Independent component analysis in the light of Information Geometry

Jean-François Cardoso

Independent component analysis in the light of Information Geometry

Jean-François Cardoso
Institut d'Astrophysique de Paris
CNRS, France

October 31, 2022

Version accepted for publication in 'Information Geometry'.
DOI: https://doi.org/10.1007/s41884-022-00073-x

Abstract

I recall my first encounter with Professor Shun-ichi Amari who, once upon a time in Las Vegas, gave me a precious hint about connecting Independent Component Analysis (ICA) to Information Geometry. The paper sketches, rather informally, some of the insights gained in following this lead.

1 Amari and Pythagoras in Las Vegas

Independent Component Analysis (ICA) of a random N-vector X consists in finding a linear transform B (an invertible $N \times N$ matrix) making the entries of $Y = BX$ 'as independent as possible'. There are (infinitely) many matrices B which can decorrelate the entries of Y and, if the data are Gaussian, decorrelation implies independence so that ICA has nothing to offer here. However, for *non Gaussian* data, independence is stronger than decorrelation and the situation is somehow the opposite: no matrix B can produce a vector $Y = BX$ with independent entries unless the distribution of X is 'special'. That special case, of course, is when $X = AS$ where S a vector of independent entries and A is some invertible matrix. Moreover, in that case, there is an essential uniqueness of ICA: if the entries of $Y = BX$ are independent, they must be those of S, possibly up to permutation and rescaling or, equivalently, B must be of the form $B = PA^{-1}$ where matrix P has one and only one non-zero entry in each row and each column [10].

In other words, the only way of restoring independence of non Gaussian random variables which have been mixed is to unmix them. From this property stems the usefulness of ICA in many applications, whenever N sensors can be assumed to receive a mixture of independent sources but the coefficients of the mixing are unknown or cannot be determined by physical modeling. ICA makes 'blind source separation' possible: this is the ability to recover mixed

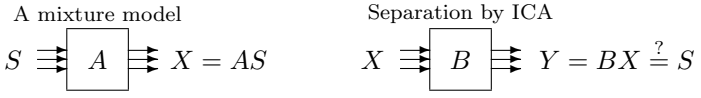

Figure 1: The hypothetical data model of linear mixture $X = AS$ and a separating matrix B trying to recover the underlying sources in S. If the entries of S are statistically independent and non Gaussian, then matrix B can restore the independence between the entries of $Y = BX$ *only* by separating the sources, that is, the entries of Y are those of S (possibly up to rescaling and permutation).

underlying sources without resorting to any prior information on the system (matrix A),

resorting only to one key assumption: non Gaussian, statistically independent sources. This ability is what made ICA such an attractive tool in many applications in which the presence of independent sources is a strong but often plausible hypothesis.

In 1995, I was not yet aware that the Earth was heading toward an environmental disaster so I shamelessly flew to Las Vegas to attend a symposium on nonlinear theory (whatever that means) where I presented a paper on the invariance of ICA. I had been working extensively on ICA for a few years but, having stumbled upon Amari's book [1], I was trying to familiarize myself with Information Geometry which I found to be a fascinating and inspiring vision. At the end of my presentation, Amari stood up and made a kind comment. I was overjoyed: the Grand Master of Information Geometry was entering the field of ICA!

After the session, Amari invited me for a drink and generously shared an idea with me. I already knew that the Kullback-Leibler divergence (KLD) $D[\,P\|\,Q] = \int \log \frac{P(x)}{Q(x)}\; dP(x)$ from a distribution P to another distribution Q gives rise to a Pythagorean theorem when used in conjunction with an finite-dimensional exponential family of distributions. Amari pointed to me that the set of N-variate distributions with independent entries, which is at the heart of ICA, can be seen as an exponential family, albeit an infinite-dimensional one[1].

