Quantifying Sustainability: Methodology for and Determinants of an Environmental Sustainability Index

Kobi Abayomi
Georgia Institute of Technology, USA
Victor de la Pena
Columbia University, USA
Upmanu Lall
Columbia University, USA
Marc Levy
CIESIN at Columbia University, USA

ABSTRACT
We consider new methods of component extraction and identification for the Environmental Sustainability Index (ESI) – an aggregation of environmental variables created as a measure of overall progress towards environmental sustainability. Principally, we propose and illustrate a parametric version of Independent Component Analysis via Copulas (CICA). The CICA procedure yields a more coherent picture of the determinants of environmental sustainability.

INTRODUCTION
Shrinkage methods - statistical dimension reductions – are important and popular alternatives to numerical models in fields as diverse as climatology, psychology and econometrics. The objective in these methods is to identify a subset of coordinates that sufficiently describe the evolution of specific state variables. From an applied perspective, the goal is to identify (possibly lower dimension) versions of multivariate data via the extraction of salient characteristics. The data may then be recast, modulo these characteristics, as input to further modeling. From a theoretical perspective, the proposition of a method for dimension reduction depends upon the declaration of characteristics that can offer a sound basis for extraction.

Breiman states – *Statistics starts with data*; improved methods can illustrate latent phenomena and uncover alternative metrics in extant data [Breiman 2001]. This statistical duality, the hysteretic iteration of statistical theory and data application, is especially instrumental in emerging fields where functional and causal representations are sparse.

Social indexes, in particular environmental indexes, seek to describe as well as predict phenomena that are often poorly measured and ill-defined. An *index* is a metric, often at administrative levels, used to characterize a latent quality.

Gross Domestic Product (GDP) and of the Dow Jones indexes are common economic indices; Pacific Decadal Oscillation (PDO) and El Nino ([Francis 1998], [Gershunov1998]), climatological indices; the National Threat Level could also be called an index. Example environmental indices are the Natural Disaster Hotspots report [CHRR-World Bank 2005]; the Human and Ecosystems Wellbeing Indexes - (HWI) and (EWI) [Prescott-Allen 2001]; and the United Nations Human Development Index - (HDI) [UNDP 2006].

A goal for these environmental indices is the extraction of salient, perhaps latent, characteristics that describe or predict the elusive and undefined sustainability concept. *A fortiori*, the identification of as yet unmeasured information can illustrate the appropriate experimental design and thus guide future measurement (See Fuentes et al. [2007] for a creative example using Bernardo's [1979] fundamental comment on information maximization as a criteria).

Independent Component Analysis (ICA) - and the special case Principal Component Analysis (PCA) - extract uncorrelated and statistically independent components - or bases - of multivariate data. In ICA the model is explicit - the observed data are mixed independent sources; in PCA, implicitly, the data are mixed multivariate Gaussian. These *component analysis procedures* are used to reduce dimension – by yielding a lower order basis – and to parse or elucidate latent factors.

Environmental data are often non-Gaussian, and frequently – characteristically – extreme value [Meyers and Ganipati 2006]. Researchers apply an array of approaches: from spatial-temporal processes [Stein 2007], to stochastic optimization [Tsai and Chen 2004], and hierarchical models [Lin, Gelman, Price,and Krantz 1999]. Environmental statisticians rely upon a suite of statistical methodologies as the underlying processes are complex (as in transport phenomena), multiple (as in wastewater treatment), or latent (as in ecology). Environmental statisticians face particular challenges in modeling environmental processes; these are typically 'out-of-control' and require more sophisticated assumptions.

While the concept of sustainability has been widely embraced, it has been defined only vaguely and has proven difficult to measure with any consensus. There is a critical need for sustainability indicators; environmental statisticians have a stake in making the broad concept of sustainability operational. Researchers can justify an increased focus by providing specific measures – which decision makers can use and the public can judge – of progress or failure.

In this chapter we illustrate the 2002 Environmental Sustainability Index and exploit its dependency structure using a new version of ICA – Copula Based Component Analysis (CICA) – to extract a reduced component set as the determinants of environmental sustainability. This approach is designed to highlight important information, suggest some focal metrics, and discredit others.

A unifying definition for an i*ndex*, in the context of this paper: a function that maps disparate multivariate data onto a scalar at administrative units. An index should be:

I. **Transparent**: The methodology use to construct the index should be clear and unambiguous. Assumptions and decisions that affect index values (`scoring') should be well stated.
II. **Reproducible:** The algorithm or method used to generate the index – the list of scores and ranks for a set of administrative units – should be replicable on similar data.
III. **Defensible:** The elements and variables of the index should map to concepts the index claims to measure.

THE 2002 ENVIRONMENTAL SUSTAINABILITY INDEX (ESI)

The 2002 Environmental Sustainability Index (ESI) was created as a measure of overall progress towards environmental sustainability and designed to permit systematic and quantitative comparison between nations [World Economic Forum 2002]. The ESI is a scaled linear combination of 64 variables of environmental concern. Environmental measures (such as oxide emissions and concentration) are included along with political indicators relevant (such as civil liberty and level of corruption) that are relevant to environmental sustainability [World Economic Forum 2001, 2002].

The 2002 ESI is defined as:

$$\mathrm{ESI} = 100 * \Phi\left(\frac{1}{|\mathsf{K}|} \sum_{k \in \mathsf{K}} \frac{1}{|J_k|} \sum_{j \in J_k} \left(\frac{Y_j - \bar{Y}_j}{S_{y_j}} \right) \right). \qquad (1)$$

Here: J_k is the index set for the variables in the kth `indicator' of the ESI: the ESI is averaged over the `indicators'; the ESI `components' are a heuristic grouping, not used in calculating the index; K is the index set for the indicators; $|\mathsf{K}|$ and $|J_k|$ are the number of indicators and number of variables in the kth indicator; $\bar{Y}_j$ is the sample mean for variable j – across countries, S_{y_j} is the sample standard deviation for variable j, Φ is the inverse standard normal distribution function. See Table 1.

