

# GEOMETRIC DEPENDENCY ANALYSIS

Samer Abdallah and Mark Plumbley

Centre for Digital Music,  
Queen Mary, University of London  
Technical Report C4DM-TR06-05  
Version 1.0 – 11th August 2006

We investigate methods for estimating and visualising the residual dependency structure in distributed representations such as those produced by independent component analysis (ICA). This process involves estimating the mutual information between many pairs of components, for which we use a method based on nonlinear correlations. The pairwise dependencies are then represented geometrically using multidimensional scaling (MDS) in such a way that strongly dependent components are brought close together. The method is applied to ICA representations of speech and music audio signals. The resulting geometric structures reflect the time-frequency structure of the speech-derived ICA representation and a more complex relationship, related to the harmonic structure of Western musical scales, between frequencies in the music-derived representation.

*Keywords:* Mutual information estimation, MDS, topographic ICA, nonlinear correlation, residual dependency.



# GEOMETRIC DEPENDENCY ANALYSIS

SAMER A. ABDALLAH AND MARK D. PLUMBLEY

**ABSTRACT.** We investigate methods for estimating and visualising the residual dependency structure in distributed representations such as those produced by independent component analysis (ICA). This process involves estimating the mutual information between many pairs of components, for which use a method based on nonlinear correlations. The pairwise dependencies are then represented geometrically using multidimensional scaling (MDS) in such a way that strongly dependent components are brought close together. The method is applied to ICA representations of speech and music audio signals. The resulting geometric structures reflect the time-frequency structure of the speech-derived ICA representation and a more complex relationship, related to the harmonic structure of Western musical scales, between frequencies in the music-derived representation.

*Keywords:* Mutual information estimation, MDS, topographic ICA, nonlinear correlation, residual dependency.

## 1. INTRODUCTION

In this paper we investigate the concept of similarity between the components of a distributed representation such as might be generated by an unsupervised learning system such as independent component analysis (ICA). In particular, we show how these relationships can be defined in terms of the statistical structure of the representation and visualised geometrically.

The concept of relationships between components of a representation is relatively commonplace. Consider, for example, the use of a discretely sampled intensity values to represent an image. The components of the representation are the individual pixels, which have neighbourhood relationships in the image plane quite apart from any concept of similarity between complete images. Similarly, the components of a 1-dimensional discrete Fourier transform can be ordered linearly frequency, while the components of a discrete wavelet transform can be arranged in a 2-dimensional time-scale space.

Of direct relevance to our work in this paper is *topographic ICA* (or TICA) [Hyvärinen et al., 2001], which also addresses the question of how to treat similarity between components by examining statistical dependencies between the component activities. Our work picks up this thread, but instead of modelling the dependency structure as a lattice with fixed and predetermined topology, we investigate the possibilities of a geometric representation.

---

Author address: Department of Electronic Engineering, Queen Mary University of London, Mile End Road, London E1 4NS, UK. E-mail {samer.abdallah,mark.plumbley}@elec.qmul.ac.uk.

Our first task is to define precisely what we mean by ‘similarity between components of a representation’ and how this differs from the also familiar but distinct concept of similarity between the objects being represented.

**1.1. Notational conventions.** The following conventions are adopted in this paper:  $\mathbb{R}^*$  denotes the set of non-negative real numbers including zero;  $M..N$  denotes the set of integers  $\{k|M \leq k \leq N\}$ ; both  $(x^{(M)}, \dots, x^{(N)})$  and  $(x^{(k)})_{k \in M..N}$  denote the sequence of values taken by  $x^{(k)}$  as  $k$  progresses through the integers  $M..N$ . The probability density function (pdf) of a random variable (rv)  $X : \Omega \rightarrow \mathcal{X}$  (i.e. an rv taking values in  $\mathcal{X}$ ) will be written as  $p_X : \mathcal{X} \rightarrow \mathbb{R}^*$ . Clearly, to be admissible as a pdf, it must satisfy  $\int_{\mathcal{X}} p_X(x) dx = 1$ . Similarly, the joint pdf of two variables  $X : \Omega \rightarrow \mathcal{X}$  and  $Y : \Omega \rightarrow \mathcal{Y}$  is  $p_{XY} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^*$ . Other pieces of notation will be defined in the text as necessary.

**1.2. Similarity in distributed representations.** Consider a set  $\mathcal{S}$  of possible observations, where each element stands for what might be thought of in some applications as a complete pattern of sensory stimulation. One way to introduce a notion of similarity here is to define a function  $d_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^*$  as a measure of difference between any two observations. We call this concept ‘P-similarity’ (for ‘*pattern* similarity’, [Abdallah, 2002]) to distinguish it from a different form of similarity, defined below, which is the main subject of this paper.

Suppose now that we have a function  $\psi : \mathcal{S} \rightarrow \mathcal{X}^{\mathcal{A}}$  which maps each element  $s$  of  $\mathcal{S}$  to a tuple  $\psi(s) = x = (x_a, x_b, \dots)$ , whose components are indexed by the elements  $a, b$ , etc. of a discrete set  $\mathcal{A}$  and take values in  $\mathcal{X}$ . Given such a distributed representation, we can ask in what sense can the units of the representation, that is, the elements of the indexing set  $\mathcal{A}$ , be considered similar or dissimilar? In other words, we can investigate the topological and metrical structure of the set  $\mathcal{A}$  as expressed by some dissimilarity function  $d_{\mathcal{A}} : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}^*$ . We call this ‘R-similarity’ [Abdallah, 2002], to indicate that it concerns similarity between *representational units*.

To give a concrete example, suppose that  $\mathcal{S}$  is some vector space over the complex numbers  $\mathbb{C}$ , with an inner product  $\langle \cdot, \cdot \rangle$  and norm  $\|s\| = \sqrt{\langle s, s \rangle}$ . In this case, a measure of P-similarity could be supplied using the induced Euclidean metric  $d_{\mathcal{S}}(s, t) = \|t - s\|$ . Suppose in addition that the function  $\psi : \mathcal{S} \rightarrow \mathbb{C}^N$  is linear; then  $\mathcal{A}$  could be defined as the set of  $N$  vectors, say  $\mathcal{A} = \{u_i \in \mathcal{S} | i \in 1..N\}$ , which determine the linear transformation. If  $x = \psi(s)$ , then for each  $\alpha \in \mathcal{A}$ , the corresponding component could be defined as  $x_{\alpha} = \langle \alpha, s \rangle$ . The function  $d_{\mathcal{A}}$  would then quantify some notion of dissimilarity between the vectors  $\alpha \in \mathcal{A}$ , not necessarily as vectors in themselves, but in regards to their role in defining the representation.

To see how  $d_{\mathcal{A}}$  might reflect something other than the Euclidean distance between the vectors in  $\mathcal{A}$ , consider the case where  $\mathcal{S}$  is a space of 1-dimensional discrete-time signals and  $\psi$  is a 1-dimensional discrete Fourier transform. The elements

of  $\mathcal{A}$  then form an orthogonal basis of  $\mathcal{S}$ . As such, they are all equally dissimilar according to the Euclidean metric  $d_{\mathcal{S}}$ , but as Fourier basis vectors labelled by a frequency, they can be linearly ordered according to those frequencies such that the ‘distance’ between two Fourier components is, for example, the difference between their frequencies, or between the logarithms of their frequencies. We implicitly acknowledge this when we plot a Fourier transform with the frequencies in order. We are justified in doing so because this usually makes the Fourier transform easier to interpret.

The earlier example of discrete 2-dimensional images can also be formalised in this way:  $\mathcal{S}$  would be the set of all possible images,  $\mathcal{A} = (1..N) \times (1..M)$  would be the set of pixel locations in an  $N \times M$  image, and  $x_{\alpha}$  would be the image intensity at the pixel addressed by  $\alpha = (i, j)$ . An appropriate choice for  $d_{\mathcal{A}}$  might be the Euclidean distance between pixels in the image plane.

**1.3. Data-driven similarity measures.** In the case of images and Fourier transforms, it was easy to identify an appropriate measure of similarity between the representational elements by examining the nature of the representation. But what, if any, arrangement would make sense for an arbitrary representation, perhaps generated by some process of unsupervised learning, where we have no intuitions to guide us?

Our approach begins with the common observation that visual images usually exhibit *spatial coherence*, which is to say that nearby image patches tend to have similar colours or patterns. More generally, spatial coherence means that pixel intensities can often be *predicted* from the intensities of their neighbours, which implies that they are not statistically independent. Thus, if, for some reason, we had lost track of the spatial arrangement of the pixels, we might aim to recover it by looking for statistical dependencies between the pixel intensities over a number of images. Turning this around, one could argue that the spatial arrangement is appropriate and useful precisely because it localises dependencies in the image: the dependencies are what ‘bind’ the image into coherent localities, over which, for example, meaningful averages can be taken. The hypothesis is that even if there were no preordained spatial arrangement, one could potentially be created by invoking this principle of localising dependencies.

These ideas were formalised and implemented in *topographic independent component analysis* [Hyvärinen et al., 2001], which proposed an explicit link between statistical dependence on the one hand and topological locality on the other. Independent component analysis (ICA) is especially relevant in this context because its specific aim is to minimise the dependencies between the output components. Hence, these *residual* dependencies represent the failure of the algorithm to achieve complete independence. Localising these residual dependencies may help in visualisation and interpretation of results, but more importantly, it can localise the information required for subsequent computations. For example, dependency localisation is used

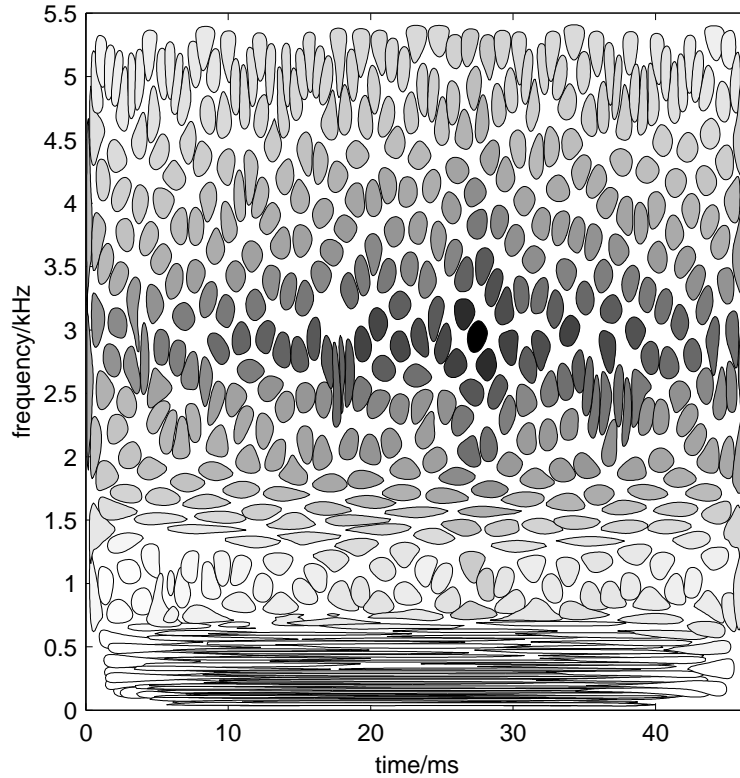


FIGURE 1. Local dependencies in a speech derived ICA basis. Each of the 512 basis vectors is plotted as a region in time-frequency, obtained by taking a contour of the Wigner-Ville distribution of that basis vector (see [Abdallah, 2002] for more details.) The shading encodes the nonlinear correlation (see § 2) between each component and the one at approximately (28 ms, 3 kHz), showing that the dependencies are local in time-frequency and that it should be possible to recover the time-frequency organisation through an analysis of the dependency structure.

as an intermediate step in multilayer ICA [Matsuda and Yamaguchi, 2003], where the localisation allows each layer to be connected only sparsely to the next. A similar principle of minimum ‘wire-length’ costs has also been suggested as an explanation for the formation of topology-preserving cortical maps in the brain [Mitchison, 1995].

The idea behind geometric dependency analysis (GDA) is to manifest the dependency structure *geometrically*, by arranging the variables that make up the distributed representation in a metric space such that the distances between variables reflects the strength of their dependence. In topographic ICA, the local dependencies are embodied as a predefined system of neighbourhoods. The topology is *given* while the algorithm determines which basis vector to put at each lattice site. In GDA however, the basis vectors are given while the algorithm determines a geometric arrangement embedded within a metric space of some given dimensionality. The representational units are free to lie on any manifold of any dimension and topology

as long as this can be embedded in the larger space; they are free to find their own intrinsic topology and dimensionality if one exists.