Indeed, consider the 'product manifold' $\mathbb{P}$ as the set of N-variate probability distributions which are the *product* of their marginal distributions or, in other words, distributions of N-vectors with independent entries. Let P_Y denote the distribution of some random N-vector Y and let $P_S = \prod_i P_{S_i} \in \mathbb{P}$ be any distribution of independent entries. By substitution, one easily finds

$$D[\,P_Y\|\,P_S] = D[\,P_Y\|\,\textstyle\prod_i P_{Y_i}] + D[\,\prod_i P_{Y_i}\|\,\prod_i P_{S_i}] \tag{1}$$

which shows that the minimum of $D[\,P_Y\|\,P_S]$ over $\mathbb{P}$ is reached for $P_S = \prod_i P_{Y_i}$. That minimum value is a well known quantity: the mutual information (between the entries) of Y, denoted

$$\mathcal{I}(Y) \stackrel{\text{def}}{=} D[\,P_Y\|\,\textstyle\prod_{i=1}^N P_{Y_i}]. \tag{2}$$

Further, since $D[\,\prod_i P_{Y_i}\|\,\prod_i P_{S_i}] = \sum_i D[\,P_{Y_i}\|\,P_{S_i}]$, the KLD from an N-variate distribution P_Y to a target distribution of independent components $P_S = \prod_{i=1}^N P_{S_i}$ admits a decomposition:

$$D[\,P_Y\|\,\textstyle\prod_i P_{S_i}] = \mathcal{I}(Y) + \sum_i D[\,P_{Y_i}\|\,P_{S_i}] \tag{3}$$

into a part $\mathcal{I}(Y)$ which does not depend on the target distribution P_S but only measures the amount of dependence between the entries of Y and a part which only measures marginal discrepancies between Y and S. Thus, Eq. 3 is the form taken by the Pythagorean theorem on $\mathbb{P}$. With this insight, Amari provided me with a point of contact between ICA and Information Geometry.

The idea of using mutual information as a criterion for ICA had already been proposed in the seminal paper of Comon [10] but the geometrical connection could offer more insights, in particular in relation to non Gaussianity. Indeed, non Gaussianity is not only required for blind separability but it can also be used as a criterion for finding a separating matrix : looking for maximally independent sources and looking for maximally non Gaussian sources are two possible routes to blind separation [10].

This short paper shows how mutual information and non Gaussianity are geometrically related. Sec. 2 follows the maximum likelihood principle for ICA, leading to mutual information. The latter is then related to non Gaussianity by another Pythagorean theorem in Sec. 3, illustrating how non Gaussianity allows to express statistical independence beyond mere decorrelation. Some consequences for ICA are sketched in Sec. 4.

2 Likelihood and Kullback matching for ICA

We start by setting up the simplest ICA model. It assumes a zero-mean random N-vector S with independent entries —the so-called 'sources'— mixed by an (invertible) $N \times N$ matrix A (the 'mixing matrix'):

$$X = AS \quad \text{with} \quad S \sim Q(S) = \textstyle\prod_{i=1}^N q_i(S_i) \qquad \text{[the basic ICA model]} \tag{4}$$

[1]We would need technical conditions to make this the more rigorous. Hereafter, we assume that all source distributions have a strictly positive density with respect to the Lebesgue measure.

where $q_1, \ldots, q_N$ are N scalar probability distributions for the sources. The complete parameter set is $\theta = (A, Q) = (A, q_1, \ldots, q_N)$. Since the aim of ICA is to recover the sources by inverting A, the source distributions q_i are considered to be nuisance parameters while A is the parameter of interest.

To gain some insights into the likelihood of the ICA model (4), we examine the average shape of the log-density. It is an easily demonstrated general fact that, for any parametric model

$$\mathbb{E}_X \log P_\theta(X) = -D[\, P_X \| \, P_\theta] - H(X) \tag{5}$$

where $H(X)$ denotes Shannon differential entropy. Since the latter does not depend on the model, the shape of the average log-likelihood landscape is controlled by $D[\, P_X \| \, P_\theta]$, showing that the maximum likelihood principle corresponds to minimizing the Kullback divergence from the data distribution P_X to the model distribution P_θ. In the following, we explore the minimization of $D[\, P_X \| \, P_\theta]$ as the guiding principle for ICA.