Environmental Systems (13 variables) Measurements on the state of natural stocks such as air, soil, and water
Environmental Stresses (15 variables) Measurements on the stress on ecosystems such as pollution and deforestation.
Vulnerability (5 variables) Measurements on basic needs such as health, nutrition, and mortality.
Capacity (18 variables) Measurements of social and economic variables such as corruption and liberty, energy consumption, and schooling rate.
Stewardship (13 variables) Measurements of global cooperation such as treaty participation and compliance.

Table 1: Components of the 2002 Environmental Sustainability Index

The ESI, like other indices of environmental concern (such as the environmental wellbeing index (EWI), and the human development index (HDI)) condenses dissimilar social and physical metrics into cohesive summaries for national level comparisons [Prescott-Allen 2001, Osberg 2002]. The goal for the ESI is to capture the most recent version of available data to get the best, most recent, snapshot. The approach is to use the most recent year available for each variable at each country.

The breadth of the ESI - 64 dissimilar variables from varied sources - presents aggregation and processing challenges: in particular missing values (missingness) and complex dependencies. Some variables are composites of information from several sources: pollutant yield divided by land area conditioned on population density, for variables in the `Environmental Systems' and `Environmental Stresses' indicators - for example. Others may be imprecise across observations: mortality and disease variables in the `Vulnerability' indicator, for instance. See Annex 1 and Annex 2 of the 2002 ESI report for elucidation [World Economic Forum 2002].

Constructing the ESI using only available cases would have severely restricted its scope; yet it was important to have a reasonability check for the imputations, In Abayomi [2008] we look at the fit of a chained equation imputation model to the completed data and we suggested post hoc diagnostics designed to account for inconsistency and missingness in multivariate data collected from multiple sources.

The ESI was calculated, using the equation in (1), on the completed – post-imputation – data. The use of the inverse standard normal distribution in (1) guaranteed scores on the range 0-100; scaling each variable by its sample standard deviation set the contribution of each in deviation units; combining variables in groups before average allowed each component to have equal contribution to the overall score. Generally, countries with higher GDP scored higher in the ESI – though the relationship is not perfect. For example, the United States scored lower than Canada, and China scored lower than Australia.

An illustration of the final, completed data ESI is in Figure 1.

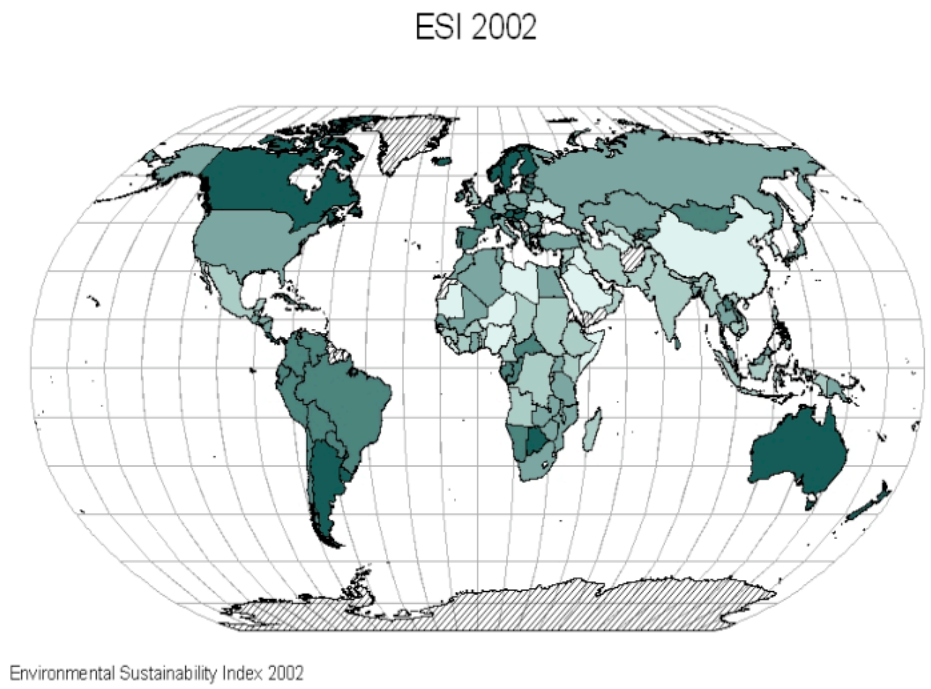

Figure 1: The 2002 ESI. Darker color indicated a higher, more 'sustainable score on the index. Canada and Norway, for example, are more 'sustainable' than China or the United States.

CICA FOR DETERMINANTS OF ENVIRONMENTAL SUSTAINABILITY

The Component Analysis Procedure

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 $\mathbf{y}$. 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 $\mathrm{Cov(y\_i,y\_j)} = 0, \quad i \neq j$, 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. See 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 (in the maximal correlation sense [see Hyvarinen 2001]) 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 **s** . The columns of **Y** are the estimates of these independent components, or signals; which the rotation B is an estimate of A^{-1}. See Figure 2, a la Cardoso [1996].