To formalise the probabilistic aspects of the representation, we now suppose that our observations  $s$  are realisations of a random variable  $S : \Omega \rightarrow \mathcal{S}$ , where  $(\Omega, \mathcal{B}, P)$  is a probability space<sup>1</sup>. This means that we can model the distributed representation defined by the function  $\psi$  as the random variable  $X = \psi(S)$  taking values in  $\mathcal{X}^{\mathcal{A}}$ , and therefore each component is a random variable, which we will write as  $X_\alpha : \Omega \rightarrow \mathcal{X}$  for all  $\alpha \in \mathcal{A}$ . Our aim is to derive a dissimilarity function  $d_{\mathcal{A}}$  such that  $d_{\mathcal{A}}(\alpha, \beta)$  is a measure of the independence between  $X_\alpha$  and  $X_\beta$ , which we propose to quantify in terms of the mutual information  $I(X_\alpha, X_\beta)$ . It has been argued elsewhere that mutual information is an appropriate and effective measure of statistical dependence [Comon, 1994, Cardoso, 1997]. When learning from the data in an unsupervised setting, all we have available is a sequence of observations  $(s^{(1)}, \dots, s^{(K)})$  so the dependency between  $X_\alpha$  and  $X_\beta$  must be estimated from the two sequences  $(x_\alpha^{(k)})_{k \in 1..K}$  and  $(x_\beta^{(k)})_{k \in 1..K}$ .

The overall procedure can be summarised as follows:

- (1) Obtain a distributed representation using ICA;
- (2) Measure the residual dependence between ICA components;
- (3) Map residual dependencies to distances;
- (4) Embed distance structure into a metric space using multidimensional scaling.

## 2. ESTIMATING MUTUAL INFORMATION USING NONLINEAR CORRELATION

Our aim in this section is to develop a method for estimating the mutual information between a pair of random variables  $X$  and  $Y$ , which are assumed to be two components of a distributed representation generated using independent component analysis (ICA). An important requirement here is that the method should be cheap enough to apply to potentially hundreds of thousands of pairs, since we aim to apply it to every pair of components in a 512-component ICA representation. In what follows, we will assume that  $X$  and  $Y$  take values in  $\mathcal{X}$  and  $\mathcal{Y}$  respectively, though in practice, these will both be subsets of  $\mathbb{R}$ . We will be making use of the standard definitions of variance, covariance and correlation:

$$\begin{aligned} \text{var}(X) &\triangleq \text{E } X^2 - (\text{E } X)^2, \\ \text{cov}(X, Y) &\triangleq \text{E } XY - (\text{E } X)(\text{E } Y), \\ \text{corr}(X, Y) &\triangleq \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}, \end{aligned} \tag{1}$$

where all three are to be considered as operators acting on random variables, as is the expectation operator  $\text{E}$ .

---

<sup>1</sup>That is,  $\Omega$  is a set of elementary events,  $\mathcal{B}$  is a  $\sigma$ -algebra over  $\Omega$  and  $P : \mathcal{B} \rightarrow \mathbb{R}$  is a probability measure.

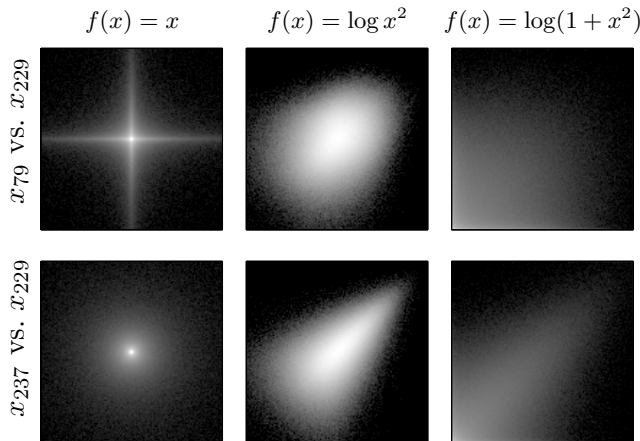


FIGURE 2. Joint histograms of  $f(x_i)$  vs.  $f(x_j)$  for three functions  $f$  and two pairs of components from a music-derived ICA representation. Both  $x_{237}$  and  $x_{229}$  (bottom row) are essentially sinusoidal components at 937 Hz, but in quadrature phase: they have a circularly symmetric distribution and are strongly dependent, though uncorrelated. In the top row,  $x_{79}$ , a sinusoidal component at 883 Hz, is almost independent of  $x_{229}$ .

As previous researchers have pointed out, [Hyvärinen et al., 2001] ICA tends to produce decorrelated components; indeed, those ICA algorithms that consist of pre-whitening followed by a rotation are constrained to do so. In addition, applications of ICA to natural sound and image ensembles [Olshausen and Field, 1996, Simoncelli, 1999, Hyvärinen et al., 2001, Abdallah, 2002] tend to produce components with certain distinctive characteristics: typically, the joint distributions of pairs of components are strongly super-Gaussian with shapes ranging from the distinctive ‘cross’ shape for independent components (fig. 2, top-left) to the circular symmetry visible in fig. 2 (bottom-left). These distributions have also been observed in wavelet analysis of natural images [Wainwright et al., 2001]. In most cases, the distributions have an approximate bilateral symmetry in both axes, which implies a low correlation between the two variables. However, circular symmetry is compatible with independence only for Gaussian random variables, and so the non-Gaussian circularly symmetric distributions observed here imply dependent components—this is the basis of the independent subspace analysis (ISA) model [Hyvärinen and Hoyer, 2000]. For these reasons, we do not expect the linear correlation  $\text{corr}(X, Y)$  to provide a good indicator of the residual dependence between  $X$  and  $Y$ .

**2.1. Activity correlation for symmetric pdfs.** Several authors have proposed that the residual dependencies be characterised as a form of ‘co-activity’, which can be captured by examining the ‘energy’ correlations  $\text{corr}(X^2, Y^2)$ , that is, the correlations between the squared values of the components [Simoncelli, 1999, Hyvärinen and Hoyer, 2000, Hyvärinen et al., 2001, Wainwright et al., 2001, Schwartz and Simoncelli, 2001].

While energy correlation might in some cases be a valid way to characterise dependence arising from co-activity, measuring energy correlations directly may not yield good quantitative estimates of the dependence: if the variables  $X$  and  $Y$  have heavy-tailed distributions, the relevant fourth-order moments ( $E X^2 Y^2$ ,  $E X^4$  and  $E Y^4$ ) may not be defined. Even if these moments are defined, they may be difficult to estimate accurately leading to energy correlation estimates with high variance.

In fact, there is a good reason why magnitude correlations are a relevant measure of dependence between two real-valued random variables with a symmetric joint density: if  $X$  and  $Y$  have a density function  $p_{XY}$  such that  $p_{XY}(x, y) = p_{XY}(-x, y)$  and  $p_{XY}(x, y) = p_{XY}(x, -y)$  for all  $(x, y) \in \mathbb{R}^2$ , then the mutual information between the absolute values of  $X$  and  $Y$  is the same as that between  $X$  and  $Y$  themselves, that is,  $I(|X|, |Y|) = I(X, Y)$ . Furthermore, it is well known that the MI is invariant to invertible transformations of the variables, so if  $f$  and  $g$  are two invertible functions and  $I(|X|, |Y|) = I(X, Y)$ , then  $I(f(|X|), g(|Y|)) = I(X, Y)$ . This is the justification behind our approach: we will aim to choose invertible nonlinear functions  $f$  and  $g$  such that we can estimate  $I(f(|X|), g(|Y|))$  on the understanding that this will be equal to  $I(X, Y)$ .

**2.2. Mutual information for Gaussian variables.** Let us return to the consideration of two general real-valued random variables  $X$  and  $Y$ . If  $X$  and  $Y$  were jointly Gaussian with correlation coefficient  $\rho_{XY} = \text{corr}(X, Y)$  then the mutual information could be expressed directly in terms of the correlation coefficient only:

$$I(X, Y) = I^{\mathcal{G}}(X, Y) \triangleq -\frac{1}{2} \log(1 - \rho_{XY}^2). \quad (2)$$

Bach and Jordan [Bach and Jordan, 2002] call  $I^{\mathcal{G}}(X, Y)$  the *Gaussian mutual information*, which is also Cardoso's *correlation*  $C(\{X, Y\})$  [Cardoso, 2003]. Even if  $X$  and  $Y$  were not jointly Gaussian, one might wish to use  $I^{\mathcal{G}}(X, Y)$  in (2) as an approximation to the mutual information between  $X$  and  $Y$ . One would expect that the approximation error to be related to the extent to which  $X$  and  $Y$  are jointly non-Gaussian. For example, the circularly symmetric but non-Gaussian densities discussed in §2 have zero correlation and hence zero Gaussian mutual information even though the variables are dependent.

One approach to this difficulty is to make use of the well-known *data processing inequality*, which states that for any functions  $f$  and  $g$ ,

$$I(f(X), g(Y)) \leq I(X, Y) \quad (3)$$

with equality when both  $f$  and  $g$  are invertible. Thus, if we could choose  $f$  and  $g$  such that  $I(f(X), g(Y))$  could be estimated efficiently, then we would obtain a lower bound on  $I(X, Y)$ . For an early application of this idea to text data see [Shannon, 1951]. Furthermore, if we could choose  $f$  and  $g$  to make  $f(X)$  and  $g(Y)$  *jointly* Gaussian, then we would have

$$I(X, Y) \geq I(f(X), g(Y)) = I^{\mathcal{G}}(f(X), g(Y)), \quad (4)$$

the Gaussian mutual information therefore providing a lower bound which can be estimated efficiently and cheaply from a sample correlation coefficient. It will generally not be possible to find a transformation which jointly Gaussianises  $(X, Y)$ , and so in the next section we investigate the conditions under which  $I^{\mathcal{G}}(f(X), g(Y))$  provides a lower bound on  $I(f(X), g(Y))$  and hence on  $I(X, Y)$ .

**2.3. Relationship between  $I$  and  $I^{\mathcal{G}}$ .** Since we would like to estimate the mutual information  $I(X, Y)$  from the Gaussian mutual information  $I^{\mathcal{G}}(X, Y) = \frac{1}{2} \log(1 - \rho_{XY}^2)$ , we would like to know if there is a strict relationship between them. This question was addressed by Cardoso [Cardoso, 2003], the main result of which we review here. We know that for uncorrelated but not independent variables we would have  $\rho_{XY} = 0$  implying  $I^{\mathcal{G}}(X, Y) = 0$ , but  $I(X, Y) \geq 0$ . It is therefore tempting to speculate that  $I(X, Y) \geq I^{\mathcal{G}}(X, Y)$  as we conjectured previously [Abdallah and Plumley, 2003]. However, information theoretic quantities are often notoriously counter-intuitive, and this particular conjecture is false. For a counterexample consider the case where  $X = Z + V_1$  and  $Y = Z + V_2$ , where  $Z$  takes values in  $\{-1, +1\}$  with equal probability and  $V_1$  and  $V_2$  are iid Gaussian with some small variance  $\sigma^2$ . Here we have  $I(X, Y) \approx 1$  bit, but as  $\sigma^2 \rightarrow 0$ ,  $\rho_{XY} \rightarrow 1$  and therefore  $I^{\mathcal{G}}(X, Y) \rightarrow \infty$ . So we know that  $I(X, Y) = I^{\mathcal{G}}(X, Y)$  if  $X$  and  $Y$  are jointly Gaussian, but otherwise the situation is not so clear.

Bach and Jordan [Bach and Jordan, 2002] showed that, if  $X$  and  $Y$  are both *marginally* Gaussian, then  $I(X, Y) \geq I^{\mathcal{G}}(X, Y)$ . As shown by Cardoso [Cardoso, 2003] we can also gain other insight into this process by considering the geometry of the Kullback-Leibler divergence (*KL-divergence*, or sometimes *I-divergence* or *cross-entropy*) between probability distributions, using concepts from *information geometry* [Amari and Nagaoka, 2001].

Given two probability density functions (pdfs)  $p$  and  $q$  over  $\mathcal{X}$ , the KL-divergence from  $p$  to  $q$  is defined as

$$D(p \parallel q) \triangleq \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \quad (5)$$

which is nonnegative:  $D(p \parallel q) \geq 0$  with equality if and only if  $p(x) = q(x)$  for all  $x \in \mathcal{X}$ , (except, possibly, for a set of points with zero total probability with respect to  $p$ ). Using this we can write

$$I(X, Y) = D(p_{XY} \parallel p_X \otimes p_Y) \quad (6)$$

where  $p_{XY}$  is the joint pdf of  $X$  and  $Y$ , and  $p_X \otimes p_Y$  is the function defined as the product of the marginals, that is,

$$\begin{aligned} p_X \otimes p_Y &: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^* \\ (p_X \otimes p_Y)(x, y) &= p_X(x)p_Y(y). \end{aligned} \quad (7)$$

By definition,  $X$  and  $Y$  are independent if and only if  $p_{XY}(x, y) = p_X(x)p_Y(y)$ , and hence  $p_{XY} = p_X \otimes p_Y$ , which implies that  $I(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent.