We can take advantage of a specific feature of the ICA model: it is a *transformation model* and the KLD is invariant under invertible transforms. Hence the KLD from the data distribution P_X to the model distribution,P_θ of AS equals the KLD from the distribution of $A^{-1}X$ to the distribution of S for any invertible matrix A. Therefore, for the ICA model (4), we have

$$D[\, P_X \| \, P_{\theta=(A,Q)}] = D[\, P_{A^{-1}X} \| \, P_{\theta=(I_N,Q)}] = D[\, P_{A^{-1}X} \| \, Q]. \tag{6}$$

Since the data X and the parameter of interest A enter only via $Y = A^{-1}X$ in Eq. (6), the message from the maximum likelihood principle is very clear: the likeliest A should make the transformed data $Y = A^{-1}X$ as close as possible to the (hypothetical) source distribution $Q = \prod_i q_i$ in the sense of minimizing the Kullback mismatch $D[\, P_Y \| \, Q]$.

Proceeding, we invoke decomposition (3) which reads here as:

$$D[\, P_{A^{-1}X} \| \, Q] = D[\, P_Y \| \, \textstyle\prod_i q_i] = \mathcal{I}(Y) + \sum_i D[\, P_{Y_i} \| \, q_i] \tag{7}$$

and shows that minimizing $D[\, P_Y \| \, Q]$ is trying to achieve a composite objective: making the entries of Y as independent as possible while also making their distributions as close as possible to the marginal targets $q_1, \ldots, q_N$.

Recall that the spirit of source separation is to proceed blindly as much as possible. Just as we impose no constraints on A, it is desirable to let the nuisance parameters $Q = \prod_i q_i$ be determined from the data themselves. This is easily done (at least, in theory!) according to Eq. (7): for any value of A, the Kullback mismatch $D[\, P_{A^{-1}X} \| \, \prod_i q_i]$ is minimized with respect to the source distribution q_i by making $D[\, P_{Y_i} \| \, q_i]$ equal to 0, *i.e.* by estimating the source distribution q_i to be the marginal distribution P_{Y_i}. Then we are left with

$$\min_{q_1, \ldots, q_N} D[\, P_Y \| \, \textstyle\prod_i q_i] = \mathcal{I}(Y). \tag{8}$$

We conclude that the maximum likelihood principle leads to the mutual information $\mathcal{I}(Y)$ as the objective of choice for ICA when nothing is known about the source distributions, in support of the original proposal of Comon [10].

3 Independence and non Gaussianity

We already mentioned that looking for components which are maximally non Gaussian is a possible route to source separation. We now give the geometric connection between these two objectives: moving away from being Gaussian and moving closer to being independent. All that is required are two applications of the Pythagorean theorem.

We start by defining a measure of non Gaussianity for a zero-mean[2] random N-vector with distribution P_Y. Denoting $\mathbb{G}$ the exponential family of all zero-mean N-variate Gaussian distributions, the Pythagorean theorem on $\mathbb{G}$ takes the form

$$D[\, P_Y \| \, \mathcal{N}(\Sigma)] = D[\, P_Y \| \, \mathcal{N}(\mathrm{Cov} Y)] + D[\mathcal{N}(\mathrm{Cov} Y) \| \, \mathcal{N}(\Sigma)]. \tag{9}$$

[2]For minimizing the notation and without any real loss of generality, all distributions are restricted to have zero mean in the following.

where Cov$Y$ is the covariance matrix of Y and where $\mathcal{N}(\Sigma)$ denotes the zero-mean N-variate normal density with covariance matrix Σ. The non Gaussianity $\mathcal{G}(Y)$ of a zero-mean random vector Y is naturally defined as the divergence from its distribution to its best Gaussian approximation, which by Eq. (9), is $\mathcal{N}(\text{Cov}Y)$:

$$\mathcal{G}(Y) \stackrel{\text{def}}{=} D[\, P_Y \| \, \mathcal{N}(\text{Cov}Y)].$$

Hence Eq. (9) shows that the divergence from a distribution to any Gaussian target has two parts: divergence from Gaussianity (independent of the target) plus divergence of covariance matrices.

