Journal of Machine Learning Research 4 (2003) 1177-1203 Submitted 10/02; Published 12/03

Dependence, Correlation and Gaussianity in Independent Component Analysis

Jean-François Cardoso CARDOSO@TSI.ENST.FR
ENST/TSI
46 Rue Barrault
75634, Paris, France

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-ichi Amari

Abstract

Independent component analysis (ICA) is the decomposition of a random vector in linear components which are "as independent as possible." Here, "independence" should be understood in its strong statistical sense: it goes beyond (second-order) decorrelation and thus involves the non-Gaussianity of the data. The ideal measure of independence is the "mutual information" and is known to be related to the entropy of the components when the search for components is restricted to uncorrelated components. This paper explores the connections between mutual information, entropy and non-Gaussianity in a larger framework, without resorting to a somewhat arbitrary decorrelation constraint. A key result is that the mutual information can be decomposed, under linear transforms, as the sum of two terms: one term expressing the decorrelation of the components and one expressing their non-Gaussianity.

Our results extend the previous understanding of these connections and explain them in the light of information geometry. We also describe the "local geometry" of ICA by re-expressing all our results via a Gram-Charlier expansion by which all quantities of interest are obtained in terms of cumulants.

Keywords: Independent component analysis, source separation, information geometry, mutual information, minimum entropy, non-Gaussianity, cumulant expansions

Introduction

Independent component analysis (ICA) is the decomposition of a random vector in linear components which are "as independent as possible." This emerging technique appears as a powerful generic tool for data analysis and the processing of multi-sensor data recordings. In ICA, "independence" should be understood in its strong statistical sense: it is not reduced to decorrelation (second-order dependence) because, for the purpose of ICA, second order statistics fail to capture important features of a data set, as shown by the fact that there are infinitely many linear transforms which decorrelate the entries of a random vector.

In ICA, the measure of choice for statistical independence is the "mutual information," defined below in Equation (1). Its use as an objective function for ICA has been proposed by Comon (1994) and this choice is strongly supported by the fact that it corresponds to the likelihood criterion when a model of independent components is optimized with respect to all its parameters: the linear transform of interest and the distributions of the underlying components (see Cardoso, 1998).

A fascinating connection exists between maximum independence and minimum entropy: if one *decides* to restrict the search for linear components to components which are *uncorrelated* and have unit variance (this later constraint is just a normalization convention), then the objective of minimizing the mutual information is equivalent to the objective of minimizing the sum of the entropies of all components (see Section 1.2).

The aim of this contribution is to explore the connections between mutual information, entropy and non-Gaussianity in a general framework, without imposing the somewhat arbitrary decorrelation constraint. Our results extend the previous understanding of these connections with, in Equation (17), the following key point: in ICA, mutual information can be decomposed in two terms, a term measuring decorrelation between the components and a term measuring their Gaussianity. As it turns out, these connections are expressively described in the light of information geometry.

This paper is organized as follows. In Section 1, we recall the traditional definitions related to the information-theoretic objectives of ICA. In Section 2, we consider the problem of approximating the distribution of a vector by a Gaussian distribution and by a product distribution (a distribution with independent entries). This process naturally introduces the relevant information-theoretic definitions for independence, decorrelation and non-Gaussianity and highlights their connections. Their relevance is discussed in the context of ICA. Section 3 illuminates these results by exhibiting their underlying geometric structure in the framework of information geometry. Section 4 describes the *local* (in the vicinity of Gaussian distributions) geometry in terms of cumulants. Additional comments conclude the paper.

1. Information-Theoretic Objectives of ICA

This section briefly recalls the basic definitions for the objective functions considered in ICA. Throughout the paper, we consider an n-dimensional random vector Y. All the sums $\sum_i \cdots$ which appear in the paper are to be read as sums $\sum_{i=1}^{n} \cdots$ over the entries of this vector, and similarly for the products $\prod_i \cdots$.

1.1 Mutual Information or "Dependence"

The *independent component analysis* (ICA) of a random vector Y, as introduced by Comon (1994), consists of finding a basis on which the coordinates of Y are "as independent as possible." A measure of statistical independence is the Kullback-Leibler divergence (KLD) from the distribution $P(Y)$ of Y to the product of its marginal distributions:

$$I(Y) \stackrel{\text{def}}{=} K\big[P(Y), \textstyle\prod_i P(Y_i)\big], \tag{1}$$

where the divergence $K\big[Q,R\big]$ from a probability distribution Q to another distribution R is defined as

$$K\big[Q,R\big] \stackrel{\text{def}}{=} \int_Y Q(y) \log \frac{Q(y)}{R(y)} dy. \tag{2}$$

It is a non-negative quantity: $K\big[Q,R\big] \geq 0$ with equality only if $Q(y)$ and $R(y)$ are identical (more precisely: if they agree everywhere but possibly on sets of zero Q-probability). Thus $I(Y) = 0$ if and only if the distribution of Y is the product of its marginals so that $I(Y)$ does define a measure of independence. For further reference, we recall that the KLD is invariant under invertible transforms:

if Y and Z are two n-dimensional vectors then

$$K\big[P(Y),P(Z)\big] = K\big[P(\mu+TY),P(\mu+TZ)\big] \tag{3}$$

for any fixed invertible $n \times n$ matrix T and any fixed shift μ.

1.2 Decorrelation and Minimum Entropy

In practical implementations of ICA, it is sometimes preferred to determine the best linear transform of Y via a two-step approach: after mean removal, the first step is to "whiten" or to "sphere" the data and the second step is to rotate them to maximize a measure of independence. The whitening step is a simple operation by which the covariance matrix R_Y of Y is made equal to the identity matrix, yielding components which are *uncorrelated* but not necessarily *independent*. The rotation step keeps the covariance of Y equal to the identity, thus preserving the whiteness, hence the decorrelation of the components. This technique is sometimes referred to as the "orthogonal approach." In the orthogonal approach, the overall transformation is often explicitly obtained as the product of a whitening matrix by an orthogonal matrix but for the purpose of analysis, there is no need to consider the specific practical implementation; an equivalent formulation is to say that the particular measure of independence (used in the second step) is to be optimized "under the whiteness constraint," that is, over all the linear transforms such that R_Y is the identity matrix.

By turning the logarithm of the ratio in (2) into a difference of logarithms, one readily finds that

$$I(Y) = \sum_i H(Y_i) - H(Y), \tag{4}$$

where $H(\cdot)$ denotes Shannon's "differential entropy" for continuous random variables:

$$H(Y) = -\int_Y P_Y(y) \log P_Y(y) dy.$$

Recall that applying a linear transform T to a vector adds a term $\log|\det T|$ to its entropy. It follows that the entropy $H(Y)$ remains constant under the whiteness constraint because all the white versions of Y are rotated versions of each other and because $\log|\det T| = 0$ if T is an orthogonal matrix.

Thus the dependence (or mutual information) $I(Y)$ admits a very interesting interpretation under the whiteness constraint since it appears by (4) to be equal, up to the constant term $H(Y)$, to the sum of the marginal entropies of Y. In this sense, ICA is a minimum entropy method *under the whiteness constraint*. It is one purpose of this paper to find out how this interpretation is affected when the whiteness constraint is relaxed.

2. Dependence, Correlation and Gaussianity

This section examines how well (in the Kullback-Leibler divergence) the distribution of a random vector can be approximated by a product distribution (that is, a distribution with independent entries), by a Gaussian distribution, or by a Gaussian-product distribution. In doing so, several key definitions are introduced.

2.1 Product Approximation and Dependence

The divergence $K\left[P(Y), \prod_i Q_i\right]$ from the distribution $P(Y)$ of a given n-vector to the product of any n scalar probability distributions $Q_1, \ldots, Q_n$ can be decomposed as:

$$K\left[P(Y), \prod_i Q_i\right] = K\left[P(Y), \prod_i P(Y_i)\right] + K\left[\prod_i P(Y_i), \prod_i Q_i\right]. \tag{5}$$

This classic property is readily checked by direct substitution of the distributions into the definition of the KL divergence; its geometric interpretation as a Pythagorean identity is given at Section 3.

The first divergence on the right hand side of (5) is the dependence (or mutual information) $I(Y)$ defined in Equation (1). The second divergence on the right hand side involves two product densities so that it decomposes into the sum of the divergences between the marginals:

$$K\left[\prod_i P(Y_i), \prod_i Q_i\right] = \sum_i K\left[P(Y_i), Q_i\right].$$

It follows that the KLD from a distribution $P(Y)$ to any product distribution can be read as

$$K\left[P(Y), \prod_i Q_i\right] = I(Y) + \sum_i K\left[P(Y_i), Q_i\right]. \tag{6}$$

This decomposition shows that $K\left[P(Y), \prod_i Q_i\right]$ is minimized with respect to the scalar distributions $Q_1, \ldots, Q_n$ by taking $Q_i = P(Y_i)$ for $i = 1, \ldots, n$ because this choice ensures that $K\left[P(Y_i), Q_i\right] = 0$ which is the smallest possible value of a KLD. Thus, the product distribution

$$P^P(Y) \stackrel{\text{def}}{=} \prod_i P(Y_i)$$

appearing in Definition (1) is the best "product approximation" to $P(Y)$ in that it is the distribution with independent components which is closest to $P(Y)$ in the KLD sense.

2.2 Gaussian Approximation and Non-Gaussianity

Besides independence, another distributional property plays an important role in ICA: normality. Denoting $\mathcal{N}(\mu, R)$ the Gaussian distribution with mean μ and covariance matrix R, the divergence $K\left[P(Y), \mathcal{N}(\mu, R)\right]$ from the distribution of a given n-vector Y to a Gaussian distribution $\mathcal{N}(\mu, R)$ can be decomposed into

$$K\left[P(Y), \mathcal{N}(\mu, R)\right] = K\left[P(Y), \mathcal{N}(\mathrm{E}Y, R_Y)\right] + K\left[\mathcal{N}(\mathrm{E}Y, R_Y), \mathcal{N}(\mu, R)\right], \tag{7}$$

where R_Y denotes the covariance matrix of Y. Just as in (5), this decomposition is readily established by substitution of the relevant densities in the definition of the KLD or, again, as an instance of a general Pythagorean theorem (see Section 3).

Decomposition (7) shows that $K\left[P(Y), \mathcal{N}(\mu, R)\right]$ is minimized with respect to μ and R for $\mu = \mathrm{E}Y$ and $R = R_Y$ since this choice cancels the (non-negative) second term in Equation (7) and since the first term does not depend on μ or R. Thus,

$$P^G(Y) \stackrel{\text{def}}{=} \mathcal{N}(\mathrm{E}Y, R_Y)$$

is, not surprisingly, the best Gaussian approximation to $P(Y)$ in the KLD sense. The KL divergence from the distribution of Y to its best Gaussian approximation is taken as the definition of the (non) *Gaussianity* $G(Y)$ of vector Y:

$$G(Y) \stackrel{\text{def}}{=} K\left[P(Y), P^G(Y)\right]. \tag{8}$$

This definition encompasses the scalar case: the non-Gaussianity of each component of Y_i is

$$G(Y_i) = K\big[P(Y_i), \mathcal{N}(\mathrm{E}Y_i, \mathrm{var}(Y_i))\big]. \tag{9}$$

For further reference, we note that the non-Gaussianity of a vector is invariant under invertible affine transforms. Indeed if vector Y undergoes an invertible affine transform: $Y \to \mu + TY$ (where T is a fixed invertible matrix and μ is a fixed vector), then its best Gaussian approximation undergoes the same transform so that

$$G(\mu + TY) = G(Y). \tag{10}$$

This is because the KLD itself is invariant under affine transforms as recalled in Equation (3).