We can also consider the KL-divergence  $D(p_{XY} \parallel p_{XY}^{\mathcal{G}})$  between  $p_{XY}$  and the pdf of a bivariate Gaussian random variable  $\mathcal{N}(\mu_{XY}, \Sigma_{XY})$  with the same second order statistics; that is, with  $\mu_{XY} = (\mathbb{E} X, \mathbb{E} Y)$  and

$$\Sigma_{XY} = \begin{pmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{pmatrix}. \quad (8)$$

The pdf of  $\mathcal{N}(\mu_{XY}, \Sigma_{XY})$  is, in the sense of KL-divergence, the ‘closest’ Gaussian pdf to  $p_{XY}$ . The operation of finding this nearest Gaussian is essentially a projection from  $p_{XY}$  on to the manifold of Gaussian densities; we will write the resulting pdf as  $p_{XY}^{\mathcal{G}}$ , which can be thought of as the projection of  $p_{XY}$  on to the manifold  $\mathcal{G}$ . Similarly, the density functions  $p_X$  and  $p_Y$  can be projected onto the manifold of univariate Gaussian densities, yielding the projections  $p_X^{\mathcal{G}}$  and  $p_Y^{\mathcal{G}}$  respectively; that is

$$p_X^{\mathcal{G}} = p_{\mathcal{N}(\mathbb{E} X, \text{var}(X))}. \quad (9)$$

The concept of projecting a pdf onto a manifold also applies to the formation of the product-of-marginals density  $p_X \otimes p_Y$ . This can be understood as the projection of  $p_{XY}$  onto the manifold of product densities  $\mathcal{P}$ , and thus written as  $p_{XY}^{\mathcal{P}}$ .

Following Cardoso [Cardoso, 2003], we call

$$G(X) \triangleq D(p_X \parallel p_X^{\mathcal{G}}) \quad (10)$$

the *non-Gaussianity* of  $X$ , and similarly  $G(X, Y) \triangleq D(p_{XY} \parallel p_{XY}^{\mathcal{G}})$  is the non-Gaussianity of the pair  $(X, Y)$ . Since it is a KL-divergence,  $G(X) \geq 0$  with equality iff  $X$  is Gaussian. Similarly  $G(X, Y) = 0$  iff the pair  $(X, Y)$  is jointly Gaussian. Finally we have

$$\begin{aligned} I^{\mathcal{G}}(X, Y) &= D(p_{XY}^{\mathcal{G}} \parallel p_X^{\mathcal{G}} \otimes p_Y^{\mathcal{G}}) \\ &= -\frac{1}{2} \log(1 - \rho_{XY}^2) \geq 0 \end{aligned} \quad (11)$$

echoing (2), as the Gaussian mutual information between  $X$  and  $Y$ .

The geometry of the Gaussian manifold  $\mathcal{G}$  and the product manifold  $\mathcal{P}$  is such that the projections commute, which means that the KL-divergence  $D(p_{XY} \parallel p_X^{\mathcal{G}} \otimes p_Y^{\mathcal{G}})$  can be decomposed as either (i) a projection onto  $\mathcal{G}$  followed by a projection onto  $\mathcal{P}$

$$\begin{aligned} &D(p_{XY} \parallel p_{XY}^{\mathcal{G}\mathcal{P}}) \\ &= D(p_{XY} \parallel p_{XY}^{\mathcal{G}}) + D(p_{XY}^{\mathcal{G}} \parallel p_{XY}^{\mathcal{G}\mathcal{P}}) \\ &= G(X, Y) + I^{\mathcal{G}}(X, Y) \end{aligned} \quad (12)$$

or alternatively as (ii) a projection onto  $\mathcal{P}$  followed by a projection onto  $\mathcal{G}$  [Csiszár, 1975, Amari, 1985, Cardoso, 2003]:

$$\begin{aligned} D(p_{XY} \parallel p_{XY}^{\mathcal{P}\mathcal{G}}) &= D(p_{XY} \parallel p_{XY}^{\mathcal{P}}) + D(p_{XY}^{\mathcal{P}} \parallel p_{XY}^{\mathcal{P}\mathcal{G}}) \\ &= I(X, Y) + G(X) + G(Y). \end{aligned} \quad (13)$$

(For a proof of this relation without appealing to the geometry of the manifolds, see [Shore and Johnson, 1981, Property 10].) These relationships can be visualised in as follows:

$$\begin{array}{ccc} p_{XY} & \xrightarrow{G(X, Y)} & p_{XY}^{\mathcal{G}} \\ \downarrow I(X, Y) & & \downarrow I^{\mathcal{G}}(X, Y) \\ p_X \otimes p_Y & \xrightarrow{G(X) + G(Y)} & p_X^{\mathcal{G}} \otimes p_Y^{\mathcal{G}} \end{array} \quad (14)$$

Since  $p_{XY}^{\mathcal{G}\mathcal{P}} = p_{XY}^{\mathcal{P}\mathcal{G}} = p_X \otimes p_Y$ , the two paths must give the same total divergence and we obtain

$$I(X, Y) = I^{\mathcal{G}}(X, Y) + G(X, Y) - G(X) - G(Y) \quad (15)$$

and hence

$$I^{\mathcal{G}}(X, Y) \leq I(X, Y) + G(X) + G(Y). \quad (16)$$

Thus, if  $X$  and  $Y$  are both *marginally* Gaussian (without necessarily being jointly Gaussian), then  $G(X) = G(Y) = 0$  and  $I^{\mathcal{G}}(X, Y)$  will indeed be a lower bound on  $I(X, Y)$ . To the extent that  $X$  and  $Y$  are ‘nearly’ Gaussian, greater or lesser confidence can be placed in  $I^{\mathcal{G}}(X, Y)$  as an approximate lower bound on the mutual information.

In what follows we will assume that this is a reasonable approach to take: building our measurements on simply the correlation coefficient  $\rho_{XY}$  is a very convenient approach, if it is valid. It is possible that we could extend the argument above to other exponential families, which would give us more flexibility in our approach [Verdú, 1996].

**2.4. Alternative mutual information estimates.** While we will estimate mutual information based on  $\rho_{XY}$ , for completeness we should mention that there are several alternative methods available to estimate mutual information from a data set, and some of these have recently been applied to ICA problems.

For example, Darbellay and Vajda [Darbellay and Vajda, 1999a] introduced a histogram-like method for estimating mutual information which adaptively partitions the input space so that smaller cells are used where required, and this was subsequently applied to the analysis of wavelet models of images [Liu and Moulin, 2001].

In the bioinformatics field, the correlation coefficient  $\rho_{XY}$  (there called *Pearson correlation*  $\hat{C}_{XY}$ ) has traditionally been used as a measure of similarity for gene-expression measurements. Steuer *et al.* [Steuer et al., 2002] observed that  $|\rho_{XY}|$  shows a good one-to-one correspondence with mutual information on this type of data, and they found no situations with low  $\rho_{XY}$  but high  $I(X, Y)$ . However, this appears to depend somewhat on the data considered, since a subsequently developed B-spline method showed improved analysis over Pearson correlation [Daub et al., 2004].

Pham [Pham, 2003] described a fast mutual information estimator applied to post-nonlinear ICA, and Learned, Miller and Fisher [Learned-Miller and Fisher, 2003] performed ICA using MI estimator based on the estimator of Vasicek [Vasicek, 1976] which estimates the entropy using the spacing between ordered samples as a guide to the (inverse) probability density.

Kraskov *et al.* [Kraskov et al., 2004] developed an estimator for mutual information from  $k$ -nearest neighbour statistics. This is used as part of a least-dependent-component analysis approach to source separation applied to a range of problems including foetal ECG analysis [Stögbauer et al., 2004] (see also [Nicolaou and Nasuto, 2005]).

### 3. MULTIDIMENSIONAL SCALING

In the previous section we have considered how to estimate the mutual information between random variables using, for example, nonlinear correlation. We now consider how to arrange these variables in a metric space in order to visualise geometrically the dependency structure. We will adopt a method based on multidimensional scaling (MDS) [Cox and Cox, 2001].

The aim of MDS is to arrange a set of points in some (usually low-dimensional) metric space given only the distances between pairs of points. These might be actual distances measured in some higher-dimensional space, or they might be arbitrary dissimilarities obtained by some other method. In either case, it may not be possible to find an arrangement of points that gives exactly the target distances, in which case, the aim of MDS is to minimise a *stress* function which gauges the discrepancy between the target and the achieved distances.

Let us suppose that  $\mathcal{A}$  is a finite set and  $d_{\mathcal{A}} : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}^*$  is a function which gives the target distance  $d_{\mathcal{A}}(\alpha, \beta)$  between any two objects  $\alpha$  and  $\beta$  in  $\mathcal{A}$ . Each  $\alpha \in \mathcal{A}$  is assigned a point  $\mathbf{r}_{\alpha} \in \mathcal{M}$  where  $(\mathcal{M}, d_{\mathcal{M}})$  is a metric space, typically a low-dimensional Euclidean space. Letting  $\mathcal{M}^{\mathcal{A}}$  denote the set of all possible mappings from  $\mathcal{A}$  to  $\mathcal{M}$ , and  $\mathbf{r}$  denote a particular mapping (and hence configuration of points), MDS consists of minimising a stress function  $J : \mathcal{M}^{\mathcal{A}} \rightarrow \mathbb{R}$

$$J(\mathbf{r}) = \sum_{\alpha, \beta \in \mathcal{A}} f_J(d_{\mathcal{M}}(\mathbf{r}_{\alpha}, \mathbf{r}_{\beta}), d_{\mathcal{A}}(\alpha, \beta)), \quad (17)$$

where  $f_J: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is a function which measures the discrepancy between a target distance and an actual distance in the metric space, and  $\mathbf{r}_\alpha$  is the image of  $\alpha$  under the mapping  $\mathbf{r}$ .

Depending on  $f_J$ , different forms of MDS found in the literature may be obtained. Klock and Buhmann [Klock and Buhmann, 1997] describe the three stress functions described below as *global*, *intermediate* and *local* normalisation of the a stress contribution from each pair of points, which is how we will refer to them. In each case, the stress contribution is the squared-difference between distances.

Torgerson’s least-squares MDS [Torgeson, 1952] uses

$$f_G(u, v) = (u - v)^2 \tag{18}$$

which therefore minimises the mean squared error between the point distances in the new space and those in the original space. This is ‘global’ normalisation since each squared difference receives the same weight in the total stress.

Sammon [Sammon, 1969] proposed the use of the stress function

$$f_I(u, v) = \frac{(u - v)^2}{v} \tag{19}$$

where the extra factor  $1/v$  means that contributions to the stress from shorter target distances are given greater weight, and hence more accurately reflected in the resulting configuration than are longer distances.

If we believe that the target distances are known only with some *relative* accuracy, such that the errors are proportional to the distances themselves, then the function

$$f_L(u, v) = \left( \frac{u - v}{v} \right)^2 \tag{20}$$

can be used, since it represents the square of the relative error in matching the target and realised distances. This is ‘local’ normalisation since each squared-difference is normalised by the square of the target distance, meaning that contributions from short distances are weighted even more strongly than in the Sammon stress function (19).

Once a stress function is chosen, one of a number of optimisation algorithms can be applied to find a minimum. We used a general purpose trust-region quasi-Newton algorithm as implemented in the Matlab optimisation toolbox function *fminunc*. This makes use of the analytical derivatives of the stress function with respect to the coordinates of the points in  $\mathcal{M}$ ; these are quite straightforward to derive and not given here. Because of the dimensionality of the optimisation (in excess of 1000 variables in our experiments), only the diagonal part of the Hessian was approximated during the optimisation.

#### 4. INFERRING AN ‘EFFECTIVE DISTANCE’ FOR MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) can generate a spatial representation of objects given only their pairwise distances, but the dependency estimation methods of § 2

yield a correlation coefficient or a mutual information. To do MDS using dependency estimates, one option would be to use *nonmetric* MDS [Kruskal, 1964], where the given dissimilarities are assumed only to be monotonically related to some distance measure. However, we wish to retain a geometric interpretation of the result, so next step is to find a reasonable mapping from the mutual information to a distance such that some interesting structure can emerge from MDS.

