L^2 DE-GAUSSIANIZATION AND INDEPENDENT COMPONENT ANALYSIS

Takeshi Yokoo, *Bruce W. Knight*†, *and Lawrence Sirovich**

takeshi.yokoo@mssm.edu
Laboratory of Applied Mathematics*
Mount Sinai School of Medicine, NY 10029, USA
Laboratory of Biophysics†
Rockefeller University, NY 10021, USA

ABSTRACT

Given an arbitrary standardized (zero mean and unit variance) probability density, we measure its departure from the standard normal density by the L^2 distance between the two density functions. In particular, we consider three different L^2 norms, each distinguished by their weight functions. We investigate the reciprocal Gaussian, uniform, and Gaussian weight functions, and present respective Hermite series representations of non-Gaussianity. We show that the L^2 metric defined with the reciprocal Gaussian weight is directly related to the moment-based approximation of differential entropy. We argue that this is a non-robust measure of non-Gaussianity, as the division by the Gaussian places heavy weights on the tails of the density. Improved robustness is achieved by using the uniform or Gaussian weight functions, both of which effectively suppresses the sensitivity to outliers in the estimations. We choose the L^2 Euclidean metric to define a measure of non-Gaussianity, and show how it leads to an L^2 de-Gaussianization algorithm for independent component analysis.

1. INTRODUCTION

Projection pursuit is an exploratory data analysis technique that searches multidimensional data for components with interesting structures. Classical literature on the subject (*e.g.* [14]) suggests that the non-Gaussianity of the projected (marginal) density may be used to quantify the “interesting-ness” of a component. Several indices for non-Gaussianity have been proposed, including the Tukey-Friedman index [14], various entropic quantities and their cumulant-approximations [21], and the Fisher information [17]. Once an index is defined, an optimization algorithm searches the data for the dimensions along which non-Gaussianity is maximal. However, these indices invariably involve estimation of arbitrary probability density functions (PDFs), a task that is known to be difficult both theoretically and computationally. For this reason, much of the research in projection pursuit has focused on the development of accurate and computationally efficient search criteria.

In recent years, a connection between non-Gaussianity and independent component analysis (ICA) has been suggested [15, 19, 18, 20]. The minimization of mutual information in multidimensional data can be shown to be equivalent to the minimization of the sum of the marginal entropies among orthogonal components. Since entropy of a standardized (zero mean and unit variance) probability density is maximal when it is Gaussian, this loosely translates to a problem of finding a coordinate system whose axes point to the dimensions of maximal non-Gaussianity. This view has an intuitive appeal, as a sum of independent, non-Gaussian random variables tends to a more Gaussian-like random variable, which is a consequence of the Central Limit Theorem.

We attempt to quantify the non-Gaussianity of an arbitrary standardized probability density by the L^2 norm of the difference between the given density and the standard normal. This can be interpreted as the square-distance, with respect to some measure, between the two functions in the space of square integrable functions. In Sec. 2 we review three types of L^2 metrics, that are characterized by different weight functions (*i.e.* different measures on the real line), and show that a suitable Hermite polynomial expansion for each metric space leads to a non-Gaussianity index based on the coefficients of the polynomial terms. Friedman [13], Hall [16], and Cook et al [11] used such non-Gaussianity indices to perform projection pursuit. In Sec. 3, we show that Friedman's formulation of non-Gaussianity is equivalent to the moment-based approximation of differential entropy for near-Gaussian PDFs. This result will motivate the use of L^2 distance based contrast functions for independent component analysis. In particular, Hall's formulation leads to an explicit derivation of the FastICA contrast [20] with a Gaussian non-linearity.

2. MEASURING NON-GAUSSIANITY

2.1. Preliminary Definitions

Let X be a standardized random variable with some probability density $f(x)$. We attempt to assess f's departure from Gaussianity by comparing it with it's Gaussian counterpart,

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \tag{1}$$

with zero mean and unit variance. If one regards f and g as elements of the function space of probability densities, the deviation of f from normality may be assessed by an L^2 metric defined with some positive measure of the real line, $\mu(x)$:

$$D^2 = \int_{-\infty}^{\infty} (f(x) - g(x))^2 \, w(x) dx, \tag{2}$$

where $w(x)$ is given by $d\mu/dx = w(x)$. This definition corresponds to the integrated square-difference between functions f and g, measured with the weight function $w(x)$. Although we leave $w(x)$ unspecified at this point, we assume that we choose w such that the integral converges for most reasonable densities. In the mathematical literature it is customary to denote the L^2 space with measure μ by $L^2(\mu)$, but in the rest of this paper we specify the same space with the derivative of the measure, and write it as $L^2(w)$.