2.3 Gaussian-Product Approximation and Correlation

Looking forward to combining the results of the previous two sections, we note that the product approximation to $P(Y)$ can be further approximated by its Gaussian approximation and vice versa. The key observation is that the same result is obtained along both routes: the Gaussian approximation to the product approximation, on one hand, and the product approximation to the Gaussian approximation, on the other hand, are readily seen to be the same distribution

$$P^{P\wedge G}(Y) \stackrel{\text{def}}{=} \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y)),$$

where $\mathrm{diag}(\cdot)$ denotes the diagonal matrix with the same diagonal elements as its argument. The KLD from $P(Y)$ to its best Gaussian-product approximation $P^{P\wedge G}(Y)$ can be evaluated along both routes.

Along the route $P(Y) \to P^P(Y) \to P^{P\wedge G}(Y)$, we use decomposition (6) with $Q_i = \mathcal{N}(\mathrm{E}Y_i, \mathrm{var}(Y_i))$. This choice ensures that $\prod_i Q_i = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$ and, in addition, by Definition (9), $K\big[P(Y_i), Q_i\big]$ is nothing but the non-Gaussianity $G(Y_i)$ of the i-th entry of Y. Thus, property (6) yields a decomposition

$$K\big[P(Y), P^{P\wedge G}(Y)\big] = I(Y) + \sum_i G(Y_i) \tag{11}$$

into well identified quantities: dependence and marginal (non) Gaussianities.

Along the route $P(Y) \to P^G(Y) \to P^{P\wedge G}(Y)$, instantiating property (7) with $\mu = \mathrm{E}Y$ and $R = \mathrm{diag}(R_Y)$ yields

$$K\big[P(Y), P^{P\wedge G}(Y)\big] = G(Y) + C(Y), \tag{12}$$

where we use Definition (8) of the non-Gaussianity and where we introduce the *correlation* $C(Y)$ of a vector Y, defined as

$$C(Y) \stackrel{\text{def}}{=} K\big[\mathcal{N}(\mathrm{E}Y, R_Y), \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))\big] = K\big[P^G(Y), P^{P\wedge G}(Y)\big]. \tag{13}$$

The scalar $C(Y)$ appears as an information-theoretic measure of the overall correlation between the components of Y. In particular, $C(Y) \geq 0$ with equality only if the two distributions in (13) are identical, *i.e.* if R_Y is a diagonal matrix. The correlation $C(Y)$ has an explicit expression as a function of the covariance matrix R_Y: standard computations yields

$$C(Y) = \frac{1}{2}\mathrm{off}(R_Y) \quad \text{where} \quad \mathrm{off}(R) \stackrel{\text{def}}{=} \log(\det(\mathrm{diag}(R))) - \log(\det(R)).$$

The function $R \rightarrow \text{off}(R)$ is sometimes used as a measure of diagonality of a positive matrix R. This correlation $C(Y)$ is invariant under the shifting or rescaling of any entry of Y. If we define a standardized vector $\tilde{Y}$ by

$$\tilde{Y} = \text{diag}(R_Y)^{-\frac{1}{2}}(Y - \text{E}Y) \quad \text{or} \quad \tilde{Y}_i = (Y_i - \text{E}Y_i)/\text{var}^{\frac{1}{2}}(Y_i), \tag{14}$$

then $C(Y) = C(\tilde{Y})$ so that the correlation actually depends on the covariance matrix only via the *correlation* matrix $R_{\tilde{Y}}$ of Y. We have $C(Y) = \frac{1}{2}\text{off}(R_{\tilde{Y}})$. In particular, if Y is a weakly correlated vector, *i.e.* $R_{\tilde{Y}}$ is close to the identity matrix, a first order expansion yields

$$C(Y) \approx \frac{1}{2} \sum_{1 \leq i < j \leq n} \rho_{ij}^2 \qquad \text{where} \qquad \rho_{ij} \stackrel{\text{def}}{=} \text{E}\tilde{Y}_i\tilde{Y}_j = \text{corr}(Y_i, Y_j). \tag{15}$$

In general, however, our definition of correlation is not a quadratic function of the correlation coefficients ρ_{ij}. For instance, in the simple case $n = 2$, one finds $C(Y) = -\frac{1}{2}\log(1 - \rho_{12}^2)$.

2.4 Dependence, Correlation and Non-Gaussianity

Two different expressions of the Kullback divergence from the distribution $P(Y)$ of a random vector to its closest Gaussian-product approximation $P^{P \wedge G}(Y)$ have been obtained at equations (11) and (12). Equating them, we find a general relationship between dependence, correlation and non Gaussianity:

$$I(Y) + \sum_i G(Y_i) = G(Y) + C(Y). \tag{16}$$

We note that all the quantities appearing in (16) are invariant under the rescaling or shifting of any of the entries of Y.

When Y is a Gaussian vector, then $G(Y) = 0$ and its marginals are also Gaussian: $G(Y_i) = 0$ for $i = 1, n$. In this case, property (16) reduces to $I(Y) = C(Y)$, that is, the obvious result that the dependence is measured by the correlation in the Gaussian case. Relation (16) also shows that, for a non-Gaussian vector, the difference $I(Y) - C(Y)$ between dependence and correlation is equal to the difference between the joint non-Gaussianity $G(Y)$ and the sum of the marginal non-Gaussianities (more about this difference in Section 3.4).

The most significant insight brought by property (16) regards the ICA problem in which the dependence $I(Y)$ is to be minimized under *linear* transforms. We saw in Equation (10) that the non-Gaussianity $G(Y)$ is invariant under linear transforms so that property (16) also yields

$$I(Y) = C(Y) - \sum_i G(Y_i) + \text{cst}, \tag{17}$$

where "cst" is a term which is constant over all linear transforms of Y (and actually is nothing but $G(Y)$). Hence, the issue raised in the introduction receives a clear and simple answer:

> Minimizing under *linear* transforms the dependence $I(Y)$ between the entries of a vector Y is equivalent to optimizing a criterion which weights in equal parts the correlation $C(Y)$ of the components and (the opposite of) their non-Gaussianities $G(Y_1), \ldots, G(Y_n)$.

In other words, *linear* components which are as independent as possible are components which are as uncorrelated and as non-Gaussian as possible, this statement being quantified by expression (17).

2.5 Objective Functions of ICA

Expression (17) nicely splits dependence into correlation and non-Gaussianity. It lends itself to a simple generalization. Let w be a positive number and consider the weighted criterion

$$\phi_w(Y) = wC(Y) - \sum_i G(Y_i) \tag{18}$$

to be optimized under linear transforms of Y. Different weights correspond to different variations around the idea of ICA:

- For $w = 1$, the minimization of ϕ_w under linear transforms is equivalent to minimizing the dependence $I(Y)$.
- For $w = 0$, the criterion is a sum of *uncoupled* criteria and the problem boils down to independently finding directions in data space showing the maximum amount of non Gaussianity. This is essentially the rhetoric of projection pursuit, initiated by Friedman and Tukey (1974) and whose connections to ICA have long been noticed and put to good use for justifying sequential extraction of sources as in the fastICA algorithm (see Hyvärinen, 1998).
- For $w \to \infty$, the criterion is dominated by the correlation term $C(Y)$. Thus, for w large enough, $\phi_w(Y)$ is minimized by linear transforms such that $C(Y)$ is arbitrarily close to 0. The latter condition leaves many degrees of freedom; in particular, since all the quantities in (18) are scale invariant, one can freely impose $\mathrm{var}(Y_i) = 1$ for $i = 1, n$. These normalizing conditions together with $C(Y) = 0$ are equivalent to enforcing the whiteness of Y. Thus, for w large enough, the unconstrained minimization of $\phi_w(y)$ is equivalent to maximizing the sum $\sum_i G(Y_i)$ of the marginal non-Gaussianities under the whiteness constraint. This objective of maximal marginal non-Gaussianity and the objective of minimum marginal entropy recalled in Section 1.2 should be identical since they both stem from the minimum dependence objective under the whiteness constraint. Indeed, it is easily seen that if $\mathrm{var}(Y_i) = 1$ then $H(Y_i) = -G(Y_i) + \frac{1}{2}\log 2\pi e$.

3. Geometry of Dependence

This section shows how the previous results fit in the framework of information geometry. Information geometry is a theory which expresses the concepts of statistical inference in the vocabulary of differential geometry. The "space" in information geometry is a space of probability distributions where each "point" is a distribution. In the context of ICA, the distribution space is the set of all possible probability distributions of an n-vector. Two subsets (or *manifolds*) will play an important role: the *Gaussian manifold*, denoted as $\mathcal{G}$, is the set of all n-variate Gaussian distributions and the *product manifold*, denoted as $\mathcal{P}$, is the set of all n-variate distributions with independent components. Statistical statements take a geometric flavor. For instance, "Y is normally distributed" is restated as "the distribution of Y lies on the Gaussian manifold" or "$P(Y) \in \mathcal{G}$."

The full theory of information geometry has a rich duality structure which has been elucidated mostly by Amari (1985). In this paper however, only simple notions of information geometry are needed: we essentially rely on the notion of "exponential families of distributions" and on an associated Pythagorean theorem in which the Kullback-Leibler divergence plays a role similar to a squared distance in Euclidean geometry. These two notions are briefly recalled below. The

geometric view point has already been used in ICA to analyze contrast functions (see Cardoso, 2000) and estimating equations (see Amari and Cardoso, 1997).

3.1 Simple Facts from Information Geometry

We start by some simple notions of information geometry. For the sake of exposition, technical complications are ignored. This will be even more true in Section 4.

Exponential families. The product manifold $\mathcal{P}$ and the Gaussian manifold $\mathcal{G}$ both share an important property which makes geometry simple: they are *exponential families* of probability distributions. An exponential family has the characteristic property that it contains the *exponential segment* between any two of its members. The exponential segment between two distributions with densities $p(x)$ and $q(x)$ (with respect to some common dominating measure) is the one-dimensional set of distributions with densities

$$p_\alpha(x) = p(x)^{1-\alpha} q(x)^\alpha e^{-\psi(\alpha)}, \qquad 0 \le \alpha \le 1,$$

where $\psi(\alpha)$ is a normalizing factor such that p_α sums to 1. Thus, the log-density of exponential families has a simple form. An L-dimensional exponential family of distributions can be parameterized by a set $\alpha = (\alpha_1, \ldots, \alpha_L)$ of real parameters as

$$\log p(x;\alpha) = \log g(x) + \sum_{l=1}^{L} \alpha_l S_l(x) - \psi(\alpha), \tag{19}$$

where $g(x)$ is some reference measure (not necessarily a probability measure), $S_l(x)$ are scalar functions of the variable and $\psi(\alpha)$ is such that $\int p(x;\alpha)dx = 1$.

The Gaussian manifold $\mathcal{G}$ is a p-dimensional exponential manifold with $L = n + n(n+1)/2$: one can take $g(x)$ to be any n-variate Gaussian distribution and

$$\{S_l(x)\}_{l=1}^{L} = \{x_i | 1 \le i \le n\} \cup \{x_i x_j | 1 \le i \le j \le n\}. \tag{20}$$

The product manifold $\mathcal{P}$ also is an exponential manifold but it is infinite dimensional since the marginal distributions can be freely chosen. Similarly to Equation (19), any distribution $p(x)$ of $\mathcal{P}$ can be given a log-density in the form

$$\log p(x) = \log g(x) + \sum_{i=1}^{n} r_i(x_i) - \psi(r_1, \ldots, r_n), \tag{21}$$

where $g(x)$ is some product distribution of $\mathcal{P}$ and each functions $r_i(\cdot)$ can be freely chosen (as long as the sum of its exponential remains finite). If $g(x)$ is the Lebesgue measure and the $r_i(x_i)$ are log-densities, then $\psi = 0$.

