Copula Based Independent Component Analysis

Kobi Abayomi*, Upmanu Lall†, Victor de la Pena‡

March 4, 2008

Abstract

We propose a parametric version of Independent Component Analysis (ICA) via Copulas - families of multivariate distributions that join univariate margins to multivariate distributions. Our procedure exploits the role for copula models in information theory and in measures of association, specifically: the use of copulae densities as parametric mutual information, and as measures of association on the rank statistics.

The copula approach offers a unified view of component analysis procedures, in particular, by parameterizing multivariate dependence. ICA then, via the copula, is a generalization of Principal Component Analysis (PCA) - where the copula model may be non-Gaussian.

Generally, the goal is to orthogonalize a measure of multivariate dispersion, yielding an orthogonal basis for a multivariate data set. The flexibility of the copula approach allows for parameterizations of non-gaussian, non-monotone dependence. Additionally, we note a possible use for the copula approach in generalized component extraction procedures (such as Canonical Correlation Analysis) and ultimately the broader class of Generalized Linear Models.

1 Introduction

Shrinkage methods - statistical dimension reductions - are important and popular alternatives to numerical models in a variety of settings: generalized (see Heckman, J. [1976]); and in particular applied settings such psychometrics and climatology (see Hotelling [1933], Clarke, R. T. [2001,2002,2003]);. The objective in these methods is to identify a lesser subset of coordinates, in a high dimensional space, that sufficiently describe the evolution of specific state variables. The theme is to supplant or replace numerical models, which can be computationally intractable, with statistical alternatives. Subsequently, the reduction of multivariate data into a vector (or scalar) of low dimension can be a precursor of an unavailable physical model by illuminating salient characteristics.

1.1 PCA and ICA...

Given multivariate data $\mathbf{x}_k$, the goal in Principal Component Analysis (PCA) is to find the linear transformation (i.e. rotation matrix), $\mathbf{y} = B\mathbf{x}$ that minimizes the off-diagonal variance of

*Department of Statistics, Duke University, kobi.abayomi@duke.edu
†Department of Environmental Engineering. ula2@columbia.edu.
‡Department of Statistics. vhd1@columbia.edu. All authors: Columbia University

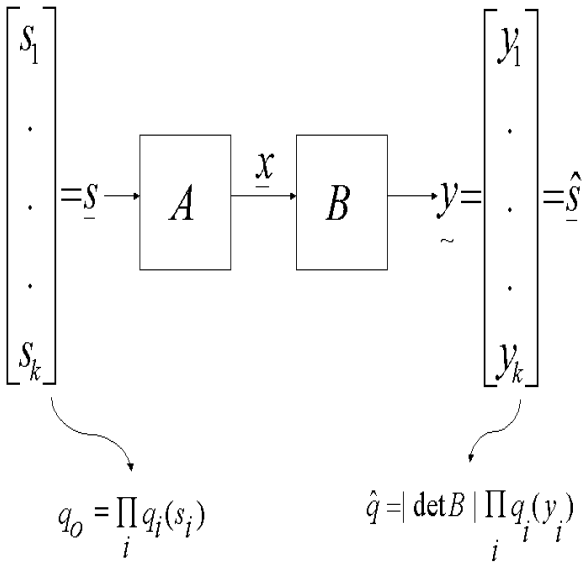

Figure 1 Diagram of Independent Component Analysis (ICA)/Blind Signal Separation (BSS) mixing and separating matrices. Typically, independent signals $\mathbf{s}$ are observed via unknown full rank rotation A as $\mathbf{x}$. The ICA/BSS procedure yields $\mathbf{y} = \hat{\mathbf{s}}$ outputs as estimates of the independent signals. The distribution of the inputs and outputs should be proportional.

$\mathbf{y}$. When $\Sigma = ((Cov(y_i, y_j)))_{i,j=1..k}$ is the covariance matrix of $\mathbf{x}_k$ the very well known result is to generate the eigenvectors for $\Sigma = \mathbf{e}^t \Lambda \mathbf{e}$ - Λ is a diagonal matrix of eigenvalues - which yields $y_i = \mathbf{e}^t \mathbf{x}$, with $Cov(y_i, y_j) = 0$, $i \neq j$, or the rotation which yields linear independence (see Johnson and Wichern [1998] for a comprehensive take).

Multivariate analysis via PCA is a venerable member of the statistical canon; PCA results are often intermediate steps in larger investigations where the component outputs may be inputs in standard predictor-response models or more generalized 'indices' of higher order measurements. Oja [1992], for example.

In Independent Component Analysis (ICA) the minimization of off-diagonal variation in $\mathbf{y}$ is strengthened to statistical independence, beyond the second order condition. Here, the goal is to find the linear transformation (i.e. rotation matrix) of $\mathbf{x}_k$, $\mathbf{y} = B\mathbf{x}$, such that the observed $y_i = b_i \mathbf{x}$ are *nonlinearly* correlated[1] of $y_j = b_j \mathbf{x}$; here the model for statistical independence is explicit. The observed data are modeled as mixed outputs $\mathbf{x} = A\mathbf{s}$, of independent sources $\mathbf{s}$. The columns of $\mathbf{Y}$ are the estimates of these independent components, or signals; which the rotation B is an estimate of A^{-1}. See Figure 1

Independent Component Analysis (ICA) can be cast as a generalization of the PCA program where more general versions of statistical independence succeed covariation and thus uncorrelatedness (Jutten and Herault [1991]). In both versions the objective is the recovery of the linear rotation A of the independent signals $\mathbf{x}$. The difference is the characterization of statistical independence or *contrast function*, and the implicit or explicit distributional assumptions on the inputs (See Cardoso [1998], Brunel et al. [2005]).

ICA extends independence beyond covariance. While zeroed covariation is sufficient for independence under the Gaussian assumption typically operant in PCA. When dependency is not appropriately captured by the second moment, covariance is an insufficient proxy for statistical

[1] In the maximal correlation sense: $\rho(y_i, y_j) = \sup_{f,g}(f(y_i), g(y_j)) = 0$ [see Hyvarinen 2001].

independence. For a simple example, take functional dependency $x_i = h(x_j) = {x_j}^2$, for example, $\mathbb{E}(x_i) = 0$. Here $Cov(x_i, x_j) = 0$ though x_i, x_j are completely statistically dependent. ICA can be seen as PCA under a more general contrast function, based on an alternate measure . In PCA we seek the linear rotation that minimizes covariance; in ICA we seek the rotation that minimizes, for example: entropy, mutual independence, higher order decorrelation, etc. (Cardoso [1998]).

In the simplest ICA models - including Blind Signal Separation (BSS) - the number of signals is equal to the number of sources: the rotation matrix is of full rank.

1.2 ...the Copula approach.

The copula families of multivariate distributions, those where a candidate joint distribution is evaluated *on* a set of univariate marginals, are functions from $\mathbb{I}^k$ to $\mathbb{I}$. As densities – full derivatives – copulas are the ratio of the joint density to the product of the univariate marginals (Nelsen [1999], Frees and Valdez [1998]). In this sense the copula representation captures the dependence within $\mathbf{x}$: the value of the copula is the proportion of dependence to full independence. This proportion is maximal when there is no gain to modelling a multivariate $\mathbf{x}$; each element x_i, separately, is sufficient. This property, in particular, recommends the copulae family as a fertile point of departure for dependence models.

We propose parametric copulae families as estimators for the dependency in $\mathbf{x}$ under broad dependence conditions. Specifically, we investigate the copula approach as a generalized 'engine' for the contrast functions - measures of statistical dependence - which characterize ICA analysis. We suggest a Copula based version of Independent Component Analysis (CICA) - where we model and rotate dependency information in $\mathbf{x}$ via copulae families.

Our version - CICA - replaces non-parametric, higher order proxies for independence with parametric examples from the copula literature. This parametric modelling appeals: 1) to the duality between information minimization within the component outputs and likelihood maximization for the rotated source model and 2) to the partitioning of the full likelihood of the outputs into model fit and dependence minimization.

We investigate ICA via copula based measures of association on partite reductions of $\mathbf{x}$ - in direct analogy to the PCA via covariance matrix we can view the ICA procedures as orthogonalizations of higher order tensors to capture non-elliptical dependence. The flexibility of partite reduction allows us to suggest appropriate copula families for non-gaussian dependence pathologies - specifically extreme value, non-monotone and inhomogenous data - within a multivariate set. Lastly, we illustrate a consistent framework for fully parameterized ICA (in a bayesian setting) perhaps.

2 Setup

2.1 Measures of dependence

Heuristically, a measure of dependence captures the strength of the joint distribution on the index (of an index set) of a multivariate vector (See Renyi [1959] for a full treatment). Well known versions of measures of dependence - *measures of association* - include: Pearson's correlation coefficient ρ, the maximal correlation, Kendall's tau ρ_τ, Spearman's rho ρ_S (see Mari, Kotz [2001]), and Linfoot's informational correlation (see Linfoot [1957] or Joe [1987]). We include mutual information (see Kullback [1959]) with these measures.

Typically these are scalar measures defined on a pair of random variables. Any multivariate generalization is often, though not necessarily, a function from $\mathbb{R}^k$ to $\mathbb{R}$ (usually $[0, 1]$ or $[-1, 1]$) where weights are apportioned across the index set of the multivariate vector. See Wolff [1980], Simon [1977], Joe [1990] and section A.1 in the appendix.