We expand the function $f(x)$ in the integral (2) in terms of Hermite polynomials, a set of orthogonal functions on the entire real line with respect to an appropriate Gaussian weight. Following the notation in Abramowitz and Stegun [1], two distinct families of Hermite polynomials for $n = 0, 1, 2\cdots$, are generated by the derivatives of the Gaussian PDF,

$$\begin{aligned} He_n(x) &= (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n} e^{-\frac{1}{2}x^2} && (3) \\ H_n(x) &= (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, && (4) \end{aligned}$$

where $H_n(x) = \sqrt{2^n} He_n(x/\sqrt{2})$. Following standard practice, we refer to the first set as Chebyshev-Hermite, and the second as Hermite polynomials. The first few polynomials are: $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2x$, $H_3(x) = 8x^3 - 12x$, and $H_4(x) = 16x^4 - 48x^2 + 12$.

Hermite polynomials satisfy an orthogonality relationship,

$$\begin{aligned} \int_{-\infty}^{\infty} He_n(x)He_m(x)g(x)dx &= \delta_{nm}n! && (5) \\ \int_{-\infty}^{\infty} H_n(x)H_m(x)g^2(x)dx &= \delta_{nm}2^{n-1}n!/\sqrt{\pi}, && (6) \end{aligned}$$

with respect to the weight functions $g(x)$ for Chebyshev-Hermite polynomials $He_n(x)$, and $g^2(x)$ for Hermite polynomials $H_n(x)$. We will review three forms of non-Gaussianity indices based on the squared functional distance, due to Friedman [13], Hall [16], and Cook [11], respectively. Each index is defined by a different form of orthogonal series expansion for the arbitrary density $f(x)$, written in terms of either Chebyshev-Hermite or Hermite polynomials.

2.2. Gram-Charlier Series and the Friedman Distance

Friedman [13] introduced a non-Gaussianity index based on the idea that, when a Gaussian random variable Z is transformed with the function $2G(Z) - 1$, where G is the Gaussian cumulative density function $G(z) = \int_{-\infty}^{z} g(\xi)d\xi$, the resulting density is uniform on $[-1, 1]$ [25]. Hence, one may apply the same transformation to X and compute the Euclidean distance (an L^2 norm with an uniform weight), of the transformed density from uniformity. Friedman defined this distance in terms of the coefficients of the Legendre polynomial expansion, but subsequently Hall [16] and Cook [11] showed that this index is equivalent to the L^2-metric between the original (untransformed) densities f and g, measured by the $L^2(1/g)$ norm,

$$D_F^2 = \int_{-\infty}^{\infty} (g(x) - f(x))^2 \frac{1}{g(x)} dx, \quad (7)$$

which we will call the Friedman distance. Due to the reciprocal Gaussian weight in the integral, one can see that this is a heavily tail-weighted definition of non-Gaussianity. One can immediately foresee a potential problem; if the distance were to be estimated from samples of X (as is often the case in real situations), the tails of $f(x)$ usually correspond to regions of the probability space that are rarely sampled. Therefore, the density estimation of f in the tail region is inherently not robust.

The orthogonal series representation of f that is particularly appealing for this $L^2(1/g)$ metric is the Gram-Charlier type-A series [12],

$$f(x) = g(x) \sum_{n=0}^{\infty} \frac{a_n}{n!} He_n(x), \quad (8)$$

where the coefficients on the polynomials are given by,

$$a_n = \int_{-\infty}^{\infty} f(x)He_n(x)dx. \quad (9)$$

Since f is a PDF, the coefficients may be viewed as the expectations, $\mathbf{E}[He_n(X)]$, and hence can be estimated from samples $\{x_1, x_2, \cdots, x_T\}$,

$$a_n = \mathbf{E}\left[He_n(X)\right] \approx \frac{1}{T} \sum_{t=1}^{T} He_n(x_t). \quad (10)$$

Because the Chebyshev-Hermite polynomials are linear combinations of the monomials, x^k; $k = 0, 1, 2, \cdots$, the coefficients of the Gram-Charlier series are determined by the moments of the random variable X. In particular, since X has zero mean and unit variance, we know that $a_0 = 1$, $a_1 = 0$, and $a_2 = 0$. The implicit assumption in this formulation is that f vanishes quickly enough as $x \to \pm\infty$ that $f(x)/\sqrt{g(x)}$ is square-integrable [26]. This restriction is responsible for the poor convergence of this series representation [12], as many commonly occurring densities do not meet this requirement [5]. However, for the time being we assume that this condition holds for the given density f.