Let us now combine the Pythagoras theorem of Eq. (1) related to independence and the Pythagoras theorem of (9) related to Gaussianity. It is interesting to do it in terms of successive approximations. When dealing with the distribution P_Y of an N-vector which is too complicated to handle, two widely used simplifying assumptions are that Y is normally distributed or that its entries are independent, that is, approximating distribution P_Y either by $P_Y^{\mathbb{G}} \stackrel{\text{def}}{=} \mathcal{N}(\text{Cov}Y)$ or by $P_Y^{\mathbb{P}} \stackrel{\text{def}}{=} \prod_i P_{Y_i}$. In geometric terms, these approximations are projections onto $\mathbb{G}$ or onto $\mathbb{P}$.

An even cruder approximation would be to use both the Gaussian and the independent approximations. Projecting either $P_Y^{\mathbb{P}}$ onto $\mathbb{G}$ or projecting $P_Y^{\mathbb{G}}$ onto $\mathbb{P}$ leads in both cases to $P_Y^{\mathbb{G}\mathbb{P}} \stackrel{\text{def}}{=} \mathcal{N}(\text{diag}(\text{Cov}Y))$. This is pictured in Fig. 2 showing the four aforementioned distributions, forming two triangles: $[P_Y \to P_Y^{\mathbb{G}} \to P_Y^{\mathbb{P}\mathbb{G}}]$ and $[P_Y \to P_Y^{\mathbb{P}} \to P_Y^{\mathbb{P}\mathbb{G}}]$. The key point is that these are two *right* triangles which *share a common hypotenuse* $[P_Y \to P_Y^{\mathbb{P}\mathbb{G}}]$. Hence, $D[\, P_Y \| \, P_Y^{\mathbb{P}\mathbb{G}}]$ has two complementary expressions, using either triangle:

$$D[\, P_Y \| \, P_Y^{\mathbb{P}\mathbb{G}}] = D[\, P_Y \| \, P_Y^{\mathbb{P}}] + D[\, P_Y^{\mathbb{P}} \| \, P_Y^{\mathbb{P}\mathbb{G}}] = D[\, P_Y \| \, P_Y^{\mathbb{G}}] + D[\, P_Y^{\mathbb{G}} \| \, P_Y^{\mathbb{P}\mathbb{G}}] \tag{10}$$

and applying each time the relevant Pythagorean relation (1) or (9).

Two of the divergences appearing in (10) are already understood: one is the mutual information $\mathcal{I}(Y) = D[\, P_Y \| \, P_Y^{\mathbb{P}}]$ measuring dependence; the other is the non Gaussianity $\mathcal{G}(Y) = D[\, P_Y \| \, P_Y^{\mathbb{G}}]$ measuring... just that. The other two divergences also have a clear statistical meaning. One is $D[\, P_Y^{\mathbb{G}} \| \, P_Y^{\mathbb{P}\mathbb{G}}] = D[\mathcal{N}(\text{Cov}Y) \| \, \mathcal{N}(\text{diagCov}Y)]$ measuring how far the covariance matrix Cov$Y$ is from its diagonal part diagCov$Y$. Hence, it measures the non diagonality of Cov$Y$ and therefore appears as the natural scalar measure of the correlation between the entries of Y. We thus define the *correlation* of a random vector as

$$\mathcal{C}(Y) = D[\, P_Y^{\mathbb{G}} \| \, P_Y^{\mathbb{P}\mathbb{G}}] = D[\mathcal{N}(\text{Cov}Y) \| \, \mathcal{N}(\text{diagCov}Y)] \tag{11}$$

The last divergence showing up in (10) is $D[\, P_Y^{\mathbb{P}} \| \, P_Y^{\mathbb{P}\mathbb{G}}]$. Being a divergence between two distributions of vectors with independent entries, it is just the sum of the pair-wise divergences between the entries. Since each of those actually is the divergence from the distribution of Y_i to its best Gaussian approximation, one has $D[\, P_Y^{\mathbb{P}} \| \, P_Y^{\mathbb{P}\mathbb{G}}] = \sum_i \mathcal{G}(Y_i)$, the sum of marginal Gaussianities. Thus Eq. (10) finally yields the desired connection between mutual information, correlation and non Gaussianity:

$$\mathcal{I}(Y) + \sum_i \mathcal{G}(Y_i) = \mathcal{C}(Y) + \mathcal{G}(Y). \tag{12}$$

The quantities defined via the KLD behave as nicely as possible: by projection onto the Gaussian manifold $\mathbb{G}$, statistical dependence — as measured by mutual information $\mathcal{I}(Y)$— reduces to correlation $\mathcal{C}(Y)$ while by projection onto $\mathbb{P}$, the (full, joint) non-Gaussianity $\mathcal{G}(Y)$ is reduced to marginal non-Gaussianity $\sum_i \mathcal{G}(Y_i)$. Incidentally, the reduction of divergence is the same for both projections since Eq. (12) also reads $\mathcal{I}(Y) - \mathcal{C}(Y) = \mathcal{G}(Y) - \sum_i \mathcal{G}(Y_i)$.