In the special case that $\mathbf{X} \sim Ell_k(\mu, \Sigma)$ - setting $\mu = 0$ without loss of generality - Σ is a *scatter* matrix which is sufficient for the dependency in $\mathbf{X}$. In these *elliptical* distributions the dependence in the conditional distributions $F_{X_i,X_j|\mathbf{X}_{-i,-j}}$ is independent of the value of $\mathbf{X}_{-i,-j}$: second order parameters fully characterize dependency across indices of the multivariate vector; second order moments, $\Sigma = ((\sigma_{ij}))$, are typically sufficient.

This condition, sufficient for Gaussian or t-distributed $\mathbf{x}$ for example, is a special case of r-independence, here $r = 2$, (see Ibragimov [2005], Joe [1997]). See section A.2 in the appendix.

If $\Theta = \hat{\Sigma}$ is an estimator for Σ under 2-independence; $\Theta = ((\theta_{ij}))$ completely specifies the scatter matrix. In the Gaussian example each θ_{ij} is the sample covariance, or a version of it. It is enough to parameterize Θ pairwise for ellipsoidal distributions; more generally Θ may be indexed over any subset $\mathbf{X}$. In this way Θ can be seen to collect dependencies in index sets of a multivariate vector. When the index set is only pairwise, the dependency is elliptical and θ_{ij} is an axis of the multivariate ellipsoid.

One approach is to parameterize and estimate partite models, using measures of dependence where the parameter θ is estimable on indices of the multivariate vector. We note that symmetry for a bivariate measure of dependence - exchangeability in the multivariate extension - is a Rényi postulate that a dependency model may not necessarily satisfy. We expand on this idea in the context of full and partite models below.

2.2 Copulas

One measure of dependency, or 'engine' for measure of dependency, is the copula functionalization (Nelsen 1999, 2006), defined via its density, on a multivariate $\mathbf{x} = (x_1, ..., x_k)$ as

$$dC(\mathbf{x}) = \frac{dF_{\mathbf{x}}(\mathbf{x})}{\prod dF_{x_i}(x_i)} \tag{1}$$

is the *multivariate copula density* for $\mathbf{x}$

here $dC(\mathbf{x})$; is the full derivative of a distribution function which takes the marginal distributions $F_{x_1}, ..., F_{x_k}$ as its arguments.[2] The copula distribution, then, is a distribution function on the space of the marginals to the unit hypercube, $(F_{X_1}, ...F_{X_k}) \mapsto \mathbb{I}^k$.

Many measures of association are copula dependent in that they can be expressed via the copula. (See Nelsen [1999, 2006], Wolff [1980], Schweizer [1981]). Kendall's tau and Spearman's rho are copula based, for example, as versions of an expected value of the copula or copula density (see section A.3 in the appendix). Any measure of association that can be expressed on a multivariate family with fixed margins is expressable via the copula. Pearson's correlation, to contrast, is an example that cannot be expressed via the copula (see section A.3 in the appendix).

2.3 Mutual information via Copula

Recall that the mutual information (see Kullback [1968]), for a multivariate $\mathbf{X}$ with distribution function $F(\mathbf{X})$is

$$MI(\mathbf{x}) \equiv \int_{\Omega} dF(\mathbf{x}) log(\frac{dF_{\mathbf{X}}}{\prod dF_{X_i}}) \tag{2}$$

where Ω is the probability space for $\mathbf{X}$. Using (1) above, this can be re-expressed as

$$MI(\mathbf{x}) \equiv \int_{\mathbb{I}^k} dC(\mathbf{u}) log(dC(\mathbf{u})) \tag{3}$$

[2]We prefer this notation in lieu of the subscript for the full derivative. We will use C for the copula distribution function.

(see Davy and Doucet [2003]). If $T \sim F, dF = f$ then $-H(T) = \mathbb{E}(f(T)log(T))$ is called the *entropy* for t (see Ash [1965]) and we write

$$MI(\mathbf{X}) = -H(\mathbf{u}) = \int_{\mathbb{I}^k} dC(\mathbf{u})log(dC(\mathbf{u})) \tag{4}$$

The mutual information then - as the expected value of the log of the copula density, can be computed, or estimated, from a parametric copula

Our approach is to: (1) use the mutual information as the measure of dependence; (2) access or estimate the mutual information via a parametric copula model. Taking the elliptical case - $\mathbf{X} \sim Ell_k(., \Sigma)$ - case for illustration, suppose

$$\mathbf{C}_\Theta = ((C_{\theta_{ij}})) \tag{5}$$

is a consistent model for the multivariate dependence in elliptically distributed $\mathbf{X}$. Then, each $C_{\theta ij}$ is an estimator for $\frac{dF_{x_i,x_j}}{dF_{x_i}dF_{x_j}}$. Here, a matrix of bivariate copulae are estimators of functional dependency. This is a natural approach given the role of the copula. Then the measures of association for each pair i, j are functions of each $C_{\theta ij}$ and the scatter matrix Σ - in this case approximated via a so-called *mutual information matrix* $MI(x_i, x_j)$

$$MI_\Theta(X_i, X_j) = ((H[C_{\theta ij}]) = ((MI_{\theta ij})) \tag{6}$$

can be characterized via these bivariate copulae.

In the PCA/ICA literature, *contrast functions* are objective functions for source separation: let $\psi(\mathbf{y}) = 0$ imply y_i and y_j are independent $\forall i \neq j$ where ψ is a particular contrast function. The minimization of this function is the PCA/ICA algorithm.

Our approach is to employ the copula, exploiting its natural role within measures of association and as a model for dependence/independence, as the engine for ICA contrast functions.

3 Copula Based Independent Component Analysis

There are two common approaches to ICA: direct minimization of a 'redundancy' measure on the outputs or information maximization between the outputs and hypothesized inputs (see Obradovic and Deco [1998]). Lee. et al. as well as Cardoso (Lee, T [2000], Cardoso [1998]) illustrate the similarity between these methods. Both of these approaches are tantamount to minimization of the mutual information in the rotated vector $\mathbf{x}$, within a full model which includes the input distribution from dependence to independence. We cast our version, which we see as a unification, in context of the copula.

We restrict our take to the simplest ICA mode: *blind source separation*, where the assumption is that

$$\mathbf{x} = A\mathbf{s} \tag{7}$$

is observed as a linear rotation B (square and full rank), of $\mathbf{s}$ independent signals. The goal is to find

$$\mathbf{y} = B\mathbf{x} \tag{8}$$

where $\mathbf{y} = \hat{\mathbf{s}}$ is an estimate of the source signals. The statistical model here has two components: the recovery of A and the estimation of the distribution of $\mathbf{s}$. As A is assumed full rank and thus invertible, the source separation problem is recast as a search for $B = \hat{A^{-1}}$. Typically, we are interested in the joint distribution of the source

$$q(\mathbf{s}) = \prod_{i=1}^{k} q_i(s_i) \tag{9}$$

represented as independent under the model. The goal of ICA then is to minimize a contrast function that recovers the independence expressed in (9). Recasting the contrast functions in context of the copula, we can view ICA as a copula based procedure.

3.0.1 via Maximum Likelihood

The Kullback-Liebler divergence between two probability density functions $f(\mathbf{t})$ and $g(\mathbf{t})$ we notate

$$\mathbb{K}(f, g) = \int_{\mathbf{t}} f(\mathbf{t}) log(\frac{f(\mathbf{t})}{g(\mathbf{t})}) \tag{10}$$

We can use the notation $\mathbb{K}(w, z)$ for the divergence between the distribution of two random vectors w and z. $\mathbb{K} \geq 0$ with equality if and only if w and z have the same distribution; $\mathbb{K}$ is not symmetric.

Using the model in (7), with q the true distribution for $\mathbf{s}$ - for a given model $(B, q = q_0)$ for the mixing matrix and distribution of the signals, the density of the observed data $\mathbf{x}$ is

$$\hat{q}(\mathbf{x}; B, q_0) = |detA|^{-1} q_0(A^{-1}\mathbf{x}) \tag{11}$$

under the parametric model $\hat{q}$ with parameters (A, q_0), since $\mathbf{s} = A^{-1}\mathbf{x}$.

For N independent samples of $\mathbf{x}$, $\underline{\mathbf{x}} = (\mathbf{x}_1, ..., \mathbf{x}_N)$, the log-likelihood is

$$log(\hat{q}(\underline{\mathbf{x}}, A, q_0)) = \sum_{i=1}^{N} log(q_0(A^{-1}\mathbf{x}_i))) - log(|detA|). \tag{12}$$

Put

$$L(\hat{q}) = L(B, q_0) = N^{-1} \sum_{i=1}^{N} log(q_0(A^{-1}\mathbf{x}_i)) - log(det|A|) \tag{13}$$

this converges to its expectation as $N \to \infty$,

$$L(\hat{q}) = \int q(\mathbf{x}) log(\hat{q}(\mathbf{x}, A, q_0)) d\mathbf{x} \tag{14}$$

again with q the (true) distribution of the observed data.

Rewrite (14) as

$$L(\hat{q}) = -\int q(\mathbf{x}) log(q(\mathbf{x})) - \int q(\mathbf{x}) log(\frac{q(\mathbf{x})}{\hat{q}(\mathbf{x}, A, q_0)}). \tag{15}$$

The right hand side is just $\mathbb{H}(\mathbf{x}) - \mathbb{K}(q(\mathbf{x}), \hat{q}(\mathbf{x}))$. Since $\mathbb{H}(\mathbf{x})$ is independent of A, maximization of the log likelihood is equivalent to minimizing the divergence between the observed density and our parametric model.

Rewrite the log-likelihood of the model for $\mathbf{x}$,

$$L(A, q_0) = H(\mathbf{x}) - \mathbb{K}(q(\mathbf{x}), \hat{q}(\mathbf{x}, A, q_0)) \tag{16}$$

maximizing the likelihood of $B = \hat{A^{-1}}$ is equivalent using

$$\frac{\partial L}{\partial B} = -\frac{\partial}{\partial B}\mathbb{K}(q(\mathbf{x}), \hat{q}(\mathbf{x}, A, q_0)) \tag{17}$$

Again, since A is invertible, the KL divergence is invariant and minimization of (17) is equivalent to (21).