Now, substituting the series (8) for f, (7) becomes

$$D_F^2 = \int_{-\infty}^{\infty} g(x)^2 \left(1 - \sum_{n=0}^{\infty} \frac{a_n}{n!} He_n(x)\right)^2 \frac{1}{g(x)} dx. \quad (11)$$

When we expand the square and use the orthogonality (5) with respect to weight $g(x)$, we see that

$$D_F^2 = (a_0 - 1)^2 + \sum_{n=1}^{\infty} a_n^2. \quad (12)$$

Noting $a_0 = 1$, $a_1 = 0$, and $a_2 = 0$, this simplifies to

$$D_F^2 = \sum_{n=3}^{\infty} a_n^2. \quad (13)$$

Thus, non-Gaussianity is defined in terms of Hermite coefficients of f, which are uniquely determined by the moments of X. Although this formulation elegantly relates the moments and the L^2 distance from a Gaussian, it also illuminates the same difficulty that was mentioned earlier. It is well known that the estimation of higher order moments $\mathbf{E}[X^k]$: $k > 2$, is difficult, because the estimates are exquisitely sensitive to large deviations. We can see that this sensitivity results because the $L^2(1/g)$ norm places heavy weights on the tails of f where outliers occur.

2.3. Gauss-Hermite Series and the Hall Distance

From the point of view of the L^2 metric space, perhaps the most natural weight is the uniform function $w(x) = 1$, which treats every point on the entire real line democratically. Hall [16] proposed such an index based on the L^2 Euclidean distance, $L^2(1)$, from the standard normal,

$$D_H^2 = \int_{-\infty}^{\infty} (g(x) - f(x))^2 \, dx, \tag{14}$$

which we call the Hall distance. If f is a square integrable function (g certainly is, since g^2 is proportional to a Gaussian with variance $1/\sqrt{2}$), this integral is convergent. In such a case, we may expand f in terms of Hermite polynomials as follows:

$$f(x) = g(x) \sum_{n=0}^{\infty} \frac{b_n}{\sqrt{\kappa_n}} H_n(x), \tag{15}$$

where

$$b_n = \frac{1}{\sqrt{\kappa_n}} \int_{-\infty}^{\infty} f(x) H_n(x) g(x) dx, \tag{16}$$

and $\kappa_n = 2^{n-1} n!/\sqrt{\pi}$ is the normalization constant. This form of Hermite expansion is sometimes called the Gauss-Hermite series (*e.g.* [27]). Unlike the Gram-Charlier series, the polynomials used here are the Hermite polynomials (not Chebyshev-Hermite) and the Gaussian weight appears in both the decomposition and the reconstruction formulae. The Gauss-Hermite coefficients defined by (16) can also be considered as the expectation values,

$$b_n = \mathbf{E}\left[\frac{1}{\sqrt{\kappa_n}} H_n(X) g(X)\right] \approx \frac{1}{T\sqrt{\kappa_n}} \sum_{t=1}^{T} H_n(x_t) g(x_t), \tag{17}$$

and thus can be estimated from the samples x_t. In particular, one expects that these coefficients are robust against outliers, as large values of $|x_t|$ are attenuated by the tails of the Gaussian.

If we substitute the series representation (15) into the L^2 metric formula (14), and use the orthogonality conditions (5), we see that the Hall distance is

$$D_H^2 = (b_0 - \sqrt{\kappa_0})^2 + \sum_{n=1}^{\infty} b_n^2. \tag{18}$$

Again, the L^2 distance is expressed as the sum of squared Hermite coefficients, with a zeroth order correction because the origin is taken to be the standard normal. In general, we do not know *a priori* the first few terms of the sum as we did in the Gram-Charlier case, because the coefficients (16) are no longer directly linked to moments. However, this is only a minor computational disadvantage considering the benefit of the robustness gained by this formulation.