4 Relevance to independent component analysis

The connection between independence, correlation and non Gaussianity of Eq. (12) makes no reference to the ICA model and is independent of it. Its impact on Independent Component Analysis is revealed by one final observation. Recall that ICA deals with linear transforms of a vector $Y = A^{-1}X$. Now, if a vector Y undergoes some (invertible) linear transform, its

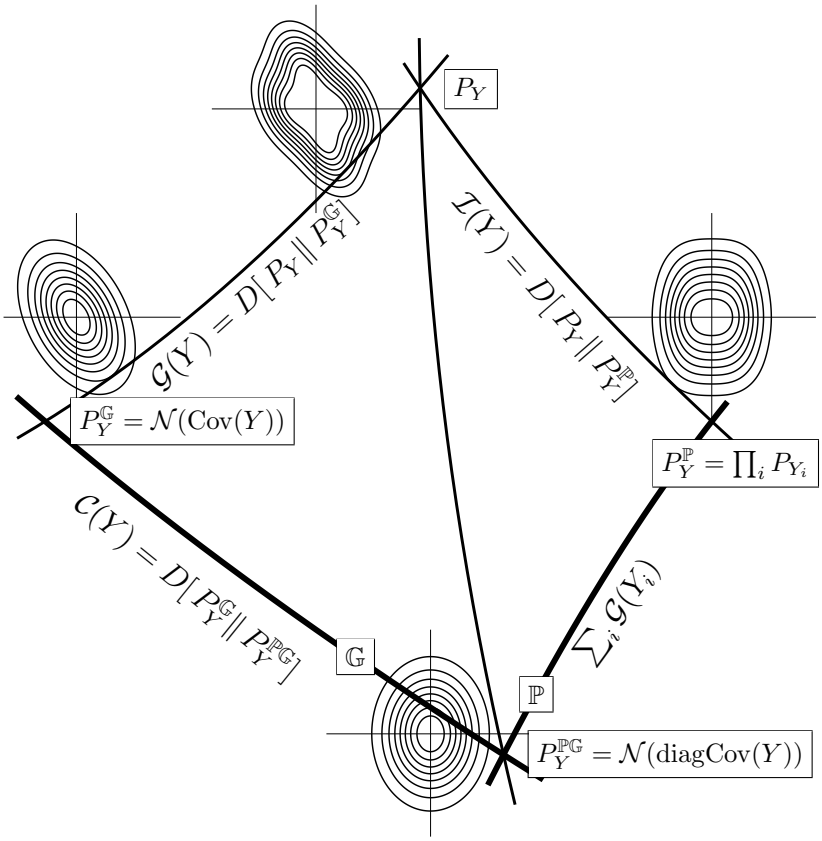

Figure 2: A probability density P_Y for a vector Y can be approximated as having independent components (approximation $P_Y^{\mathbb{P}}$) or as being Gaussian (approximation $P_Y^{\mathbb{G}}$) or both as $P_Y^{\mathbb{PG}}$. These approximations correspond to projections onto exponential manifolds. Those densities form two 'right triangles', each giving rise to a Pythagorean theorem, and sharing a common hypotenuse, thus relating the 'lengths' of the other sides, leading to Eq. (12) The lengths of all sides have a clear and simple statistical meaning, allowing to connect independence, correlation and non Gaussianity in a single information-geometric picture.

