Entropy and Dependence Estimation

Barnabás Póczos
Department of Computing Science,
University of Alberta, Canada

Eötvös Loránd University
Neural Information Processing Group

July 2, 2009
Contents

• Entropy estimation
  – Spacing
  – Leonenko and Kozahenko
  – Euclidean Graphs + Combinatorial Optimization

• Kernel Mutual Information

• Copula framework

• Copula methods for ICA
  – Schweitzer Wolf dependence

• Copula methods for MI estimation

• Results
Applications of Mutual Information

- Information theory
- Feature selection
- Clustering
- Image registration
- Independent Component Analysis
- Independent Subspace Analysis
- Optimal experiment design
- Structure learning
- Causality detection
Independent Component Analysis

and

Independent Subspace Analysis
ICA Cost Functions

Let \( y = Wx \), \( y = [y_1; \ldots; y_M] \), and let us measure the dependence using Shannon's mutual information:

\[
J_{ICA_1}(W) = I(y_1, \ldots, y_M) = \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \cdots p(y_M)} dy,
\]

Let \( H(y) = H(y_1, \ldots, y_m) = -\int p(y_1, \ldots, y_m) \log p(y_1, \ldots, y_m) dy \).

\( H(Wx) = H(x) + \log |\det W| \), thus

\[
I(y_1, \ldots, y_M) = \int p(y_1, \ldots, y_M) \log \frac{p(y_1, \ldots, y_M)}{p(y_1) \cdots p(y_M)}
\]
\[
= -H(y_1, \ldots, y_M) + H(y_1) + \ldots + H(y_M)
\]
\[
= -H(x_1, \ldots, x_M) - \log |\det W| + H(y_1) + \ldots + H(y_M).
\]

\( H(x_1, \ldots, x_M) \) is constant, \( \log |\det W| = 0 \), thus

\[
J_{ICA_2}(W) = H(y_1) + \ldots + H(y_M)
\]
ISA Cost Functions

Mutual Information: \[ I(y^1, \ldots, y^m) = \int \log \frac{p(y)}{p(y^1) \cdots p(y^m)} \, dy \]

Shannon-entropy: \[ H(y) = -\int p(y) \log p(y) \, dy \]

Assume \( y = Wx \). Then

\[
\begin{align*}
H(y) &= H(y^1, \ldots, y^m) = H(Wx) = H(x) + \log |W| \\
I(y^1, \ldots, y^m) &= -H(x) - \log |W| + \sum_{i=1}^{m} H(y^i) \\
I(y^1, \ldots, y^m) &= -H(y^1, \ldots, y^m) + \sum_{j=1}^{m} \sum_{i=1}^{d} H(y^j_i) - \sum_{j=1}^{m} I(y^j_1, \ldots, y^j_d) \\
H(y^j) &= H(y^j_1, \ldots, y^j_d) = \sum_{i=1}^{d} H(y^j_i) - I(y^j_1, \ldots, y^j_1)
\end{align*}
\]

and we get the following ISA cost functions:
ISA Cost Functions

\[ J_{ISA_1}(W) \doteq I(y^1, \ldots, y^m) \]

\[ J_{ISA_2}(W) \doteq H(y^1) + \ldots + H(y^m) \]

\[ J_{ISA_3}(W) \doteq \sum_{j=1}^{m} \sum_{i=1}^{d} H(y^j_i) - \sum_{j=1}^{m} I(y^j_1, \ldots, y^j_d) \]

\[ J_{ISA_4}(W) \doteq I(y^1_1, \ldots, y^m_d) - \sum_{j=1}^{m} I(y^j_1, \ldots, y^j_d) \]
Entropy Estimation
1D Spacing Methods

- Vasicek 1976, Test for normality
- Used in RADICAL, E. L. Miller
Multi-dimensional Entropy Estimations, Method of Kozahenko and Leonenko

Let \( \{z(1), \ldots, z(n)\} \) denote \( n \) i.i.d. samples drawn from the distribution of \( z \in \mathbb{R}^d \).

Let \( \mathcal{N}_{1,j} \) be the nearest neighbour of \( z(j) \) in the sample set.

Then the nearest neighbor entropy estimation:

\[
\hat{H}(z) = \frac{1}{n} \sum_{j=1}^{n} \log(n\| \mathcal{N}_{1,j} - z(j) \|) + \ln(2) + C_E,
\]

where \( C_E = -\int_0^\infty e^{-t} \ln(t) dt \) is the Euler-constant.