A Pythagorean theorem. Exponential families play a central role in statistics as well as in information geometry where they behave to some extent like "flat manifolds," (see Amari, 1985). In particular, they give rise to a Pythagorean theorem illustrated by Figure 1. This theorem states that for an exponential family $\mathcal{E}$ of distributions and for any distribution P (not necessarily in $\mathcal{E}$), there is a unique distribution Q of $\mathcal{E}$ such that the divergence from P to any distribution R of $\mathcal{E}$ decomposes as

$$K\big[P,R\big] = K\big[P,Q\big] + K\big[Q,R\big]. \tag{22}$$

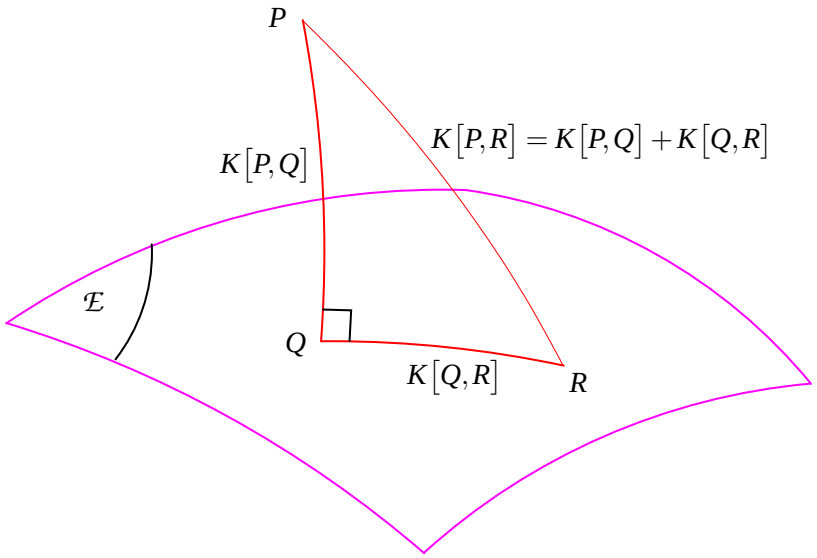

Figure 1: The Pythagorean theorem in distribution space. The distribution Q closest to P in an exponential family $\mathcal{E}$ is the "orthogonal projection" of P onto $\mathcal{E}$. The divergence $K\left[P,R\right]$ from P to any other distribution R of $\mathcal{E}$ can be decomposed into $K\left[P,R\right] = K\left[P,Q\right] + K\left[Q,R\right]$, *i.e.* a normal part and a tangent part.

This decomposition shows that Q is the closest distribution to P in $\mathcal{E}$ in the KLD sense since $K\left[Q,R\right]$ reaches its minimum value of 0 for $Q = R$. Thus, Q can be called the "orthogonal projection" of P onto $\mathcal{E}$ and decomposition (22) should be read as a Pythagorean theorem in distribution space with the KL divergence being the analogue of a squared Euclidean distance.

An important remark is that all distributions P projecting at point Q in an exponential manifold characterized by functions $S_l(x)$ as in (19) are such that

$$\mathrm{E}_P S_l(x) = \mathrm{E}_Q S_l(x) \quad l = 1, \ldots, p.$$

In particular, all distributions projecting at a given point on $\mathcal{G}$ have the same first and second order moments according to (20) and all distributions projecting at a given point on $\mathcal{P}$ have the same marginal distributions. This is the information-geometric interpretation of the results recalled in Sections 2.1 and 2.2.

On "lengths." We offer a word of warning about the Pythagorean theorem and information geometry in general: strictly speaking, the KL divergence $K\left[P,Q\right]$ from distribution P to distribution Q should *not* be interpreted as a (squared) "length" because the divergence is not a distance (it is not symmetric in general). Even when dealing with segments from P to Q, the divergence does not result (in general) from integrating an infinitesimal length along some path from P to Q in distribution space. In the following, the word "length" is sometimes used nonetheless, but always between quotes to remind the reader that the Pythagorean theorem actually is a point-to-point relation.

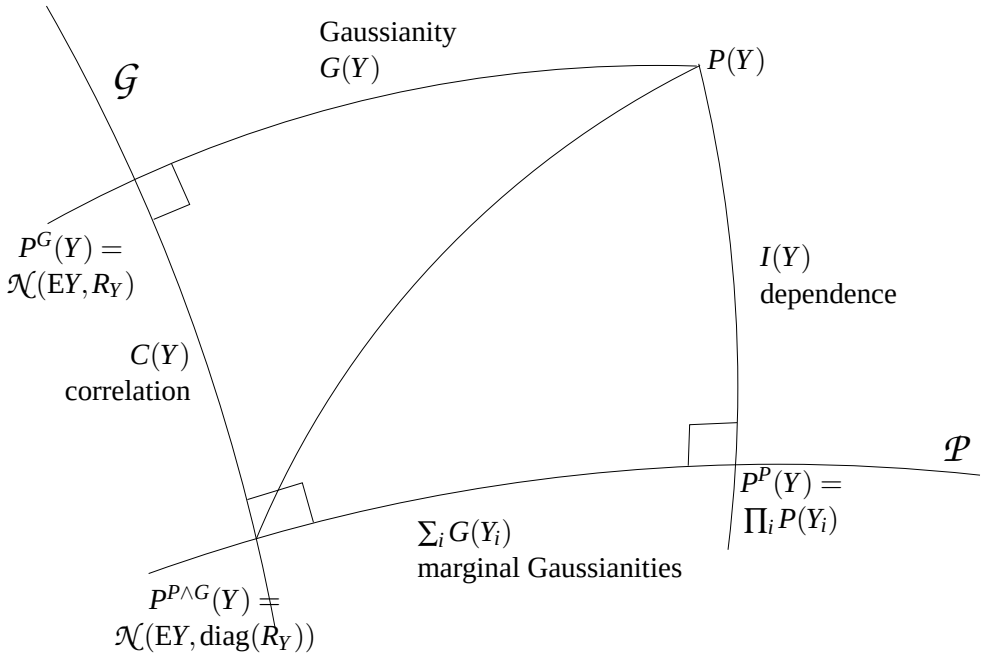

Figure 2: The geometry of decomposition (16) in distribution space (information geometry). The property $I(Y) + \sum_i G(Y_i) = G(Y) + C(Y)$ is found by applying a Pythagorean theorem to two right-angled triangles with a common hypotenuse: the segment from $P(Y)$ to its best Gaussian-product approximation $\mathcal{N}(\mathrm{diag}(R_Y))$.

3.2 Geometry of Dependence and Non-Gaussianity

Information geometry allows to represent the various relations encountered so far in a single synthetic picture (see Figures 2 and 3). Again, each "point" in the figure is a probability distribution of an n-vector. The product manifold $\mathcal{P}$, and the Gaussian manifold $\mathcal{G}$ are shown as one-dimensional manifolds but their actual dimensions are (much) larger, indeed.

As seen in Section 2, the minimum KL divergence from an arbitrary distribution $P(Y)$ to the product manifold and to the Gaussian manifold is reached at points $P^P(Y) = \prod_i P(Y_i)$ and $P^G(Y) = \mathcal{N}(\mathrm{E}Y, R_Y)$ respectively. These two distributions are the "orthogonal projections" of $P(Y)$ onto $\mathcal{P}$ and $\mathcal{G}$ respectively. The dependence $I(Y)$ and the non-Gaussianity $G(Y)$ of Y are read as divergences from $P(Y)$ to its projections and the corresponding edges are labelled accordingly. The two projections of $P(Y)$ on each manifold are further projected onto each other manifold, reaching the point $P^{P\wedge G}(Y) = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$ in each case. The corresponding edges are labelled as $C(Y)$ and $\sum_i G(Y_i)$.

The Pythagorean theorem can be applied to both triangles, taking $P = P(Y)$ and $R = P^{P\wedge G}(Y) = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$ and appropriate values for $\mathcal{E}$, Q as follows. On the Gaussian side, we take $\mathcal{E} = \mathcal{G}$ and $Q = P^G(Y) = \mathcal{N}(\mathrm{E}Y, R_Y)$; then the Pythagorean decomposition (22) corresponds to Equation (12). On the product side, we take $\mathcal{E} = \mathcal{P}$ and $Q = P^P(Y) = \prod_i P(Y_i)$; then the Pythagorean decomposition (22) corresponds to Equation (11).

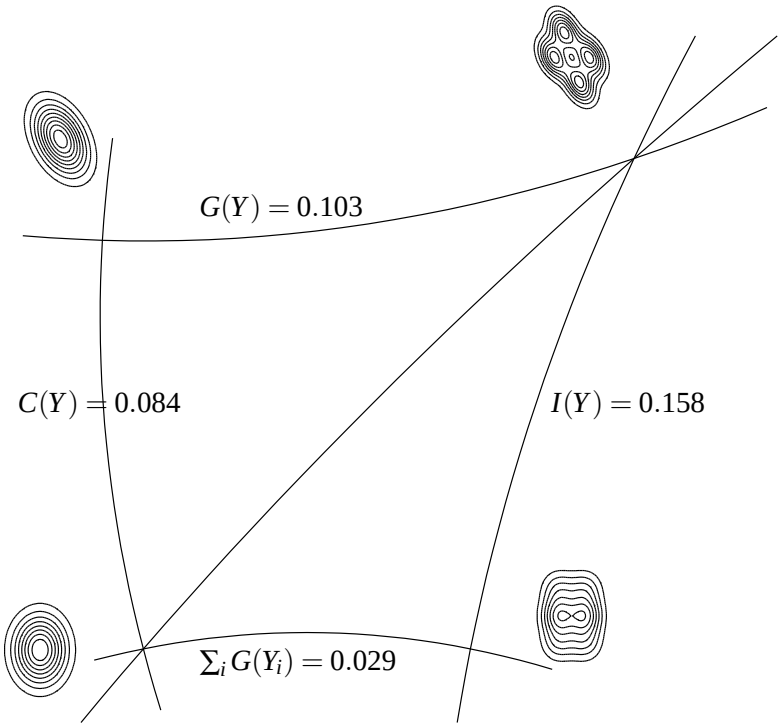

Figure 3: An example of the two-way decomposition. Kullback-Leibler divergences are expressed in "nats" and not in "bits," *i.e.*, they are computed using the natural logarithm.

In summary, our key result (16) relating dependence, correlation and Gaussianity admits a simple geometric picture: the quantities of interest are the "lengths" of the edges of two "right-angled" triangles; these "lengths" are related because the two triangles share the same hypotenuse.

The generic picture is displayed on Figure 2; a specific example of a 2-dimensional distribution is shown on Figure 3. For $P(Y)$ we take the distribution of $Y = CS$ where C is an 2×2 linear transform and the entries of the 2×1 vector S are independent with bimodal (marginal) distributions. Contour plots for the densities are shown as well as the specific values of the divergences. A weak (and futile) effort has been made to have the squared lengths of the edges on the figure be roughly proportional to the corresponding KL divergences.

3.3 Some Comments

Dependence and marginal Gaussianity as projections. The dependence I is defined as the divergence from P to P^P and the correlation C as the divergence from P^G to $P^{P \wedge G}$. Thus, both ends of the segment $P \to P^P$ which defines dependence are projected on the Gaussian manifold to the corresponding ends of the segment $P^G \to P^{P \wedge G}$ which defines correlation. In short, the "dependence segment" projects on $\mathcal{G}$ to the "correlation segment." This is perfectly in line with the fact, already noted, that dependence reduces to correlation in a Gaussian model. Just as unsurprisingly, the joint

Gaussianity $G(Y)$ reduces to marginal Gaussianity $\sum_i G(Y_i)$ when projecting onto the manifold $\mathcal{P}$ of independent components.