Gaussian approximation undergoes the *same* transform. Therefore, by invariance of the KLD, the non Gaussianity $\mathcal{G}(Y) = D[\,P_Y \,\|\, P_Y^{\mathbb{G}}]$ is constant under linear transforms. Therefore, in a linear search for independent components, one has

$$\mathcal{I}(Y) = \mathcal{C}(Y) - \sum_i \mathcal{G}(Y_i) + \text{constant} \qquad (\text{for any } Y = BX). \tag{13}$$

Therefore, *making the entries of Y as independent as possible amounts to make them as uncorrelated and as non Gaussian as possible*, in the sense of Eq. (13) *i.e.* giving *equal weight* to decorrelation and to non Gaussianity.

The mutual information $\mathcal{I}(Y)$ is conceptually simple but quite a challenge to estimate from data because density estimation is hard in the multidimensional case, and downright impossible in practice as soon as the dimension N is larger than a few units. It is remarkable how relation (13) breaks down this complexity: the correlation $\mathcal{C}(Y)$ is a simple function of a covariance matrix while each one of the marginal non Gaussianities $\mathcal{G}(Y_i)$ only depends on the distribution of a *scalar* variable. The only challenging term in (13) is hidden in the constant and need not to be explicitly evaluated if one is only concerned with minimizing the mutual information.

This raises the algorithmic issue of actually minimizing the mutual information. Since $\mathcal{I}(Y)$ itself, as a separation criterion, was obtained as a solution of the minimization prob-

lem (8), one approach is to alternate minimizations of $D[\,P_Y \| \prod_i q_i]$ with respect to A (changing $Y = A^{-1}X$) and with respect to the source distributions $q_1, \ldots, q_N$. It was shown in [3] that the non-parametric estimation of the source densities can be theoretically achieved without loss of statistical efficiency with respect to the case when the source densities are known in advance. This property has a geometric origin: the Fisher-orthogonality at point $Q = \prod_i q_i$ between the product manifold $\mathbb{P}$ and the N^2-dimensional 'system manifold' $\mathbb{S} \stackrel{\text{def}}{=} \{P_{CS} \,|\, C \in \mathrm{GL}(N), S \sim Q\}$ which is the manifold of all distributions of all invertible mixtures of S when $P_S = Q$.

In practice, a non-parametric estimation of the mutual information, or of the marginal non Gaussianities or of the source densities could carry too much of a burden in many real applications. What happens if adopting the opposite option: choosing in advance some model densities q_i and keeping them fixed in the ICA likelihood? Actually, the stationary points of $D[\,P_Y \| \prod_i q_i]$ with respect to linear transforms of Y have a simple expression: they are characterized by $\mathbb{E}\{\psi_i(Y_i)Y_j\} = 0$ $(1 \leq i \neq j \leq N)$ where $\psi_i = -q_i'/q_i$ is the *score function* of density q_i. These non linear decorrelation conditions are fulfilled if the entries of Y are independent because then $\mathbb{E}\{\psi_i(Y_i)Y_j\} = \mathbb{E}\{\psi_i(Y_i)\}\mathbb{E}\{Y_j\}$ for $i \neq j$ and the last factors $\mathbb{E}Y_j$ cancel for zero-mean sources. Hence, independent sources in Y are stationary points of $D[\,P_Y \| \prod_i q_i]$ *regardless* of the choice of the source models q_i!

However, to find separated sources by minimizing $D[\,P_Y \| \prod_i q_i]$, one needs more than a stationary point: one needs a minimum. Whether or not $D[\,P_{A^{-1}X} \| \prod_i q_i]$ is at a local minimum with respect to variations of A depends on the guessed q_i distributions being 'not too wrong', a condition which can receive a quantitative expression in terms of the correlation between the true and guessed score functions ψ_i [4, 6]. This robustness property could be traced back to a geometric property: the orthogonality of $\mathbb{P}$ and $\mathbb{S}$.