This estimation is means-square consistent, but not robust. Let us try to use more neighbors!
Multi-dimensional Rényi’s Entropy Estimations

Let us apply Rényi’s-entropy for estimating the Shannon-entropy:

\[ H_\alpha = \frac{1}{1-\alpha} \log \int f^\alpha(z) \, dz \]

\[ \lim_{\alpha \to 1} H_\alpha = -\int f(z) \log f(z) \, dz \]

Let us use
- *K-nearest neighbors*
- *geodesic spanning trees*

for estimating the multi-dimensional Rényi’s entropy.
(It could be much more general...)
Beardwood - Halton - Hammersley
Theorem for kNN graphs

Let \{z(1), \ldots, z(n)\} denote \(n\) i.i.d. samples drawn from the distribution of \(z \in \mathbb{R}^d\).

Let \(\mathcal{N}_{k,j}\) be the \(k\) nearest neighbours of \(z(j)\) in the sample set.

Let \(\gamma = d - d\alpha\), then

\[
\frac{1}{1-\alpha} \log \left( \frac{1}{kn^\alpha} \sum_{j=1}^{n} \sum_{v \in \mathcal{N}_{k,j}} \|v - z(j)\|^{\gamma} \right) \to H_\alpha(z) + c,
\]
as \(n \to \infty\)

Lots of other graphs, e.g. MST, TSP, minimal matching, Steiner graph...etc could be used as well.
Multi-dimensional Entropy Estimations Using Geodesic Spanning Forests

- Build first an *Euclidean neighbourhood graph*:
  - use the edges of the $k$ nearest nodes to each node $z(p)$

- Find *geodesic spanning forests* on this graph:
  - (minimal spanning forests of the Euclidean neighbourhood graph)
Euclidean Graphs

Let \( \{z(1), \ldots, z(n)\} \) denote \( n \) i.i.d. samples drawn from distribution \( z \in \mathbb{R}^d \)

- Euclidean neighbourhood graph

\[
E = \{e : e(p, q) = z(p) - z(q) \in \mathbb{R}^d, z(q) \in N_{k,p}\}
\]

- Weight of minimal (\( \gamma \)-weighted) Euclidean spanning forest:

\[
L_\gamma(z) = \min_{T \in \mathcal{T}} \sum_{e \in T} \|e\|^{\gamma}
\]

Where \( \mathcal{T} \) is the set of all \( \gamma \)-weighted Euclidean spanning forests

Let \( \gamma = d - d\alpha \), then

\[
\frac{d}{\gamma} \log \frac{L_\gamma(z)}{n^\alpha} \to H_\alpha(z) + c, \text{ as } n \to \infty
\]
Estimation of the Shannon-entropy
Examples
(J. A. Costa and A. O. Hero)

Uniform on unit square: $n = 400$ samples

4-NNG on 2D uniform: $\gamma = 1$

MST on 2D uniform: $\gamma = 1$
Dependence Estimation Using Kernel Methods
Kernel covariance (KC)
A. Gretton, R. Herbrich, A. Smola, F. Bach, M. Jordan

Let \( x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y} \) stochastic variables.
We want to measure their dependence.

\[
J_{KC} \doteq \sup_{f \in \mathcal{F}^x, g \in \mathcal{F}^y} \left| E \{ [f(x) - Ef(x)][g(y) - Eg(y)] \} \right|
\]

\[
J_{KC}^{\text{emp}} \doteq \sup_{f \in \mathcal{F}^x, g \in \mathcal{F}^y} \left| \frac{1}{n} \sum_{l=1}^{n} \left\{ [f(x_l) - \frac{1}{n} \sum_{j=1}^{n} f(x_j)][g(y_l) - \frac{1}{n} \sum_{j=1}^{n} g(y_j)] \right\} \right|
\]

where \( x_1, \ldots, x_n, \text{ and } y_1, \ldots, y_n \) \( n \) pieces of i.i.d. samples from \( x, y \) variables,
and \( \mathcal{F}^x, \mathcal{F}^y \) are sets of real valued functions.

The calculation of the supremum over function sets is extremely difficult. Reproducing Kernel Hilbert Spaces make it easier.
RKHS construction for $x, y$ stochastic variables.