Thus correlation and marginal Gaussianity can be respectively understood as projection of the dependence on the Gaussian manifold and as projection of joint Gaussianity on the product manifold.

If we were dealing in Euclidean geometry, we would conclude that the projected segments have a shorter "length." Actually, the variation of "length" is the same for both $G(Y)$ and $I(Y)$ since Equation (16) also reads

$$I(Y) - C(Y) = G(Y) - \sum_i G(Y_i), \tag{23}$$

showing that the loss from full dependence $I(Y)$ to correlation $C(Y)$ equals the loss from joint Gaussianity $G(Y)$ to the sum $\sum_i G(Y_i)$ of marginal Gaussianities. However, we are dealing with a curved geometry here: it is not true that projecting onto an exponential family always reduces the KL divergence.[1] This is not even true in the specific case under investigation, as seen next.

A failed conjecture. Even though correlation is the projection of dependence onto the Gaussian manifold, the conjecture that $I(Y) \geq C(Y)$ is not true in general. A simple counter-example[2] is for a two-dimensional Y distributed as $Y = \begin{bmatrix} 1 \\ 1 \end{bmatrix} S + \sigma N$ where S is a binary random variable taking values 1 and -1 with equal probability and N is a $\mathcal{N}(0, I)$ variable independent of S. It is not difficult to see that for small enough σ, the dependence $I(Y)$ is close to 1 bit while the correlation $C(Y)$ diverges to infinity as $-\log\sigma^2$. Thus $C(Y) > I(Y)$ for small enough σ, disproving the conjecture. Subsection 3.4 investigates this issue further by developing a finer geometric picture.

Location-scale invariance. Figure 2 and following show the Gaussian manifold $\mathcal{G}$ and the product manifold $\mathcal{P}$ intersecting at a single point. In fact $\mathcal{G} \wedge \mathcal{P}$ is a $2n$-dimensional manifold since it contains all uncorrelated Gaussian distributions (n degrees of freedom for the mean and n degrees of freedom for the variance of each component). These degrees of freedom correspond to location-scale transformations: $Y \rightarrow \mu + DY$ with μ a fixed (deterministic) arbitrary n-vector and with D a fixed arbitrary invertible diagonal $n \times n$ matrix. If Y undergoes such a transform, so do its product, Gaussian and Gaussian-product approximations so that the quantities $I(Y)$, $C(Y)$, $G(Y)$, and $G(Y_i)$ are invariant under location-scale transform (because of the invariance (3) of the KL divergence).

Thus the whole geometric picture is strictly invariant under location-scale changes of Y: these additional dimensions need not be represented in our figures. Each figure is a "cut" with constant location-scale parameters across the whole distribution space. All these cuts, which are orthogonal (in some sense) to the location-scale directions, show the same picture.

Orthogonality. So far, we have used the term "orthogonal" mostly in "orthogonal projection." Figures like 2 are decorated with the symbol of a right angle at point $P^P(Y)$ for instance, even though we have not defined the segment from $P(Y)$ to $P^P(Y)$ which would be "orthogonal" to the manifold at this point. The intuitive reader may also feel that the "double Pythagoras' used above may depend on $\mathcal{P}$ and $\mathcal{G}$ being "orthogonal" at $P^{P \wedge G}(Y)$.

Actually, the product manifold $\mathcal{P}$ and the Gaussian manifold $\mathcal{G}$ do meet orthogonally along their intersection, the location-scale family discussed above. Unfortunately, in order to discuss this

1. We are grateful to Shun-ichi Amari for providing a simple counter-example.
2. We are grateful to Mark Plumbley for this counter-example.

point, one needs to define the notion of tangent plane to a statistical manifold and the notion of orthogonality in information geometry. To keep this article flowing, explaining these notions is deferred to appendix A which only contains a brief introduction to the topic. The notion of tangent plane is also used in Section 4.

Note that there is no need to prove that $\mathcal{P}$ and $\mathcal{G}$ intersect orthogonally to reach the conclusions of previous sections: one only has to realize that the product-Gaussian approximation and the Gaussian-product approximation to $P(Y)$ "happen" to be identical.

3.4 Capturing Second-order and Marginal Structures

We have discussed the best approximation of an n-variate distribution by a Gaussian distribution or by an independent distribution. We now take a look at combining these two aspects.

Figures like 2 or 3 are misleading if they suggest that the distributions $P(Y)$, $P^G(Y)$, $P^P(Y)$ and $P^{P\wedge G}(Y)$ are somehow "coplanar." It is in fact interesting to introduce the "plane" spanned by the product manifold $\mathcal{P}$ and the Gaussian manifold $\mathcal{G}$. To this effect, consider the family of probability distributions for an n-vector with a log-density in the form

$$\log p(y) = \log \phi(y) + \sum_i a_i y_i + \sum_{ij} a_{ij} y_i y_j + \sum_i r_i(y_i) - \psi, \tag{24}$$

where $\phi(y)$ is the n-variate standard normal distribution $\mathcal{N}(0,I)$, where a_i $(i = 1,n)$ and a_{ij} $(i,j = 1,n)$ are real numbers, where $r_i(\cdot)$ $(i = 1,n)$ are real functions and where ψ is a number such that $\int p(y) = 1$. This number depends on the a_i's, on the a_{ij}'s and on the $r_i(\cdot)$'s. By construction, the set of probability distributions generated by Definition (24) forms an exponential family. It contains both the Gaussian manifold $\mathcal{G}$ (characterized by $r_i(\cdot) = 0$ for $i = 1,n$) and the product manifold $\mathcal{P}$ (characterized by $a_{ij} = 0$ for $1 \le i \ne j \le n$) (compare to (19-20) and to (21) respectively). It is actually the smallest exponential family containing both $\mathcal{P}$ and $\mathcal{G}$; it could also be constructed as the set of all exponential segments between any distribution of $\mathcal{P}$ and any distribution of $\mathcal{G}$. Thus, this manifold can be thought of as the "exponential span" of $\mathcal{P}$ and $\mathcal{G}$ and for this reason, it is denoted as $\mathcal{P} \vee \mathcal{G}$.

Let $P^{P\vee G}(Y)$ denote the projection of $P(Y)$ onto $\mathcal{P} \vee \mathcal{G}$, *i.e.*

$$P^{P\vee G}(Y) = \arg \min_{Q\in\mathcal{P}\vee\mathcal{G}} K\big[P(Y), Q(Y)\big].$$

Since $\mathcal{P} \vee \mathcal{G}$ is an exponential family, we have yet another instance of the Pythagorean theorem:

$$\forall R \in \mathcal{P} \vee \mathcal{G} \quad K\big[P,R\big] = K\big[P,P^{P\vee G}\big] + K\big[P^{P\vee G},R\big]. \tag{25}$$

It follows that the distribution R of $\mathcal{P}$ which minimizes $K\big[P,R\big]$ is the same as the distribution R of $\mathcal{P}$ which minimizes $K\big[P^{P\vee G},R\big]$. Thus $P^{P\vee G}(Y)$ projects onto $\mathcal{P}$ at the same point as $P(Y)$, namely at point $P^P(Y) = \prod_i P_i(Y_i)$. Similarly, $P^{P\vee G}(Y)$ projects onto $\mathcal{G}$ at the same point as $P(Y)$, that is, at point $P^G(Y) = \mathcal{N}(\mathrm{E}Y, R_Y)$. It follows that $P^{P\vee G}(Y)$ does have the same marginal distributions *and* the same first and second order moments as $P(Y)$.

Our geometric picture can now be enhanced as shown on Figure 4. The figure displays the new distribution $P^{P\vee G}$ and also the "right angles" resulting from the projection of $P(Y)$ onto $\mathcal{P} \vee \mathcal{G}$, onto $\mathcal{P}$ and onto $\mathcal{G}$ and those resulting from the projection $P^{P\vee G}$ onto $\mathcal{G}$ and $\mathcal{P}$.

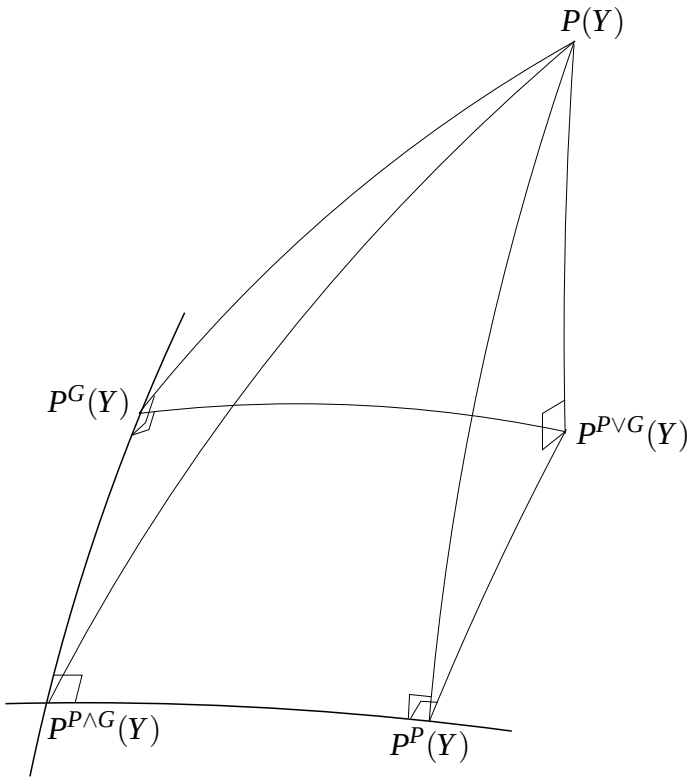

Figure 4: Projection onto the "exponential plane" $\mathcal{P}\vee\mathcal{G}$ spanned by the product manifold $\mathcal{P}$ and the Gaussian manifold $\mathcal{G}$. The projected distribution $P^{P\vee G}(Y)$ further projects on $\mathcal{P}$ at $P^P(Y)=\prod_i P(Y_i)$ and on $\mathcal{G}$ at $P^G(Y)=\mathcal{N}(\mathrm{E}Y,R_Y)$ as $P(Y)$ itself does.

Instantiating the Pythagorean decomposition (25) with $R(Y)=P^P(Y)=\prod_i P(Y_i)$ and with $R(Y)=P^G(Y)=\mathcal{N}(\mathrm{E}Y,R_Y)$ respectively yields

$$I(Y) = K\left[P,P^P\right]=K\left[P,P^{P\vee G}\right]+K\left[P^{P\vee G},P^P\right]=K\left[P,P^{P\vee G}\right]+I(P^{P\vee G}), \quad (26)$$

$$G(Y) = K\left[P,P^G\right]=K\left[P,P^{P\vee G}\right]+K\left[P^{P\vee G},P^G\right]=K\left[P,P^{P\vee G}\right]+G(P^{P\vee G}). \quad (27)$$

These identities are to be interpreted as follows: distribution $P^{P\vee G}$ is the simplest approximation to $P(Y)$ in the sense that its log-density contains only terms as those in (24); it captures its marginal structure and its first and second order structure; it is however less dependent and more Gaussian than the original distribution $P(Y)$ with the same quantity $K\left[P,P^{P\vee G}\right]$ "missing" from the dependence and from the (non) Gaussianity. An explicit example is displayed in Figure 5.