Following maximum likelihood is equivalent to finding the matrix B such that $\mathbf{y} = B\mathbf{x}$ is as close as possible to the sources, in $\mathbb{K}$ 'distance'.

3.0.2 via Infomax

Bell and Sejnowski [1995] show that maximizing information transfer over a nonlinear transform $G(\mathbf{y})$ and rotation matrix A is equivalent to maximizing the joint entropy of the non-linear outputs. Take

$$G_i(s) = \int_{-\infty}^{s} q_i(t)dt \tag{18}$$

as a distribution function for the source s_i. $G_i(s_i)$ is a non-linear transform.

As well, note

$$\mathbb{I}(\mathbf{x}) = \int_{\mathbf{X}} p(\mathbf{x}) log(\frac{p(\mathbf{x})}{\prod_{i=1}^{k} p_i(x_i)})d\mathbf{x} \tag{19}$$

as the multivariate mutual information for a multivariate $\mathbf{x}$.

The joint entropy of the outputs via their - nonlinear - transforms G is

$$\mathbb{H}(G(\mathbf{y})) = \sum_{i=1}^{k} \mathbb{H}(G(y_i)) - \mathbb{I}(G(\mathbf{y})) \tag{20}$$

The joint entropy is maximized when $I(G(\mathbf{y})) = 0$ and $G(y_i)$ is uniform. Recovery of A is via the minimization of (20)

$$\frac{\partial \mathbb{H}(G(\mathbf{y}))}{\partial A} = \frac{\partial}{\partial A}(-\mathbb{K}(G^*(\mathbf{y}), G(\mathbf{y}))) \tag{21}$$

where $\mathbb{K}(.,.)$ is the Kullback-Liebler divergence from above and G^* is the multivariate - maximal entropy - uniform distribution. For invertible, 'well-picked', transforms $G(\mathbf{y})$, (21) is equivalent to the divergence between the estimate $\hat{p}$ and the observed distribution p in (17), above.

3.0.3 K-L distance: Marginal fit + Independence

We treat the distribution of the sources, $q = q_0$, as fixed in the above approaches, though the full parameter space is (B, q_0). In (17) and (21) the maximization is with respect to B.

Optimization over B alone is problematic if q_0 is far from the true distribution; in these cases we should minimize the divergence $\mathbb{K}(p, \hat{p}(B, q))$ over q as well.

Following the outline in Cardoso [1998], denote $\mathbf{y}^*$ a random, mutually independent vector with margins distributed equivalently with $\mathbf{y}$. It is well known (See Kullback [1959]) that

$$\mathbb{K}(\mathbf{y}, \mathbf{s}) = \mathbb{K}(\mathbf{y}, \mathbf{y}^*) + \mathbb{K}(\mathbf{y}^*, \mathbf{s}) \tag{22}$$

for $\mathbf{s}$ independent. The K-L distance between the outputs and the sources can be partitioned as: (1) the K-L distance between a distribution and the closest independent distribution,

$\mathbb{K}(\mathbf{y}, \mathbf{y}^*)$, and (2) the K-L distance between the true and hypothesized distributions for the sources.

The full program is achieved by taking $\mathbf{s} = \mathbf{y}^*$, which minimizes the marginal mismatch, thus the right most term in (22). The program is completed by minimizing the mutual information in the chosen distribution, the middle term in (22). The first is minimization with respect to the choice of q_0; the second is minimization for the mixing matrix B.

A common preliminary to many ICA programs is the removal of Gaussianity by pre-'whitening' the observed data, often via PCA. See the Appendix.

3.1 ...via the copula model

To recap: given a full model $(\hat{q}; B = \hat{A}^{-1}, q_0)$ - $\hat{p}$ the distribution of the data $\mathbf{x}$, A the estimate of the inverse of the mixing matrix, q_0 the hypothesized distribution of the sources...

1. Match distributions between input ($\mathbf{s}$) and output ($\mathbf{y}$): 'Move' everything - $(\hat{p}; A = \hat{B}^{-1}, q_0)$ - until the distributions between the ICA output and the hypothesized inputs match. Simply, this is a search for the model with minimum Kullback divergence between inputs and outputs. *This is the maximum likelihood approach*: maximizing the likelihood of the outputs is equivalent to minimizing the negative entropy of the outputs. When q_0 is well fixed this is equivalent to minimizing the mutual information in the outputs...
2. ...Match structure within output: 'Move' just $(\hat{p}; A = \hat{B}^{-1})$ - until the mutual information within the outputs is minimized. This is the argument in section 2.4.3. This, *the mutual information approach*, exploits the equivalence of statistical independence and zeroed mutual information. The below can be seen as an approximation...
3. ...Orthogonalize a dependency (i.e. 'redundancy') measure and fit marginal distributions separately: This procedure is applied by explicitly using PCA to whiten the data, and then an $r-$ variate 'redundancy' measure to rotate the whitened data. Oja [1997] calls this the *the non linear PCA approach* - which relies on the partitioning of model fit and independence outlined in the argument in section 3.0.3. This version has been shown (see Lee et al. [2000]) to be an approximate of the generalized information theoretic procedure.

In practice $q = q_0$, the distributions of the sources is fixed and/or treated as a nuisance parameter. The ICA program is then just a search for A - the best estimate of the mixing matrix.

The standard approach is to estimate the above objectives using higher order, non-parametric statistics, often cumulants (see Comon [1994]). We discuss this in section 3.2.

The multivariate copula density is the ratio of the full density to the full density under independence - the product of the marginal densities. The development here is to replace the non parametric estimating equations with parametric versions, via the copula - exploiting the natural role of the copula in mutual information.

The program, in the order of the above enumeration, is to estimate the mutual information of the outputs using a copula model for: (1) where the marginal distributions q and mixing matrix A are parameterized; (2) where the marginal distributions are fixed and only the mixing matrix A is parameterized; (3) where the marginal distributions are accessed only via a mutual information (or other measure of association) array - A is parameterized.

3.1.1 MLE via copula, Infomax via copula

We can view the MLE, and thus infomax, arguments as an extension of (24). Here we access the maximum likelihood principle via the parametric copula estimate of the mutual information. The likelihood in (17) and (21) is, in terms of the copula,

$$\frac{\partial L}{\partial B} = \frac{\partial}{\partial B} N^{-1} \sum_{i=1}^{N} C_\theta(\underline{u}_i) \tag{23}$$

where $\mathbf{u}$ are fixed as $u_i = G(y_i)$, the univariate distributions of $\mathbf{y}$. Then the last term in (20) is

$$\mathbb{I}(\mathbf{u}) = \mathbb{I}(G(\mathbf{y})) = \mathbb{I}(G(A\mathbf{x})) = \mathbb{E}(log(dC(G_1, ..., G_k))) = \mathbb{E}(log(dC(\mathbf{u}))) \tag{24}$$

Again, the first term in the derivative of (20) is invariant to G.[3] So, (21) is equivalent to finding B such that $\mathbb{I}(G(B\mathbf{x})) = \sum_{i=1}^{k} \mathbb{H}(G(y_i))$, that is

$$\mathbb{E}_B(log(dC(G_1, ...G_k))) = \mathbb{E}(log(dC(\mathbf{u}))) = \sum_{i}^{k} \mathbb{E}_B(log(G_i)) = \mathbb{E}_B(log(u_i)) \tag{25}$$

or the rotation that yields the independence copula for a parametric copula family C_Θ.

3.1.2 Mutual information via copula

To cast the partitioning in (22) in terms of the copula we write the middle term as

$$\min_B \mathbb{I}(\mathbf{y}; B) = \min_B \mathbb{E}_B(log(dC_\Theta(\mathbf{u}))) \tag{26}$$

and the final term in (22)

$$\min_\Theta [C_\Theta(\mathbf{u}) - \prod_{i=1}^{k}(u_i)]. \tag{27}$$

That is we minimize the mutual information via the copula via rotation A after minimizing the distance between parametric copula and independent marginals.

3.1.3 K-L distance - via Copula: Independence + Marginal fit

Set $\mathbf{u}^* = G(\mathbf{y}^*)$ where $\mathbf{y}^*$ is still a random, mutually independent vector with margins distributed equivalently with $\mathbf{y}$. Thus, $\mathbf{u}^*$ is independent with margins distributed as $\mathbf{y}$. In analogy with section 3.0.3

$$\mathbb{K}(\hat{\mathbf{u}}, \mathbf{u}) = \mathbb{K}(\hat{\mathbf{u}}, \mathbf{u}^*) + \mathbb{K}(\mathbf{u}^*, \hat{\mathbf{u}}) \tag{28}$$

with $\hat{\mathbf{u}}$ the estimate of the true sources output from a copula based procedure and $\mathbf{u}$ the true distribution of the sources. The K-L distance between the outputs and the sources is then: (1) the fit of the outputs to independence $\mathbb{K}(\hat{\mathbf{u}}, \mathbf{u}^*)$; and (2) the fit of the marginals of the outputs to the true source distributions.

The full program is achieved by taking $\mathbf{u} = \mathbf{u}^*$, which minimizes the marginal mismatch, thus the right most term in (28). The program is completed by minimizing the mutual information in the chosen distribution, the middle term in (28). The first is minimization with respect to hypothesized source distribution; the second is minimization for the mixing matrix.

[3] G is treated as a nuisance parameter here. In general a poorly 'picked' G will invalidate infomax.