Let us choose $K^x(\cdot, \cdot) \in \mathbb{R}^{d_x \times d_x} \rightarrow \mathbb{R}$, $K^y(\cdot, \cdot) \in \mathbb{R}^{d_y \times d_y} \rightarrow \mathbb{R}$ kernel functions.

These kernels define the following Reproducing Kernel Hilbert Spaces:

\[
\mathcal{F}^x \doteq \{ f : f = \sum_{j=1}^{\infty} \psi_j \Phi^x_j(\cdot), \sum_{j=1}^{\infty} \frac{\psi_j^2}{\lambda^x_j} < \infty \},
\]

\[
\mathcal{F}^y \doteq \{ f : f = \sum_{j=1}^{\infty} \psi_j \Phi^y_j(\cdot), \sum_{j=1}^{\infty} \frac{\psi_j^2}{\lambda^y_j} < \infty \},
\]

where $\Phi^x_j(\cdot)$, $\Phi^y_j(\cdot)$, $\lambda^x_j$, $\lambda^y_j$ are eigenfunctions and eigenvalues corresponding to the $K^x(\cdot, \cdot)$, $K^y(\cdot, \cdot)$ Hilbert spaces.

Denote the unit balls by

\[
\mathcal{F}^x_\ast \doteq \{ f : f \in \mathcal{F}^x, \| f \|_{\mathcal{F}^x_\ast} \leq 1 \}
\]

\[
\mathcal{F}^y_\ast \doteq \{ f : f \in \mathcal{F}^y, \| f \|_{\mathcal{F}^y_\ast} \leq 1 \}.
\]
Kernel covariance (KC)

\[ f(x) = \langle f, K^x(\cdot, x) \rangle_{\mathcal{F}^x} \text{ and } f(\cdot) = \sum_{j=1}^{n} c_j K^x(\cdot, x_j) + f_{\perp}(\cdot), \text{ thus} \]

\[ f(x_i) = \langle f, K^x(\cdot, x_i) \rangle_{\mathcal{F}^x} = \langle \sum_{j=1}^{n} c_j K^x(\cdot, x_j) + f_{\perp}(\cdot), K^x(\cdot, x_i) \rangle_{\mathcal{F}^x} = \sum_{j=1}^{n} c_j K^x(x_j, x_i). \]

And what is more, after some calculation we get, that

\[ [f(x_1) - \frac{1}{n} \sum_{i=1}^{n} f(x_i), \ldots, f(x_n) - \frac{1}{n} \sum_{i=1}^{n} f(x_i)] = c^T \tilde{K}^x \]

\[ [g(y_1) - \frac{1}{n} \sum_{i=1}^{n} g(y_i), \ldots, g(y_n) - \frac{1}{n} \sum_{i=1}^{n} g(y_i)] = d^T \tilde{K}^y \]

Where \( K^x = \{K(x_i, x_j)\}_{i,j}, \ H \doteq I_n - \frac{1}{n} 1_n 1_n^T, \) and \( \tilde{K}^x \doteq H K^x H \)

Thus, for the estimation of \( J_{KC}^{emp} \) we have to calculate the maximum of \( c^T \tilde{K}^x \tilde{K}^y d \) over \( c, d \in \mathbb{R}^n \) subject to \( c^T \tilde{K}^x c = 1, \ d^T \tilde{K}^y d = 1. \)
Dependence Estimation Using Copula Methods
COPULA

- Function that couples multivariate distribution function to their marginal functions
- Multivariate distribution function whose marginals are uniform on \([0,1]\)

Copulas are useful for:

- Studying scale free measures of dependence
  - nonparametric independence test
  - financial risk management
- Constructing families of multivariate distributions
- Sampling from multivariate distributions
- Studying Chapman-Kolmogorov equations in Markov processes
Brief History

• **Hoeffding 1940**, studying distributions with \([-1/2,1/2]^2\) support and uniform marginals

“Had Hoeffding chosen the unit square \([0,1]^2\) instead of \([-1/2,1/2]^2\) he would have discovered copulas…” (Schweizer)

• **Frechet 1951**, independently rediscover many of the same results ) Frechet-Hoeffding bounds.

• **Feron 1956**, introduce the word “copula” (grammatical term that links subject and predicate)

• **Sklar 1959**, Sklar’s theorem.