With the construction of $P^{P\vee G}$, we now have four distributions $P^{P\wedge G}$, P^G, P^P and $P^{P\vee G}$ which are coplanar in an information-geometric sense. Also, by construction, the polygon formed by these four distributions has "right angles" at points $P^{P\wedge G}$, P^G and P^P. Thus, if we were dealing with an Euclidean geometry, this polygon would be a rectangle. This is definitely not the case in information geometry. However, a curved geometry can be approximated by a flat (Euclidean) geometry over

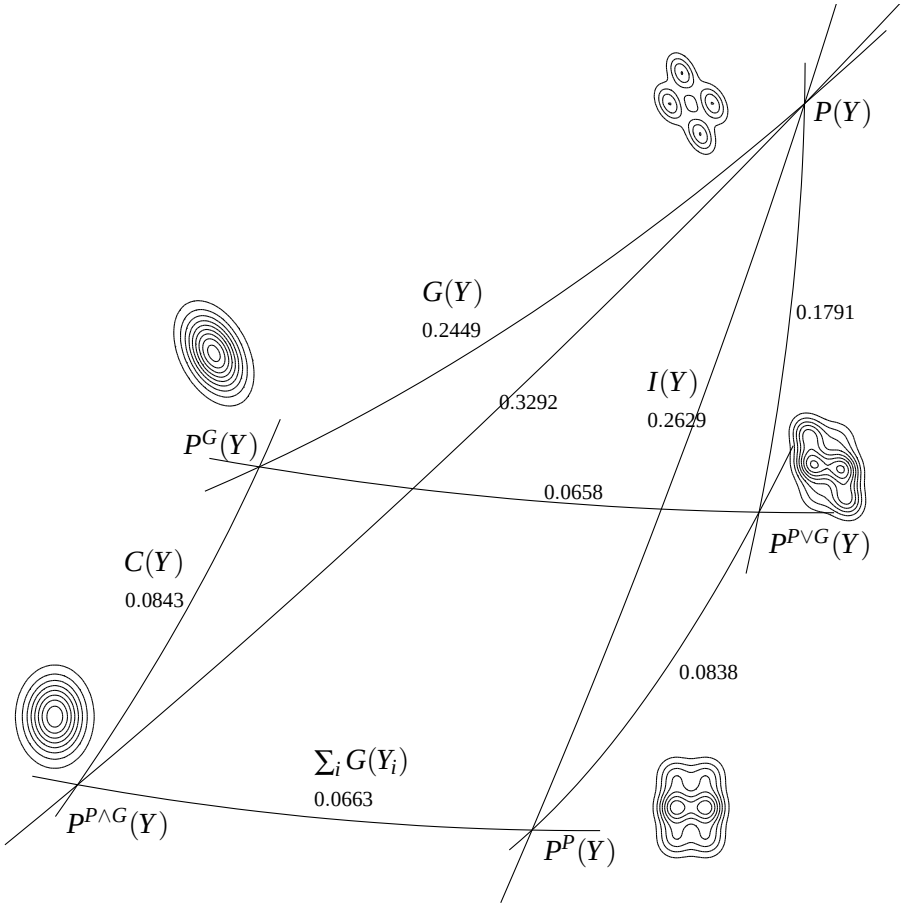

Figure 5: Projection onto $\mathcal{P} \vee \mathcal{G}$. The distribution $P(Y)$, with Gaussianity $G(Y) \approx 0.2449$ nats and dependence $I(Y) \approx 0.2629$ nats is approximated by $P^{P\vee G}(Y)$ which has the same mean, same covariance and same marginal distributions as Y but reduced Gaussianity and dependence: $G(P^{P\vee G}) \approx 0.0658$ nats and $I(P^{P\vee G}) \approx 0.0838$ nats. The reduction of both quantities is $K\left[P, P^{P\vee G}\right] \approx 0.1791$ nats.

small enough neighborhoods. This is the case here if the divergence of $P(Y)$ from $P^{P\wedge G}(Y)$ is small enough. In this approximation, points P, P^G, P^P and $P^{P\vee G}$ can be assimilated to the vertices of a rectangle, in which case the opposite edges of the rectangle approximately have identical "lengths."

$$K\left[P^{P\vee G}, P^G\right] \approx K\left[P^P, P^{P\wedge G}\right] \quad i.e. \quad G(P^{P\vee G}) \approx \sum_i G(Y_i),$$
$$K\left[P^{P\vee G}, P^P\right] \approx K\left[P^G, P^{P\wedge G}\right] \quad i.e. \quad I(P^{P\vee G}) \approx C(Y).$$

These approximations, combined with (26-27) then yields, for a weakly correlated, weakly non-Gaussian vector Y:

$$\begin{aligned} I(Y) &\approx C(Y) + K\left[P, P^{P\vee G}\right], \\ G(Y) &\approx \sum_i G(Y_i) + K\left[P, P^{P\vee G}\right]. \end{aligned}$$

Thus, in the limit of weakly correlated, weakly non-Gaussian distributions, correlation is smaller than dependence and marginal non-Gaussianity is smaller than joint non-Gaussianity by the same *positive* amount $K\left[P, P^{P\vee G}\right]$ which measures the loss of approximating $P(Y)$ by the simplest distribution with same marginal and Gaussian structure.

Unfortunately, as shown by the counter-example of Section 3.3, this property is not true in full generality that $I(Y) \geq C(Y)$ or, equivalently by (23), that $G(Y) \geq \sum_i G(Y_i)$.

The next section studies in more detail the "local picture" —*i.e.* $P(Y)$ is close to $P^{P\wedge G}(Y) = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$— by relying on a cumulant expansion of the KL divergence.

4. Cumulants and the Local Geometry

Cumulants and other higher order statistics have been a popular and sometimes effective tool to design tractable objective functions for ICA (see for instance Comon (1994), Cardoso (1999)). They can be introduced on an *ad hoc* basis or as approximations to information-theoretic objective functions such as those discussed above. This section shows how the story of previous sections is told in the language of cumulants.

We look at the *local* geometry *i.e.* when the distribution of Y is in the neighborhood of $P^{P\wedge G}(Y) = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$, that is, when Y is weakly correlated and weakly non-Gaussian. In this limit, the various statistical manifolds can be approximated by their respective tangent planes and the Kullback divergence can be approximated by a quadratic measure: the geometry becomes Euclidean. This section is organized as follows: we recall the notion of tangent plane to a statistical model and how the Kullback divergence becomes a quadratic distance in the tangent plane. Next, the tangent plane is equipped with an orthogonal basis: the (multi-dimensional) Hermite polynomials stemming from the Gram-Charlier expansion. On this basis, the coordinates of a distribution are its cumulants. Combining all these results, we give cumulant-based expansions for the dependence $I(Y)$, the Gaussianity $G(Y)$ and the correlation $C(Y)$ allowing us to complement the global non-Euclidean picture of previous sections with a more detailed local Euclidean picture.

Tangent plane. We give an informal account of the construction of the tangent plane to a statistical model. Let $n(x)$ and $p(x)$ be probability densities. Then, the random variable

$$e(x) = \frac{p(x)}{n(x)} - 1 \tag{28}$$

has zero-mean under $n(x)$ since $\int n(x)e(x)dx = \int (p(x) - n(x))dx = 1 - 1 = 0$. If $p(x)$ is "close to $n(x)$," then $e(x)$ is a "small" random variable. Conversely, if $e(x)$ is a "small function" with zero-mean under $n(x)$, then $n(x)(1+e(x))$ is a density which is close to $n(x)$. With this construction, the vector space of random variables with zero mean and finite variance can be identified to the tangent plane $\mathcal{T}_n$ to the set of probability distributions at point $n(x)$. An infinitesimal vector $e(x)$ in this space is mapped to a distribution $n(x)(1+e(x))$ which is infinitesimally close to $n(x)$.

If $p(x)$ and $q(x)$ denote two distributions close to $n(x)$, the second-order expansion of the KLD in powers of $e_p(x) = p(x)/n(x) - 1$ and of $e_q(x) = q(x)/n(x) - 1$ is found to be

$$K\big[p,q\big] \approx K_n\big[p,q\big] \stackrel{\text{def}}{=} \frac{1}{2}\int \left(\frac{q(x)}{n(x)} - \frac{p(x)}{n(x)}\right)^2 n(x)dx = \frac{1}{2}\mathrm{E}_n\{(e_q(x) - e_p(x))^2\}. \tag{29}$$

This is a quadratic form in $e_p - e_q$: the tangent space is equipped with this Euclidean metric, which is the approximation of the KLD in the vicinity of n.

Gram-Charlier expansion. A (multivariate) Gram-Charlier expansion of $p(x)$ around a reference distribution $n(x)$ is an expansion of $p(x)$ in the form

$$e_p(x) = \frac{p(x)}{n(x)} - 1 = \sum_i \eta_i h^i(x) + \frac{1}{2!}\sum_{ij} \eta_{ij} h^{ij}(x) + \frac{1}{3!}\sum_{ijk} \eta_{ijk} h^{ijk}(x) + \cdots, \tag{30}$$

where the η's are coefficients and the h's are fixed functions of x which depend on the choice of the reference point $n(x)$. For this expression of the Gram-Charlier expansion and multivariate Hermite polynomials, see McCullagh (1987).

For our purposes, we only need to consider a simple case in which the reference distribution is the standard n-variate normal density:

$$n(x) = \phi(x) \stackrel{\text{def}}{=} (2\pi)^{-n/2}\exp(-\|x\|^2/2). \tag{31}$$

The h functions in the Gram-Charlier expansion are the multidimensional Hermite polynomials:

$$\begin{aligned} h^i(x) &= x_i, & (32)\\ h^{ij}(x) &= x_i x_j - \delta_{ij}, & (33)\\ h^{ijk}(x) &= x_i x_j x_k - x_i\delta_{jk}[3], & (34)\\ h^{ijkl}(x) &= x_i x_j x_k x_l - x_i x_j \delta_{kl}[6] + \delta_{ij}\delta_{kl}[3], & (35) \end{aligned}$$

and so on, where δ_{ij} is the Kronecker symbol. The bracket notation means that the corresponding term should be repeated for all the relevant index permutations. For instance: $x_i\delta_{jk}[3]$ stands for $x_i\delta_{jk} + x_j\delta_{ki} + x_k\delta_{ij}$ and $x_i x_j\delta_{kl}[6]$ stands for all the partitions of the set (i,j,k,l) into three blocks according to the pattern $(i|j|kl)$. There are 6 such partitions because there are 6 ways of picking two indices out of four: $x_i x_j\delta_{kl}[6] = x_i x_j\delta_{kl} + x_i x_k\delta_{jl} + x_i x_l\delta_{jk} + x_j x_k\delta_{il} + x_j x_l\delta_{ik} + x_k x_l\delta_{ij}$. Similarly, $\delta_{ij}\delta_{kl}[3] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$.

