Samples.

Number of Samples	Standard Deviation	Index E_1
1000	0.033	0.18
10,000	0.013	0.07
100,000	0.007	0.038

10 Examples

In this section, we compare geometric ICA algorithms with other ICA algorithms: the Extended Infomax algorithm (Lee, Girolami, & Sejnowski, 1999), which is based on the classical ICA algorithm (Bell & Sejnowski, 1995), and the FastICA algorithm (Hyvärinen & Oja, 1997). As geometric algorithms, we use the ordinary geometric algorithm from section 4 (Puntonet & Prieto, 1995) and the FastGeo algorithm from above. Calculations were made on a PIII-850 PC using Matlab 6.0.

In our first example, we consider a mixture of two Laplacian signals. The results of the different algorithms are shown in Table 2. For each algorithm,

Table 2: Comparison of Time per Run and Cross-talking Error of ICA Algorithms for a Random Mixture of two Laplacian Signals.

Algorithm	Elapsed Time [s]	Index E_1
Extended Infomax	11.1	0.072±0.002
FastICA (pow3=default)	0.068	0.076±0.004
FastICA (tanh)	0.11	0.052±0.001
FastICA (gauss)	0.12	0.048±0.001
Ordinary Geometric	>60	0.18±0.10
FastGeo	0.84	0.110±0.071

Note: Means and standard deviations were taken over 1000 runs (100 runs for Extended Infomax and Ordinary Geo) with 10,000 samples and uniformly distributed mixing matrix elements.

Table 3: Comparison of Time per Run and Cross-Talking Error of ICA Algorithms for a Random Mixture of Two Sound Signals with 22,000 Samples.

Algorithm	Elapsed Time [s]	Index E_1
Extended Infomax	41.2	0.058±0.002
FastICA (pow3=default)	0.14	0.050±0.005
FastICA (tanh)	0.24	0.022±0.001
FastICA (tanh)	0.26	0.019±0.001
Ordinary Geometric	>60	0.49±0.29
FastGeo	0.89	0.136±0.087

Note: Means and standard deviations were taken over 1000 runs (100 runs for Extended Infomax and Ordinary Geo) with uniformly distributed mixing matrix elements.

we measure the mean elapsed CPU time per run and the mean cross-talking error E_1 with its standard deviation. Both the Extended Infomax and the FastICA algorithm perform best in terms of accuracy, and in terms of computational speed, FastICA lives up to its name, being followed by FastGeo, then InfoMax, and then the ordinary geometric algorithm, always by one scale difference. The last algorithm lacks accuracy and also shows some convergence problems, whereas FastGeo lies between the geometric algorithm and FastICA and Infomax regarding accuracy.

The second example deals with real-world data: two audio signals (one speech and one music signal; see Table 3). The results are similar to the Laplacian toy example. FastICA outperforms the other algorithms in terms of speed. The accuracy of Extended Infomax and FastICA is comparable, and FastGeo is slightly (factor 4) worse but faster than the Extended Infomax. The ordinary geometric algorithm is both slower and less accurate, mainly because of convergence problems.

11 Higher Dimensions

So far we have explicitly considered two-dimensional data sets only. In real-world problems, however, the sensor signals are usually high-dimensional data sets, for example, EEG data with 21 dimensions. Therefore, it would be satisfactory to generalize geometric algorithms to higher dimensions. The ordinary geometric algorithm can be easily translated to higher-dimensional cases, but one faces serious problems in the explicit calculations. In order to approximate higher-dimensional pdfs, it becomes necessary to have an exponentially growing number of samples available, as will be shown.

The number of samples in a ball B^{d-1} of radius ϑ on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ divided by the number of samples on the whole S^{d-1} can be calculated as follows if we assume a uniformly distributed random vector.

Let $B^d := \{x \in \mathbb{R}^d \mid |x| \leq 1\}$ and $S^{d-1} := \{x \in \mathbb{R}^d \mid |x| = 1\}$. The volume of B^d can be calculated by $vol(B^d) = \frac{\pi^{\frac{d}{2}}}{(\frac{d}{2})!} = c_d$. It follows for $d > 3$:

$$\frac{\text{Number of samples in ball}}{n} = \frac{n\frac{\text{vol}(B^{d-1})\vartheta^{d-1}}{\text{vol}(S^{d-1})}}{n} \leq \frac{\vartheta^{d-1}c_{d-1}}{c_{d+1}} = \frac{\vartheta^{d-1}d}{\pi}.$$

Obviously, the number of samples in the ball decreases by $\vartheta^{d-1}d$ if $\vartheta < 1$, which is the interesting case. To have the same accuracy when estimating the medians, the decrease must be compensated by an exponential growth in the number of samples. For three dimensions, we have found a good approximate of the demixing matrix by using 100,000 samples. In four dimensions, the reconstructed mixing matrix could not be reconstructed correctly, even with larger numbers of samples.

A different approach for higher dimensions has been taken by Bauer et al. (2000), where A has been calculated using $\frac{d(d-1)}{2}$ projections of X from $\mathbb{R}^d$ onto $\mathbb{R}^2$ along the different coordinate axes and reconstructed the multidimensional matrix from the two-dimensional solutions. However, this approach works satisfactorily only if the mixing matrix A is close to the unit matrix up to permutation and scaling.