The most important requirement for the distance measure is that it be a monotonically decreasing function of the mutual information, so that more dependent units are brought closer together than less dependent units. It is also reasonable to require that the distance tend to infinity as the mutual information tends to zero. Beyond that, the choice is to some extent arbitrary. In our work, we have focussed on a particular mapping from mutual information to distance obtained by considering a hypothetical model which is describe below.

Our starting point is the hypothesis that the notions of similarity and distance are related to the idea of ‘significant’ versus ‘insignificant’ differences and hence to the concept of ‘noise’ as that part of a signal which does not signify anything of importance. Consider, for example, the relationship between spherical Gaussian noise, mean-square error measures, and Euclidean distance: if a process includes Gaussian noise, then mean-square error measures the likelihood that some difference can be accounted for by noise, and Euclidean distance is an appropriate measure of distance. These ideas are discussed at greater length in [Abdallah, 2002], but the main conclusion we use here is that statistical structure resulting from a Gaussian noise process should lead to the familiar Euclidean distance measure.

With that thought in mind, consider an hypothetical model, which, for the sake of argument, one can think of as an imaging model. In this model, an ‘image’ consists of a set of lattice sites (‘pixel’ postions) in a Euclidean space at which there may be a certain number of particles (‘photons’). The noise process consists in each particle undergoing a random displacement to a nearby lattice site according to a certain probability distribution. If the number of particles at each site is initially independent, this process will induce observable dependencies between neighbouring lattice sites. If many such images are observed and the displacement distribution is known, it should be possible to recover the geometry of the lattice by examining these induced dependencies.

On passing to the continuous limit in both the lattice density and the number of particles at each position, the image becomes a continuous field  $u : \mathbb{R}^E \rightarrow \mathbb{R}$  over the  $E$ -dimensional space  $\mathbb{R}^E$ , and the noise process becomes, on the average, a convolution with a kernel (or ‘point spread’ function)  $\phi : \mathbb{R}^E \rightarrow \mathbb{R}$  representing the distribution of displacements. The final image is therefore  $x = \phi * u$ , where  $*$  denotes convolution in  $E$  dimensions. Let us further suppose that the noiseless images are Gaussian and spatially uncorrelated and the kernel is also Gaussian:  $\phi(\mathbf{r}) \propto \exp(-\|\mathbf{r}\|^2)$ . Under these conditions, we can predict the correlation between

(a) Multivariate distributions

Pareto	$\text{Par}_N(\alpha) \sim 1 + \frac{\text{iid}_N(\text{Exp})}{\mathcal{G}(\alpha, 1)}$	$\alpha \in \{0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5, 8\}$
Burr	$\text{Burr}_N(c, \alpha) \sim [\text{Par}_N(\alpha) - 1]^{1/c}$	$\alpha \in \{0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5, 8\}$ $c \in \{1, 2, 6\}$
Weibull	$\text{Weibull}_N(c, \alpha) \sim [\log \text{Par}_N(\alpha)]^{1/c}$	$\alpha \in \{0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5, 8\}$ $c \in \{1, 2, 6\}$
Student's t	$\text{T}_N(v) \sim \frac{\text{iid}_N(\mathcal{N}(0, 1))}{\sqrt{\mathcal{G}(v/2, 2/v)}}$	$v \in \{0.5, 0.75, 1, 2, 3, 5, 10\}$

(b) Nonlinear functions

$\lambda t.t^2$	$\lambda t. t $	$\lambda t. t ^{0.25}$	$\lambda t. t ^{0.5}$	$\lambda t.\log t^2$	$\lambda t.\log(1+t^2)$
-----------------	-----------------	------------------------	-----------------------	----------------------	-------------------------

FIGURE 3. Specifications of (a) multivariate dependent distributions and (b) nonlinear functions used in this article.  $\text{Exp} : \Omega \rightarrow \mathbb{R}^*$  denotes a univariate exponential random variable, that is,  $p_{\text{Exp}}(x) = e^{-x}$ ,  $\mathcal{G}(a, b)$  denotes a Gamma random variable with shape parameter  $a$  and scale parameter  $b$ , and  $\text{iid}_N(X)$  denotes an  $N$ -component random tuple constructed from  $N$  independent copies of  $X$ . Operators such as  $\log$ , addition and exponentiation are to be understood as acting componentwise when applied to random tuples. (The functions are written using the  $\lambda$ -calculus notation, where  $\lambda x.y$  denotes the function obtained by treating the variable  $x$  as a formal parameter in expression  $y$ . For example,  $\lambda t.t^2$  denotes the function that squares its argument.)

values at different lattice sites as a function of the distance between them. Since the autocorrelation function of the noiseless images,  $R_{uu}$ , is zero at all shifts except zero, and  $x$  is obtained by shift-invariant filtering, the autocorrelation function of the observed images is  $R_{xx} = R_{uu}(\mathbf{0})(\phi * \phi^\dagger)$  where  $\phi^\dagger(\mathbf{r}) = \phi(-\mathbf{r})$ . For the Gaussian kernel, we find that  $R_{xx}(\mathbf{r}) \propto R_{uu}(\mathbf{0}) \exp(-\frac{1}{2}\|\mathbf{r}\|^2)$  and hence the correlation coefficient  $\rho_{\alpha\beta} = \text{corr}(x(\alpha), x(\beta))$  between two sites  $\alpha$  and  $\beta$  separated by the vector  $\mathbf{r} = \beta - \alpha$  is  $\rho_{\alpha\beta} = R_{xx}(\mathbf{r})/R_{xx}(\mathbf{0}) = \exp(-\frac{1}{2}\|\mathbf{r}\|^2)$ , where  $\|\mathbf{r}\|$  is nothing other than the Euclidean distance in  $\mathbb{R}^E$  between  $\alpha$  and  $\beta$ .

We suggest that this is a plausible relationship between correlation and distance for Gaussian random variables. Furthermore, using (2), we can express the correlation coefficient and hence the distance in terms of the mutual information, yielding an expression which we propose as a reasonable *definition* of dissimilarity between continuous random variables as a function of their mutual information. Returning to the notation used in the introduction, where  $X_\alpha$  is the random variable labelled

by  $\alpha \in \mathcal{A}$ , and  $d_{\mathcal{A}}$  is the desired dissimilarity function on  $\mathcal{A}$ , we obtain

$$d_{\mathcal{A}}(\alpha, \beta) \triangleq \sqrt{-\log(1 - \exp[-2I(X_{\alpha}, X_{\beta})])}. \quad (21)$$

While this is a somewhat speculative proposal obtained by heuristic means, it does display the appropriate asymptotic behaviour in that  $d_{\mathcal{A}}(\alpha, \beta) \rightarrow 0$  as  $I(X_{\alpha}, X_{\beta}) \rightarrow \infty$  and  $d_{\mathcal{A}}(\alpha, \beta) \rightarrow \infty$  as  $I(X_{\alpha}, X_{\beta}) \rightarrow 0$ . While it is symmetric in its two arguments, it does not qualify as a metric since it does not necessarily satisfy the triangle inequality, and it is possible to have  $d_{\mathcal{A}}(\alpha, \beta) = 0$  with  $\alpha \neq \beta$  in the case that  $X_{\alpha}$  and  $X_{\beta}$  are deterministically related. However, it has proved to give sensible and interesting results, as we will show in the following experiments.

## 5. EVALUATION WITH ARTIFICIAL DATA

In the experiments below, we evaluate the performance of the nonlinear correlation based mutual information estimation method, using data drawn from several bivariate distributions and using several nonlinear functions. In each case, the analytically derived true mutual information is compared with estimates derived from the nonlinear correlation. We also estimate the non-Gaussianity of the marginals using a histogram based method in order to verify the inequality in (16).

When we come to analysing audio data, we cannot know the true mutual information. We therefore judge the performance of the nonlinear correlation method against an adaptive-binning histogram-based method based on that originally proposed by Darbellay and Vajda [Darbellay and Vajda, 1999b] (see appendix § A for further details). In order to gauge its accuracy and reliability, we also apply the histogram-based method to the artificially generated data.

Data was sampled from a selection of bivariate distributions related to the multivariate Pareto distribution [Devroye, 1986, p. 601], including multivariate forms of the Burr and Weibull distributions as defined in fig. 3(a). These are easy to sample from and the mutual information is expressible analytically [Darbellay and Vajda, 2000]. The different members of the family are obtainable from each other by invertible transformations of the variables, and hence, by the data processing inequality (3), share the same expression for the mutual information. In the bivariate case, e.g.  $(X, Y) \sim \text{Par}_2(\alpha)$ , this reduces to

$$I(X, Y) = \frac{1 - \alpha}{\alpha(1 + \alpha)} + \log \frac{1 + \alpha}{\alpha}. \quad (22)$$

We also used a family of circularly symmetric multivariate Student's t distributions  $T_2(v)$  with different degrees of dependence (controlled by the degrees of freedom parameter  $v$ ) since this family also has an analytically expressible mutual information and approximates some of the circularly symmetric heavy-tailed joint distributions resulting from ICA of audio data (such as the one illustrated in the bottom-left of fig. 2). In the case that  $(X, Y) \sim T_2(v)$ , the mutual information is

$$I(X, Y) = \gamma\left(\frac{v}{2}\right) - 2\gamma\left(\frac{v+1}{2}\right) + \gamma\left(\frac{v+2}{2}\right) \quad \text{where} \quad \gamma(u) = \log \Gamma(u) - u\Psi(u) \quad (23)$$

**Bivariate Burr:  $\text{Burr}_2(c, \alpha)$**

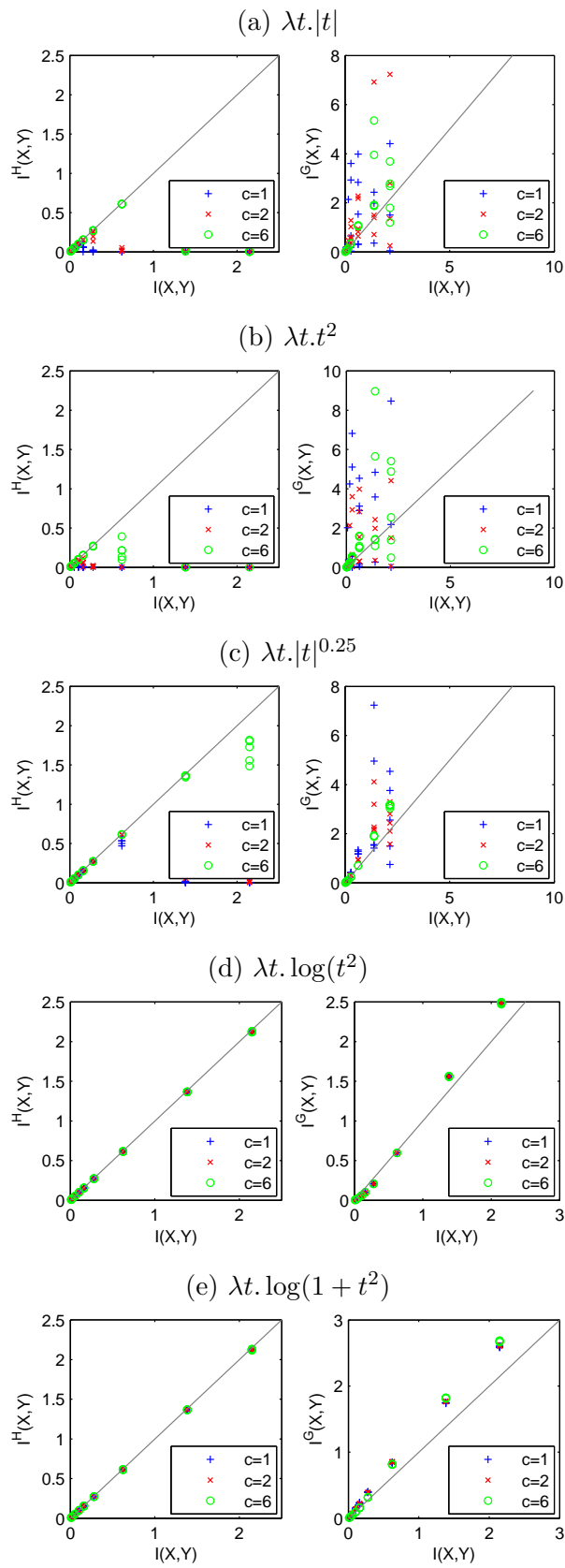


FIGURE 4. Tests with bivariate Burr distributions. See text for details.