2.4. Fourier-Hermite Series and the Cook Distance

A related but different definition of the L^2 distance was proposed by Cook et al [11], that measures the square distance with a Gaussian weight,

$$D_C^2 = \int_{-\infty}^{\infty} (g(x) - f(x))^2 \, g(x) dx. \tag{19}$$

which we call the Cook distance. The Chebyshev-Hermite series expansion of f for this $L^2(g)$ norm is,

$$f(x) = \sum_{n=0}^{\infty} \frac{c_n}{n!} He_n(x), \tag{20}$$

where the coefficients are again the expectations,

$$\begin{aligned} c_n &= \int_{-\infty}^{\infty} f(x) He_n(x) g(x) dx & (21) \\ &= \mathbf{E}\left[He_n(X) g(X)\right] & (22) \end{aligned}$$

In this representation, f is expressed as a linear combination of the Chebyshev-Hermite polynomials, and the Gaussian weight appears only in the decomposition. This formulation is valid even for cases when $f(x)$ itself is not square-integrable, as long as the product function $f(x)g(x)$ is square integrable. In order to obtain the Cook distance in terms of the Hermite coefficients, we also expand g as,

$$g(x) = \sum_{n=0}^{\infty} \frac{d_n}{n!} He_n(x), \tag{23}$$

where the coefficient d_n can be analytically obtained [1, 11],

$$d_{2n} = \frac{(-1)^n \sqrt{(2n)!}}{n! 2^{2n+1} \sqrt{\pi}}, \; d_{2n+1} = 0; \; n = 0, 1, 2, \cdots. \tag{24}$$

Inserting (20) and (23) into (19) to evaluate the Cook distance of f, we see that

$$D_C^2 = \int_{-\infty}^{\infty} \left(\sum_{n=1}^{\infty} \frac{c_n - d_n}{n!} He_n(x) \right)^2 g(x) dx. \tag{25}$$

and by using the orthogonality condition, we have

$$D_C^2 = \sum_{n=0}^{\infty} (c_n - d_n)^2. \tag{26}$$

Evidently, this non-Gaussianity measure has the opposite effect of Friedman's definition with reciprocal Gaussian weight; Cook distance attempts to put more weight around the mean of the distribution rather than on the tails. When the Hermite coefficients are estimated from samples, such an active weighting of the center makes sense if the mean of the distribution coincides with the mode where the probability density is the greatest, so that the estimates of the coefficients are based on the most frequently observed values. However, when the mean and the mode (or multiple modes) are very far from one another, which may be the case with highly non-Gaussian densities, this center-weighted metric faces the same difficulty as the tail-weighted metric of $L^2(1/g)$.

A minor aesthetic issue with this particular representation is that, although the sum (26) is exact and the nonzero d's decay faster than exponentially with n, the non-Gaussianity information is not expressed as compactly as the Gram-Charlier or Gauss-Hermite representations, where the subtraction of the Gaussian only occurs in the first term of the sum.

3. INDEPENDENT COMPONENT ANALYSIS CONTRASTS

3.1. A Data Model for Independent Component Analysis

In the previous section, we reviewed three L^2 distance methods for characterizing a density function's departure from normality. In order to illustrate a clear link between the above development and independent component analysis (ICA), we briefly review the ICA model and the relevant terminology.

Suppose we have a set of K mutually independent, real-valued stationary stochastic processes, $\{S_1, S_2, \cdots, S_K\}$, with continuous probability densities $p_1(s_1), \cdots, p_K(s_K)$, respectively. We assume that each random variable S_k has zero-mean and unit variance. For notational convenience, we denote the random vector $[S_1, S_2, \cdots, S_K]^T$ as $\mathbf{S}$, but we emphasize that we are interested in the "set" $\{S_k\}$, and not concerned with the order of the elements in $\mathbf{S}$. We may think of each S_k as a random number generator that outputs a number at times $t = 1, \cdots, T$. We denote the sampled vector at time t by $\mathbf{s}(t) = [s_1(t), \cdots, s_K(t)]^T$.[1] At each time instance, we linearly mix the source signals by an unknown non-singular $K \times K$ matrix $\mathbf{M}$,

$$\mathbf{y}(t) = \mathbf{M}\,\mathbf{s}(t), \tag{27}$$

such that $y_i(t) = \sum_j m_{ij} s_j(t)$. The objective of independent component analysis is: given T samples from the mixtures $\mathbf{y}(t)$, estimate the original source signals $\mathbf{s}(t)$ without the knowledge of the mixing matrix $\mathbf{M}$. In other words, we need to compute the de-mixing $K \times K$ non-singular matrix $\mathbf{W}$,

$$\mathbf{W}\,\mathbf{y}(t) = \mathbf{x}(t), \tag{28}$$

such that the set of the transformed variables $\{x_k(t)\}$ is an estimate of the original set $\{s_k(t)\}$.