The η coefficients in (30) are the differences between the cumulants of $p(x)$ and the cumulants of $n(x)$. Specifically, if the distribution of x is $p(x)$, we denote

$$\kappa^p_{i_1\ldots i_r} = \mathrm{cum}(x_{i_1},\ldots,x_{i_r}), \tag{36}$$

and the η's are given by

$$\eta_{i_1\ldots i_r} = \kappa^p_{i_1\ldots i_r} - \kappa^n_{i_1\ldots i_r}. \tag{37}$$

Upon choosing $n(x) = \phi(x)$, we have $\kappa^n_i = 0$, $\kappa^n_{ij} = \delta_{ij}$ and all the cumulants of higher order of $n(x) = \phi(x)$ cancel so that the η coefficients for a given distribution $p(x)$ simply are

$$\eta_i = \kappa^p_i, \quad \eta_{ij} = \kappa^p_{ij} - \delta_{ij}, \quad \eta_{ijk} = \kappa^p_{ijk}, \quad \eta_{ijkl} = \kappa^p_{ijkl}, \quad \ldots. \tag{38}$$

Hermite basis. By Equation (28), a distribution $p(x)$ "close" to $\phi(x)$ corresponds to a "small" vector $e_p(x) = p(x)/\phi(x) - 1$. This is a zero-mean (under ϕ) random vector; via the Gram-Charlier expansion (30), this vector can be decomposed on the Hermite polynomials with coefficients $\{\eta_i, \eta_{ij}, \eta_{ijk}, \ldots\}$ of Equation (38). Thus we can see the numbers $\{\eta_i, \eta_{ij}, \eta_{ijk}, \ldots\}$ as the coordinates of $p(x)$ on the Hermite basis. There is a catch though because the Gram-Charlier expansion (30) shows a highly regular form but suffers from a small drawback: it is redundant. This is because both the Hermite polynomials and the parameters $\{\eta_i, \eta_{ij}, \eta_{ijk}, \ldots\}$ are invariant under any permutation of their indices so that most of the terms appears several times in (30). For instance, the term associated to, say, the indices (i, j, k) actually appears $3! = 6$ times if all indices in (i, j, k) are distinct. Thus a non-redundant set of polynomials is the following Hermite basis:

$$\bigcup_{r\geq 1} \left\{ h^{i_1 \ldots i_r}(x) | 1 \leq i_1 \leq \cdots \leq i_r \leq n \right\}. \tag{39}$$

This set of polynomials is orthogonal: two Hermite polynomials of different orders are uncorrelated under $n(x) = \phi(x)$ while the correlation of two Hermite polynomials of same order is:

$$\int h^{i_1 i_2 \cdots i_r}(x) h^{j_1 j_2 \cdots j_r}(x) \phi(x) dx = \delta_{i_1 j_1} \delta_{i_2 j_2} \cdots \delta_{i_r j_r} [r!]. \tag{40}$$

Again, the bracket notation $[r!]$ means a sum over all index permutations (and there are $r!$ of them at order r). If two index sequences $i_1 i_2 \cdots i_r$ and $j_1 j_2 \cdots j_r$ are ordered as in (39) then either they are identical or they correspond to orthogonal polynomials according to (40).

Small cumulant approximation to the Kullback divergence. Combining the orthogonality property (40) and the expression (37) of the coefficients as a difference of cumulants, the quadratic expression (29) of $K_n[\cdot, \cdot]$ takes, at point $n(x) = \phi(x)$, the form:

$$\begin{aligned} K_\phi[p, q] &= \frac{1}{2} \sum_{r\geq 1} \frac{1}{r!} \sum_{i_1, \ldots, i_r} \left(\kappa^q_{i_1, \ldots, i_r} - \kappa^p_{i_1, \ldots, i_r} \right)^2 \\ &= \frac{1}{2} \left\{ \sum_i \left(\kappa^q_i - \kappa^p_i \right)^2 + \frac{1}{2!} \sum_{ij} \left(\kappa^q_{ij} - \kappa^p_{ij} \right)^2 + \frac{1}{3!} \sum_{ijk} \left(\kappa^q_{ijk} - \kappa^p_{ijk} \right)^2 + \cdots \right\}. \end{aligned} \tag{41}$$

A restricted version of (41) was given by Cardoso (1998) without justification, including only terms of orders 2 and 4 as the simplest non-trivial approximation of the Kullback divergence for symmetric distributions (in which case, odd-order terms vanish). Obtaining (41) is not completely straightforward because it requires taking into account the symmetries of the cumulants: see appendix B for an explicit computation.

The manifolds. The tangent planes at $n(x) = \phi(x)$ to the product manifold $\mathcal{P}$ and to the Gaussian manifold $\mathcal{G}$ are linear spaces denoted $\mathcal{P}_L$ and $\mathcal{G}_L$ respectively. They happen to have simple basis in terms of Hermite polynomials. Again, if $p(x)$ is close to $\phi(x)$, then $e_p(x) = p(x)/n(x) - 1$ is a small random variable. We use a first order expansion: $\log p(x) = \log \phi(x) + \log(1 + e_p(x)) \approx \log \phi(x) + e_p(x)$ which makes it easy to identify log-densities and Gram-Charlier expansions for $\mathcal{P}_L$ and $\mathcal{G}_L$.

For $\mathcal{G}_L$, if $p(x)$ is a Gaussian distribution, then $\log p(x)$ is a second degree polynomial in the x_i's. Thus an orthogonal basis for $\mathcal{G}_L$ is

$$\{h^i | 1 \leq i \leq n\} \bigcup \{h^{ij} | 1 \leq i \leq j \leq n\}, \tag{42}$$

and $\mathcal{G}_L$ is characterized by the fact that the η coordinates of order strictly greater than 2 are zero.

For $\mathcal{P}_L$, we note that if $p(x)$ is a distribution of independent entries, then $\log p(x)$ (or $\log p(x)/\phi(x)$) is a sum of functions of individual entries. Thus the relevant Hermite polynomials for $\mathcal{P}_L$ are

$$\bigcup_{i=1}^{n} \left\{ h^i(x), h^{ii}(x), h^{iii}(x), h^{iiii}(x), \ldots \right\}. \tag{43}$$

The polynomial sets (42) and (43) have some elements in common, namely the set

$$\bigcup_{i=1}^{n} \left\{ h^i(x), h^{ii}(x) \right\}, \tag{44}$$

which corresponds to these distributions which are both normal and of independent components, *i.e.* the Gaussian distributions with diagonal covariance matrix.

A four-way decomposition. The bases (42) and (43) for $\mathcal{G}_L$ and $\mathcal{P}_L$ and (44) for their intersection hint at a four-way decomposition of the index set. Let $\mathbf{i} = (i_1, \ldots, i_r)$ denote a r-uple of indices and let $\mathbf{I}$ denote the set of all r-uples for all $r \geq 1$. Expansion (30) is a sum over all $\mathbf{i} \in \mathbf{I}$. In view of (42), a meaningful decomposition of $\mathbf{I}$ is as $\mathbf{I} = \mathbf{I}^l \cup \mathbf{I}^h$ where $\mathbf{I}^h$ contains the r-uples for $r \geq 3$ (high order) and $\mathbf{I}^l$ contains the r-uples of order $r = 1$ or $r = 2$ (low order). The basis (42) for $\mathcal{G}_L$ is made of Hermite polynomials with indices in $\mathbf{I}^l$. In view of (43), another decomposition is $\mathbf{I} = \mathbf{I}^a \cup \mathbf{I}^c$ where $\mathbf{I}^a$ contains the r-uples of any order with identical indices and $\mathbf{I}^c$ contains the r-uples of any order with indices not all identical. Superscripts l and h refer to **l**ow and **h**igh orders while superscripts a and c refer to **a**uto-cumulants and to **c**ross-cumulants. The basis (43) for $\mathcal{P}$ is made of Hermite polynomials with indices in $\mathbf{I}^a$.

Combining high/low with cross/auto features, the set of all r-uples splits as:

$$\mathbf{I} = \mathbf{I}^{la} \cup \mathbf{I}^{lc} \cup \mathbf{I}^{ha} \cup \mathbf{I}^{hc}, \tag{45}$$

where $\mathbf{I}^{la} = \mathbf{I}^l \cap \mathbf{I}^a$, $\mathbf{I}^{ha} = \mathbf{I}^h \cap \mathbf{I}^a$, and so on. For instance, basis (44) for uncorrelated Gaussian distributions is made of Hermite polynomials with indices in (la).

We thus identify four orthogonal directions for moving away from $\mathcal{N}(0, I)$: a move in the (la) direction changes the mean and the variance of each coordinate while preserving Gaussianity and decorrelation; a move in the (lc) direction also preserves Gaussianity, doing so by introducing correlation between coordinates without changing their mean or their variances; a move in the (ha) changes the marginal distributions (with constant means and variances) but preserves the mutual independence between coordinates; finally the direction (hc) corresponds to any change orthogonal to those previously listed, *i.e.* changes preserving both the marginal structure and the first and second order moment structure.

In our geometric picture, the distributions of interest are obtained as projections onto the product manifold $\mathcal{P}$, onto the Gaussian manifold $\mathcal{G}$, onto their span $\mathcal{P} \vee \mathcal{G}$ and onto their intersection $\mathcal{P} \wedge \mathcal{G}$. In the local approximation, these orthogonal projections boil down to zeroing the relevant coordinates in the four-way decomposition of the Hermite basis. This can be summarized by the following table:

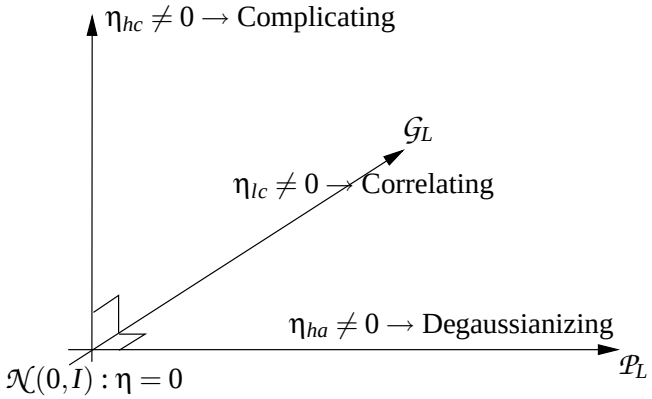

Figure 6: Three orthogonal directions away from $\mathcal{N}(0,I)$ in the local geometry. The origin $\mathcal{N}(0,I)$ has coordinates $\eta = (\eta_{la}, \eta_{lc}, \eta_{ha}, \eta_{hc}) = (0,0,0,0)$. For non-zero coordinates η_{lc}, the distribution labelled by η becomes correlated while remaining Gaussian. Similarly, introducing non-zero coordinates η_{ha} de-Gaussianizes the marginal distributions, preserving the independence between components. Finally, introducing non-zero coordinates η_{hc} introduces a complicated dependency structure which is captured neither by the correlation structure nor by the marginals.

$P(Y)$	η_{la}	η_{lc}	η_{ha}	η_{hc}
$P^{P\vee G}(Y)$	η_{la}	η_{lc}	η_{ha}	0
$P^{G}(Y) = \mathcal{N}(\mathrm{E}Y, R_Y)$	η_{la}	η_{lc}	0	0
$P^{P}(Y) = \prod_i P_i(Y_i)$	η_{la}	0	η_{ha}	0
$P^{P\wedge G}(Y) = \mathcal{N}(\mathrm{E}Y, \mathrm{diag}R_Y)$	η_{la}	0	0	0
$\mathcal{N}(0,I)$	0	0	0	0

and is illustrated on Figure 6. In this figure, the "uninteresting direction" (la) which corresponds to location-scale changes is not displayed.