• **Schweizer-Wolf 1981**, Schweizer-Wolf’s \(\sigma\) for measure of dependence.
Copula

Bivariate \textit{copula} $C$ is a multivariate distribution (cdf) defined on a unit square with uniform univariate marginals:

\[
C : [0, 1]^2 \rightarrow [0, 1] \\
C(u, 1) = u, \ C(1, v) = v, \ \forall u, v \in [0, 1] \quad \text{Uniform margins} \\
C(u, 0) = 0, \ C(0, v) = 0, \ \forall u, v \in [0, 1] \quad \text{Grounded}
\]

\[
C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \quad \text{2-increasing} \\
\forall 0 \leq u_1 \leq u_2 \leq 1, \ 0 \leq v_1 \leq v_2 \leq 1
\]
Frechet-Hoeffding bounds

\[ M(u, v) = \min(u, v) \quad W(u, v) = \max(u + v - 1) \quad \Pi(u, v) = uv \]

\[ \max(u + v - 1) = W(u, v) \leq C(u, v) \leq M(u, v) = \min(u, v) \]

Figs. from Nielsen
Sklar’s Theorem

[Sklar, 59]

The copula couples joint distribution to its univariate margins.

\[ F(x, y) = C(F_X(x), F_Y(y)), \quad x, y \in \mathbb{R} \]

\[ C(u, v) = F(F_X^{-1}(u), F_Y^{-1}(v)), \quad u, v \in [0, 1] \]

\[ U = F_X(X), \quad V = F_Y(Y), \quad \Rightarrow (U, V) \sim C(\cdot, \cdot) \]
Independence Copula ($\Pi$)

\[ X \perp Y \iff F(X, Y) = F(X)F(Y) \]

\[ \prod(u, v) = F\left(F_X^{-1}(u), F_Y^{-1}(v)\right) = F_X\left(F_X^{-1}(u)\right)F_Y\left(F_Y^{-1}(v)\right) = uv \]

\[ X \perp Y \iff C(U, V) = \prod(U, V) \]
Copula properties

• Invariance under strictly monotone transformations
  \[
  \alpha \uparrow, \beta \uparrow \Rightarrow C_{\alpha(X)\beta(Y)}(u, v) = C_{XY}(u, v)
  \]
  \[
  \alpha \uparrow, \beta \downarrow \Rightarrow C_{\alpha(X)\beta(Y)}(u, v) = u - C_{XY}(u, 1 - v)
  \]
  \[
  \alpha \downarrow, \beta \uparrow \Rightarrow C_{\alpha(X)\beta(Y)}(u, v) = v - C_{XY}(1 - u, v)
  \]
  \[
  \alpha \downarrow, \beta \downarrow \Rightarrow C_{\alpha(X)\beta(Y)}(u, v) = u + v - 1 + C_{XY}(1 - u, 1 - v)
  \]
  Important when we need dependence measure that is invariant under strictly monotone rescaling the variants.

• Order statistics can be expressed with copulas
  \[
P(\max(X, Y) \leq t) = C(F_X(t), F_Y(t))
  \]
  \[
P(\min(X, Y) \leq t) = F_X(t) + F_Y(t) - C(F_X(t), F_Y(t))
  \]
Basic Properties

• Frechet-Hoeffding bounds

\[ \max(u + v - 1) = W(u, v) \leq C(u, v) \leq M(u, v) = \min(u, v) \]

• Lipschitz continuity

\[ |C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1| \]

• Convex linear combination of copulas

\[ \theta \in [0, 1] \Rightarrow (1 - \theta)C_1 + \theta C_2 \text{ is copula function.} \]
Concordance

X, Y variables are **concordant** if large (small) values are tend to be associated with large (small) values of the other.

Let \((x_1, \ldots, x_N), (y_1, \ldots, y_N)\) be samples from X and Y.

We say that \((x_i, y_i)\) and \((x_j, y_j)\) are

- **concordant** if \((x_i - x_j)(y_i - y_j) > 0\)
- **discordant** if \((x_i - x_j)(y_i - y_j) < 0\)
Measure of Concordance

**Definition (Concordance)**

κ association between X,Y continuous variables with copula C, s.t.