Perhaps the most theoretically pleasing approach to solving an ICA problem is to find an optimal transformation $\mathbf{W}$ such that the transformed components are as mutually independent as possible. There is an abundance of papers on various theoretical and practical considerations on the estimation of independence, based on maximum likelihood [2, 24, 8], informax [4], mutual information and minimum marginal entropy principles [10]. These ideas lead to different optimization criteria called "contrasts", but they have subsequently been shown to be closely related to one another [8, 23, 22].

In this paper, we focus on the marginal entropy contrast. In the interest of space, we forgo its derivation as such can be found in many places in review literature (*e.g.* [6, 10]), and we simply state that our minimization criterion is the sum of marginal entropies of $\mathbf{X}$,

$$C(\mathbf{X}) = \sum_{k=1}^{K} \mathsf{H}(X_k), \tag{29}$$

under the whiteness constraint $\mathbf{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{I}$, where $\mathsf{H}(X_k)$ is the differential entropy of the k-th entry of $\mathbf{X}$. The whiteness constraint is a direct consequence of our assumption that the sources are independent with zero mean and unit variance.

3.2. Differential Entropy and the L^2 Distance

In this section, we indicate the conditions under which the $L^2(1/g)$ distance is linearly related to the moment approximation of the differential entropy. Differential entropy of any random variable X with a continuous PDF is defined as the expectation of the logarithm of the density,

$$\mathsf{H}(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{30}$$

Following Amari et al [2] we take the Gram-Charlier (type-A) expansion (8) of f, and noting $a_0 = 1$, $a_1 = 0$, and $a_2 = 0$, we write it as

$$f(x) = g(x) \left(1 + \sum_{n=3}^{\infty} \frac{a_n}{n!} He_n(x) \right). \tag{31}$$

Substituting (31) in (30), we have,

$$\begin{aligned} \mathsf{H}(X) &= -\int_{-\infty}^{\infty} g(x) \left(1 + \sum_{n=3}^{\infty} \frac{a_n}{n!} He_n(x) \right) && (32) \\ &\cdot \log \left\{ g(x) \left(1 + \sum_{n} \frac{a_n}{n!} He_n(x) \right) \right\} dx && (33) \end{aligned}$$

If f is sufficiently close to normality such that $f(x) \approx g(x)(1 + \epsilon(x))$, where $\epsilon(x) \ll 1$ for all x, this expression easily simplifies to

$$\mathsf{H}(X) = \frac{1}{2} \left(\log 2\pi + 1 - \sum_{n=3}^{\infty} a_n^2 \right) + O(\epsilon^3), \tag{34}$$

by expanding $\log(1 + \epsilon)$ in Taylor series and using the orthogonality of the Chebyshev-Hermite polynomials. This result shows that the quadratic approximation of $\mathsf{H}(X)$ for near-Gaussian X is a function of the Hermite coefficients of order 3 and higher, and thus determined by the moments of f. Comparison with (13) shows that this approximation of differential entropy is linearly related to the Friedman distance (13), and so their extrema coincide exactly. Hence, when the departure of f from normality is small, entropy minimization is identical to the maximization of non-Gaussianity in the $L^2(1/g)$ metric space.