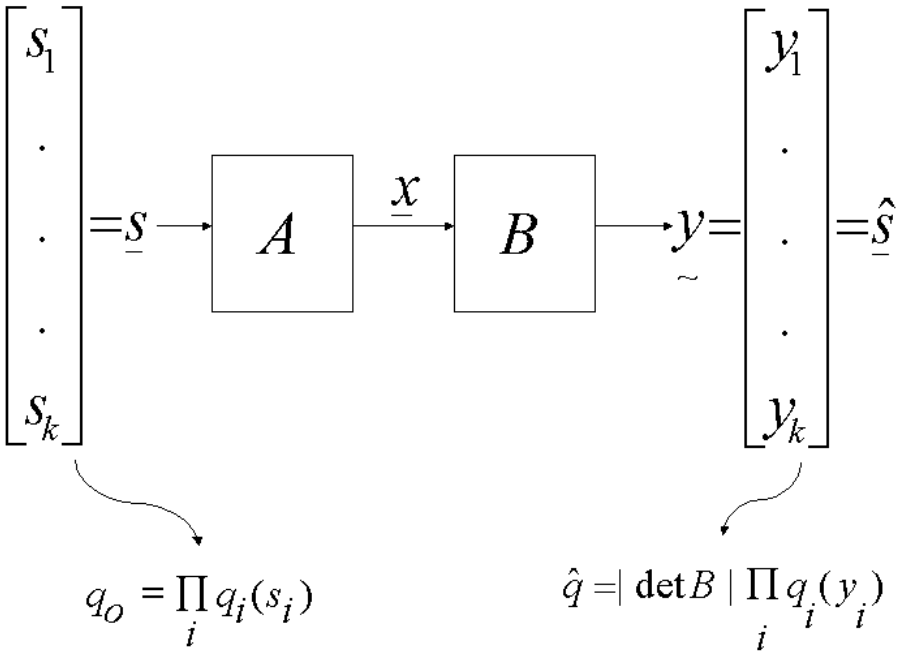

Figure 2: Diagram of Independent Component Analysis (ICA) mixing and separating matrices. Typically, independent signals **s** are observed via unknown full rank rotation A as **x** . The ICA/BSS procedure yields **y** = $\hat{\mathbf{s}}$ outputs as estimates of the independent signals. The distribution of the inputs and outputs should be proportional.

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, **x** . The difference is the characterization of statistical independence or contrast function, and the implicit or explicit distributional assumptions on the inputs (See Cardoso [1993], 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 independence. For a simple example, take functional dependency $x_i = h(x_j) = x_j^2$, for example, $\mathsf{E}(x_i) = 0$. 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 de-correlation, etc. (Cardoso [1996]).

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.

The Copula Approach

A *copula* is a multivariate distribution on marginal distribution functions --- a distribution function on a $k-$ dimensional cube --- and holds the dependency of the full joint distribution. In illustration: take two random variables

$X_1 \sim F_{X_1}, X_2 \sim F_{X_2}$.

A copula is a function that takes the 'grades' as arguments --- the pair (U,V) are the 'grades' of (X_1,X_2) --- and returns a joint distribution function

$C(U,V) = F_{X_1,X_2}$,

with marginals F_{X_1}, F_{X_2}. In a simple illustration, the Gumbel-Hougard copula --- $C_\theta(u,v) = (u+v-1) + (1-u)(1-v) * e^{-\theta \ln(1-u)\ln(1-v)}$ --- is easily derived from the bivariate exponential distribution: $H_\theta(x,y) = 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)}$. Notice if $\theta = 0$, then $C_\theta(u,v) = uv$ ...the independence copula.

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 (see Nelsen [1999]). 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 use parametric copulae families as estimators for the dependency in $\mathbf{x}$ under broad dependence conditions. Specifically, the copula approach offers a generalized 'engine' for the contrast functions - measures of statistical dependence - which characterize ICA analysis. This yields a copula based version of Independent Component Analysis (CICA) - where we model and rotate dependency information in $\mathbf{x}$ via copulae families.

This version - CICA - replaces non-parametric, higher order proxies for independence with parametric examples from the copula literature. This parametric modeling 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.

Here, we can construct 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. This is a consistent framework for fully parameterized ICA.

Copulas in ICA

The copula measure of dependency is defined via its density, on a multivariate $\mathbf{x} = (x_1,...,x_k)$,as

$$dC(\mathbf{x}) = \frac{dF_{\mathbf{x}}(\mathbf{x})}{\prod\prod dF_{x_i}(x_i)} \tag{2}$$

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. The copula distribution, then, is a distribution function on the space of the marginals to the unit hypercube, $(F_{X_1},...F_{X_k}) \mapsto \mathsf{I}^k$.

The mutual information (see Kullback [1959]), for a multivariate $\mathbf{X}$ with distribution function $F(\mathbf{X})$ is

$$MI(\mathbf{x}) \equiv \int_{\Omega} dF(\mathbf{x}) log(\frac{dF_{\mathbf{X}}}{\prod\prod dF_{X_i}}) \tag{3}$$

where Ω is the probability space for $\mathbf{X}$. Using equation (2) above, this can be re-expressed as

$$MI(\mathbf{X}) \equiv \int_{\mathsf{I}^k} dC_\theta(\mathbf{u}) log(dC_\theta(\mathbf{u})) = MI(\mathbf{u}) = \mathsf{E}(log(dC_\theta(\mathbf{u}))) \tag{4}$$

When $T \sim F, dF = f$ then $-H(T) = \mathsf{E}(f(T) log(T))$ is called the *entropy* for t (see Ash 1965) and here,

$$MI(\mathbf{X}) = -H(\mathbf{u}) = \int_{\mathsf{I}^k} dC(\mathbf{u}) log(dC(\mathbf{u})) \tag{5}$$

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

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 $\forall i \neq j$ --- then ψ is a particular contrast function. The minimization of functions of these types is the essence of the PCA/ICA algorithm.

Essentially, this approach demonstrates a role for the copula as the apparatus for these contrast functions, which exploits its natural appearance in measures of association, here the mutual information, and as a model for dependence/independence. This is choosing the mutual information as the engine for the ICA contrast function. This is a special case the component analysis problem via minimization of a \ *parametric* probability distance. This yields symmetry with the principles of likelihood maximization and employs a decomposition of the *Kullback-Liebler* distance.

Kullback-Liebler as dependence distance

The Kullback-Liebler [Kullback 1959] divergence between two probability density functions $f(\mathbf{t})$ and $g(\mathbf{t})$ we notate

$$\mathsf{K}(f,g) = \int_{\mathbf{t}} f(\mathbf{t}) log(\frac{f(\mathbf{t})}{g(\mathbf{t})}) \tag{6}$$

between two probability density functions, $f(\mathbf{t})$ and $g(\mathbf{t})$.

The mutual information is a special instance of the Kullback-Liebler (K-L) probability distance between independence and dependence.