**Bivariate Weibull: Weibull<sub>2</sub>(c, α)**

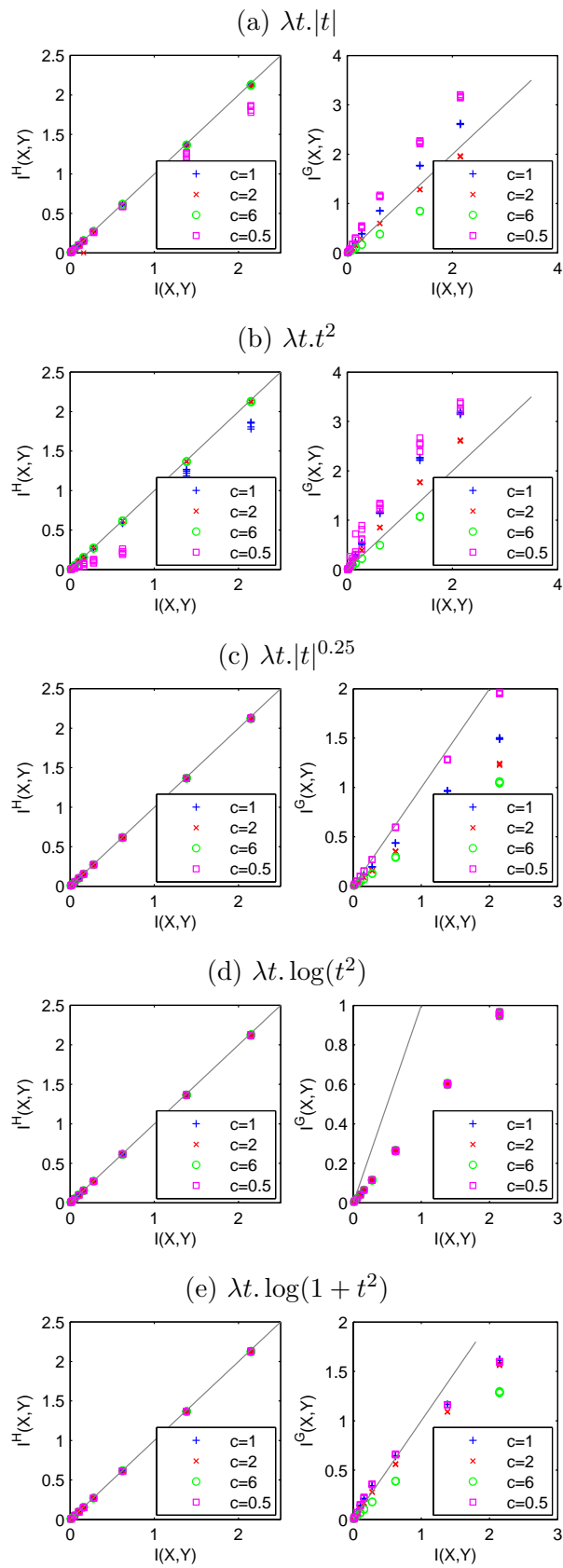


FIGURE 5. Tests with bivariate Weibull distributions. See text for details.

and  $\Gamma$  is the gamma function and  $\Psi$  is the digamma function.

For each family of distributions and each combination of parameters, 5 independent draws of 100,000 samples each were taken. Then, each of the functions in fig. 3(b) was applied and the Gaussian mutual information  $I^G(X, Y)$  estimated from the measured correlation coefficient, according to (2). The mutual information was also estimated from the same nonlinearly-transformed data using the histogram-based method (appendix §A); this will be denoted by  $I^H(X, Y)$ . The non-Gaussianities  $G(X)$  and  $G(Y)$  were estimated using an identity derivable from (10):

$$\begin{aligned} G(X) &= H(\mathcal{N}(\mathbb{E} X, \text{var}(X))) - H(X) \\ &= \frac{1}{2}(1 + \log 2\pi \text{var}(X)) - H(X). \end{aligned} \tag{24}$$

that is, the non-Gaussianity of  $X$  is the difference between the entropy of  $X$  and the entropy of a Gaussian with the same variance. The variances were estimated directly from the data while the entropies were estimated from 1-dimensional histograms of the marginal distributions of  $X$  and  $Y$ . Finally, the true mutual information was computed for each distribution using (22) and (23).

Some of these results are plotted and compared in figures 4 to 6. Each 5-by-2 grid of plots pertains to one family of random variables while each row shows the results obtained using one nonlinear function. Each scatter-plot shows the aggregate results for each of the 5 independent draws for each combination of the parameters listed in fig. 3(a). The first column shows the histogram-based estimates of mutual information while the second shows the nonlinear correlation-based estimates. The results for the bivariate Pareto distribution are not shown due to lack of space, but they are qualitatively similar to the results for the Burr distribution with  $c = 1$ .

Several observations can be made about these results. They confirm that using the function  $\lambda t \cdot |t|$  (which is equivalent to the identity function for the non-negative Pareto, Burr and Weibull variables), it is difficult to get good estimates of mutual information when the data distribution is very heavy tailed, which is the case for the bivariate Pareto, Burr and Student's  $t$  distributions (see figures 4(a) and 6(a)). In these cases, both the histogram and the nonlinear correlation based estimates of mutual information are not reliable. This situation is exacerbated by squaring the values using the function  $\lambda t \cdot t^2$  (plot (b) in figures 4–6), which makes the tails of the distribution even heavier. In contrast, as more ‘compressive’ nonlinearities such as  $\lambda t \cdot |t|^{0.25}$  and  $\lambda t \cdot \log t^2$  are used (plots (c)–(e) in figures 4–6), the histogram-based mutual information estimates become quite accurate while the nonlinear correlation-based estimates have much less variance and increase monotonically with the true mutual information.

For the less heavy-tailed bivariate Weibull distributions, the less compressive nonlinearities (first three rows in fig. 5) result in reasonable MI estimates. This is consistent with the results from the heavy tailed distributions, since a multivariate

**Bivariate Student's t:  $T_2(v)$**

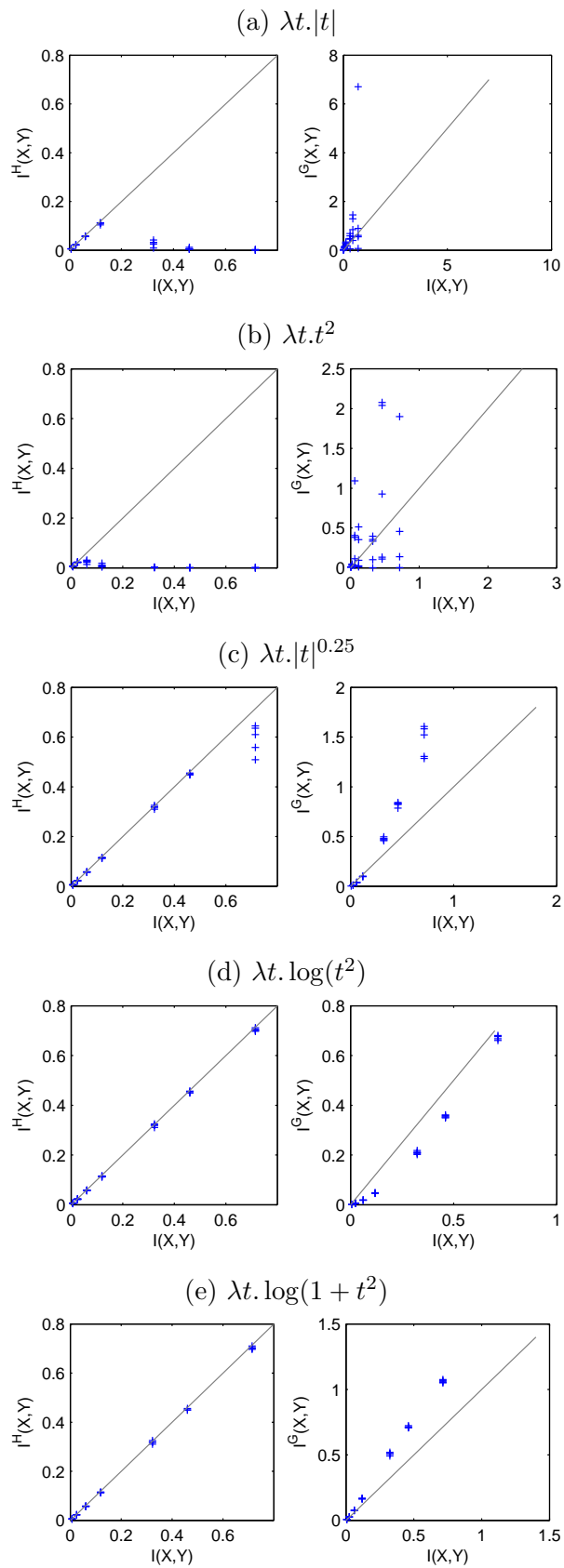


FIGURE 6. Tests with bivariate Student's t distributions.

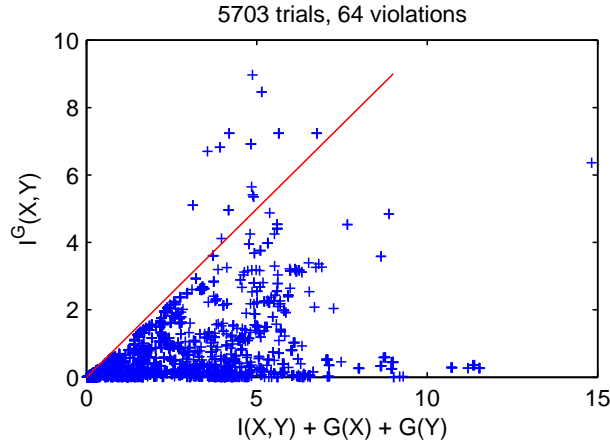


FIGURE 7. Illustration of inequality (16). In 5703 trials using nonlinearly transformed data drawn from several random variables, both the Gaussian mutual information  $I^G(X, Y)$  and the non-Gaussianities  $G(X)$  and  $G(Y)$  were estimated from the data. In 64 trials (the points above the diagonal line), the inequality was violated. In every one of these cases, the data distribution is so heavy-tailed that the true non-Gaussianity is undefined, and thus the numerical estimates cannot be expected to be accurate.

Weibull random variable already incorporates a logarithmic transformation from a Pareto random variable, as shown in fig. 3(a). The results for the circularly symmetric Student’s  $t$  data show a clear improvement in the quality of the estimates when logarithmic compression is used. These are especially relevant for our subsequent experiments, since empirically, the family of bivariate Student’s  $t$  distributions are a good match for some of the circularly symmetric distributions we obtain from ICA, e.g., the lower-left histogram of fig. 2.

Fig. 7 shows a comparison of the two quantities on either side of the inequality of (16). We can see that the relation is largely verified, except for a few points above the diagonal line. The explanation for this can be found in the limited resolution of the histograms used to estimate the entropies of  $X$  and  $Y$ , which results in a systematic under-estimation of the non-Gaussianities  $G(X)$  and  $G(Y)$  of the distributions. In fact, in every case, the true non-Gaussianities of the underlying random variables (which happen to be analytically expressible in these cases) are undefined due to their infinite variance, so the non-Gaussianity cannot in fact be estimated in any meaningful way. It is possible that more sophisticated methods of entropy estimation would see fewer violations of the inequality.

## 6. EXPERIMENTS WITH AUDIO DATA

**6.1. Extraction of ICA basis functions.** The first stage in the evaluation with recorded audio data is to use ICA to form the distributed representation we are going to analyse. The primary motivation for using ICA here is to find a representation in which the activities of the different components are as independent as possible with

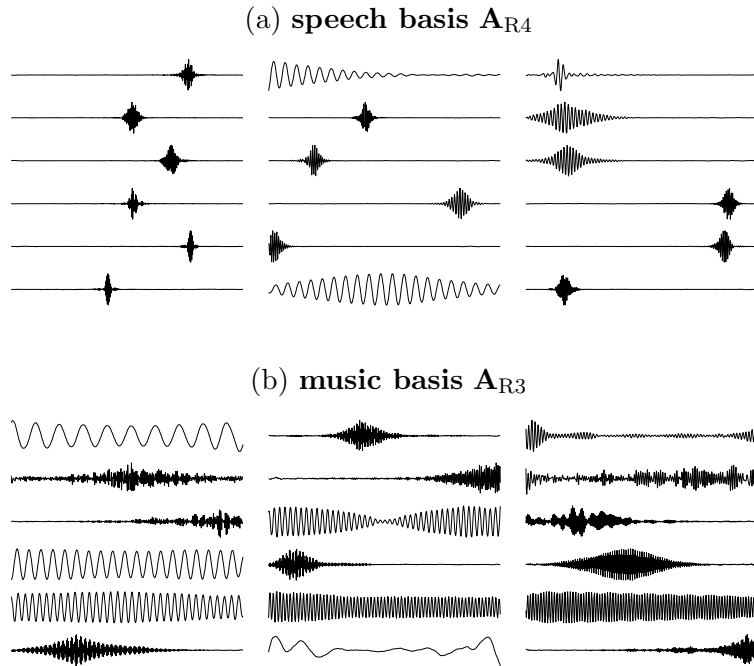


FIGURE 8. Some of the basis vectors obtained by ICA of (a) speech and (b) music. (Each basis has 512 vectors in total.)

the class of transformations available. Given such a representation, we can usefully examine the structure of the remaining dependencies, since this may shed some light on the meaning of the resulting representation and suggest appropriate processing strategies to apply next. (However, the method of geometric dependency analysis can in principle be applied to any distributed representation regardless of how it is defined.)