The robustness of ICA with respect to the source model is illustrated by the Infomax algorithm [5] which is an important example since it triggered a lot of interest for ICA in neurosciences. Infomax uses a fixed, popular non-linear function $\psi_i(s) = \tanh(s)$ and tries to solve $\mathbb{E}\{\tanh(Y_i)Y_j\} = \delta_{ij}$. Since $\tanh(s)$ is the score function for a density $q(s) \propto 1/\cosh(s)$ which has much heavier tails than a Gaussian distribution, infomax will usually operate successfully in uncovering sources with heavy-tailed distributions even if their density is not exactly proportional to $1/\cosh(s)$ (albeit at the cost of some unavoidable loss of statistical efficiency). Using $\psi_i(s) = \tanh(s)$ is implicitly like trying to fit a model of heavy-tailed, or *sparse* sources. We have seen that the best criterion $\mathcal{I}(Y)$ does not specifically want sparse sources but rather non Gaussian sources and being sparse is just a particular way of being non Gaussian. In presence of sources with densities of various kinds, both heavy-tailed and light-tailed, it becomes necessary to develop source-adaptive methods, in the spirit of mutual information and of its decomposition (13) in terms of decorrelation and non-Gaussianity.

A final comment is in order regarding the so-called 'orthogonal' ICA methods. This popular approach to ICA relies on the idea that source separation can proceed in two steps: in a first easy step, the data are 'whitened' (decorrelated and normalized to unit variance) and in a second step they are rotated, hence preserving decorrelation [11]. In other words, an orthogonal method seeks a separating matrix in the form $B = U\,\mathrm{Cov}(X)^{-1/2}$ where matrix U is constrained to be a rotation ($UU^\dagger = I_N$). Such a construction strictly enforces the decorrelation of $Y = BX$, *i.e.* it guarantees $\mathcal{C}(Y) = 0$. Hence, it can be seen as a variant of mutual information which would put an infinite weight on the objective of decorrelation, leaving only the degrees of freedom in U to express independence beyond decorrelation by maximizing the marginal non Gaussianities $\sum_i G(Y_i)$. Some loss of statistical efficiency is expected in the orthogonal approach since, as per Eq. (13), mutual information (which derives from the maximum likelihood principle) wants to give equal weight to the objectives of decorrelation and of 'degaussianization'.

5 Conclusion

This paper focused on the geometrical connection illustrated by Fig. 2 and on some of its consequences, so quite a few points were left unaddressed, in particular in relation to ICA as a *transformation* model. That the parameter of interest A lives in the multiplicative group $\mathrm{GL}(N)$ has some nice consequences in terms of statistical and algorithmic performance. In

particular, the natural gradient [2] of Amari takes a very simple form in ICA, where it becomes a 'relative gradient' [9]. But there is more geometry to ICA and the interested reader is referred to [8] or [7] for more on this topic.

Information geometry offers a wonderful source of inspiration for scientists who like to think in terms of pictures, graphs, sketches. The connection between dependence, correlation, and non Gaussianity presented in this paper can easily be demonstrated without resorting to information geometry but I would never have uncovered it without geometric thinking. I am grateful to Professor Amari for starting it.

Declarations

Data Availability: Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.
Conflict of interest: The author states that there is no conflict of interest.

References

[1] Shun-Ichi Amari. *Differential-Geometrical Methods in Statistics*. Number 28 in Lecture Notes in Statistics. Springer, Heidelberg, 1985.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[3] Shun-Ichi Amari and Jean-François Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Trans. on Sig. Proc.*, 45(11):2692–2700, November 1997. Special issue on neural networks.

[4] Shun-Ichi Amari, T.-P. Chen, and A. Cichocki. Stability analysis of adaptive blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[5] A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1004–1034, 1995.

[6] Jean-François Cardoso. On the stability of source separation algorithms. *Journal of VLSI Signal Processing Systems*, 26(1/2):7–14, April 2000.

[7] Jean-François Cardoso. *Unsupervised adaptive filters*, volume 1, chapter Entropic contrasts for source separation: geometry and stability, pages 139–190. John Wiley & sons, Simon Haykin editor, Hoboken, NJ, 2000.

[8] Jean-François Cardoso. Dependence, correlation and non Gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, December 2003.

[9] Jean-François Cardoso and Beate Laheld. Equivariant adaptive source separation. *IEEE Trans. on Sig. Proc.*, 44(12):3017–3030, December 1996.

[10] P. Comon. Independent component analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics.

[11] Pierre Comon and Christian Jutten, editors. *Handbook of Blind Source Separation. Independent Component Analysis and Applications*. Academic Press (Elsevier), Amsterdam, 2010.