3.3. Maximization of the Euclidean Distance

If one viewed ICA algorithms as de-Gaussianization methods, one could adopt ICA contrast functions based on other definitions of L^2 metrics, such as the Hall and the Cook distances (14, 19). For reasons stated in Sec 2, we choose to use the Euclidean metric $L^2(1)$ to define a non-Gaussianity index. Note that each component X_k is a standardized random variable, $\mathbf{E}[\mathbf{X}] = 0$, and $\mathbf{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{I}$. A natural extension of the L^2 metric to an ICA contrast is then given by the sum of $L^2(1)$ non-Gaussianity indices of X_k across all K dimensions,

$$D_H^2(\mathbf{X}) = \sum_{k=1}^{K} D_H^2(X_k), \tag{35}$$

[1] We denote random variables by upper case, and samples by lower case characters. Vectors and matrices are distinguished from scalars by bold face characters.

where

$$D_H^2(X_k) = (b_0(X_k) - \sqrt{\kappa_0})^2 + \sum_{n=1}^{\infty} b_n(X_k)^2. \qquad (36)$$

In particular, if we truncate the sum (36) by taking only the zero-th order terms for each X_k, we can show

$$\begin{aligned} D_H^2(\mathbf{X}) &\approx \sum_{k=1}^{K} (b_0(X_k) - \sqrt{\kappa_0})^2 & (37) \\ &\approx \frac{1}{\kappa_0} \sum_{k=1}^{K} (\mathbf{E}[g(X_k)] - \mathbf{E}[g(Z)])^2 . & (38) \end{aligned}$$

Here, X_k is the standardized random variable with an unknown density f_k, Z is a standard Gaussian random variable and g is the standard Gaussian PDF. This truncated form of the multidimensional $L^2(1)$ distance is proportional to an ICA contrast due to Hyvärinen [18], and the fixed-point iteration implementation called FastICA was introduced in [20].

4. DISCUSSION

The role of the projection pursuit index is to quantify the departure from normality, so it is natural to use a Gaussian PDF as a reference function from which non-Gaussianity is measured. The derivatives of the Gaussian density function provide a natural set of orthonormal functions, the Hermite functions, for describing deviations from normality. We showed a linear relationship between the $L^2(1/g)$ non-Gaussianity metric and the approximate entropy of a PDF, but it is not clear if the other two metrics, $L^2(1)$ and $L^2(g)$, have similar implications in the context of information theory. Because $1/g(x)$, 1, and $g(x)$ weigh the real line differently, the optima of non-Gaussianity in these metric spaces do not precisely coincide with the maxima of statistical independence. However, the same criticism applies to minimization algorithms of the marginal entropy (34), as the approximation is only valid for near-Gaussian PDF. For a very non-Gaussian density, its entropy estimate by (34) is a quadratic extrapolation from a standard Gaussian function, which could be very far from the true value of entropy. In practice, this is often not a serious issue, as the data imposes a finite dimensional constraint on the function space, and for the purpose of source separation, approximations often suffice for many kinds of data. Nevertheless, the output of these algorithms should be examined with scrutiny and preferably be validated by other means than the independence criterion, such as some *a priori* knowledge of the nature or the predicted structures of the source signals.

In the interest of space, we do not present an algorithm for L^2 degaussianization or simulation results. It is possible to implement an $L^2(1)$ de-gaussianization algorithm by approximating the distance as in Sec. 3.3 and applying a fixed-point iteration algorithm, as done by [20]. Alternatively, one can also follow [10] and [6] and successively apply Jacobi plane rotations to yield a sequence of orthogonal matrices, maximizing the ICA contrast between a pair of orthogonal components at each iteration. Another promising approach for simultaneous maximization of L^2 distances in multiple dimensions takes advantage of the algebraic structure of the Stiefel manifold, a vector space of orthogonal matrices. ICA implementations in the Stiefel manifold has appeared in recent publications such as [3], and [9], and our L^2 de-Gaussianization scheme could be implemented as a conjugate-gradient algorithm in the Stiefel manifold.

5. CONCLUSION

In recent years, several authors have shown that independent component analysis can be implemented as special cases of projection pursuit. Minimization of marginal entropies can be identified with projection pursuit of orthogonal components with an entropic index [15]. Hyvarinen [18] introduced an abstraction of the entropic measure as some characterization of non-Gaussianity, obtained by projecting a PDF onto different sets of orthogonal "measuring" functions. In these examples, projection pursuit indices were used to construct the ICA contrast function, which was to be optimized to achieve the separation of the independent components.

We present three theoretical results that we hope will shed some new light to the relationship between projection pursuit and ICA. (i) We demonstrate that maximization of square-distance from a Gaussian PDF in the $L^2(1/g)$ metric space is equivalent to minimization of differential entropy (its moment approximation around a Gaussian, to be precise), and hence motivate ICA implementations based on direct L^2 de-Gaussianization algorithms. (ii) We point out the inherent difficulty of entropy estimation from finite samples, but if we view entropy as one example of the L^2 non-Gaussianity indices, we can devise more robust non-Gaussianity measures by choosing better-behaved weight functions. (iii) We provide a classical functional analysis framework in which the FastICA algorithm [20] can be derived and generalized.