3.2 Discussion of approach

The copula perspective allows parameterization of dependence, in particular mutual information, via the copula model. The copula entropy $H(\mathbf{u})$ is the mutual information of the outputs - transformed by their univariate cdf's, $u_i = F(y_i)$. The maximization of this entropy is equivalent to the minimization of the mutual information between the inputs. Further, the entropy is maximal when the first term of (20) - or the fit of the marginals to the true source distribution - is large and the second - or the mutual information across the margins - is small. These results are fertile point of departure for a fully parameterized component model.

In such a general model: the source distributions $u_i = \int_0^{s_i} \hat{q}(t)dt$ - are the arguments of the copula $C_\Theta = (\mathbf{u})$; the rotation matrix $B = \hat{A}^{-1}$ is a function of Θ. A full likelihood, then, would cover the space of source distributions, copula parameters, and subsequently all linear rotations - *including* those on the possibly non-linear pre-image space for the source distributions.

There are drawbacks to this approach:

- A model which includes the source distributions as parameters may be over-parameterized, especially if the rotation matrix is not of full rank.
- There are no natural multivariate families - i.e. families with nice properties that accept ranges of dependence. Unusual and undesirable dependencies in subsets - conditional distributions - of the marginals (see Joe [1987, 2001]). Parametrically heterogenous models may be excluded.
- The parameter space can possibly increase exponentially with k - on the same order as the power set.

Versions of this approach where the joint dependence is captured in a single parameter (multivariate or scalar) may not be applicable for the ICA problem. Consider the minimization in (26) and (27): if $\Theta^\perp$ is the copula parameter at independence then $\lim_{\Theta \to \Theta^\perp} C_\Theta(\mathbf{u}) = \prod_i^k u_i$ and the rotation matrix at $\Theta^\perp$ is unidentifiable. A possible application is in Multivariate Coordinate Analysis (MCA) [Oja 1997] or Canonical Correlation Analysis (CCA) [Hardoon, Szedmak, Shawe-Taylor 2004] where the outputs need not be independent.[4] We note as an aside that a fully Bayesian approach would be natural for a fully parameterized - source and mixing matrix - model.

Lastly, we note, as important aside: 1) that common algorithms for ICA employ a version of PCA whitening as a pre-processing step [although implicitly; Hyvarinen 1999]; 2) higher order cross moments are necessarily computed at low dimension [Cardoso and Comon 1996; Cardoso and Souloumiac 1993]; 3) a disadvantage of reliance upon statistical moments is highlighted in an example by Mccullagh [1994] [see also Lindsay 2000] where distributions perturbed (via sine wave) are indistinguishable in moments.

4 Simulated Applications

In the examples (figures) below we apply the CICA procedure to various settings in a demonstration of the flexibility of the method, and comparison to extant ICA algorithms. In the classical ICA program mixing matrix A is full rank: in the simulations below we investigate the performance of CICA in full and deficient rank examples.

1. Full rank ($k = 3$) Gumbel-Hougard type copula to rotated Generalized Extreme Value (GEV) type distribution (see Stephenson [2003], Appendix). This example illustrates the

[4] A full copula model would determine the gradient descent procedure yielding - possibly - interesting components everywhere but independence.

utility of the CICA procedure in what we believe will be the most common CICA setting: extreme value/non-gaussian data sets.

2. Full rank ($k = 3$) two-parameter ($BB6$) copula on independent sources — $S_1 \sim U(-1,1)^2,\ S_2 \sim Gumbel(0,1),\ S_3 \sim \chi^2$ — with known mixing. This example is a test of the ability of the CICA procedure to recover exactly the independent sources.
3. Full rank ($k = 2$) FGM type copula on known, mixed Gaussian and Laplacian sources — standard demonstration examples for fastICA procedure. This example compares the CICA procedure with a common ICA (fastICA) algorithm.
4. Two-parameter archimedean ($BB6$) copula[5] on on deficient rank ($m = 4, k = 5$) mixed independent sources — $S_1 \sim U(-1,1)^2,\ S_2 \sim Gumbel(0,1),\ S_3 \sim \chi^2$ — with known mixing; two non-independent sources are included $S_4 \sim Z+N_1,\ S_5 \sim Z+N_2$. $Z \sim Exp(1)$ and $N_1 \sim N(0,.5),\ \perp\ N_2 \sim N(0,1)$.

In Abayomi [2008b] we apply a partite version of the CICA procedure on the Environmental Sustainability Index (ESI) — a high dimension data set — for a fifth example. In this version of the CICA procedure, we approximate a full model by using bivariate copula to estimate a scatter-mutual information matrix: $\hat{\Sigma} = ((MI_{X_i,X_j}))$. In outline:

Set $\mathbf{y} = \mathbf{RWx}$, with $\mathbf{W}$ a 'whitening' matrix - the product of a PCA - and $\mathbf{R}$ the CICA rotation. This allows diagonalization of the final mutual information matrix via well known procedures like Singular Value Decomposition (SVD) - see Appendix.

- Treating the univariate distributions $u_i = F_{X_i}(x_i)$ as observed.
- Bivariate Mutual Information(s), or, $E(log(dC(u_i, v_i)))$ are the elements of our scatter matrix
- Construct scatter/kernel matrix
- Choose copula families at each bivariate pair: $C(\mathbf{u}) = \eta_{\theta_1,\theta_2}(\eta^{-1}_{\theta_1,\theta_2}(u) + \eta^{-1}_{\theta_1,\theta_2}(v))$. $\Gamma_{C_\Theta} = ((C_{\theta_{ij}}(F_{w_i x_i}(w_i x_i), F_{w_j x_j}(w_j x_j))))_{i,j=1..k}$
- Find orthogonalization of Γ_{C_Θ}; via SVD $\lambda_1, ..., \lambda_k$ are the eigenvalues and $\mathbf{e}_1, ..., \mathbf{e}_k$ are the eigenvectors.
- Yield $y_k = b_k \mathbf{x}_k = r_k w_k x_k$ with $y_i \perp y_j$ via C_Θ

For $i = j$, the mutual information is just the entropy: we compute the entropy using the empirical distribution, the copula fitting is unnecessary.

4.1 CICA on Multivariate Extreme Value (GEV) distribution

We believe the most likely application for the CICA procedure will be on multivariate non-gaussian/extreme value data. We simulated non-independent multivariate Generalized Extreme Value (GEV) data using the algorithms outlined in Stephenson [2003] for logistic type multivariate extreme value distributions.

$$G(x_1, x_2, x_3) = \exp\{-({x_1}^{\frac{1}{\alpha}} + {x_2}^{\frac{1}{\alpha}} + {x_3}^{\frac{1}{\alpha}})^\alpha\}$$

where $\alpha \in [0, 1]$ is a dependency parameter and $x_i = [1 + \frac{\xi_i(t_i - \mu_i)}{\sigma_i}]^{-\xi_i}$ — with (μ_i, σ_i, ξ_i) the location, scale, and shape parameters of the ith univariate GEV distribution.

We used a trivariate Gumbel-Hougard type copula[6] as the 'dependency gradient'; the parameter $B = \hat{A}^{-1}$ was estimated using $\mathbf{R}$'s non-linear constrained optimizer [Lange 2001]. This

[5] See Abayomi [2008], Appendix.

[6] From a 'frequentist' perspective, the *a priori* choice of the copula function, i.e. 'dependency gradient', to optimize/use as score function is arbitrary.

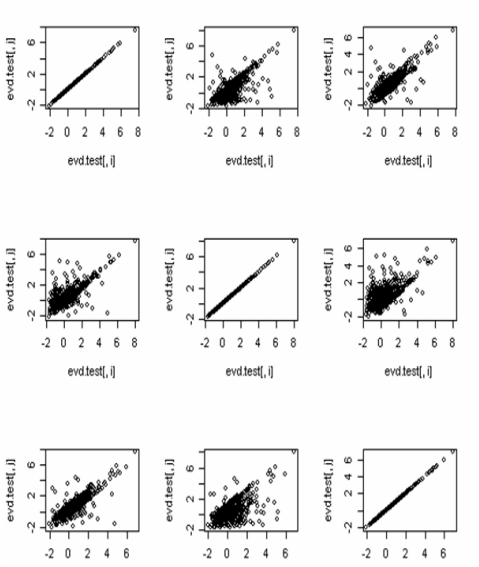

Figure 2 Scatterplot matrix of simulated extreme value distribution, $k = 3$, $n = 100$ [Stephenson 2002]. The joint distribution is minimally dependent and the marginals are characteristically non-Gaussian.

is, directly, the optimization of the score/estimating equations for $B = \hat{A}^{-1}$, represented in the analogous equations (17), (21), (24) (and the first term on the RHS of (28)) on the empirical cdf's of the univariate margins. See Figures 2, 3, 4.

The outputs of the CICA procedure are approximately independent: the parameter value for the trivariate GH copula is near independence for each family.

For a last 'test' of independence, Gaussian, and t type copulas fit to the outputs of the CICA procedure are yielded parameter estimates at independence, as well.

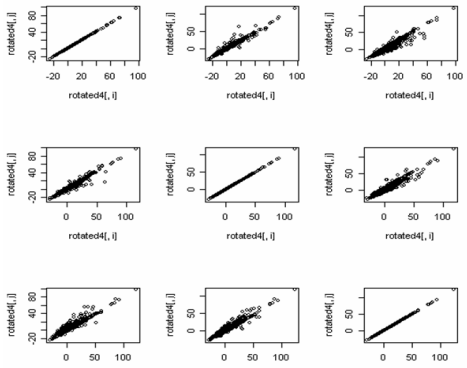

Figure 3 Scatterplot matrix of full rank 'rotation' of simulated multivariate extreme values, $k = 3$, $n = 100$.