If $\mathbf{X}_k$ is $k-$ dimensional multivariate with density function dF and marginal distributions $dF_1,...,dF_k$ then

$$\mathsf{K}(dF,\prod_{i=1}^{k} dF_i) = MI(\mathbf{X}) \tag{7}$$

A classic property of (7) is its decomposability

$$\mathsf{K}(\mathbf{y},\mathbf{s}) = \mathsf{K}(\mathbf{y},\mathbf{y}^*) + \mathsf{K}(\mathbf{y}^*,\mathbf{s}). \tag{8}$$

with $\mathbf{y}^*$ a random vector with independent entries and margins distributed as $\mathbf{Y}$; $\mathbf{S}$ is an independent vector.

In the component analysis procedure --- with $\mathbf{y}$ the outputs and $\mathbf{S}$ the unobserved sources --- the total distance between the model and the outputs is decomposed into the deviation from independence of the outputs $\mathsf{K}(\mathbf{Y},\mathbf{y}^*)$ and the mismatch of the marginal distributions $\mathsf{K}(\mathbf{Y}^*,\mathbf{s})$.

$$\left\{ \begin{array}{c} Total \\ Mismatch \end{array} \right\} = \left\{ \begin{array}{c} Deviation\ from \\ Independence \end{array} \right\} + \left\{ \begin{array}{c} \text{Marginal} \\ Mismatch \end{array} \right\} \tag{9}$$

Setting $\mathbf{u}^* = G(\mathbf{y}^*)$ --- G our best estimate for the marginal distributions of $\mathbf{y}$ --- 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 $\mathbf{y}$. Then the KL distance is:

$$\mathsf{K}(\hat{\mathbf{u}},\mathbf{u}) = \mathsf{K}(\hat{\mathbf{u}},\mathbf{u}^*) + \mathsf{K}(\mathbf{u}^*,\hat{\mathbf{u}}) \tag{10}$$

with $\hat{\mathbf{u}}$ the estimate of the true sources.

The CICA algorithm: Full Model, via Estimating Equations

This approach yields *estimating equations*, equations for the parameters of the component analysis model. In this *full* CICA method – we derive estimating equations for the mixing parameter B in the model $\mathbf{Y} = B\mathbf{X}$ by minimizing the KL distance (i.e. maximizing the likelihood).

Under fixed assumptions about the distribution of the sources, two terms are minimized: the true objective, the mutual information, expressed via the copula; the mismatch of the marginal distributions to the assumed distributions.

Write the independence term as

$$\min_B \mathbf{MI}(\mathbf{y};B) = \min_B \mathsf{E}(log(dC_\Theta(\mathbf{u}))) \tag{11}$$

and the marginal fit term as

$$\min_\Theta [C_\Theta(\mathbf{u}) - \prod_{i=1}^{k} (u_i)]. \tag{12}$$

That is, minimize the mutual information via the copula via rotation $B = \hat{A}^{-1}$ after minimizing the distance between parametric copula and independent marginals. Since A is invertible, the KL divergence is invariant; maximization of the model likelihood – under independence – is equivalent to minimizing equation (13), below.

$$\frac{\partial \mathsf{H}(G(\mathbf{y}))}{\partial A} = \frac{\partial}{\partial A}(-\mathsf{K}(G^{*}(\mathbf{y}), G(\mathbf{y}))) \tag{13}$$

This is the same as maximizing the *score*, equation (14)

$$\frac{\partial L}{\partial B} = -\frac{\partial}{\partial B}\mathsf{K}(q(\cdot), \hat{q}(\cdot, B)) \tag{14}$$

via the marginal distributions

$$\frac{\partial L}{\partial B} = -\frac{\partial}{\partial B}\mathsf{K}(\hat{\mathbf{u}}, \mathbf{u}) \tag{15}$$

using the copula model. The estimates for B are yielded by partial derivatives, or score maximization $\partial L / \partial B$ --- either through gradient descent or analytically. See Figure 3.

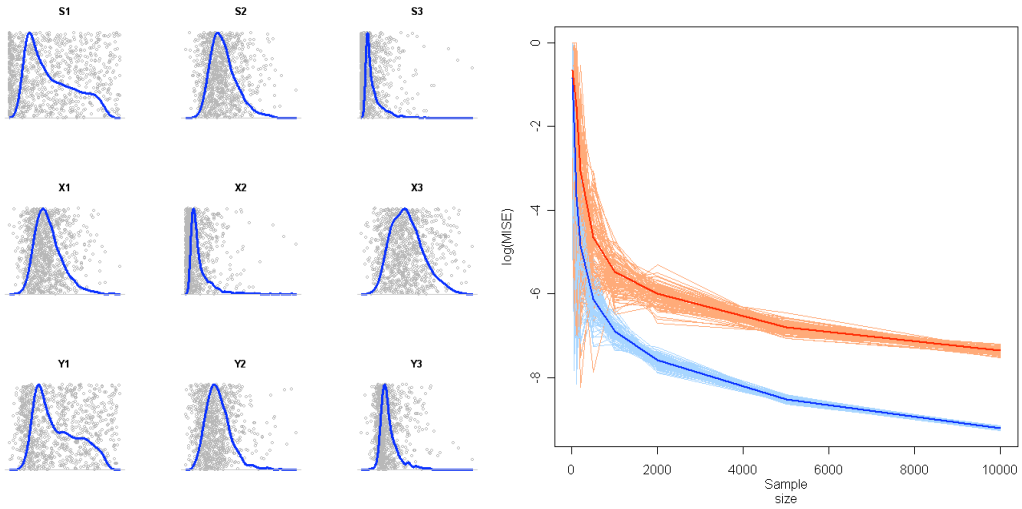

Figure 3: Left Hand Panel: CICA model applied to Gumbel-Hougard dependency gradient; Right Hand Panel: Log Mean Integrated Squared Error (MISE) of typical ICA (fastICA) and CICA models. The first row are the source distributions, all non-normally distributed: $S_1 \sim (U(-1,1))^2$, $S_2 \sim Gumbel(0,1)$, $S_3 \sim \chi^2$. 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.(b) Log Mean Integrated Squared Error (MISE) of fastICA (see hyvarinen 1999) and CICA model applied to mixed Gaussian and Laplacian sources (via Gumbel-Hougard copula) - $S1$ and $S2$ above. 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.