- $-1 \leq \kappa_{X,Y} \leq 1; \kappa_{X,X} = 1; \kappa_{X,-X} = -1$
- $\kappa_{X,Y} = \kappa_{Y,X}$
- $X, Y$ independent $\Rightarrow \kappa_{X,Y} = \kappa_{\Pi} = 0$
- $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$
- $C_1 \preceq C_2 \Rightarrow \kappa_{C_1} \leq \kappa_{C_2}$
- $C_N \to C$ pointwise $\Rightarrow \lim_{N \to \infty} \kappa_{C_N} = \kappa_C$

**Theorem:** Y is increasing function of X $\Rightarrow \kappa_{X,Y} = \kappa_M = 1$

Y is decreasing $\Rightarrow \kappa_{X,Y} = \kappa_W = -1$

$\alpha, \beta$ monotone increasing $\Rightarrow \kappa_{\alpha(X),\beta(Y)} = \kappa_{X,Y}$

**Theorem:** Kendall’s t, Spearman’s $\rho$, Gini’s $\gamma$, Blomquist’s $\beta$ are measure of Concordance
Kendall’s $\tau$

c = #concordant pairs, d = #discordant pairs

Sample version

$$t = \frac{c-d}{c+d} = \frac{c-d}{N(N-1)/2} = P_{emp}(concordance) - P_{emp}(discordance)$$

Population version

Let $(X_1, Y_1) \sim F, (X_2, Y_2) \sim F$, where $F(x, y) = C(F_X(x), F_Y(y))$

$$\tau_{XY} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Theorem, Kendall’s $\tau$ with copulas:

$$\tau = 4 \iint_{I^2} C(u, v) \, dC(u, v) - 1$$

$$= 4E[C(U, V)] - 1, \text{ where } (U, V) \sim C$$

$$= 1 - 4 \iint_{I^2} \frac{\partial C(u, v)}{\partial u} \frac{\partial C(u, v)}{\partial v} \, dudv$$
Spearman’s $\rho$

**Sample version:**

Pearson’s linear correlation between the ranks.

**Population version**

Let us have 3 2-dimensional i.i.d. variables:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3) \sim F(x, y) = C(F_X(x), F_Y(y))$$

$$\rho = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0])$$

**Theorem, Spearman’s $\rho$ with copulas:**

$$\rho = 12 \iint_{I^2} uv dC(u, v) - 3$$

$$= \frac{E[UV] - E[U]E[V]}{\sqrt{\text{var}(U)} \sqrt{\text{var}(V)}}, \text{ where } (U, V) \sim C$$

$$= 12 \iint_{I^2} (C(u, v) - uv) dudv$$

33
Measure of Dependence

Problem with concordance:

\[ X \perp Y \implies \kappa = 0 \]

\( \delta \) association between \( X, Y \) continuous variables with copula \( C \) is a measure of dependence if

\[ \begin{align*}
\ast & \quad \delta_{X,Y} = \delta_{Y,X} \\
\ast & \quad 0 \leq \delta_{X,Y} \leq 1 \\
\ast & \quad \delta_{X,Y} = 0 \iff X, Y \text{ independent} \\
\ast & \quad \delta_{X,Y} = 1 \iff \text{strictly monotone function of each other.} \\
\ast & \quad \alpha, \beta \text{ strictly monotone function} \implies \delta_{\alpha(X), \beta(Y)} = \delta_{X,Y} \\
\ast & \quad C_N \to C \text{ pointwise} \implies \lim_{N \to \infty} \delta_{C_N} = \delta_C
\end{align*} \]
Schweizer-Wolff Measures of Dependence

$L_p$-norm between the copula for the distribution and the independence copula.

\[(L_p \text{ norm}) \delta = \left( \frac{\Gamma(2p+3)}{2\Gamma^2(p+1)} \iiint \mathbb{I}_2 |C(u,v) - \Pi(u,v)|^p \, dudv \right)^{1/p} \]

**Special cases:**

\[(L_1 \text{ norm}) \sigma = 12 \iiint \mathbb{I}_2 |C(u,v) - \Pi(u,v)| \, dudv \]
\[\rho = 12 \iiint (C(u,v) - \Pi(u,v)) \, dudv \]

\[(L_2 \text{ norm}) \phi = \left( 90 \iiint \mathbb{I}_2 |C(u,v) - \Pi(u,v)|^2 \, dudv \right)^{1/2} \]

\[(L_\infty \text{ norm}) \lambda = 4 \sup_{\mathbb{I}_2} |C(u,v) - \Pi(u,v)| \]

**Multivariate L_1 version:**

\[\sigma_M = \frac{2^M(M+1)}{2^M-(M+1)} \int_{\mathbb{I}^M} |C(u) - \prod_{i=1}^M u_i| \, du \]
From ranks to copula