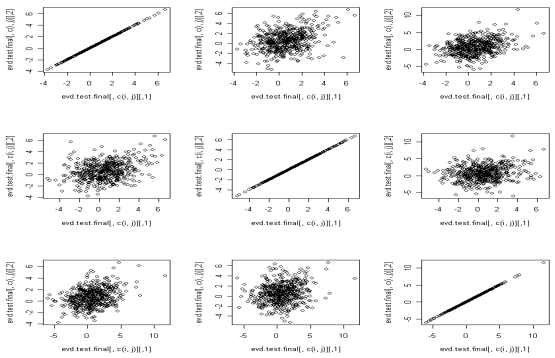

Figure 4 Scatterplot matrix of CICA outputs or 'source estimates', $k = 3$, $n = 100$: $y_i = B_i x_i$; using Gumbel-Hougard type copula as 'dependency gradient'. The bivariate scatter plots appear to have independent distributions within each plot; additionally, Gaussian, t, and Gumbel-Hougard [Nelsen 1999] copula fit to outputs have parameter estimates at or near independence.

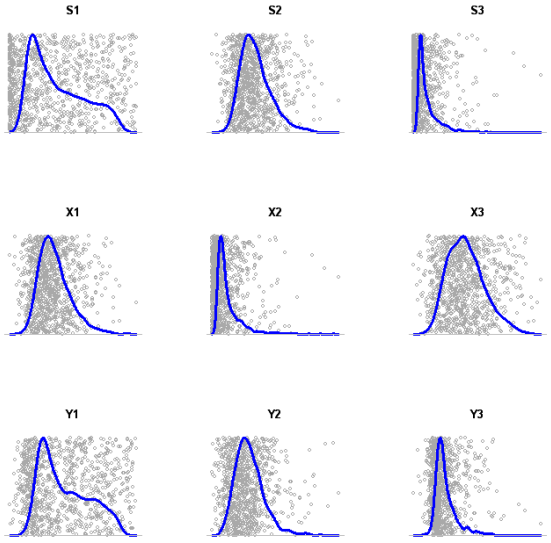

Figure 5 CICA model applied via Gumbel-Hougard type 'dependency gradient'. The first row are the source distributions, all non-normally distributed: $S_1 \sim (U(-1,1))^2$, $S_2 \sim Gumbel(0,1)$, $S_3 \chi^2$. The second row are the 'data' observed after a full rank rotation. The third row are the outputs - estimated sources. The data are plotted in dark gray; estimated density is superimposed in blue.

4.2 CICA on known, mixed independent sources

For a second example we applied the CICA procedure to independent data from known source distributions after pre-multiplication by a full-rank mixing matrix. See Figure (5). CICA recovered the independent sources: density and scatter plots of the outputs match the known sources.

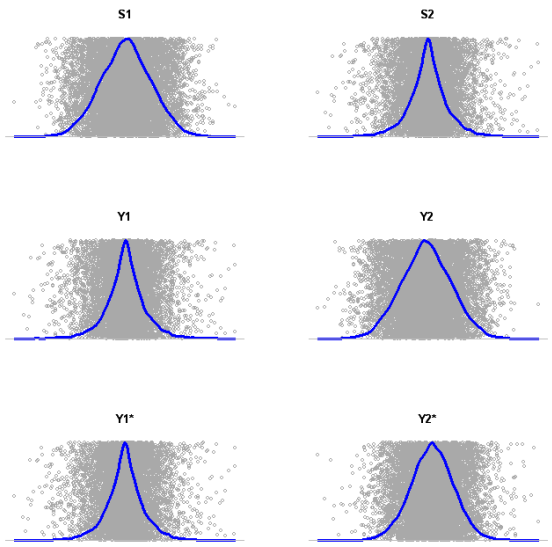

Figure 6 fastICA [Hyvarinen 1999] and CICA model applied to mixed Gaussian and Laplacian sources - $S1$ and $S2$ in the first row of the graph, $n = 10000$. The first row are the source distributions. The second row are the fastICA estimates. The third row are the CICA estimates. The data are plotted in dark gray; estimated density is superimposed in blue. Both algorithms appear to recover the sources - though the order of the outputs is swapped. In general, ICA recovery is permutation invariant.

4.3 fastICA vs. CICA, sample size comparison

Here, we applied the FGM version of CICA on an example used in the well known fast-ICA [Hyvarinen 1999] procedure — a common ICA algorithm.[7] See figure 6.

The performance of the two algorithms is similar — with respect to the closeness of the estimates to the true source distributions — when the entire data (10,000 observations) are used. As we noted above, we expect the (semi) parametric CICA procedure to outperform non-parametric fast-ICA type procedures on smaller sample sizes (Mccullagh [1994] [see also Lindsay 2000], mentioned above). Figure (7) plots the mean $\mathcal{L}^2$ distance between the true and estimated and densities: $n^{-1}\sum_{n=1}^{N}(q(y_n) - \hat{q}_n(y_n))^2$, for increasing values of n. The CICA algorithm appears to perform slightly better on this metric at lesser sample sizes.

Since the CICA procedure minimizes

$$\mathbb{K}(\hat{u}, u^*) = -H(\mathbf{u})$$

via equation (24) write $\mathbb{K}(\hat{u}, u^*) = \mathbb{K}_\theta$; since H is complete for choice of θ [Kullback 1968]. Then the mean integrated squared error

$$E_{\mathbf{u}}[E_\theta(\mathbb{K}_{\hat{\theta}} - \mathbb{K}_\theta)^2] \sim E_\theta(\hat{\theta} - \theta)^2 = O(\frac{1}{n})$$

of the copula based estimator is superior to

[7] An alternative procedure - Mutual Information Least Independent Component Analysis MILCA [Stogbauer, et al. 2004] - employs bivariate estimation of mutual information, similar to the partite version of CICA applied in Abayomi [2008b]; mutual information in MILCA is estimated semi-parametrically (via binning).

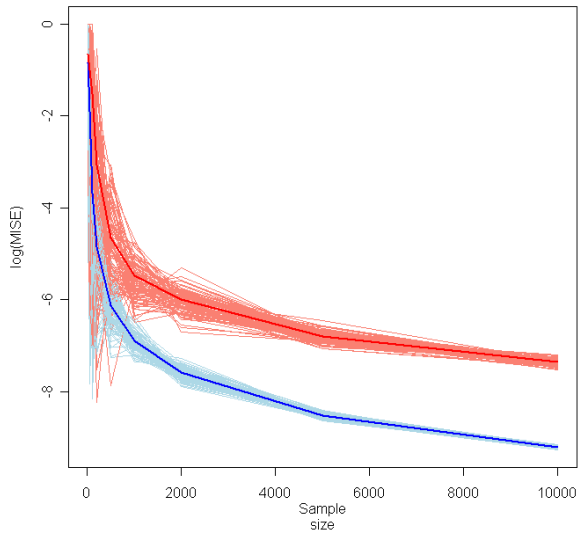

Figure 7 Log Mean Integrated Squared Error (MISE) of fastICA [Hyvarinen 1999] and CICA model applied to mixed Gaussian and Laplacian sources (via Gumbel-Hougard copula) - $S1$ and $S2$ in Figure (6). The MISE is $n^{-1}\sum_{n=1}^{N}(q(y_n)-\hat{q}_n(y_n))^2$ as N ranges from 10 to 10000. MISE for CICA is in blue, fastICA is in red. The $y-$axis is plotted on log scale to highlight the difference: the distance between the two curves is on the order $O(n^{-1/5})$. The CICA procedure has a marginally better error rate, and less variability over (100) random draws at each sample size. The mean MISE curves are plotted in darker color.

$$E_{\mathbf{x}}[E_q(\hat{q}(\mathbf{x})-q(\mathbf{x}))^2]=O(\frac{1}{n^{4/5}})$$

the kernel (averaging) based estimator used typically; $\hat{q}$ is the kernel estimator of the source distribution. In the CICA approach, conditioning on the univariate marginals allows us to consider the KL distance $\mathbb{K}$ parametrically.

In general, parametric models have a faster (in sample size) convergence rate than non-parametric models[8] [van der Vaart 1998]. The CICA procedure is semi-parametric, since the marginal parameters are true nuisance parameters. The mutual information (via the copula) is invariant to univariate changes in location, scale — we expect the CICA algorithm to have close to a parametric convergence rate.

Figure (7) is comparison of $\mathcal{L}^2$ error rate for the CICA and fastICA algorithm on the mixed Gaussian and Laplacian source example in Figure (6) for random draws of samples of size ranging from 10 to 10000. On average, the CICA algorithm seems to outperform the ICA algorithm though the variation in MISE is slightly higher for CICA.