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. In the particular case of elliptical dependence we can write the density of the copula as

$$dC_{\Theta}(\mathbf{u}) = \phi(\frac{1}{2}\mathbf{u}^T\Sigma^{-1}\mathbf{u}) = \phi(t) \tag{16}$$

with $\Theta = \Sigma$ the 'scatter' 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 full CICA procedure we minimize the expected log of the above via equation (11) 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 (equation (9)) is ignored. Alternately, note that PCA via 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-linear space or `dependency gradient'.

Lastly, note under ordinary ICA --- any transform of the margins is arbitrary and identifying the contrast gradient from the entropy is difficult. In CICA, via equation (9), the second term on the RHS is identifiable from the first term --- allowing for Gaussianity in the sources.

The advantages of this approach over non-parametric component analysis models are:

1) Flexible choice of non-linear transformations $\mathbf{u}$.
2) Superior convergence of parametric estimators and stability of parametric estimators on small datasets.
3) Specification, 'tuneability' and interpretability of dependency.

The main drawback of this method, especially on high dimension data, is the computational difficulty of the score maximization, equations (13) through (15). This full algorithm requires a non-linear optimization procedure on the full dimension of the data simultaneously.

The CICA algorithm: Partite Model, Determinants of the ESI

Essentially, the full method is simultaneously minimizing the mutual information and marginal fits. In the fully parameterized setting, the joint entropy of the outputs is:

$$\mathsf{H}(\mathbf{u}) = \sum_{i=1}^{k} \mathsf{H}(u_i) - \mathsf{I}(\mathbf{u}) \tag{17}$$

and the full method is equivalent to the maximization of the above equation. Notice that $\sum_{i=1}^{k} \mathsf{H}(u_i)$ is maximized when the u_i are uniform - when the hypothesized marginal distributions for the components are well `fit'. $\mathsf{I}(\mathbf{u})$ is minimized when the components are independent.

An alternate, though philosophically consonant, approach is to:

1) Still exploit the empirical distribution, setting $u_i = \hat{F}_i^n(x_i)$, treating the univariate marginals as observed or unparameterized, but…
2) Fit copulas pairwise, say, and minimize $\mathbf{I}(\mathbf{u})$ by diagonalization of a *mutual information matrix*

$$MI_{\Theta}(X_i, X_j) = ((dC_{\theta_{ij}}(\mathbf{u}) log(dC_{\theta_{ij}}(\mathbf{u}))]) = ((MI_{\theta ij})) \quad (18)$$

This approach permits dependencies that may be restricted or inaccessible in many multivariate copulas, where the index sets for the dependence parameters must be hierarchical or nested (see Joe 1997, Simon 1986). As well, the number of families of bivariate copula is much larger than the those for multivariate copula – as many bivariate copula cannot be extended into greater dimensions [Joe 1997]. Partite copula estimation, in this manner via bivariate pairs, models the k-independent marginal dependence without the restrictions inherent in k-independent full joint models, with the sacrifice that $\mathbf{I}(\mathbf{u})$ not estimated beyond the second order.

Set $\mathbf{y} = \mathbf{RWx}$, with $\mathbf{W}$ a `whitening' matrix - the PCA transformation - and $\mathbf{R}$ the ICA transformation. This allows diagonalization of the final mutual information matrix via ordinary, quick, Singular Value Decomposition (SVD). The partite algorithm is:

I. Compute $\mathbf{Wx}$, the PCA output or whitened data.
II. Estimate univariate distributions $u_i = \hat{F}_i^n(\mathbf{W}_i\mathbf{x})$, $v_j = \hat{F}_j^n(\mathbf{W}_i\mathbf{x})$ via the empirical CDF.
III. Choose copula families at each bivariate pair: $C(\mathbf{u}) = \eta_{\theta_1,\theta_2}(\eta^{-1}_{\theta_1,\theta_2}(u) + \eta^{-1}_{\theta_1,\theta_2}(v))$.
IV. The bivariate mutual information, or, $E(log(dC(u_i, v_j)))$ are the elements of the `scatter' matrix.
V. Construct `scatter' matrix $\Gamma_{C_\Theta} = ((C_{\theta_{ij}}(u_i, u_j)))_{i,j=1..k}$
VI. Compute SVD of Γ_{C_Θ}, $\lambda_1, ..., \lambda_k$
VII. Yield $y_k = b_k\mathbf{x}_k = r_k w_k x_k$ with $y_i \perp y_j$, \: $\forall i, j$ via C_Θ

When the bivariate copula are well fit: $C_{\hat{\theta}_{ij}} \rightarrow MI(C_{\hat{\theta}_{ij}}) \geq 0$ for all i, j. Thus $\mathbf{R} = ((MI(C_{\hat{\theta}_{ij}})))$ is positive semi-definite, by exchangeability and the Singular Value Decomposition of $\mathbf{R}$ yields an orthogonal basis, with respect to the mutual information.