Divergences and objective functions. In accordance with decomposition (45), the small cumulant approximation $K_\phi\big[p,q\big]$ of Equation (41) decomposes into four components:

$$K_\phi\big[p,q\big] = K_\phi^{la}\big[p,q\big] + K_\phi^{lc}\big[p,q\big] + K_\phi^{ha}\big[p,q\big] + K_\phi^{hc}\big[p,q\big] \tag{46}$$

where, for instance, $K_\phi^{la}\big[p,q\big]$ is the sum (41) with indices restricted to the set $\mathbf{I}^{la}$ and similarly for the other terms. Now the divergence from $P(Y)$ to its best uncorrelated Gaussian approximation $P^{P\wedge G}(Y)$ reads

$$\begin{aligned} K\big[P(Y), \mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))\big] &= K\big[P(\tilde{Y}), \mathcal{N}(0,I)\big] && (47)\\ &\approx K_\phi\big[P(\tilde{Y}), \mathcal{N}(0,I)\big] && (48)\\ &= D^{la}(\tilde{Y}) + D^{lc}(\tilde{Y}) + D^{ha}(\tilde{Y}) + D^{hc}(\tilde{Y}), && (49) \end{aligned}$$

where equality (47) stems from the invariance of the KL divergence under the standardization process (14) which turns Y into $\tilde{Y}$ and turns $\mathcal{N}(\mathrm{E}Y, \mathrm{diag}(R_Y))$ into $\mathcal{N}(0,I)$; where approximation (48)

is the cumulant-based approximation (29) to the KL divergence and where (49) defines the four components of this approximation, with the obvious notation

$$D^{la}(\tilde{Y}) = K_\phi^{la}\big[P(\tilde{Y}), \mathcal{N}(0,I)\big], \quad D^{lc}(\tilde{Y}) = K_\phi^{lc}\big[P(\tilde{Y}), \mathcal{N}(0,I)\big],$$

and similarly for (ha) and (hc). Let us examine these components.

First, the entries of the standardized vector $\tilde{Y}$ have by construction the same mean and variance as $\mathcal{N}(0,I)$. Therefore the distributions $P(\tilde{Y})$ and $\mathcal{N}(0,I)$ have the same cumulants in the (la) set, so that

$$D^{la}(\tilde{Y}) = 0.$$

In the three other index sets (indexed by *ha*, *lc*, *hc*), the cumulants of $\mathcal{N}(0,I)$ cancel either because they are cross-cumulants of second order or because they are cumulants of order higher than 2. Thus, according to (41), these sums involve only the normalized cumulants $\tilde{\kappa}$:

$$\tilde{\kappa}_{i_1\cdots i_r} \stackrel{\text{def}}{=} \text{cum}(\tilde{y}_1, \ldots, \tilde{y}_r), \tag{50}$$

arranged as

$$\begin{aligned}
D^{lc}(\tilde{Y}) &= \frac{1}{4}\sum_{i\neq j}\tilde{\kappa}_{ij}^2, \\
D^{ha}(\tilde{Y}) &= \frac{1}{2}\sum_{r\geq 3}\frac{1}{r!}\sum_i \tilde{\kappa}_{i\cdots i}^2 = \sum_i D_i^{ha}(\tilde{Y}) \quad \text{where} \quad D_i^{ha}(\tilde{Y}) = \frac{1}{2}\sum_{r\geq 3}\frac{1}{r!}\tilde{\kappa}_{i\cdots i}^2, \\
D^{hc}(\tilde{Y}) &= \frac{1}{2}\sum_{r\geq 3}\frac{1}{r!}\sum_{i_1\cdots i_r\neq} \tilde{\kappa}_{i_1\cdots i_r}^2,
\end{aligned}$$

where $\sum_{i_1\cdots i_r\neq}$ denotes a sum over all r-uples in which the indices are not all identical.

We can now identify term-to-term the cumulant-based approximations with their exact Kullback-based expressions. Similarly to (29), we denote $I_\phi(Y)$, $G_\phi(Y)$,..., the expressions for $I(Y)$, $G(Y)$,..., obtained in the expansion around ϕ and we have

$$\begin{aligned}
C(Y) &\approx C_\phi(Y) = D^{lc}(\tilde{Y}), \\
I(Y) &\approx I_\phi(Y) = D^c(\tilde{Y}) = D^{lc}(\tilde{Y}) + D^{hc}(\tilde{Y}) = C_\phi(Y) + D^{hc}(\tilde{Y}), \\
G(Y) &\approx G_\phi(Y) = D^h(\tilde{Y}) = D^{ha}(\tilde{Y}) + D^{hc}(\tilde{Y}) = \textstyle\sum_i G_\phi(Y_i) + D^{hc}(\tilde{Y}), \\
G(Y_i) &\approx G_\phi(Y_i) = D_i^{ha}(\tilde{Y}),
\end{aligned}$$

which can be summarized as

$$K_\phi\big[P(Y), P^{P\wedge G}(Y)\big] = \overbrace{\underbrace{D^{lc}(\tilde{Y})}_{C_\phi(Y)}}^{I_\phi(Y)=D^c(\tilde{Y})} + \overbrace{\underbrace{D^{hc}(\tilde{Y}) + D^{ha}(\tilde{Y})}_{G_\phi(Y)=D^h(\tilde{Y})}}^{\sum_i G_\phi(Y_i)}.$$

We recover the key property (16) in the form

$$I_\phi(Y) + \sum_i G_\phi(Y_i) = G_\phi(Y) + C_\phi(Y). \tag{51}$$

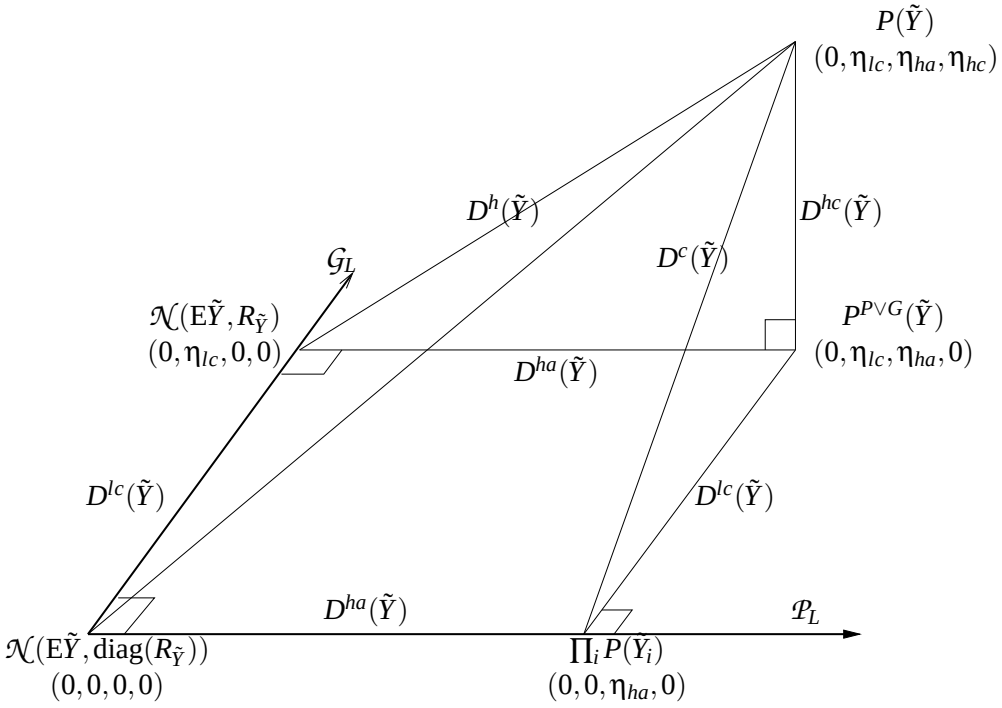

Figure 7: The local Euclidean geometry, tangent to the curved information geometry of Figure 4. The coordinates are the normalized cumulants of various types.

However, we now have an orthogonal decomposition in an Euclidean geometry:

$$I_\phi(Y) - C_\phi(Y) = G_\phi(Y) - \sum_i G_\phi(Y_i) = D^{hc}(\tilde{Y}) \geq 0. \tag{52}$$

That is, for weakly correlated and weakly non-Gaussian vectors, dependence cannot be smaller than correlation and joint Gaussianity cannot be smaller than marginal Gaussianity, reaching the same conclusions as in Section 3.4. The "missing part" $D^{hc}(\tilde{Y})$ is

$$D^{hc}(\tilde{Y}) = K_\phi\left[P, P^{P\vee G}\right] \approx K\left[P, P^{P\vee G}\right],$$

containing only high-order, cross-cumulants, *i.e.* it is due to the part of the log-density of Y which cannot be represented neither by linear-quadratic form in Y, neither by non-linear functions of Y_i, nor by any linear combination of both kinds.

We note in the passing that $C_\phi(Y) = \frac{1}{4}\sum_{i\neq j}\tilde{\kappa}_{ij}^2$, which is identical to (15).

5. Conclusion

We have developed the information geometry of dependence for an n-vector, bringing together various objective functions used in ICA. This geometry admits a tangent representation in terms of

cumulants and Hermite polynomials which is derived via formal first- and second-order expansions. We have not tried to find conditions by which this derivation could be made rigorous but the results are perfectly in line with the insights gained at Section 3.4 and with common wisdom about ICA objectives.

A note on scaling. The entropy of a *continuous* variable (as opposed to a discrete variable) cannot be directly related to the notion of information because it is not invariant under invertible transforms. To give it some meaning, one must introduce some extraneous quantity like an hypothetical "resolution" or the "level" of some additive Gaussian noise. Such a trick is needed to apply the infomax principle to ICA, as done by Bell and Sejnowski (1995), or to relate non Gaussianity to entropy. Thus, using differential entropy in a noise-free context should always be taken with a grain of salt. In contrast, quantities like $I(Y)$, $C(Y)$, $G(Y)$, defined in terms of the Kullback-Leibler divergence which itself is defined independently of a reference scale or of a reference measure. We note, in particular that the whole picture of Figure 2 is invariant with respect to the rescaling of any entry of Y.

Joint and marginal structures. This paper has described the interplay between dependence, correlation and Gaussianity, in particular in the ICA framework (17). Expression (17) shows a striking feature: on one hand, the dependence $I(Y)$ is a property of the *joint* distribution of Y; on the other hand, Equation (17) offers a decomposition of the dependence such that all the joint structure is captured in the correlation $C(Y)$ with a remaining term (the marginal Gaussianities) depending only on the marginals of the distribution. It thus seems that all the dependencies between the entries of Y are captured by $C(Y)$ which, however, is only a measure of second order independence. An explanation of this apparent paradox is that the constant term in (17), equal to $G(Y)$, is constant *only* under linear transforms of Y. It is only because ICA restricts itself to linear transforms that we can forget about the dependencies which affect the value of $G(Y)$ and which depend on the joint distribution of Y, indeed. It is fair to say that *by restricting itself to linear transforms, ICA allows the non-Gaussian part of the dependence to express itself* only *in the marginal distributions*.

Other data models for ICA. An important application of ICA is to the blind separation of sources. This paper has focused onto the objective functions which arise in ICA theory when the sources are modeled as i.i.d. non-Gaussian processes. However, blind source separation can also be achieved by resorting to Gaussian models, providing some temporal (or spatial) structure of the source distributions can be exploited. Two popular approaches are to rely on the spectral diversity of stationary sources or on time diversity of non-stationary sources. Within these two kinds of model, the information-geometric view is strikingly similar. In particular, similar to the key relation (17), mutual information appears as the correlation minus the sum of marginal "non-properties:" divergence from spectral whiteness or from stationarity of the source processes (see Cardoso, 2001).

Acknowledgments

I am grateful to Shun-ichi Amari for general inspiration and for suggesting to investigate cumulant-based expansions of ICA information geometry of ICA and to Mark Plumbley for providing the counter example of Section 3.3. Also, the comments of the anonymous reviewers have helped clarify the original manuscript.

Appendix A. Orthogonality Between the Gaussian and the Product Manifold

We explain in which sense the Gaussian manifold $\mathcal{G}$ and the product manifold $\mathcal{P}$ intersect orthogonally. Two smooth manifolds intersect orthogonally at a given point if their tangent planes at this point are orthogonal. Thus, to discuss the orthogonality between $\mathcal{P}$ and $\mathcal{G}$ at a point $\mathcal{N}(\mu,R)$ where R is diagonal, we must introduce the notion of tangent planes and of orthogonality between vectors of tangent planes. This appendix is only intended for the curious reader (but see also Section 4 for using the notion of tangent plane) and is not necessary for understanding the rest of the paper; it is also both technical and lacking in rigor. For a mathematical treatment of tangent planes in non-parametric models, see e.g., Bickel, Klaassen, Ritov, and Wellner (1993).