4.4 CICA on deficient rank, known, mixed independent sources

In this last example we compare the CICA procedure with fastICA on deficient rank ($m=4$) mixing ($k=5$) of known independent sources. The independent sources are as in section 3.4.2: $S_1 \sim U(-1,1)^2$, $S_2 \sim Gumbel(0,1)$, $S_3 \sim \chi^2$. Two non-independent sources are included

[8] $O(n^{1/2})$ vs. $O(n^{-4/5})$ for the mean $\mathcal{L}^2$ distance

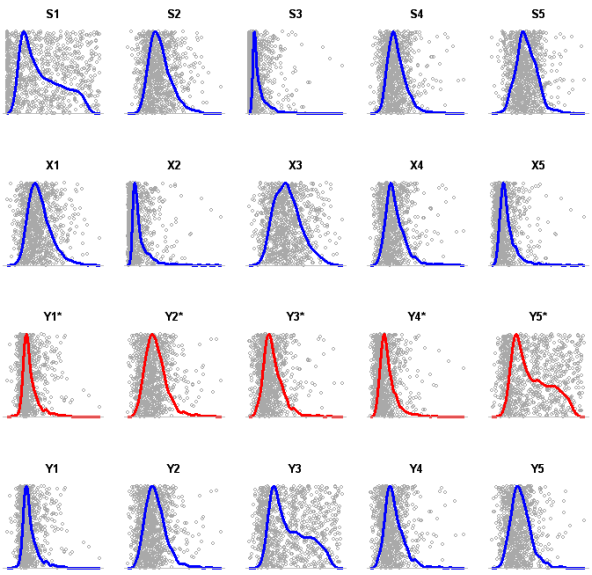

Figure 8 CICA model applied to deficient rank, $m = 4$, sources. $S_1 \sim U(-1,1)^2$, $S_2 \sim Gumbel(0,1)$, $S_3 \sim \chi^2$. Two non-independent sources are included $S_4 \sim Z + N_1$, $S_5 \sim Z + N_2$. $Z \sim Exp(1)$ and $N_1 \sim N(0,.5)$, $\perp$ $N_2 \sim N(0,1)$ are independent Gaussian noise, $n = 1000$. The first row are the source distributions. The second row are the mixed inputs. The third row are the CICA estimates, the outputs. The data are plotted in dark gray; estimated density is superimposed in blue. The presence of non-independent sources S_4 and S_5 appears to affect the estimation of the independent S_1, S_2, and S_3. Notice especially the perturbation in the density curve of Y_3 versus its source S_1. In general, ICA recovery is permutation invariant.

$S_4 \sim Z + N_1$, $S_5 \sim Z + N_2$. $Z \sim Exp(1)$ and $N_1 \sim N(0,.5)$, $\perp$ $N_2 \sim N(0,1)$ are independent Gaussian noise. See Figure (8)

Violations of source independence appear to have a measurable effect on the CICA outputs; still the source distributions are recognizable. In other settings, the presence and divination of deficient rank sources is called model selection under *sparsity* and is an area of active research.

5 Discussion

On higher dimension data, in particular the example in Abayomi [2008b], we propose the partite model outlined in section 3.3.1 and the final paragraph of section 3.4. The mutual information, or other copula based measure of dependence, array in this modelling need not be a matrix. Higher orders of dependency may be determined via partial measures of dependence and represented in an array of higher dimension. Orthogonalization procedures for tensors may be applicable (see Leibovici and Sabatier [1998]) in a natural extension of the matrix singular value decomposition. The scatter matrix (tensor of order 2) can be extended to a scatter array (tensor of order greater than 2) to include 3-dependence (or higher) [see Ibragimov 2005 again].

For an extension to a full model, in analogy with this partite procedure, a version of this is to express the copula as a convex sum or convex integral where the mixing parameter is summed - integrated over. For example

$$C_{\Theta=(\theta_1,...,\theta_k)}(\mathbf{u}) = \int_{\Theta} u_1^{\theta_1} \cdots u_k^{\theta_k} d\Theta \tag{29}$$

and

$$C_{\Theta=(\theta_1,...,\theta_k)}(\mathbf{u}) = \theta_1 \sum_{i=1}^{k} u - \theta_2 \sum_{i,j} C_1(u_i, u_j) - \cdots (-1)^k \theta_{k-1} C_k(u_1, ..., u_k) \tag{30}$$

with $\sum_i \theta_i = 1$ [Wolff 1986] are simple versions - in general $C(\mathbf{u}) = \int_{\Theta} C^*(u_{\theta_1}, ..., u_{\theta_k}) d\Theta$ is a copula where Θ is a mixing parameter and C^* is some distribution function (parameter) [see Marshall and Olkin 1988]. See the Appendix.

5.1 Unification of PCA/ICA via the Gaussian copula

CICA, or ICA via the copula, yields a unifying framework in which PCA procedures can be cast. Take the more general case of 2-dependence/elliptical dependence (see Appendix). In these cases we can write the density of the copula as

$$dC_{\Theta}(\mathbf{u}) = \phi(\frac{1}{2}\mathbf{u}^T \Sigma^{-1} \mathbf{u}) = \phi(t)$$

with $\Theta = \Sigma$ the 'scatter' tensor of order 2 — i.e. a matrix — for multivariate $\mathbf{x}_k$, and where $\phi(T) \sim o(t^2)$. The Gaussian copula is a member of this family. In the CICA program we minimize the expected log of the above via equation (26), for any copula expressed 'dependency gradient'.

It is direct to note that the PCA program is a special case — the copula density matches the above, i.e. is Gaussian or elliptical — where the marginal mismatch via equation (22) is ignored. Alternately, note that PCA vs singular value decomposition (SVD) is a quadratic optimization, consonant with the expression of the elliptical copula density.

CICA then, via the Gaussian copula, is a generalization of PCA type procedures where Θ is a more general non-affine manifold or 'dependency gradient'.

Lastly, note the first term on the RHS of (16): this is the entropy of the inputs, and by invariance of entropy to invertible transforms, the entropy of the inputs, determined solely by the source distributions. Entropy is maximal for Gaussian sources; the second term on the RHS of (16) is *the* estimating equation for the mixing matrix $B = \hat{A}^{-1}$. Under ordinary ICA — any transform of the margins is arbitrary and identifying the contrast gradient from the entropy is difficult. In CICA, via equation (22), the second term on the RHS is identifiable from the first term — allowing for Gaussianity in the sources.

5.2 Extensions: latent models; non-independent components; Bayesian CICA

PCA and ICA type models are, in practice, used to reduce the dimension of multivariate data $\mathbf{x}_k$ to a set $(l_1, ..., l_m) \approx h(x_1, ..., x_k)$ with $k << m$, and h a linear function. The factor analysis program yields $\mathbf{l}_m$ as the first m components of a PCA/ICA type output. In PCA the components are ordered by the magnitude of the associated eigenvalues; in generalized ICA (and CICA) the component ordering is unspecified.[9] The output $\mathbf{l}_k$ is a version of latent modelling: a special case where the mechanism is linear yields the PCA/ICA models; when

[9] In the partite CICA model, it is possible to order the CICA outputs by associated eigenvalue of the SVD decomposition. The interpretation is less clear, though, as the 'independence' of the outputs is permutation symmetric. Further work on multi-stage CICA — where the components are extracted singly — is necessary.

$k << m$ the models are rank-deficient. In Abayomi [2008b] we explore this dimension reduction on environmental data.

The latent model approach to CICA may be immediately applicable as a version of an imputation algorithm. Factor analysis via CICA yields latent variables that capture (possibly non-gaussian) dependence in the multivariate data. A CICA type procedure may be used as an imputation[10] model *a priori* by using CICA factor output as covariates in a chained equation regression type imputation algorithm (see Chapter 1). More generally, copula-based estimating equations can yield imputations (of quantiles) in, for example, an expectation-maximization framework.

Component analysis under non-independence/dependence is an immediate extension of the CICA perspective. The estimating equations yielded by (17), (21), and (24) need only be constrained for varying Θ at values of dependence.

A Bayesian version of CICA arises from proposing a prior on Θ, thus seeding the CICA procedure with a distribution on the dependency for any model. If a parameterization of Θ can encompass a broad class of dependency models — similar to the representation of Generalized Extreme Value (GEV) distributions.[11] In absence of such a parameterization, versions of 'dependency-gradient' averaging could be implemented via Bayesian Model Averaging (BMA).

Most promising, perhaps, is the possibility of extending the copula based estimation for the ICA model to a more general copula, or dependency based, estimation for broad classes of Generalized Linear Models (GLM). ICA is the intuitive first application: the contrast function/estimating equations are solely copula dependent. The Kullback-Liebler (KL) distance is just the Mutual Information for these models. More generally, discovery of copula-'similar' representations of KL distance will yield estimating equations for the broader class of linear models; we look to reveal estimating equations which can exploit dependency.

[10]...or 'complete data' model. Any model used for imputation is a complete data model for the missingness mechanism. See Chapter 1 for comments

[11]GEV distributions may indexed by values of a parameter γ. This representation encompasses Weibull, Frechet, and Gumbel type distribution in one model see Gumbel [1958]

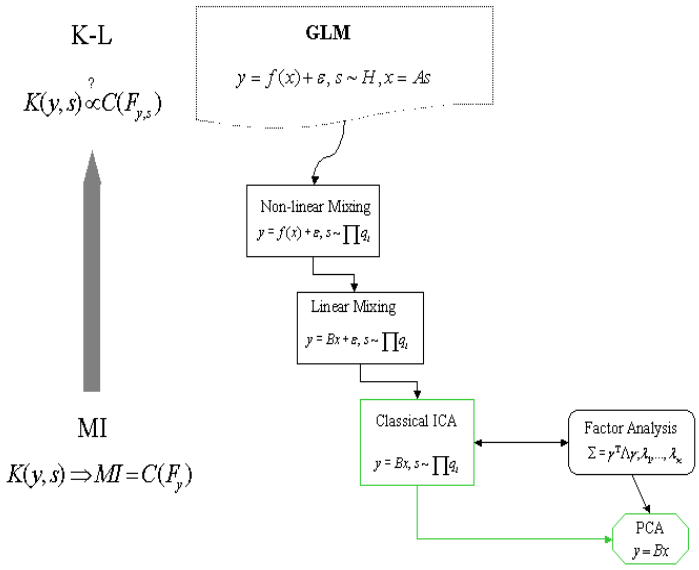

Figure 9 Illustration of extension of CICA to copula based estimation of General Linear Models (GLM). The CICA procedure is the natural first example of a class of linear models which can be defined via copula based estimating equations. The Kullback-Liebler distance has mutual information as a special case: the ICA problem is the characteristic setting and the copula families, as models for dependence/independence, are the intuitive engines for this estimation. Generalizing mutual information 'distance' (for ICA) to KL 'distance' for GLMs requires a recast to copula-'similar' families with fixed margins. The green are models covered in this dissertation.