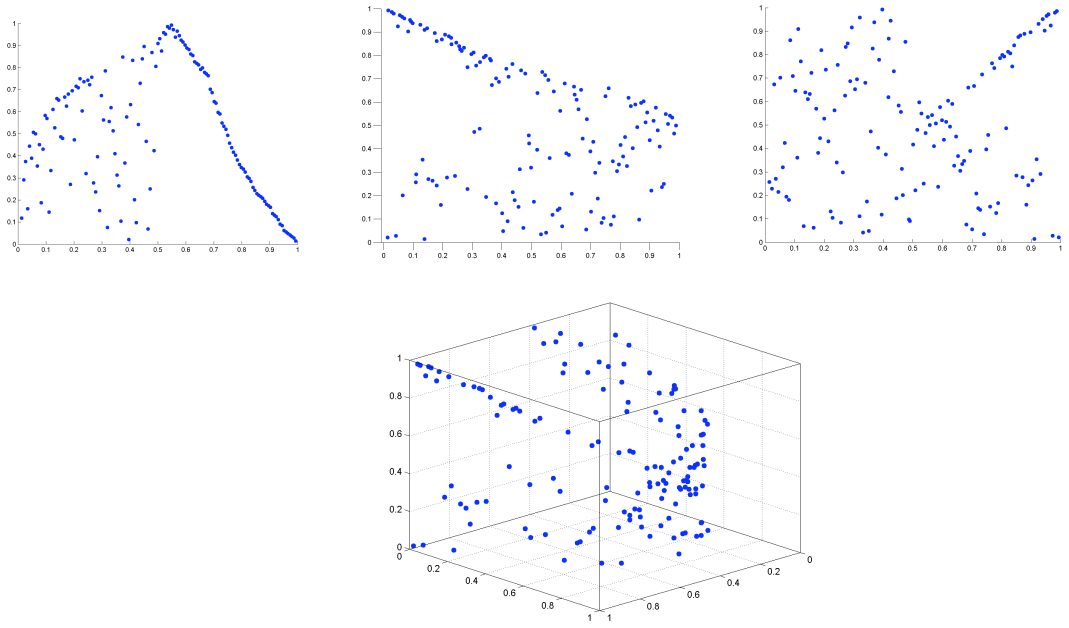

Figure 4: Upper panels: Plots of first three PCA components. Each $u_i = \hat{F}_i^n(x_i)$; the data as transformed by their empirical CDF's. Copulas are estimated, separately, on each bivariate pair. Lower panel: 3D Plot of PCA_1 vs. PCA_2 vs. PCA_3; the bivariate plots are the planes in the 3D plot. The appearance of extreme dependence in the bivariate plots - figures (a) and (b) especially, is a feature of the imputation procedure. Compare these with the bivariate diagnostic plots in Abayomi et al. [2008].

The partite approach permits copula model selection at each index in the partitioned index set; we fit bivariate copula to the $\binom{64}{2}$ pairs. The candidate copulae at each pair are two-parameter extensions of *Laplace* type copulae, a subset of the *Archimedean* family for copulas (see Abayomi [2008a]). Two-parameter families have the advantage of allowing multiple types of dependence, including some non-monotone dependence. Archimedean copulas are exchangeable and have a direct generating function representation [Joe 1997]. These properties are attractive for this partite approach: we trade for model flexibility, in a sense, at each of the bivariate margins with a full model on the entire data.

The exchangeability of the Archimedean family, with the non-negativity of the (copula) mutual information, yields a positive semi definite mutual information matrix Γ_{C_Θ} which can be orthogonalized via ordinary SVD methods. See Figure 5.

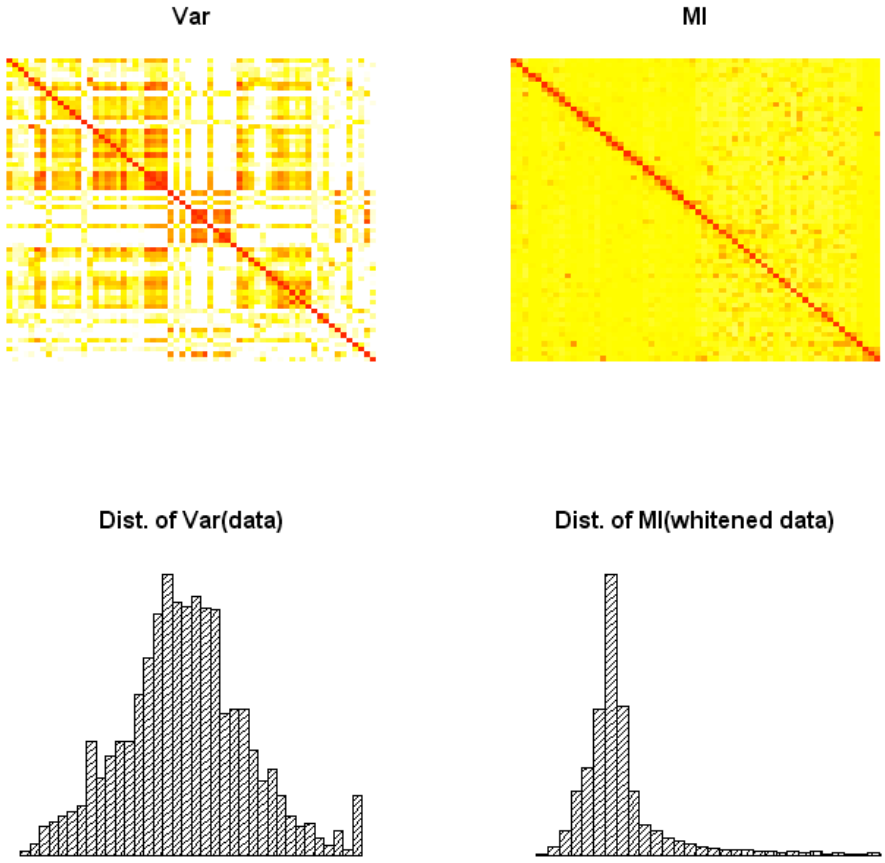

Figure 5: Image plots of covariance and mutual information matrices of ESI data - both matrices scaled for comparison. Mutual information matrix calculated via copula on `whitened' (PCA output) data. Darker color indicates greater covariance/mutual information. Image plot of MI illustrates remaining variation/information. Histogram of MI reveals the same – PCA alone ignores remaining non-Gaussian information. The MI matrix features high information about the diagonal; this is supported by the proximate listing of similar variables in the ESI data.