The discrete-time audio signal, sampled at 11025 Hz, was initially divided into a sequence of overlapping frames of  $N = 512$  samples, with an overlap of 256 samples. This sequence of frames was then subjected to a form of gain normalisation to improve the stability of the subsequent ICA algorithm. Further details can be found elsewhere [Abdallah and Plumbley, 2001]. The end result of the ICA stage was two  $N \times N$  basis matrices  $\mathbf{A}_{R4}$  and  $\mathbf{A}_{R3}$ , both encoding 512 waveforms (see fig. 8) trained using many hours of audio captured from BBC Radio 4 (mainly speech) and BBC Radio 3 (mainly Western classical music). Most of the waveforms in the two bases showed some degree of localisation in time-frequency; for example, each of the 512 ‘blobs’ in fig. 1 is the approximate time-frequency extent of one of the speech-derived basis vectors in  $\mathbf{A}_{R4}$ .

Next, for the purposes of estimating residual dependencies, shorter signals were captured from each of the two radio stations. These were also divided into sequences of overlapping frames of  $N$  samples and subjected to a time-dependent gain control. Letting the frames in this sequence be denoted by the vectors  $s^{(k)} \in \mathcal{S} = \mathbb{R}^N$  with  $k$  ranging from 1 to the length  $K$  of the sequence, then the ICA representation is obtained by referring these vectors to one of the ICA bases, which for simplicity

we will refer to as  $\mathbf{A}$  on the understanding that  $\mathbf{A}_{R4}$  or  $\mathbf{A}_{R3}$  should be used as appropriate for the original signal. Thus, we have a representation function  $\psi_{\mathbf{A}} : \mathbb{R}^N \rightarrow \mathbb{R}^N$  and for a frame  $s$ , we have  $x = \psi_{\mathbf{A}}(s) = \mathbf{A}^{-1}s$ . (Note that our use of  $s$  and  $x$  to denote the observation and its encoding respectively, which we retain in keeping with the discussion of §1, is the reverse of the usual convention in the ICA literature.)

The tuple of coefficients  $x$  will have  $N$  components corresponding with the  $N$  basis vectors, which are encoded in the  $N$  columns of the matrix  $\mathbf{A}$ . If we denote the  $i^{\text{th}}$  column of  $\mathbf{A}$  by  $A_{:i}$ , then we can define our tuple-indexing set  $\mathcal{A}$  as the basis itself, that is, the set of basis vectors  $\mathcal{A} = \{A_{:i} | i \in 1..N\}$ . Using this notation, we can, for example, write  $s = \sum_{\alpha \in \mathcal{A}} \alpha x_{\alpha}$ , which expresses  $s$  as a weighted sum of the basis vectors.

As set out in §1, our aim at this point is to estimate the mutual information between the random variables  $X_{\alpha}$  and  $X_{\beta}$ , of which  $(x_{\alpha}^{(k)})_{k \in 1..K}$  and  $(x_{\beta}^{(k)})_{k \in 1..K}$  are assumed to be realisations, for all pairs  $\alpha, \beta \in \mathcal{A}$ . Joint histograms of two such pairs of components are illustrated in the first column of fig. 2.

**6.2. Estimation of residual dependencies.** The mutual information between each pair of components was estimated using both the method of nonlinear correlation (with each of the nonlinearities listed in §5) and using the histogram-based method. The histogram-based method was applied to the data *after* each of the nonlinearities had been applied, since the results of §5 show that the distribution of the data has a strong effect on its accuracy.

The benefit of using the nonlinear correlation method became apparent here: the histogram method involved compiling some 130,000 2-D histograms and took about 4 days to compute for a 90 minute signal, whereas the nonlinear correlation method took approximately 5 minutes. (All computations were done in Matlab on a Macintosh Dual G5.)

As an indication of some of the latent structure beginning to emerge in these dependency matrices, fig. 9 shows nonlinear correlation matrix obtained from one of the music signals as represented by the music-derived basis  $\mathbf{A}_{R3}$ . In this case, many of the basis vectors are quite localised in frequency and can be characterised by a centre frequency (see [Abdallah and Plumbley, 2001]). The rows and columns of the matrix in fig. 9 have been reordered by frequency, revealing correlations along diagonal lines which is indicative of dependencies between harmonically related frequencies.

Once the dependency estimates have been transformed into distances using (21), an initial characterisation of the results can be obtained by histogramming the values in one distance matrix. This allows us to see, for example, whether the distances are clustered around a particular value or spread out. A few such histograms are shown in fig. 10, which illustrates several points. Firstly, it shows that if no nonlinearity is used (top row), the distances are distributed in an apparently featureless way about

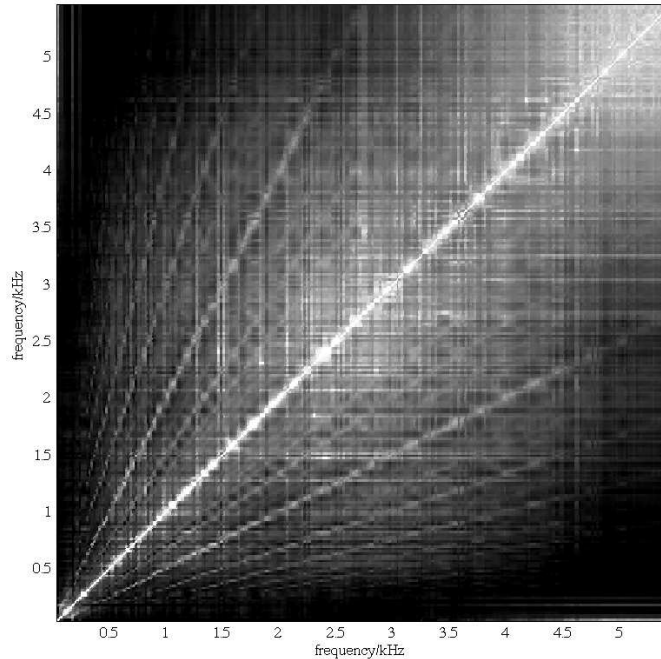


FIGURE 9. Matrix of nonlinear correlation coefficients obtained with the music-derived basis  $\mathbf{A}_{R3}$ , with rows and columns reordered and scaled by centre frequency to show strong correlations between frequencies at small whole number ratios. There is also a slight 12-cycle per octave ‘ripple’ effect visible, which is probably due to the semitone quantisation of the Western musical scale.

a larger value than in the other cases, which is consistent with the idea that linear correlation yields low estimates of mutual information for the type of data we are looking at here.

Secondly, greater width of the distributions (c) and (d) when compared with (e) and (f) is consistent with the reduced variance of mutual information estimates when using a compressive nonlinearity such as  $\lambda t \cdot \log(1 + t^2)$ .

Thirdly, comparing (c) with (d) and (e) with (f) shows that the dynamic gain control both shifts and broadens the distribution of distances, potentially revealing more ‘structure’ in the set of distances. This is consistent with the idea that the gain control tends to factor out any tied amplitude variations across all components which would otherwise result in higher mutual information estimates. Since this would simply have the effect of drawing all components closer together in the final geometry, it is, arguably, uninteresting; reducing the effect could allow more interesting geometric structure to emerge, such as the bimodality of the distribution in fig. 10(f).

Further insight into the relationships between the different dependency estimation methods can be obtained by plotting the distances obtained by one method against

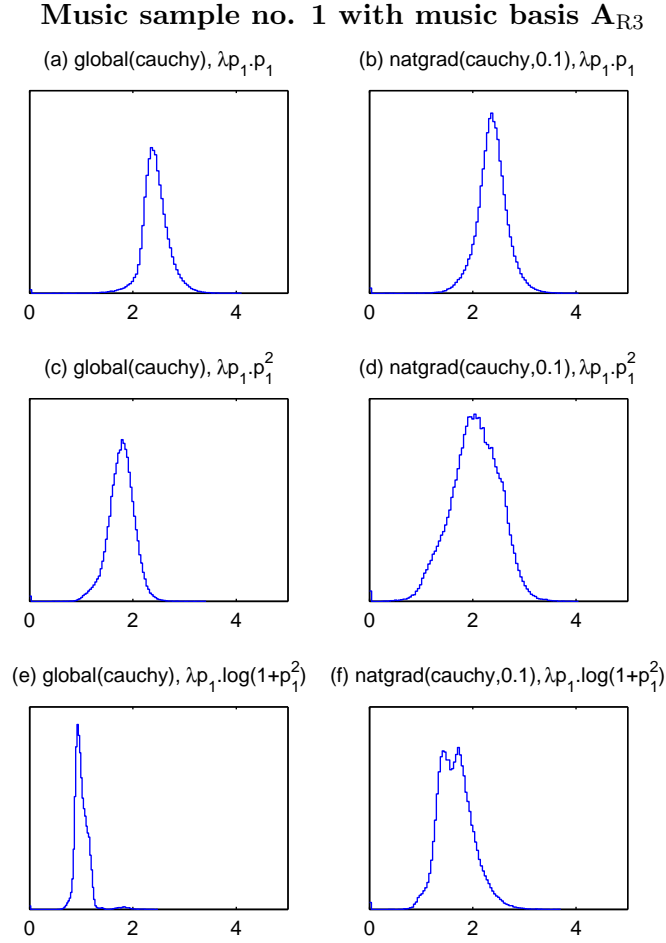


FIGURE 10. Histograms of the pairwise distances between ICA components derived from music data without (first column) and with (second column) dynamic gain normalisation, and using three different nonlinearities to compute the nonlinear correlation.

the corresponding distances obtained by another. Lack of space precludes the inclusion of many such plots, but fig. 11 shows the effect of the nonlinearity by plotting distances obtained using the method of nonlinear correlation against those obtained using the histogram-based MI estimation method. It shows how the more compressive nonlinearities in the bottom row give much more accurate estimates than those in the top row for distances below about 1.5, which corresponds to a MI of 0.08 bits and above.

Similarly, fig. 12 illustrates the effect of the dynamic gain normalisation on the measured dependencies, while fig. 13 shows how consistent the estimated dependency structures are across different samples from the same ensemble, in this case, using two independent recordings from the same radio station. Overall, these plots show that the shorter distances, corresponding with high mutual information, are much more accurately and consistently estimated than larger distances. This ties in with the discussion surrounding the ‘local normalisation’ of MDS stress function (20),

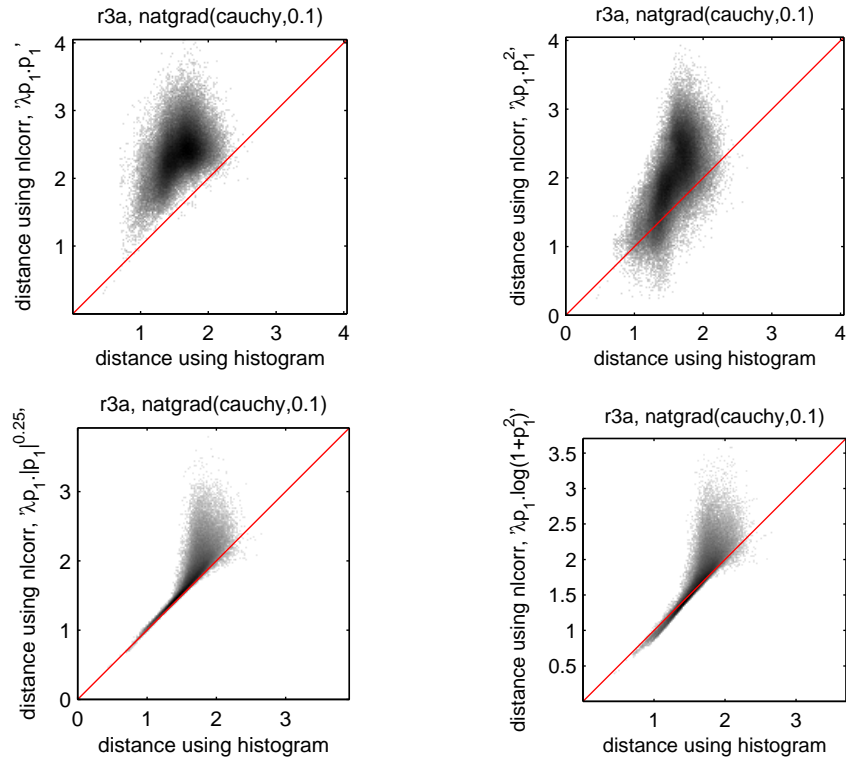


FIGURE 11. Scatter plots of distances computed using nonlinear correlation using several nonlinear functions (vertical axes), against those computed from 2D joint histograms (horizontal axes). These were obtained from the first music sample with dynamic gain adaptation. Note: the top-left plot show the result of using the linear correlation directly.