Ranks and copulas

- Invariant under monotonic transformations
- Not very sensitive to outliers (e.g. empirical cov is very sensitive)
- The Copula is the CDF of the population distribution over ranks

\[ c_N \left( \frac{i}{N}, \frac{j}{N} \right) = \begin{cases} 1/N & \text{if } \exists (x(i), y(j)) \\ 0 & \text{otherwise} \end{cases} \]
Empirical Copula Distribution

[Deheuvels, 79]

\[ z_1^1 < z_2^1 < \cdots < z_N^1 = \text{sorted values of } x_1, x_2, \ldots, x_N \]

\[ z_1^2 < z_2^2 < \cdots < z_N^2 = \text{sorted values of } y_1, y_2, \ldots, y_N \]

\[
C_N \left( \frac{i}{N}, \frac{j}{N} \right) = \frac{\# \text{ of } (x_k, y_k) \text{ s.t. } x_k \leq z_1^i \text{ and } y_k \leq z_2^j}{N}
\]
Using the Empirical Copula

Useful for computing sample versions of functions on copulas

\[
\rho = 12 \int_0^1 \int_0^1 (C(u, v) - \Pi(u, v)) \, du \, dv
\]
\[
r = \frac{12}{N^2 - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( C_N \left( \frac{i}{N}, \frac{j}{N} \right) - \frac{i}{N} \times \frac{j}{N} \right)
\]

\[
\sigma = 12 \int_0^1 \int_0^1 |C(u, v) - \Pi(u, v)| \, du \, dv
\]
\[
s = \frac{12}{N^2 - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| C_N \left( \frac{i}{N}, \frac{j}{N} \right) - \frac{i}{N} \times \frac{j}{N} \right|
\]

\[
\tau = 2 \int_0^1 \int_0^1 \int_0^1 \int_0^1 \left[ c(u, v)c(u', v') - c(u, v')c(u', v) \right] \, du \, dv \, du' \, dv'
\]

\[
t = \frac{2N}{N - 1} \sum_{i=2}^{N} \sum_{j=2}^{N} \sum_{p=1}^{i-1} \sum_{q=1}^{j-1} \left[ c_N \left( \frac{i}{N}, \frac{j}{N} \right)c_N \left( \frac{p}{N}, \frac{q}{N} \right) - c_N \left( \frac{i}{N}, \frac{q}{N} \right)c_N \left( \frac{p}{N}, \frac{j}{N} \right) \right]
\]
Contribution: New ICA Contrast

- Using a measure of dependence based on copulas, joint distributions over ranks.

**Properties**
- Does not require density estimation
- Non-parametric

**Advantages**
- Very robust to outliers
- Frequently performs as well as state of the art algorithms (and sometimes better)
- Contrast can be used with ISA (... as opposed to Radical)
  - Can be used to estimate dependence
- Easy to implement (code publicly available)

**Disadvantages**
- Somewhat slow (although not prohibitively so)
- Needs more samples to demix near-Gaussian sources
**Schweizer-Wolff ICA (SWICA)**

**Inputs:** $X$, a $2 \times N$ matrix of signals, $K$, number of evaluation angles

For $\theta=0, \pi/(2K), \ldots, (K-1)\pi/(2K)$

- Compute rotation matrix
  $$W(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

- Compute rotated signals $Y(\theta) = W(\theta)X$.
- Compute $s(Y(\theta))$, a sample estimate of $\sigma$
  - Sort
  - Compute $s$

Find best angle $\theta_m = \arg\min_{\theta} s(Y(\theta))$

**Output:** Rotation matrix $W=W(\theta_m)$, demixed signal $Y=Y(\theta_m)$, and estimated dependence measure $s=s(Y(\theta_m))$
Amari Error

- Measures how close a square matrix is to a permutation matrix

$$B = WA$$

demixing  mixing

$$r(B) = \frac{1}{2M(M-1)} \sum_{i=1}^{M} \left( \frac{\sum_{j=1}^{M} |b_{ij}|}{\max_{i} |b_{ij}|} - 1 \right) + \frac{1}{2M(M-1)} \sum_{j=1}^{M} \left( \frac{\sum_{i=1}^{M} |b_{ij}|}{\max_{i} |b_{ij}|} - 1 \right)$$

$$r(B) \in [0, 1], \quad r(B) = 0 \Leftrightarrow B \text{ is a permutation matrix}$$
Synthetic Marginal Distributions