We do not need to discuss tangent planes to manifolds in full generality because, in information geometry, tangent planes have a special representation: the vectors of a tangent plane are zero-mean random variables and orthogonality between two "vectors" is the decorrelation between the two associated random variables. This process is now described, starting with the easy case of finite-dimensional families of distributions and then showing how it may be extended to non-parametric families.

Let P be a point in $\mathcal{M}$ a L-dimensional statistical manifold, that is, distribution $p(y) = p(y;\theta^\star)$ for some value $\theta^\star$ of θ which is a vector of L real parameters. The tangent space of $\mathcal{M}$ at P is the linear space

$$\left\{ \sum_{i=1}^{L} a_i l_i(y) \right\} \quad \text{with} \quad l_i(y) = \left. \frac{\partial \log p(y;\theta)}{\partial \theta_i} \right|_{\theta=\theta^\star} \tag{53}$$

where each a_i is an arbitrary coefficient. Thus, the score functions $l_i(y)$ form a basis of the tangent plane and the a_i are the coordinates of a vector of the tangent plane on this basis. In the tangent plane, the metric is given by the Fisher information matrix: if two vectors have coordinates $\{a_i\}$ and $\{b_i\}$, their scalar product is $\sum_{ij} F_{ij} a_i b_j$ where $F_{ij} = \mathrm{E}(l_i(y) l_j(y))$. It follows that if $f(y)$ and $g(y)$ are two vectors of the tangent plane, their scalar product simply is $\mathrm{E}(f(y)g(y))$; in particular, orthogonality means decorrelation.

Similarly, we can obtain tangent vectors without resorting explicitly to parameters by noting that, if Q is close to P in $\mathcal{M}$, it has parameters $\theta = \theta^\star + \delta\theta$ where $\delta\theta$ is a small vector of $\mathbb{R}^L$ so that, at first order,

$$\sum_{i=1}^{L} \delta\theta_i l_i(y) = \sum_{i=1}^{L} \frac{\partial \log p(y;\theta^\star)}{\partial \theta_i} (\theta_i - \theta_i^\star) \approx \log \frac{p(y;\theta)}{p(y;\theta^\star)} = \log \frac{q(y)}{p(y)} \approx \frac{q(y)-p(y)}{p(y)}.$$

This is the key for understanding how a non-parametric model is mapped locally to a linear space: a distribution Q infinitesimally close to P is mapped to the random variable $\varepsilon(y) = (q(y)-p(y))/p(y)$ which has zero-mean under P and infinitesimally small variance. The tangent plane is the linear spaced spanned by these random variables. Conversely, an infinitesimal random variable $\varepsilon(y)$ with zero-mean under P is a vector by which P is shifted to Q, infinitesimally close to P, with density $q(y) = p(y)(1+\varepsilon(y))$.

With these ideas, the tangent plane to the Gaussian manifold is easily found since it is finite dimensional. The tangent plane to the product manifold can be obtained by looking for these infinitesimal random variables $\varepsilon(y)$ which preserve independence at first order when a product distribution $p(y)$ is changed into $p(y)(1+\varepsilon(y))$. The tangent planes to $\mathcal{P}$ and $\mathcal{G}$ at $P = \mathcal{N}(\mu,R)$ (where

$R = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$) are (skipping the computations)

$$\begin{aligned} \mathcal{T}_{\mathcal{G}} &= \left\{ \sum_i a_i (y_i - \mu_i) + \sum_{i \le j} a_{ij} (y_i y_j - \sigma_i^2 \delta_{ij}) \right\}, \\ \mathcal{T}_{\mathcal{P}} &= \left\{ \sum_i f_i(y_i) \mid \mathrm{E}\, f_i(y_i) = 0 \right\}. \end{aligned}$$

In other words, the tangent plane to $\mathcal{G}$ is the set of all zero-mean linear-quadratic functions of y and the tangent plane to $\mathcal{P}$ is the set of all zero-mean component-wise functions of y.

It can be seen that these spaces share some functions: all the zero-mean, linear-quadratic, component-wise functions of y. Let us denote by $\mathcal{T}_{\mathrm{LS}}$ this intersection and decompose both $\mathcal{T}_{\mathcal{G}}$ and $\mathcal{T}_{\mathcal{P}}$ as direct sums of $\mathcal{T}_{\mathrm{LS}}$ and an orthogonal complement. By definition, this is

$$\mathcal{T}_{\mathcal{P}} = \tilde{\mathcal{T}}_{\mathcal{P}} \oplus \mathcal{T}_{\mathrm{LS}} \quad \mathcal{T}_{\mathcal{G}} = \tilde{\mathcal{T}}_{\mathcal{G}} \oplus \mathcal{T}_{\mathrm{LS}},$$

with the following subspaces:

$$\begin{aligned} \mathcal{T}_{\mathrm{LS}} &= \left\{ \sum_i a_i (y_i - \mu_i) + \sum_i a_{ii} (y_i^2 - \sigma_i^2) \right\}, \\ \tilde{\mathcal{T}}_{\mathcal{G}} &= \left\{ \sum_{i<j} a_{ij} y_i y_j \right\}, \\ \tilde{\mathcal{T}}_{\mathcal{P}} &= \left\{ \sum_i r_i(y_i) \mid E\, r_i(y_i) = E(y_i - \mu_i) r_i(y_i) = E(y_i^2 - \sigma_i^2) r_i(y_i) = 0 \right\}. \end{aligned}$$

The space $\mathcal{T}_{\mathrm{LS}}$ is so denoted because it is the tangent plane to $\{\mathcal{N}(\mu, R) \mid R \text{ diagonal}\}$ which is the intersection of $\mathcal{P}$ and $\mathcal{G}$: as already noted, this intersection corresponds to changes in the location-scale directions. The three spaces $\mathcal{T}_{\mathcal{G}}$, $\mathcal{T}_{\mathcal{P}}$, and $\mathcal{T}_{\mathrm{LS}}$ are mutually orthogonal. This is easily seen by checking that any two vectors taken from two of these spaces are uncorrelated. In summary $\mathcal{P}$ and $\mathcal{G}$ intersect orthogonally along their intersection.

Appendix B. Cumulant Expansion.

We derive expression (41) for the local approximation $K_\phi\big[p, q\big]$ to the Kullback-Leibler divergence. According to the Gram-Charlier expansion (30), we have

$$e_q(x) - e_p(x) = \sum_{r \ge 1} \frac{1}{r!} d_r(x) \quad \text{with} \quad d_r(x) = \sum_{i_1, \ldots, i_r} (\kappa^q_{i_1, \ldots, i_r} - \kappa^p_{i_1, \ldots, i_r}) h^{i_1, \ldots, i_r}(x). \tag{54}$$

Here, as elsewhere, we denote $\sum_{i_1, \ldots, i_r} = \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_r=1}^n$. Then we have

$$K_\phi\big[p, q\big] = \frac{1}{2} \mathrm{E}_\phi (e_q(x) - e_p(x))^2 = \frac{1}{2} \mathrm{E}_\phi \left(\sum_{r \ge 1} \frac{1}{r!} d_r(x) \right)^2 = \frac{1}{2} \sum_{r \ge 1} \frac{1}{r!^2}\, \mathrm{E}_\phi \left(d_r(x) \right)^2. \tag{55}$$

The first equality is the Definition (29) of the quadratic approximation of the KLD; the second equality is by inserting (54); the third equality results from expanding the square and from the fact

that Hermite polynomials of different orders are uncorrelated under ϕ *i.e.* $\mathrm{E}_\phi\{d_r(x)d_{r'}(x)\} = 0$ for $r \neq r'$. We show below that

$$\mathrm{E}_\phi \left(\sum_{i_1,\ldots,i_r} S_{i_1,\ldots,i_r} h^{i_1,\ldots,i_r}(x) \right)^2 = r! \sum_{i_1,\ldots,i_r} S^2_{i_1,\ldots,i_r} \tag{56}$$

for any set $\{S_{i_1,\ldots,i_r}\}$ of numbers which are invariant under any permutation of indices. Combining (55) and (56) with $S_{i_1,\ldots,i_r} = \kappa^q_{i_1,\ldots,i_r} - \kappa^p_{i_1,\ldots,i_r}$ gives

$$K_\phi\big[p,q\big] = \frac{1}{2} \sum_{r\geq 1} \frac{1}{r!} \sum_{i_1,\ldots,i_r} \left(\kappa^q_{i_1,\ldots,i_r} - \kappa^p_{i_1,\ldots,i_r} \right)^2, \tag{57}$$

which is the desired result (41). It remains to prove (56), as follows:

$$\begin{aligned}
& \mathrm{E}_\phi \left(\sum_{i_1,\ldots,i_r} S_{i_1,\ldots,i_r} h^{i_1,\ldots,i_r}(x) \right)^2 \\
= & \sum_{i_1,\ldots,i_r} \sum_{j_1,\ldots,j_r} S_{i_1,\ldots,i_r} S_{j_1,\ldots,j_r} \mathrm{E}_\phi \left(h^{i_1,\ldots,i_r}(x) h^{j_1,\ldots,j_r}(x) \right) \\
= & \sum_{i_1,\ldots,i_r} \sum_{j_1,\ldots,j_r} S_{i_1,\ldots,i_r} S_{j_1,\ldots,j_r} \left(\delta_{i_1 j_1} \cdots \delta_{i_r j_r} [r!] \right) \\
= & \sum_{i_1,\ldots,i_r} \sum_{j_1,\ldots,j_r} S_{i_1,\ldots,i_r} S_{j_1,\ldots,j_r} \left(\sum_\sigma \delta_{i_1\sigma(j_1)} \cdots \delta_{i_r\sigma(j_r)} \right) \\
= & \sum_\sigma \sum_{i_1,\ldots,i_r} \sum_{j_1,\ldots,j_r} S_{i_1,\ldots,i_r} S_{j_1,\ldots,j_r} \delta_{i_1\sigma(j_1)} \cdots \delta_{i_r\sigma(j_r)} \\
= & \sum_\sigma \sum_{j_1,\ldots,j_r} S_{\sigma(j_1),\ldots,\sigma(j_r)} S_{j_1,\ldots,j_r} \\
= & \sum_\sigma \sum_{j_1,\ldots,j_r} S^2_{j_1,\ldots,j_r} = r! \sum_{i_1,\ldots,i_r} S^2_{i_1,\ldots,i_r},
\end{aligned}$$

where $\sum_\sigma$ denotes a sum over all permutations of indices. QED.

References

S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Number 28 in Lecture Notes in Statistics. Springer-Verlag, 1985.

S.-I. Amari and J.-F. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, November 1997.

A. J. Bell and T. J. Sejnowski. Blind separation and blind deconvolution: an information-theoretic approach. In *Proceedings of the International Conference on Acoustings, Speech, and Signal Processing*, 3415–3418, 1995.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametrics models*. The Johns Hopkins University Press, 1993.

J.-F. Cardoso. Blind signal separation: Statistical principles, *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation, 90* (8): 2009–2025, 1998.

J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11 (1):157–192, 1999.

J.-F. Cardoso. Unsupervised adaptive filters. In Haykin, S., editor. *Entropic contrasts for source separation: Geometry and stability, 1*, 139–190. John Wiley & Sons, 2000.

J.-F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA 2001*, 2001.

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

J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, 23*, 881–890, 1974.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, 273–279. MIT Press, 1998.

P. McCullagh. *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.