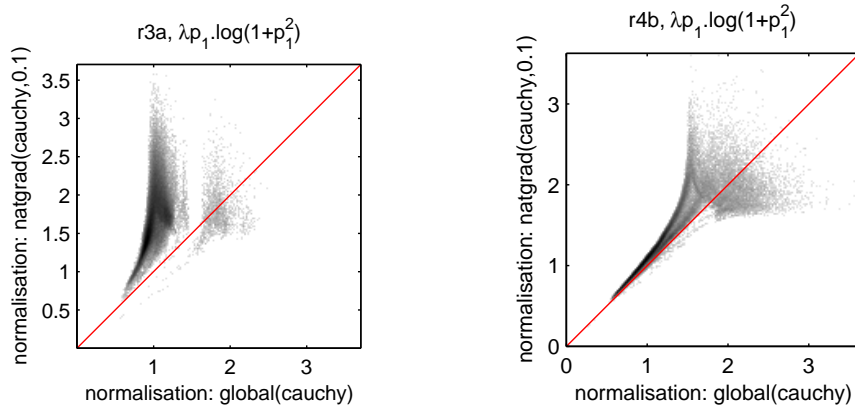


FIGURE 12. Scatter plots of distances computed with without dynamic gain adaptation against those obtained without, for music (left) and speech (right) signals. The effect is more pronounced in the case of the music signal, where distances that are clustered around 1 when no adaptation is used are spread between 1.5 and 3 when adaptive normalisation is used.

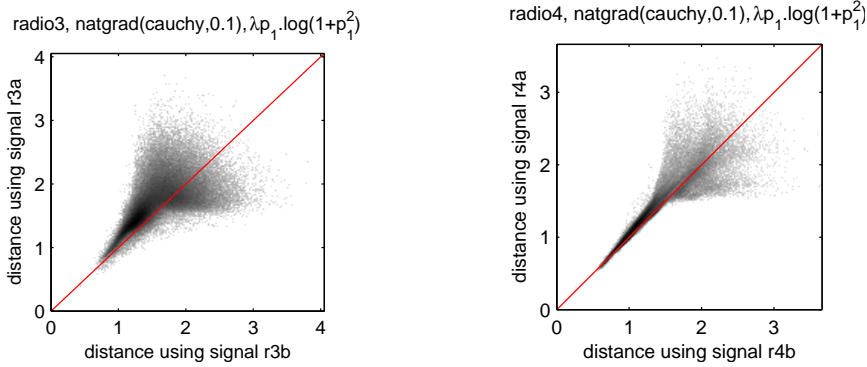


FIGURE 13. Comparison of distances computed with two independent samples from the same radio stations: (left) music recorded from BBC Radio 3; (right) speech recorded from BBC Radio 4. These show that dependency structure of the speech-derived basis  $\mathbf{A}_{R4}$  emerges fairly consistently from the two samples, whereas the dependency structure of the music-derived basis  $\mathbf{A}_{R3}$  is more affected by the particular signal used to estimate it.

which assumes that errors in estimating large distances are greater than those for short distances.

**6.3. Multidimensional scaling.** The final step is to apply MDS to the dissimilarity matrices obtained from the dependency analysis. The three stress functions (18), (19) and (20) were optimised in 2, 3, and 4 dimensions. In each case, the optimiser was run four times starting with 4 different initial configurations. One of these was the result of running Torgerson’s Classical MDS [Torgeson, 1952] on the distance matrix, while the other three were independent random configurations of points.

Again, lack of space precludes the inclusion of many of the results. In all cases, the four independent MDS runs produced comparable results with only small variations in the final stress. The three stress functions (20), (19) and (18) consistently produced similar but slightly distorted configurations; fig. 14 is a typical example of the effect.

Fig. 15 illustrates how the accuracy of the mutual information estimates obtained with different nonlinear functions affects the regularity of the geometry. In the case of the speech derived ICA basis  $\mathbf{A}_{R4}$ , the basis vectors are generally quite well localised in time and frequency (fig. 1). It is illuminating to colour each point according to the nominal frequency or time position of the corresponding basis vector. Figures 14 and 15 show that, for the speech data set, GDA has recovered what is essentially a time-frequency representation. If the MDS is carried out in three dimensions, the points remain approximately localised on a curved two-dimensional time-frequency manifold.

Some results for the music data are shown in fig. 17. These were obtained using approximately 6 hours of audio from BBC Radio 3. The overall structure is shaped

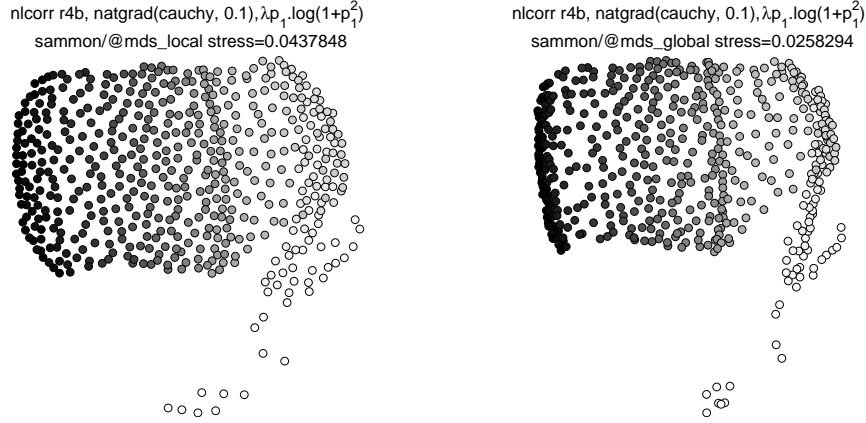


FIGURE 14. Comparison of MDS using two different stress functions: (left) local normalisation (20), and (right) global normalisation (18). The more even spacing of the configuration on the left is due to the greater weight given to short distances in the local normalisation. The colouring of the points corresponds to estimated centre frequency for each basis vector (darker means higher frequency) and shows how the horizontal dimension of the embedding space equates roughly to frequency.

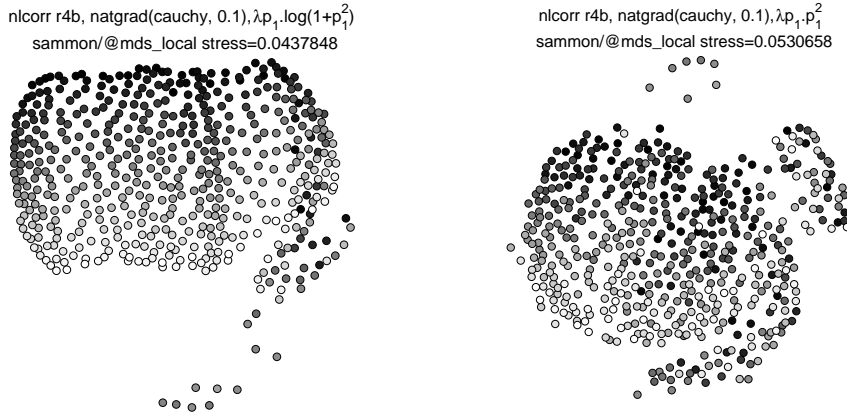


FIGURE 15. Comparison of MDS using two nonlinear correlation measures. In this case, the colouring of the points corresponds to the time localisation of the basis vectors (darker means later) and shows how the vertical dimension equates roughly to time. Note how the mapping is far more regular using the compressive nonlinearity  $\lambda t. \log(1 + t^2)$ .

approximately like a frustrum of a cone with a ‘tail’ of points near the broad end. As the grey scale of fig. 17 shows, position along the axis of the cone corresponds with the centre frequency of the basis vectors, with high-frequencies at the narrow end.

The geometry of the representation can be examined in more detail by looking at the pattern of activity generated when sinusoidal tones are input. A probe signal consisting a sinusoidal chirp sweeping from 55 Hz to 5.5 kHz over 16 s was processed

---

```
nlcorr r4a, natgrad(cauchy, 0.1),  $\lambda p_1 \cdot \log(1+p_1^2)$ 
sammon/@mds_inter stress=0.0187033039063
```

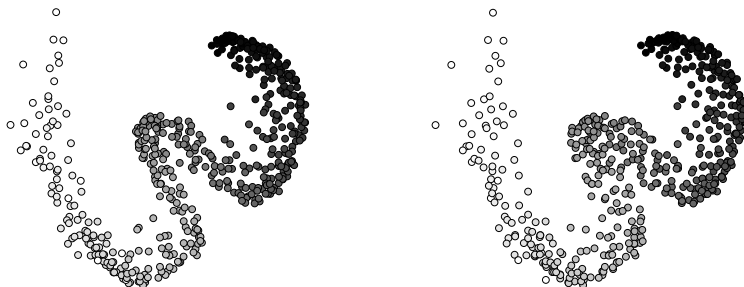


FIGURE 16. Stereo pair of configuration of speech basis  $\mathbf{A}_{R4}$  in three dimensions. The left eye image is on the left. The points lie on what is essentially a two-dimensional curved manifold viewed not quite edge-on here, corresponding to the arrangements of figs. 14 and 15, with frequency along the visible edge (colour coded) and time going perpendicularly into the page.

in the same way as the audio signals used to estimate the dependency structure, resulting in a sequence of patterns of activity in the now geometrically structured ICA representation. The ‘centre of mass’ of each pattern was computed and the progress of this point plotted as the probe tone sweeps up. The structure of the path (shown in fig. 18) is rather difficult to see in a static illustration, but in animated form, one can clearly see the centre of activity moving repeatedly across the conical structure from the surface, through the middle and out to the surface again, as well as moving gradually along the axis towards the narrow end of the frustrum.

This traversal occurs once per semitone, such that units responding to the pitches of the 12 semitone-per-octave Western scale tend to lie near the surface, whilst the interior units tend respond to the ‘out-of-tune’ frequencies. In addition, the pattern of pitches on the surface of the frustrum, when viewed along the axis, form an approximate circle of fifths—a fifth is the interval between two pitches whos fundamental frequencies are in the ratio 2:3, and is an important component of harmony in Western tonal music. A schematic of the arrangement is shown in fig. 19.

Overall, the arrangement reflects three aspects of musical pitch: (a) position on a continuous frequency scale corresponding to distance along the axis of the frustrum, (b) quantisation by semitones corresponding to distance from the axis, and (c) relationships between frequencies in the ratios 2:3 and 3:4 are reflected in the circle of fifths arrangement around the circumference of the frustrum

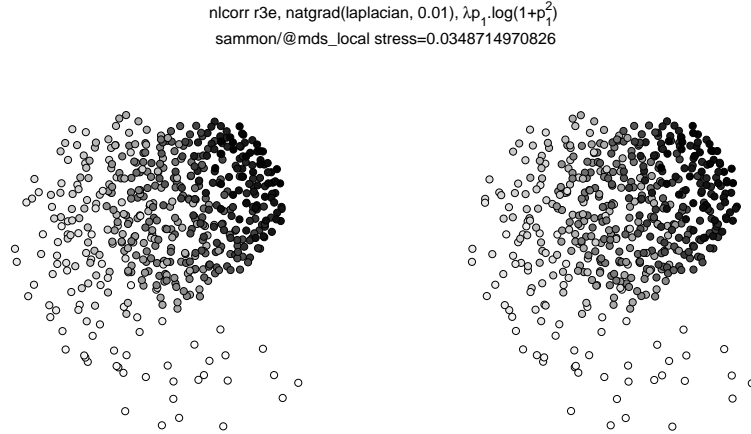


FIGURE 17. Stereo pair of configuration of music basis  $\mathbf{A}_{R3}$  in three dimensions. Here, the colour scale encodes the estimated centre frequency of each basis vector. The representation is approximately a frustum of cone, with its axis pointing right, slightly up, and slightly out of the page.