[Bach and Jordan 02]
ICA Comparison

(M=2, N=1000, 1000 repetitions)
bold $\in \{\text{best, best+0.2}\}$

<table>
<thead>
<tr>
<th>pdf</th>
<th>SWICA</th>
<th>FastICA</th>
<th>RADICAL</th>
<th>KernelICA</th>
<th>JADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.50</td>
<td>4.43</td>
<td>2.79</td>
<td>3.04</td>
<td>3.82</td>
</tr>
<tr>
<td>b</td>
<td>2.98</td>
<td>6.25</td>
<td>3.18</td>
<td>3.12</td>
<td>4.65</td>
</tr>
<tr>
<td>c</td>
<td>1.71</td>
<td>2.37</td>
<td>1.74</td>
<td>1.52</td>
<td>1.71</td>
</tr>
<tr>
<td>d</td>
<td>9.23</td>
<td>7.14</td>
<td>5.18</td>
<td>5.90</td>
<td>5.71</td>
</tr>
<tr>
<td>e</td>
<td>1.47</td>
<td>5.55</td>
<td>1.57</td>
<td>1.38</td>
<td>4.11</td>
</tr>
<tr>
<td>f</td>
<td>1.51</td>
<td>4.04</td>
<td>1.63</td>
<td>1.49</td>
<td>2.77</td>
</tr>
<tr>
<td>g</td>
<td>1.49</td>
<td>1.81</td>
<td>1.46</td>
<td>1.38</td>
<td>1.43</td>
</tr>
<tr>
<td>h</td>
<td>4.14</td>
<td>5.82</td>
<td>4.61</td>
<td>4.52</td>
<td>4.07</td>
</tr>
<tr>
<td>i</td>
<td>10.53</td>
<td>9.40</td>
<td>9.46</td>
<td>11.16</td>
<td>6.49</td>
</tr>
<tr>
<td>j</td>
<td>1.57</td>
<td>6.40</td>
<td>1.49</td>
<td>1.39</td>
<td>4.79</td>
</tr>
<tr>
<td>k</td>
<td>2.94</td>
<td>6.28</td>
<td>3.08</td>
<td>2.74</td>
<td>4.38</td>
</tr>
<tr>
<td>l</td>
<td>4.54</td>
<td>10.20</td>
<td>5.06</td>
<td>4.99</td>
<td>7.22</td>
</tr>
<tr>
<td>m</td>
<td>1.94</td>
<td>3.95</td>
<td>1.59</td>
<td>1.43</td>
<td>2.78</td>
</tr>
<tr>
<td>n</td>
<td>3.40</td>
<td>5.45</td>
<td>2.30</td>
<td>1.93</td>
<td>3.83</td>
</tr>
<tr>
<td>o</td>
<td>4.42</td>
<td>4.47</td>
<td>3.57</td>
<td>3.24</td>
<td>3.12</td>
</tr>
<tr>
<td>p</td>
<td>2.02</td>
<td>3.83</td>
<td>1.76</td>
<td>1.52</td>
<td>2.73</td>
</tr>
<tr>
<td>q</td>
<td>3.92</td>
<td>16.93</td>
<td>2.65</td>
<td>2.18</td>
<td>13.12</td>
</tr>
<tr>
<td>r</td>
<td>4.10</td>
<td>5.95</td>
<td>3.69</td>
<td>3.33</td>
<td>4.10</td>
</tr>
<tr>
<td>rand</td>
<td>2.84</td>
<td>5.88</td>
<td>2.73</td>
<td>2.52</td>
<td>4.68</td>
</tr>
</tbody>
</table>
Corrupted Music
Outliers: Multidimensional Comparison

M=4, N=2000, 1000 repetitions

M=8, N=5000, 100 repetitions
Unmixing Image Sources with Outliers (3% outliers)

Original  Mixed  SWICA  FastICA

Independent Subspace Analysis (ISA, The Woodstock Problem)

Sources

\[ s^1 \in \mathbb{R}^D \]
\[ s^2 \in \mathbb{R}^D \]
\[ s^M \in \mathbb{R}^D \]

Observation

\[ x^1 \in \mathbb{R}^D \]
\[ x^2 \in \mathbb{R}^D \]
\[ x^M \in \mathbb{R}^D \]