6 Appendix

6.1 Measures of dependence

Standard versions of measures of dependence are on a bivariate vector.

Kendall's tau is a measure of the *concordance* between random variables. Two independent pairs of random variables, (X, Y) and (X', Y') are called *concordant* if:

$$\mathbf{P}((X - X')(Y - Y')) \geq 0 \tag{31}$$

and discordant otherwise. The index ρ_τ is defined

$$\rho_\tau = \mathbf{P}((X - X')(Y - Y') \geq 0) - \mathbf{P}((X - X')(Y - Y') \leq 0) \tag{32}$$

Spearman's rho, ρ_S is defined

$$\rho_S(X, Y) = 3(\mathbb{P}((X_1 - X_2)(Y_1 - Y_3) \geq 0) - \mathbb{P}((X_1 - X_2)(Y_1 - Y_3) \leq 0) \tag{33}$$

where $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$ have common distribution function F, and are independent. Spearman's rho, then, is the difference in concordance between a pair $(X_1, Y_1) \sim F$ and an independent pair $(X_2, Y_3) \sim \perp$. Spearman's rho is the Pearson's correlation on the marginal distribution functions.

Both ρ_S and ρ_τ can be calculated on the ranks, and as such have copula based representations.

$$\tau_{X,Y} = \tau_C = 4\mathbf{E}(C(U,V)) - 1 \tag{34}$$

$$\rho_S(X,Y) = 12\mathbf{E}(C(U,V)) - 3 \tag{35}$$

Multivariate extensions of either are generally functions from $\mathbb{R}^k \to \mathbb{R}$, i.e. they take a full multivariate copula and return a scalar. For example Wolff [1989] introduced

$$\rho_S^k = \frac{\int_{\mathbb{I}^k} C_\Theta(\mathbf{u})d\mathbf{u} - \int_{\mathbb{I}^k} \Pi(\mathbf{u})d\mathbf{u}}{\int_{\mathbb{I}^k} M(\mathbf{u})d\mathbf{u} - \int_{\mathbb{I}^k} \Pi(\mathbf{u})d\mathbf{u}} \tag{36}$$

where $M(\mathbf{u})$ and $\Pi(\mathbf{u})$ are the Fréchet upper bound copula and independence copula respectively. Any possible closed form for ρ_S^k is dependent upon the parameterization Θ. Variations on this theme weight dependence in the multivariate structure and then the measure of association is a weighted sum of bivariate (or higher) scalar measures of association.

In general, sample versions are calculated as 'averages' (often non-parametric kernels (see Schweizer [1981], Schimd [2007] or Joe [1990]) over lesser dimensions.

An alternate approach is to compute scalars at lower dimensions - on index sets, say - and present the group as an array. Simon [1977] computes the first C_2^k Kendall's taus on the bivariate index set in k and generates higher order versions using the inclusion-exclusion principle. Redundancies are explored as flags for dimension reduction.

6.2 Elliptical and Archimedean Copulae

A copula family $C_{\Theta=\Sigma}(\mathbf{u})$ is called elliptical if the associated density

$$dC(\mathbf{u}) = \phi(\frac{1}{2}\mathbf{u}^T\Sigma^{-1}\mathbf{u}) = \phi(t) \tag{37}$$

can be expressed as a function of a quadratic form. The matrix Σ is positive definite and is known as the *characteristic* or *scatter* array.

For any $\mathbf{X}$ with scatter array $\Sigma = (\Sigma_{(i,j)}, \Sigma_{(i,j),(-i,-j)}, \Sigma_{(-i,-j),(i,j)}, \Sigma_{(-i,-j)})$,' the variance of $X_i, X_j | X_{(-i,-j)}$, i.e. the conditional covariance - or dependency - for X_i, X_j; i, j a pair, (i, j) all pairs, $(-i, -j)$ all pairs save i, j, is:[12]

$$\Sigma_{i,j} - \Sigma_{(i,j),(-i,-j)}\Sigma^{-1}_{(-i,-j)}\Sigma_{(-i,-j),(i,j)} \tag{38}$$

.

In the case $\mathbf{X} \sim Ellip_k$, 2-dependence, $\Sigma^{-1}_{(-i,-j)}\Sigma_{(-i,-j),(i,j)} = \mathbf{I}$ and thus the conditional variance $X_i, X_j | X_{(-i,-j)}$ is just the pairwise matrix $\Sigma_{i,j}$.

Briefly, a copula family is called *archimedean* if the copula for $\mathbf{u}_k = F_{X_K}$[13] can be expressed:

$$C(\mathbf{u}_k) = \phi^{-1}(\sum_{i=1}^{k} \phi(u_i)) \tag{39}$$

[12]This *bivariate assumption* is, implicitly, held in the specification of Σ as a rectangular matrix. Suppose, this supposition is unjustifiable for one pair say, where $x_{i'}$ is some latent intermediary, $i' \in \mathcal{I}$. This means that the appropriate parameterization is $\theta_{ij} = (\theta_{ij1}, ...\theta_{iji'}, ...)$. Then $\mathbf{K}$ is an $I \times J \times I'$ cube. Etc. Etc.

[13]Let $\mathbf{u}_k = F_{X_K} = (u_1, ..., u_k) = (F_{X_1}, ..., F_{X_k})$

where $\phi \in \mathcal{L}_\infty^*$ [14] is a function from $\mathbb{I}^K$ to $\mathbb{R}^+$. See Nelsen [1999] and Joe [1997].[15]
Take $\phi_\theta(t) = (-log(t))^\theta$, then the above construction yields

$$C(\mathbf{u}) = exp\{-[\sum_{i=1}^{k}(-log(u_i))^\theta]^{-1/\theta}\} \tag{40}$$

the *Gumbel-Hougard* family, for example. Here, the dependency is held in parameter θ; $\theta = 1$ yields the independence - the product copula.

6.3 Multivariate Copulae

Joe [1997, 2001] highlights some the inherent difficulties in the construction of multivariate copulas (with dimensions greater than 2). Non-elliptical and non-archimedean distributions have large parameter spaces - in general the dimension of the parameter space is bounded by the dimension of the power set of k. A multivariate exponential survival distribution, for example, $\overline{F}(\mathbf{x}) = exp\{\sum_{s\in S}\lambda_s max(x_i s_i)\}$, where S is bounded in dimension by $2^{|S|}$.

One approach is to generate multiple parameter copulae via a 'mixing' approach outlined by Marshall and Olkin [1986]. The general theorem states that any distribution function can be generated as a mixing distribution: Let $H_1, ..., H_k$ be univariate distributions functions; let G be a $k-$variate distribution function with such that $Pr(Z_1 > 0, ..., Z_k > 0) = \overline{G}(0, ..., 0) = 1$ and marginals G_i; let ϕ and ϕ_{θ_i} be Laplace transforms of G and G_i; $\theta = (\theta_1, ...\theta_k)$. Then, for any K a k-variate distribution function with uniform marginals (K a copula type distribution function)

$$H(\mathbf{x}) = \int \cdots \int K(exp\{-\phi_{\theta_1}^{-1}H_1(x_1)\}, ..., exp\{-\phi_{\theta_k}^{-1}H_k(x_k)\})dG(\theta) \tag{41}$$

is a $k-$variate distribution function.
Let K_δ have the form of a 1 parameter Archimedean copula as in (39)

$$K_\delta(u, v) = \phi_\delta(\phi_\delta^{-1}(u) + \phi_\delta^{-1}(v)) \tag{42}$$

with ϕ_δ a Laplace transform parameterized by δ. Then

$$C_{\theta,\delta}(u, v) = \psi_\theta(-log\phi_\delta[\phi_\delta^{-1}(e^{-\psi_\theta^{-1}(u)}) + \phi_\delta^{-1}(e^{-\psi_\theta^{-1}(v)})]) = \eta_{\theta,\delta}(\eta_{\theta,\delta}^{-1}(u) + \eta_{\theta,\delta}^{-1}(v)) \tag{43}$$

where $\eta_{\theta,\delta}(s) = \psi_\theta(-log\phi_\delta(s))$ in a natural two parameter extension of the Archimedean copula generator representation. Junker and May [2005] provide a succinct recap by setting ϕ_δ as an Archimedean generator, $g_{\theta_1} : [0, 1] \to [0, 1]$ and $f_{\theta_2} : [0, \infty) \to [0, \infty)$ strictly concave and convex functions. Then $\phi_\delta \circ g_{\theta_1}$ and $f_{\theta_2} \circ \phi_\delta$ are both two parameter Archimedean copulas, as well.

This generator composition can be extended to k dimensions, yielding copulae with $k - 1$ parameters. These parameters, though, are constrained to be decreasing, which restricts the shape of the 'dependency gradient', in analogy with the positive definitiveness constraint for the multivariate Gaussian. There is no automatic multivariate extension for broad classes of dependency — copula models are bound in shape, similar to ordinary joint distribution functions. A reasonable approach — at this point — is to model lower dimensions with various copula and compose/average across these lower dimension models.

[14] Let $\mathcal{L}_\infty^*$ be the class of infinitely differential decreasing functions with alternating signs. This is the class of completely monotone functions.