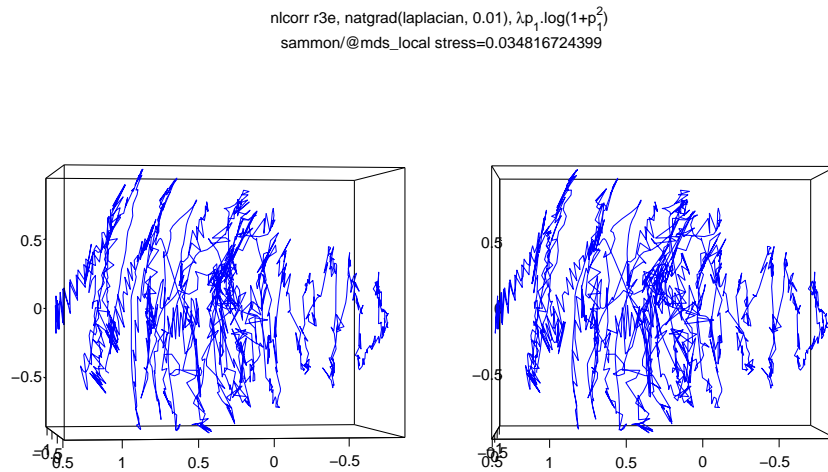


FIGURE 18. Stereo pair of the path of the ‘centre of activity’ in the music derived representation as a sinusoidal probe tone is swept up from 351 Hz up to 5.5 kHz. As well as a gradual progression along the axis (horizontal) of the conical structure as the frequency of the probe tone goes up, the centre of activity also crosses back and forth once every semitone, tracing out an approximate ‘spiral of fifths’ on the surface of the frustum.

## 7. CONCLUSIONS

In this paper we investigated the idea of a geometric representation of the dependency structure in a distributed representation, whereby strongly dependent components are placed close together in a geometric structure embedded in a low-dimensional Euclidean space.

The mutual information between pairs of random variables was estimated using the method of nonlinear correlation analysis, which was found to give good results

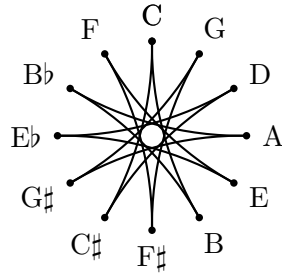


FIGURE 19. An idealised schematic of the pattern of activation (viewed along the axis of the conical structure) generated as a sinusoidal probe tone is gradually swept up in frequency. The pattern of frequencies around the rim matches the fundamental frequencies of the pitches making up the circle of fifth; that is, each step around the rim of the cone corresponds to a frequency ratio of 2:3.

for the heavy-tailed distributions we examined. The mutual information estimates obtained using compressive nonlinear functions such as  $\lambda t \cdot |t|^{0.25}$  or  $\lambda t \cdot \log(1+t^2)$  were found to be more accurate and less variable across samples than those obtained using no nonlinearity or non-compressive functions like  $\lambda t \cdot t^2$ . Good estimates of mutual information can also be obtained using adaptive-binning histogram methods, but at far greater computational expense.

When independent component analysis was applied to speech and music audio signals, the resulting components are not completely independent, but have a non-trivial residual dependency structure. This was estimated using the method of nonlinear correlations and manifested geometrically using multidimensional scaling. The accuracy of the mutual information estimates was reflected directly in the regularity of the geometric structures so obtained (see fig. 15).

In the case of the speech-derived ICA system, the dependency analysis revealed what is essentially a time-frequency manifold, in agreement with what one might intuitively expect from looking at the time-frequency distributions of the basis vectors. What is interesting is that both the basis and the geometric configuration were developed in an entirely data-driven way, suggesting that a two-dimensional time-frequency representation is in some way intrinsic to speech signals, at least when examined at this relatively low level.

In contrast, the the music based representation was quite different, and while still being organised partially in terms of frequency, had a more complex structure that reflects the harmonic and scalar structure of Western music, and indicating that a time-frequency representation may be too simplistic a way of looking at music signals at this level.

The method of geometric dependency analysis is closely related to topographic ICA [Hyvärinen et al., 2001], which also uses residual dependency to define similarity relationship in a distributed representation. One important difference is that

topographic ICA requires a predefined topology, into which the ICA basis is fitted, whereas GDA allows the topology to emerge from the data. The speech basis, for example, maintains its 2-D topology even in a 3-D embedding space. In addition, GDA may have applications in independent subspace analysis [Hyvärinen and Hoyer, 2000]: if the components fall into mutually independent groups with strong intra-group dependencies, then they should appear as widely separated clusters in the MDS embedding space. A clustering analysis, based on the MDS solution or directly on the distance matrix, could then be used to identify the subspaces.

Another difference between GDA and topographic ICA is that GDA can also be applied post-hoc to any distributed representation as long as a suitable method for estimating the residual dependencies can be found.

Although our initial application of GDA has been primarily as a visualisation tool—to visualise the dependency structure of a distributed representation—the motivation behind GDA is to create a space in which *locality* has a data-derived meaning and reflects some intrinsic structure of the data being analysed. An example of this is probe tone analysis of §6.3, which involved computing a ‘centre of activity’ in the geometric space. Another example is separation of convolutively mixed audio signals [Abdallah and Plumbley, 2004] by using GDA to find two subspaces corresponding to the two mixed signals. Our future work will investigate other types of computation which may benefit from this concept of intrinsic locality.

## 8. ACKNOWLEDGEMENTS

This work is partially supported by EPSRC grants GR/S82213/01, GR/S75802/01, EP/C005554/1 and EP/D000246/1.

### APPENDIX A. HISTOGRAM-BASED ESTIMATION OF MUTUAL INFORMATION

In the experiments of §5 and §6, we used a histogram-based mutual information estimation method modelled on the adaptive binning method of Darbellay and Vajda [Darbellay and Vajda, 1999b]. In this appendix we describe how and why our method differs from the original.

The method of Darbellay and Vajda, (which they refer to as CI, for conditional independence) requires the complete data set as a collection of pairs of values, and proceeds by partitioning the input space in such a way as to leave approximately equal numbers of observations in each of the newly introduced partitions. It is entirely unaffected by separable nonlinear transformations of the sort we use to measure nonlinear correlations.

However, we wish to work with potentially huge data sets without continually rescanning through the data, which might be arriving in a continuous stream. Hence, we summarise the sequence of observations with a relatively fine-grained histogram, which then acts as a surrogate for the actual data: each bin represents a certain

number of observations spread evenly over the bin. Each partitioning step can then be accomplished relatively cheaply without requiring access to the original data.

Unfortunately, this does introduce a dependence on the granularity of the histogram and any nonlinear functions applied to the variables, since in pathological cases with very heavy-tailed distributions, even one or two very large values can result in most of the observations ending in just a few bins. This means that the structure of the distribution on scales less than the bin size is lost and the MI estimates can be very erroneous. This is why the histogram-based method fails in the first two or three rows of figures 4–6, where the data being histogrammed is very heavy-tailed indeed. When a compressive nonlinearity is used, as in the last three rows of the above mentioned figures, the method is quite reliable and largely insensitive to the particular nonlinearity used.

## REFERENCES

- [Abdallah, 2002] Abdallah, S. A. (2002). *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*. PhD thesis, Department of Electronic Engineering, King’s College London, UK.
- [Abdallah and Plumbley, 2001] Abdallah, S. A. and Plumbley, M. D. (2001). If edges are the independent components of natural images, what are the independent components of natural sounds? In Lee, T.-W., Jung, T.-P., Makeig, S., and Sejnowski, T. J., editors, *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, December 9-13, 2001, pages 534–539.
- [Abdallah and Plumbley, 2003] Abdallah, S. A. and Plumbley, M. D. (2003). Geometric ICA using nonlinear correlation and MDS. In *Proceedings of the Fourth International Symposium on Independent Component Analysis (ICA2003)*, Nara, Japan, pages 161–166.
- [Abdallah and Plumbley, 2004] Abdallah, S. A. and Plumbley, M. D. (2004). Application of geometric dependency analysis to the separation of convolved mixtures. In *5th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2004)*, Grenada, Spain.
- [Amari and Nagaoka, 2001] Amari, S. and Nagaoka, H. (2001). *Methods of Information Geometry*. American Mathematical Society / Oxford University Press.
- [Amari, 1985] Amari, S.-I. (1985). *Differential-Geometric Methods in Statistics*. Lecture Notes in Statistics, vol. 28. Springer.
- [Bach and Jordan, 2002] Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.
- [Cardoso, 1997] Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114.
- [Cardoso, 2003] Cardoso, J.-F. (2003). Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203.
- [Comon, 1994] Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314.
- [Cox and Cox, 2001] Cox, T. and Cox, M. A. A. (2001). *Multidimensional Scaling*. Chapman & Hall/CRC, London.
- [Csiszár, 1975] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimisation problems. *The Annals of Probability*, 3(1):146–158.

- [Darbellay and Vajda, 1999a] Darbellay, G. A. and Vajda, I. (1999a). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321.
- [Darbellay and Vajda, 1999b] Darbellay, G. A. and Vajda, I. (1999b). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Information Theory*, 45(4):1315–1321.
- [Darbellay and Vajda, 2000] Darbellay, G. A. and Vajda, I. (2000). Entropy expressions for multivariate continuous distributions. *IEEE Trans. Information Theory*, 46(2):709–712.
- [Daub et al., 2004] Daub, C. O., Steuer, R., Selbig, J., , and Kloska, S. (2004). Estimating mutual information using b-spline functions — an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5:118.
- [Devroye, 1986] Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag.
- [Hyvärinen and Hoyer, 2000] Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720.
- [Hyvärinen and Hoyer, 2000] Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720.
- [Hyvärinen et al., 2001] Hyvärinen, A., Hoyer, P., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558.
- [Hyvärinen et al., 2001] Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13:1527–1558.
- [Klock and Buhmann, 1997] Klock, H. and Buhmann, J. M. (1997). Multidimensional scaling by deterministic annealing. In Pelillo, M. and Hancock, E. R., editors, *Energy Minimisation Methods in Computer Vision and Pattern Recognition, Proc. Int. Workshop EMMCVPR'97*, volume 1223 of *Lecture Notes in Computer Science*, pages 246–260.
- [Kraskov et al., 2004] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69:066138.
- [Kruskal, 1964] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:1–28,115–129.
- [Learned-Miller and Fisher, 2003] Learned-Miller, E. G. and Fisher, J. W. (2003). ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295.
- [Liu and Moulin, 2001] Liu, J. and Moulin, P. (2001). Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Transactions on Image Processing*, 10(11):1647–1658.
- [Matsuda and Yamaguchi, 2003] Matsuda, Y. and Yamaguchi, K. (2003). Linear multilayer ICA integrating small local modules. In *4th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2003)*, Nara, Japan.
- [Mitchison, 1995] Mitchison, G. J. (1995). A type of duality between self-organizing maps and minimal wiring. *Neural Computation*, 7(1):25–35.
- [Nicolau and Nasuto, 2005] Nicolaou, N. and Nasuto, S. (2005). Mutual information for EEG analysis. In *Proceedings of the 4th IEEE EMBSS UKRI PG Conference on Biomedical Engineering and Medical Physics*, pages 23–24.
- [Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339.
- [Pham, 2003] Pham, D. T. (2003). Fast algorithm for estimating mutual information, entropies and score functions. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan*, pages 17–22.

- [Sammon, 1969] Sammon, J. W. (1969). A non-linear mapping for data structure analysis. *IEEE Trans. Comput.*, C18(5):401–409.
- [Schwartz and Simoncelli, 2001] Schwartz, O. and Simoncelli, E. P. (2001). Natural sound statistics and divisive normalisation in the auditory system. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 166–172. MIT Press.
- [Shannon, 1951] Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.
- [Shore and Johnson, 1981] Shore, J. E. and Johnson, R. W. (1981). Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, IT-27(4):472–482.
- [Simoncelli, 1999] Simoncelli, E. P. (1999). Modeling the joint statistics of images in the wavelet domain. In *Proceedings of the 44th Annual Meeting of the SPIE*, pages 188–195, Denver.
- [Steuer et al., 2002] Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18 (Suppl. 2)(90002):S231–S240.
- [Stögbauer et al., 2004] Stögbauer, H., Kraskov, A., Astakhov, S. A., and Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Physical Review E*, 70:066123.
- [Torgeson, 1952] Torgeson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419.
- [Vasicek, 1976] Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(1):54–59.
- [Verdú, 1996] Verdú, S. (1996). The exponential distribution in information theory. *Problems of Information Transmission*, 32(1):86–95. English translation: Original in *Problemy Peredachi Informatsii*, vol. 32, no.1, pp.100-111, Jan.-Mar. 1996 (In Russian).
- [Wainwright et al., 2001] Wainwright, M. J., Schwartz, O., and Simoncelli, E. P. (2001). Natural image statistics and divisive normalisation: Modeling nonlinearities and adaptation in cortical neurons. In Rao, R., Olshausen, B., and Lewicki, M., editors, *Statistical Theories of the Brain*, pages 203–222. MIT Press, Cambridge, MA.