In analogy with the covariance/correlation matrix in a PCA procedure, we use the mutual information matrix Γ_{C_Θ} - estimated via the bivariate copulae - as a representation of the multivariate scatter. In PCA the covariance for each bivariate is estimated via $\mathbf{x}^T\mathbf{x}$; in this version of CICA the mutual information for each bivariate is estimated via the copula density:

$$\sum_{n=1}^{n=142} dC_{\hat{\theta}}(u_i^n, u_j^n) log(dC_{\hat{\theta}}(u_i^n, u_j^n)) \qquad (20)$$

, where each $u_i = \hat{F}_i^n(w_i x_i)$ is the order statistic of the `whitened' data, $\mathbf{W}$ the `whitening' matrix. ICA methods typically optimize the mutual information, negentropy (distance from Gaussianity), or high order

sample moments (usually cumulant) via gradient descent or other iterative procedure. In this version of CICA we substitute the iterative optimization with SVD orthogonalization.

Copulae at each pair are selected - separately - via maximum likelihood from bivariate two-parameter (Archimedian) Laplace families in Joe [1997]. Additionally, each copula model is rotated 0, 90, 180 or 270 degrees.

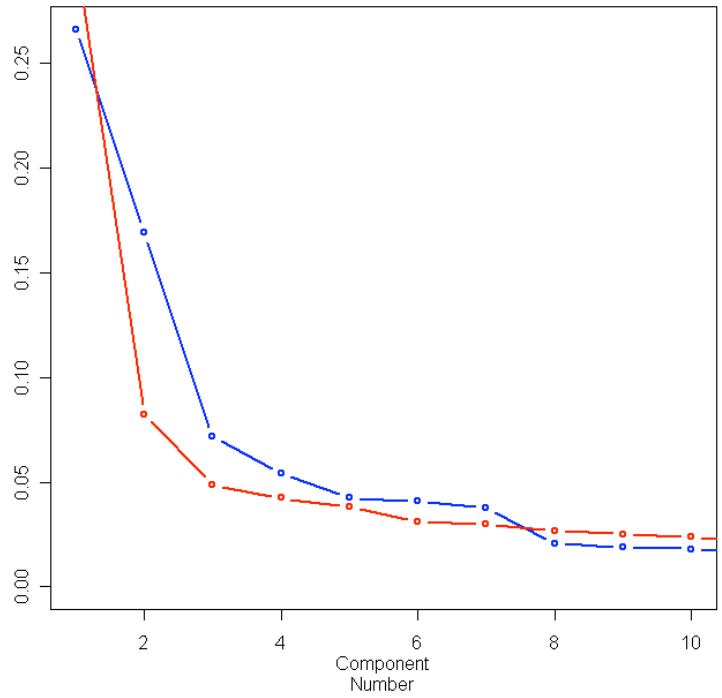

Figure 6: Scree plot [Catell 1966]: The y-axis is $\lambda_i / \sum_i \lambda_i$, where λ_i the i^{th} largest eigenvalue of the Singular Value Decomposition (SVD). The graph is an illustration of the `variation' explained up to the i^{th} component. The red line is the scree graph for PCA components on the ESI dataset; the blue line is for the CICA components. The area under each curve is the percentage of total 'variation' - then - at each component. Seven (7) components for the PCA and CICA graphs are, respectively, 57.6 and 68.3 of the total`variation'.

The Scree plot in Figure 6 [Catell 1966] suggests that the majority of `variation' (68 percent) - as approximated by the cumulative eigenvalues of the SVD – is explained at about seven components. This can be interpreted as an indicator that a majority of the variation – Gaussian as well as non-Gaussian, by the CICA procedure – in the ESI can be explained by a reduced amount of information.

The factor loadings [Wherry1984] – the coefficient weights the CICA rotation assigns to the variables of the ESI – allude to the importance of air quality in the first independent component, at least, in concert with water quality, childhood mortality and level of economic subsidy. This is illustrated in the list of variables with the greatest loadings or coefficient weights, in table 2.

Variable Name	Component 1	Component 2	Component 3
SO2	1	33	54
NO2	2	24	42
TSP	3	16	33
ISO14	4	43	35
WATCAP	5	43	35
IUCN	6	23	25
CO2GDP	7	52	61

Table 2: Variables listed in order of first CICA component loading – magnitude of absolute value – and subsequent order of loading in components 2 and 3. The first component is dominated by stressor linked to air and water quality.

The collection of traditional `stock' and non-traditional `social capacity' variables in the first three CICA components is interesting, especially so when contrasted with the loadings generated by PCA alone. The variables identified by the CICA method offer a more coherent illustration of the drivers of variation in the ESI; the divergence in the CICA factors from the PCA one is a proxy for the additional, non-Gaussian, information or variability in the ESI. This difference is due to the ability of the CICA algorithm to capture dependence information in the data beyond multivariate normality. Table 3 lists the CICA loadings vs. the PCA loadings for the first component.

CICA	PCA
SO2	NUKE
NO2	BODWAT
TSP	TFR
ISO14	FSHCAT
WATCAP	PESTHA
IUCN	WATSUP
CO2GDP	GRAFT

Table 3: First component CICA loadings vs. PCA loadings for ESI. Air and water effluent, and treaty membership dominate the first component. Conversely, the first PCA component is less cohesive. The CICA loadings – the first order determinants of the ESI – suggest that the major drivers of sustainability are pollution (air and water) and capacity.

THE FUTURE FOR (ENVIRONMENTAL) INDEXING

Any index is essentially a – linear or non-linear – collection of (almost always) non-independent variables for the purpose of projecting a multidimensional concept onto a univariate scale of comparison. The scale of comparison – the range of the index – though arbitrary, is completely determined by the scheme for index construction and the characteristics of the underlying data. A useful index must be thoughtfully constructed; consumers of the index, perhaps intuitively, typically focus on relative rankings rather than absolute score. This is certainly true for development indices – where relative performance can drive international aid.

In a direct sense, the projection of the multivariate data onto the univariate scale is **the** definition of the index. When this projection is well known or easily predictable, the scheme for construction is straightforward: construct the index, i.e. weight the variables, to minimize a loss between the index and its predictable value.