Estimation

\[ y^1 \in \mathbb{R}^D \]
\[ y^2 \in \mathbb{R}^D \]
\[ y^M \in \mathbb{R}^D \]

\[ A \in \mathbb{R}^{DM \times DM} \]
\[ W \in \mathbb{R}^{DM \times DM} \]

Find \( W \), recover \( Wx \)
Independent Subspace Analysis
Summary

• New contrast based on a measure of dependence for distribution over ranks
  – Robust to outliers
  – Comparable performance to state of the art algorithms
  – Can handle a moderate number of sources (M=20)
  – Can be used to solve ISA

• Future work
  – Further acceleration of SWICA
  – What types of sources does SWICA do well on and why?
Copula Methods for Mutual Information Estimation
Non-parametric MI estimation using copulas

Mutual information is the negentropy of the copula:

\[ I(X_1, \ldots, X_M) = \int f(x_1, \ldots, x_M) \log \frac{f(x_1, \ldots, x_M)}{f(x_1) \cdots f(x_M)} \, dx_1 \cdots dx_M \]

\[ = \sum_{i=1}^{M} H(X_i) - (X_1, \ldots, X_M) \]

\[ = \int c(x_1, \ldots, x_M) \log c(x_1, \ldots, x_M) \, dx_1 \cdots dx_M \]

Rényi’s information is the negative Rényi’s entropy of the copula:

\[ I_\alpha(X_1, \ldots, X_M) = \frac{1}{1 - \alpha} \log \int \left( \frac{\prod_{i=1}^{M} f(x_i)}{f(x_1, \ldots, x_M)} \right)^\alpha f(x_1, \ldots, x_M) \, dx_1 \cdots dx_M \]

\[ H_\alpha(X_1, \ldots, X_M) = \frac{1}{1 - \alpha} \int f^\alpha(x_1, \ldots, x_M) \, dx_1 \cdots dx_M \]
Algorithms

Algorithm for (Rényi’s) MI estimation:

- Estimate the copula using empirical copula
- Estimate its (Rényi’s) negentropy

Other methods:

2. k-Nearest Neighbours for entropy estimation
3. Kernel Density Estimation
4. Histogram (bin) based estimations
Consistency in 2D

(b) Rotated uniform

(a) Independent beta
Consistency in 2D

(c) Gauss

(d) Gauss copula (with Gamma marginals)
Rotated 2D sources

(a) Gauss

(b) Gauss copula

(c) t-copula

(d) Gauss + outliers

(e) Gauss copula + outliers

(f) t-copula + outliers
Rotated 2D uniform distribution

\[ I_{X_1, X_2} = \frac{[a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + 1]}{-2 \cos(\alpha) + \sin(\beta) \sin(\alpha) \sqrt{2}}, \]

\[ \beta = \pi/4 + \alpha \]
\[ a_1 = -\cos(\alpha) \sin(\alpha) \log(2) \]
\[ a_2 = -2 \cos(\alpha) \sin(\alpha) \log\left(\frac{\sin(\alpha) - \cos(\alpha)}{-2 \cos(\alpha) + \sin(\beta) \sin(\alpha) \sqrt{2}}\right) \]
\[ a_3 = \log\left(\frac{\sin(\alpha) - \cos(\alpha)}{-2 \cos(\alpha) + \sin(\beta) \sin(\alpha) \sqrt{2}}\right) \]
\[ a_4 = -2 \sin(\alpha) \cos(\alpha) \sqrt{2} \log(\sin(\beta))/\sin(\beta) \]
\[ a_5 = \cos(\alpha) \sin(\alpha) \]
\[ a_6 = \log(2)/2 \]
\[ a_7 = 2 \log(\sin(\beta)) \]
\[ a_8 = 2 \log(\sin(\beta)) \cos(\alpha)^2 \]
Rotated 2D uniform distribution in the presence of outliers
4D sources

(a) 4D Gaussians without outliers

(b) 4D Gaussians + outliers
Image Registration

(b) Without outliers

(c) With outliers
ISA

(a) Original
(b) Mixed

(c) Estimated
(d) Hinton
ISA in the Presence of Outliers

![Box plots comparing Amari error for Copula, kNN, JFD, KCCA, and KGV methods.](image-url)
Conclusion

Use Copulas,
They can be useful...

*Thanks for the Attention!*