Our observations suggest that ICA-like algorithms based on the L^2 metrics may provide more robust and perhaps more accurate source separation. Although it is unclear if the $L^2(1)$ and $L^2(g)$ metrics have direct connection to information theoretic definitions of independence, algorithms based on these robust metrics may be useful when other algorithms fail, or when only a rough initial estimate is desired, at which point a more theoretically pleasing algorithm can be applied to find the exact optima of independence.

Acknowledgments

I would like to express my gratitude to Dr. A. Cichocki for the inspiration for this work and generous help with various kinds of benchmark data. Also, many thanks to Ehud Kaplan, Alex Casti, Robert Uglesich, and Andrew Sornborger for helpful advice. This research was supported by DARPA MDA 972-01-0028.

6. REFERENCES

[1] M. Abramowitz and I. Stegun. Handbook of Mathematical Functions. Dover Publications, Inc., New York, 1972

[2] S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind source separation. In Advances in Neural Information Processing Systems 8, pages 757-763. MIT Press, Cambridge, MA, 1996.

[3] F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. J. Machine Learning Res. 3 1-48, 2001

[4] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7:1129-1159, 1995.

[5] S. Blinnikov and R. Moessner. Expansions for nearly Gaussian distributions. Astron. Astrophys. Suppl. Ser. 130, 193-205, 1998

[6] J. F. Cardoso. High-Order Contrasts for Independent Component Analysis. Neural Computation 11: 157-192, 1999.

[7] J. F. Cardoso, C.N.R.S. and E.N.S.T. Blind Signal Separation: Statistical Principles. Proc. IEEE 9(10): 2009-2025, 1998

[8] J. F. Cardoso. Infomax and maximum likelihood for source separation. IEEE Letters on Signal Processing, 4(4):112-114, 1997

[9] R. M. Clemente, C. G. Puntonet, J. I. Acha. "A Conjugate Gradient Method and Simulated Annealing for Blind Separation of Sources", Lecture Notes in Computer Science, Vol. 2085, pp. 810-817, June 2001.

[10] P. Comon. Independent component analysis - a new concept? Signal Processing, 36:287-314, 1994.

[11] D. Cook, A. Buja, J. Cabrera. Projection Pursuit Indexes Based on Orthonormal Function Expansions. J. of Computational and Graphical Statistics, 2(3): 225-250. 1993

[12] H. Cramer. Mathematical Methods of Statistics. Princeton Univ. Press, Princeton, 1957

[13] J. H. Friedman. Exploratory projection pursuit. J. Amer. Statist. Assoc. 82 249-266, 1987.

[14] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. IEEE Trans. of Computers, c-23(9):881-890, 1974.

[15] M. Girolami and C. Fyfe, Negentropy and Kurtosis as Projection Pursuit Indices Provide Generalised ICA Algorithms, in Proc. NIPS, Aspen, CO. Dec. 7, 1996

[16] P. Hall. Polynomial Projection Pursuit, Annals of Statistics, 17:589-605, 1989

[17] P. J. Huber, Projection Pursuit, Annals of Statistics, 13: pp. 435-475, 1985.

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

[19] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. Neural Computation, 9(7):1483-1492, 1997.

[20] A. Hyvärinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks, 10(3):626–634, 1999.

[21] M. C. Jones and R. Sibson, What is projection pursuit? Journal of the Royal Statistical Society, 150: pp. 1-18, 1987

[22] J. P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: a factorial code maximized information transfer. NETWORK, 5:565-581, 1994

[23] D. Obradovic and G. Deco. An information theory based learning paradigm for linear feature extraction. Neurocomputing, 12:203-221, 1996

[24] D. T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans. Signal Proc. 45(7): 1712-1725, 1997

[25] A. Papoulis. Probability, Random Variables, and Stochastic Processes. McGraw-Hill, 3rd edition, 1991.

[26] L. Sirovich. Introduction to Applied Mathematics. New York: Springer-Verlag, 1988.

[27] R. P. van der Marel and M. Franx. A new method for the identification of non-Gaussian line profiles in elliptical galaxies. Astrophys. J., 407 525, 1993