In general, let the value of the index, for one of a collection of administrative units, i, be θ_i. Data arrive as $\mathbf{X} = (X_1,...,X_k) \sim f_{\mathbf{X}}$, a collection of ratings/scores with some multivariate, non-independent,

distribution $f_{\mathbf{X}}$. A full (linear) indexing scheme would yield: the scores for each unit; the explicit, perhaps endogenously determined weights; and confidence intervals for the index scores as well as the variable weights. That is: $\theta_i = \sum_{j=1}^{K} c_j X_j$ - the scores for each unit; the vector , $\mathbf{c}^T$ the weighting scheme chosen for the index; confidence intervals for the scores, $\mathsf{P}(\theta_i \in (L_i, U_i)) = 1 - \alpha$, and weights, $\mathsf{P}(c_j \in (l_i, u_i)) = 1 - \alpha$.

Choosing the appropriate weighting scheme and generating confidence intervals for each scalar θ_i are separable tasks. On the other hand, the confidence intervals are of course affected by the choice of weighting scheme, even when the weights themselves are arbitrary in the sense that they are subject to an exogenous constraint chosen by the indexers.

Essentially this couples the task of definition and prediction for the indexer: assignment of the weights is the specification of the index, but the specification of the index as a proxy for measurable idea must influence the estimation of the weights. Disentangling these tasks is heuristically, computationally and theoretically non-trivial.

The author, in upcoming work on an index designed to measure progress towards the United Nations Development Program Millenium Development Goals (UNDP-MDG), addresses the joint prediction and specification problem (UNDP 2009-2010, in progress).

CONCLUSION

This chapter illustrates a generalization of Principal Component Analysis using copula families of distributions. This method, Copula Based Component Analysis (CICA), an alternative to non-parametric Independent Component Analysis (ICA) procedures, offers demonstrably more descriptive results on an index of environmental sustainability – the 2002 Environmental Sustainability Index (ESI).

The CICA method accesses non-Gaussian dependence via an information theoretic technique on the special case of linear mixing models, the component analysis family. This approach is a useful *post hoc* procedure for index construction: the goal in indexing is the, hopefully parsimonious, description of a multidimensional concept with a univariate value.

Most useful indexes, perhaps ironically, are designed to measure concepts and quantities that are not predictable, have not yet been measured, and are undefined. Environmental sustainability is certainly of that type; humanitarian and social development goals are as well generally ill defined. The statistician's role in these settings is substantial: it is, perhaps perversely ironic to avoid exact elucidation of statistical assumptions and methodology when they are dictated by the broad context of the desired measurements.

Environmental statisticians have a stake in making the broad concept of sustainability operational: by providing specific measures by which decision makers and the public can judge progress, researchers can justify increased focus.

ACKNOWLEDGEMENTS

Kobi Abayomi thanks Lynne Butler, the Haverford College Mathematics Department, and the Consortium for Faculty Diversity. Kobi Abayomi also thanks Jim Berger, Dalene Stangl, the Statistical and Applied Mathematical Sciences Institute (SAMSI) and Duke University. Much of this chapter was completed and revised in pre and post doctoral fellowships at Haverford and Duke/SAMSI.

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., de la Peña, V., and Lall, U. (2008a) ``Copula Based Independent Component Analysis'' *Working Paper.* Georgia Institute of Technology.

Bernardo, J. (1979) ''Expected Information as Expected Utility'' *Annals of Statistics.* 7, 3, 686-690.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3): 17.

Brunel, N. P., Wojiciech and Derrode, Stephane (2005). ''Copulas in Vectorial Hidden Markov Chains for Multicomponent Image Segmentation''. *ICASSP 2005*.

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

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

Catell, R. B. (1966). *Handbook of Multivariate Experimental Psychology*. Chicago, Rand McNally.

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: 22.

Fuentes, M. C., A Holland, D (2007). "Bayesian entropy for spatial sampling design of environmental data." *Environmental and Ecological Statistics*. June 2006.

Gershunov, A., and T.P. Barnett (1998). "Interdecadal modulation of ENSO teleconnections." *Bulletin of the American Meteorological Society* 79: 12.

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

Joe, Harry (1997) *Multivariate Models and Dependence Concepts.* CRC.

Johnson, R. A., and Wichern, D. W. (1998). *Applied Multivariate Data Analysis*. Upper Saddle River, N.J.: Prentice Hall.

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

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

New York.

Lin, C. G., A. Price, P. Krantz, D. (1999). ”Analysis of local decisions using hierarchical modeling, applied to home random measurement and remediation (with discussion).” *Statistical Science* 14: 33.

Meyers, W. P., Ganapati (2006). *Environmental and ecological statistics*, Volume 2. New York, Springer Science and Business.

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

Osberg L., Sharpe, A. (2002) ‘’An Index of Economic Well-Being for Selected OECD Countries.’’ *Review of Income and Wealth*. 48, 3.

Prescott-Allen, R. (2001). *The Wellbeing of Nations*, Island Press.

Stein, M. L. (2007). ”Seasonal variations in the spatial-temporal dependence of total column ozone.” *Environmetrics* 18: 16.

Tsai, J. C. C., V. C. P. Chen, M. B. Beck, and J. Chen (2004). ”Stochastic Dynamic Programming Formulation for a Wastewater Treatment Decision-Making Framework.” *Annals of Operations Research, Special Issue on Applied Optimization Under Uncertainty* 13: 13.

United Nations Development Program (UNDP) (2005) Human Development Report. Available at: http:// *hdr.undp.org/en/media/HDR06-complete.pdf*

Wherry, Robert J. (1984) Contributions to Correlational Analysis. Orlando, Fl. Academic Press

World Bank-SEDAC Report (2005) “Natural Disaster Hotspots: A Global Risk Analysis” *Technical Report*.

World Economic Forum (2001, 2002). *Environmental Sustainability Index*. Global Leaders for Tomorrow Environment Task Force, World Economic Forum and Yale Center for Environmental Law and Policy and Yale Center for Environmental Law and Policy and Center for International Earth Science Information Network. Davos, Switzerland and New York. Available at: sedac.ciesin.columbia.edu/es/esi/archive.html