12 Conclusion

The geometric ICA algorithm has been studied in a concise theoretical framework. The fixed points of the geometric ICA learning algorithm have been examined in detail. We have introduced a GCC, which has to be fulfilled by the fixed points of the learning algorithm. We further showed that it is also fulfilled by the mixed unit vectors spanning the sensor signal space. Hence, geometric ICA can solve the BSS problem. We then gave two conjectures for the unimodal case where the fixed-point property is expected to be very rigid.

We then presented a new algorithm for linear geometric ICA (FastGeo) based on histograms that is both robust and computationally much more efficient than the ordinary geometric ICA algorithm. The accuracy of the algorithm concerning the estimation of the relevant medians of the underlying data distributions, when varying both the mixing matrices and the sample numbers, has been explored quantitatively, showing a rather robust performance of the algorithm.

When comparing FastGeo with classical ICA algorithms and the ordinary geometric one, we noticed that FastGeo performs only slightly worse than the classical ones in terms of accuracy and better than the ordinary geometric one. In terms of speed, FastGeo falls between FastICA and the Extended Infomax and is much faster than the first geometric approaches, which also suffer from severe convergence problems. Furthermore, the fact that geometric algorithms and especially FastGeo are very easy to implement makes FastGeo a good choice even in comparison with the classical ICA algorithms in practical two-dimensional applications.

We also considered the problem of high-dimensional data sets with respect to the geometrical algorithms and discussed how projections to low-dimensional subspaces could solve this problem for a special class of mixing matrices.

Simulations with nonsymmetrical and nonunimodal distributions have shown promising results so far, indicating that the new algorithm will perform well with almost any distribution. This is the subject of ongoing research in our group.

In future work, besides treating nonsymmetric sources S, the two conjectures will have to be proven in full, as well as the Kohonen proof of convergence to be translated into the above model. In addition, the histogram-based algorithm could be extended to the nonlinear case similar to Puntonet et al. (1999), using multiple centered spheres for projection on the surface on which the projected data histograms could then be evaluated. Finally, we are experimenting with the FastGeo algorithm for the overcomplete case, where we are currently able to detect three or more sources in only two mixtures. This can be useful for higher-dimensional cases as in section 11.

Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Graduiertenkolleg "Nonlinearity and Nonequilibrium in Condensed Matter." We thank Christoph Bauer and Tobias Westenhuber for suggestions and comments on the ordinary geometric algorithm and Michaela Lautenschlager and Stefanie Ulsamer for their helpful comments in trying to prove convergence.

References

Amari, S., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 757–763). Cambridge, MA: MIT Press.

Bauer, C., Puntonet, C., Rodriguez-Alvarez, M., & Lang, E. (2000). Separation of EEG signals with geometric procedures. In C. Fyfe (Ed.), *Engineering of intelligent systems* (Proc. EIS'2000) (pp. 104–108). Millet, Alberta, Canada: ICSC Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129–1159.

Benaim, M., Fort, J.-C., & Pagés, G. (1998). Convergence of the one-dimensional Kohonen algorithm. *Adv. Appl. Prob., 30*, 850–869.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing, 36*, 287–314.

Cottrell, M., & Fort, J.-C. (1987). Etude d'un processus d'autoorganisation. *Annales de l'Institut Henri Poincaré, 23*(1), 1–20. (in French)

Cottrell, M., Fort, J. C., & Pagès, G. (1994). Two or three things that we know about the Kohonen algorithm. In M. Verleysen (Ed.), *Proc. ESANN'94, European Symp. on Artificial Neural Networks* (pp. 235–244). Brussels: D Facto Conference Services.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks, 10*(3), 626–634.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation, 9*, 1483–1492.

Jutten, C., Hérault, J., Comon, P., & Sorouchiary, E. (1991). Blind separation of sources, parts I, II, and III. *Signal Processing, 24*, 1–29.

Lee, T.-W., Girolami, M., & Sejnowski, T. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation, 11*, 417–441.

Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. In D. Touretzky (Ed.), *Advances in neural information processing systems, 1*. San Mateo, CA: Morgan Kaufmann.

Linsker, R. (1992). Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation, 4*, 691–702.

Prieto, A., Prieto, B., Puntonet, C., Canas, A., & Martin-Smith, P. (1999). Geometric separation of linear mixtures of sources: Application to speech signals. In J. F. Cardoso, Ch. Jutten, & Ph. Loubaton (Eds.), *Independent component analysis and signal separation* (Proc. ICA'99) (pp. 295–300). Gières, France: Imp. de Ecureuils.

Puntonet, C., Alvarez, M., Prieto, A., & Prieto, B. (1999). Separation of speech signals for nonlinear mixtures. *Lecture Notes in Computer Science, 1607*, 665–673.

Puntonet, C., & Prieto, A. (1995). An adaptive geometrical procedure for blind separation of sources. *Neural Processing Letters, 2*.

Ritter, H., & Schulten, K. (1988). Convergence porperties of Kohonen's topology conserving maps: Fluctuations, stability, and dimension selection. *Biological Cybernetics, 60*, 59–71.

Received December 14, 2001; accepted July 3, 2002.