[15] The *laplace* transform, $\phi(s) = \mathcal{L}(b(t)) = \int_{\mathbb{R}^+} e^{-sb(t)}dt$, holds a special place as a generator for mixing distributions of the archimedean type. In general, ϕ need only be in $\mathcal{L}_\infty^*$, *ibid.*

Full models, i.e. one copula family on a high dimension, constrain the 'dependency gradient' and may not capture varied dependency on lower orders. On the other hand, partite models on lower dimensions must be intelligently combined to yield consistent results on high dimensions.

We apply the partite approach by fitting two parameter copula families at bivariate margins in Abayomi [2008b] and apply the full model via 'dependency gradient' here.

6.4 PCA Whitening

Recall that the PCA program reduces the search for B with the constraint $\mathbb{E}(y_i y_j) = 0$ - the off diagonal covariance is zero.

Following the illustration in Cardoso [1998] and Lee [2000]: under the assumption in (7), further assume that $\mathbb{E}(ss^t) = I$; s is said to be 'white'. In general $\mathbf{y}$ is said to be 'whitened' if

$$\mathbb{E}(\mathbf{y}\mathbf{y}^t - I) = 0 \tag{44}$$

Let $\mathbf{z} = Wx$ such that $\mathbf{z}$ is whitened. Then set $A = RW$, that is $\mathbf{Y} = RW\mathbf{x}$. The procedure is to 'whiten' or decorrelate the source data before searching for rotation matrix R. R, then, is the basis on decorrelated data.

Some version of decorrelation pre-analysis is - almost uniformly - applied to most ICA algorithms.

6.4.1 Whitened PCA via Copula

The copula approach illustrates the similarity between ICA and PCA: PCA is a special case of ICA. In the notation of (21) and (25), general $G(\mathbf{y})$ and $G_i(y_i)$ are multivariate and univariate normal - $C(\mathbf{u})$ is the gaussian copula.

Here we 'whiten' $\mathbf{x}$ via ordinary PCA: $\mathbf{z} = W\mathbf{x}$. We then compute a version of (6) on 'whitened' $\mathbf{z}$: $\mathbb{I}_\Theta(z_i, z_j) = ((\mathbb{H}(C_{\theta_{ij}}(z_i, z_j))))$. The eigenvectors of $\mathbb{I}_\Theta(z_i, z_j)$ - $\mathbf{e}^t\Lambda\mathbf{e}$ - yield $\mathbf{y} = \mathbf{e}^t\mathbf{z} = \mathbf{e}^tW\mathbf{x}$.

Tipping and Bishop [1997] demonstrate that the subspace spanned by maximum likelihood is congruent to the subspace spanned in Probabilistic Principal Component Analysis (PPCA) - an extension of PCA where the source distributions are explicitly defined. The generalization of their result - the suggestion that eigen-decomposition of a maximum-likelihood matrix (and thus mutual information matrix, by the argument in section 2) is sufficient for CICA analysis - is a topic for future research.

7 References

Abayomi, K Gelman, A and Levy, M. (2008) "Diagnostics for Multivariate Imputation." *Journal of the Royal Statistical Society-C.* **57**, Part 3, 1-19.

Abayomi, K "CICA on the ESI". (2007) Unpublished Manuscript. Columbia University, NY.

Alfonsi, A. a. B., Damiano (2004). "New families of copulas based on periodic functions." Marnee-la-vallee, France and Milano, Italy, CERMICS, Ecole National des Ponts et Chaussees and Credit Models, Banca IMI, San Paolo Group.

Arststein, S. B., Keith; Barthe, Franck; and Naor, Assaf (2004). "Solution of Shannon's Problem on the Monotonicity of Entropy." *Journal of the American Mathematical Society.*

Bell, A. a. S., T (1995). "An Information-Maximization Approach to Blind Separation and Blind Deconvolution." Neural Computation 7: 1129-1159.

Biau, G. a. W., M. (2005). "A note on minimum distance estimation of copula densities." *Statistics and Probability Letters* 73: 105-114.

Cardoso, J and Comon, P (1996) Independent Component Analysis, a survey of some algebraic methods. In *Proceedings of ISCAS'96.* 2:93-96.

Cardoso, J and Souloumiac, A (1993) Blind beamforming for non-Gaussian signals. *IEE Proceedings F.* 140-6:362-370.

Carmona, R (2004). *Statistical Analysis of Financial Data, with an implementation in Splus.* Springer, New York.

Clarke, R. T. (2001), Surface Water and Climate - Separation of year and site effects by generalized linear models in regionalization of annual floods (Paper 2000WR900370), Water resources research, 37, 979 (978 pages).

Clarke, R. T. (2002), Surface Water and Climate - Estimating trends in data from the Weibull and a generalized extreme value distribution (DOI 10.1029/2001WR000575), Water resources research, 38, 25 (21 pages).

Clarke, R. T. (2003), Hydrology and Land Surface Studies - 24. Frequencies of future extreme events under conditions of changing hydrologic regime
(DOI 10.1029/2002GLO16214), Geophysical research letters, 30.

Comon, P. "Independent component analysis, a new concept?" Signal Proccessing, vol. 36, no. 3, pp. 287-314, Apr. 1994.

Davy, M. D., A. (2003). "Copulas: a new insight into positive time-frequency distributions." Signal Processing Letters, IEEE 10(7): 215-218.

de la Pena, V.H., Ibragimov, R and Sharakhmetov, Sh. (2003). Characterizations of joint distributions, copulas, information, dependence and decoupling, with applications to time series. *Erich L. Lehmann Symposium - Optimality, IMS Lecture Notes - Monograph Series, (J. Rojo, Ed.)* In Press.

Fermanian, J.-D. (2005). "Goodness-of-fit test for copulas." Journal of Multivariate Analysis 95: 119-152.

Francis, R.C., S.R. Hare, A.B. Hollowed, and W.S. Wooster, (1998): Effects of interdecadal climate variability on the oceanic ecosystems of the Northeast Pacific. *Fisheries Oceanography*, **7**, 1-21.

Genest, C. a. Q., Jean-Francois and Remillard, Bruno (2006). "Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation." Scandinavian Journal of Statistics 33(**2**): 337-366.

Gershunov, A., and T.P. Barnett (1998): Interdecadal modulation of ENSO teleconnections. *Bulletin of the American Meteorological Society*, **79**, 2715-2726

Gumbel, E.J. (1958) Statistics of Extremes. Columbia University Press.

Hotelling, H (1933) "Analysis of a Complex of Statistical Variables into Principal Components" *Journal of Educational Psychology*, **24**, 417-441, 498-520.

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

Hyvarinen, A. a. K., Juha and Oja, Erkki (2001). *Independent Component Analysis.* New York, Wiley.

Ibragimov, Rustam (2005). Copula Based Dependence Characterizations And Modeling For Time Series. *Harvard Institute of Economic Research.* Discussion paper number 2094.

Joe, H. (1997). Multivariate Models and Dependence Concepts. London, UK, Chapman and Hall.

Johnson, R. and W., Dean (1998). *Applied Multivariate Analysis.* Upper Saddle River, NJ, Prentice-Hall.

Junker, M and May, A (2005) 'Measurement of aggregate risk with copulas.' *Econometrics Journal* **8**. 428-454.

Hérault, J and Jutten, J. (1986) Space or time adaptive signal processing by neural network models. In *Neural Networks for Computing: AIP Conference Proceedings 151*, (Edited by J.S. Denker), American Institute for Physics, New York.

Jutten, C and Hérault, J. (1991) Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24**:1-10.

K. Lange. (2001) Numerical Analysis for Statisticians. Springer.

Lebovici, D and Sabatier, R. (1998) A Singular Value Decomposition of a $k-$Way Array for Principal Component Analysis of Multiway Data, $PTA-k$. Linear Algebra and its Applications 269:307-329.

Lindsay, B and Basak, Prasanta (2000) "Moments Determine the Tail of a Distribution (But Not Much Else)" *The American Statistician.* **54**,4:248-251.

Mari, D and Kotz, S. (2001) Correlation and Dependence. Imperial College Press. London.

McCullagh, P (1994) "Does the Moment Generating Function Characterize a Distribution?" *The American Statistician*, **48**, 208.

Kullback, Solomon (1959). Information Theory and Statistics. John Wiley and Sons, New York.

Nelsen, R (1999): An introduction to Copulas. Springer, New York.

Obradovic, D. a. D., G (1998). "Information Maximization and Independent Component Analysis: Is There a Difference." Neural Compuation 10.

Oja, E. (1992) "Principal Components, minor components, and linear neural networks." *Neural Networks*, **5**:927-935.

Oja, E. (1997). "The nonlinear PCA learning rule in independent component analysis." Neurocomputing 17: 26.

Scarsini, M. a. V., Achilles (1993). "Bivariate Distributions With Nonmonotone Dependence Structure." *Journal of the American Statistical Association* 88(421): 338-344.

Simon, Gary 'Multivariate Generalization of Kendall's Tau with Application to Data Reduction.' *Journal of the American Statistical Association*, **72**, 358: 367-376.

Sklar, A. (1973). "Random variables, distribution functions, and copulas." Kybernetica 9: 12.

Stephenson, A. (2002) "Simulating Multivariate Extreme Value Distributions of Logisitic Type" PhD. Dissertation. Lancaster University.

Stogbauer, H, Andrzejak, et al. (2004) 'Independent Component Analysis and Blind Signal Separation'. *Lecture Notes in Computer Science.* **3195**: 209-216

van der Vaart, A.W. (1998) Asymptotic Statistics. Springer-Verlag.

Whelan, N. (2004). "Sampling from Archimedean copulas." *Quantitative Finance* **4**: 339-352.
Whitt, W. (1976). "Bivariate Distributions with Given Marginals." *The Annals of Statistics* 4(6): 9.