ICA and ISA Using Schweizer-Wolff Measure of Dependence

Sergey Kirshner and Barnabás Póczos
Alberta Ingenuity Centre for Machine Learning,
Department of Computing Science,
University of Alberta, Canada

ICML 2008
Poster Session II (Paper #551)

July 8, 2008
Independent Component Analysis (ICA, The Cocktail Party Problem)

Sources

Mixing

Observation

Estimation

1

2

3

d

\( s \)

\( A \in \mathbb{R}^{d \times d} \)

\( x = As \)

\( y = Wx \)
Independent Subspace Analysis (ISA, The Woodstock Problem)

Sources

\( s^1 \in \mathbb{R}^m \)

\( s^2 \in \mathbb{R}^m \)

\( s^d \in \mathbb{R}^m \)

Observation

\( x^1 \in \mathbb{R}^m \)

\( x^2 \in \mathbb{R}^m \)

\( x^d \in \mathbb{R}^m \)

Estimation

\( y^1 \in \mathbb{R}^m \)

\( y^2 \in \mathbb{R}^m \)

\( y^d \in \mathbb{R}^m \)

\( A \in \mathbb{R}^{dm \times dm} \)

\( W \in \mathbb{R}^{dm \times dm} \)

Find \( W \), recover \( Wx \)
Solving ICA

• Sources can be recovered only up to scale and permutation

• ICA solution:
  – Removing mean, $E[y]=0$
  – Whitening, $E[yy^T]=I$
  – Finding a multidimensional rotation optimizing an objective function
    • Sequence of 2-d Jacobi (Givens) rotations
Objective Function/Contrast

• Need to measure independence (ICA)
  – Mutual information (MI) estimation
    • Kernel-ICA [Bach & Jordan, 2002]
  – Entropy estimation
    • RADICAL [Learned-Miller & Fisher, 2003]
    • FastICA [Hyvarinen, 1999]
  – ML estimation
    • KDICA [Chen, 2006]
  – Moment-based methods
    • JADE [Cardoso, 1993]
  – Correlation-based methods
    • [Jutten and Herault, 1991]

• Need to measure dependence (ISA)
  – Grouping signals recovered by ICA
    • Cardoso’s conjecture [Cardoso, 1998]
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
    • 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
Ranks

- Invariant under monotonic transformations
  - Implied non-linearity
- Not very sensitive to outliers
Rank Correlations

- Spearman’s $\rho = \text{linear correlation of the ranks}$
- Kendall’s $\tau = (\text{number of concordant pairs} \ \text{minus}\ \text{number of discordant pairs}) \ \text{divided by twice the number of all pairs}$
- Range $[-1,1]$
- Independence implies $\rho=0$ and $\tau=0$
Are Rank Correlations Useful for ICA?

Inaccurate and noisy!
Probability Integral Transform
Distribution Over Ranks
Copula

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

$$C : [0, 1]^2 \rightarrow [0, 1]$$

$$C_U (u) = C (u, 1) = u, \quad C_V (v) = C (1, v) = v, \quad \forall u, v \in [0, 1]$$

$$U = F_X (X), \quad V = F_Y (Y)$$

$$X = F_X^{-1} (U), \quad Y = F_Y^{-1} (V)$$

$$F (x, y) = C (F_X (x), F_Y (t)), \quad x, y \in \mathbb{R}^d$$

$$C (u, v) = F (F_X^{-1} (u), F_Y^{-1} (v))$$
Sklar’s Theorem

[Sklar, 59]

\[ \text{Sklar’s Theorem} \]
Useful Properties of Copulas

• Preserve concordance between the variables

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

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

• Preserve mutual information

\[ I(X; Y) = \iint_{I^2} c(u, v) \log c(u, v) \, dudv = -H(U, V) \]

• Can be viewed as a canonical form of a multivariate distribution for the purpose of the estimation of multivariate dependence
Independence Copula

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

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

\[ X \independent Y \iff C(U, V) = \Pi(U, V) \]
Schweizer-Wolff Measures of Dependence

[Schweizer and Wolff, 81]

- $L$-norm between the copula for the distribution and the independence copula
  - Range is $[0,1]$
  - 0 if and only if variables are independent

- Examples
  - $\sigma$ ($L_1$-norm)
    \[
    \sigma = 12 \iint_{I^2} |C(u,v) - \Pi(u,v)| \, du \, dv
    \]
  - $\rho$ ($L_\infty$-norm)
    \[
    \rho = 12 \iint_{I^2} (C(u,v) - \Pi(u,v)) \, du \, dv
    \]
  - $\kappa$ ($L_\infty$-norm)
    \[
    \kappa = 4 \sup_{I^2} |C(u,v) - \Pi(u,v)|
    \]
Empirical Copula

[Deheuvels, 79]

\[ z_1^1 < z_2^1 < \ldots < Z_1^N = \text{sorted values of } x_1, x_2, \ldots, x_N \]
\[ z_1^2 < z_2^2 < \ldots < Z_2^N = \text{sorted values of } y_1, y_2, \ldots, y_N \]

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

• Useful for computing sample versions of functions on copulas

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

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

\[
s \approx \frac{12}{B^2 - 1} \sum_{i=1}^{B} \sum_{j=1}^{B} \left| C_N \left( \frac{i}{B}, \frac{j}{B} \right) - \frac{i}{B} \times \frac{j}{B} \right|
\]
Is Schweizer-Wolff $\sigma$ Useful for ICA?
Schweizer-Wolff ICA (SWICA)

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

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

- Compute rotation matrix
  \[ W(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \]
  \[ O(1) \]

- Compute rotated signals $Y(\theta) = W(\theta)X$.
  \[ O(N) \]

- Compute $s(Y(\theta))$, a sample estimate of $\sigma$
  
  - Sort into $B$ bins
  
  - Compute $B$-bin approximation to $s$
    \[ O(N\log B) \]
    \[ O(NB^2) \]

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

\[ O(K) \]

\[ O(KN\log KB) \]

**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

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

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

[Bach and Jordan 02]
ICA Comparison (d=2, N=1000, 1000 repetitions)
Outliers
(d=2, N=1000, 10000 repetitions)

![Graph showing the comparison of Amari error for different ICA methods with varying number of post-whitening outliers.](image-url)
Outliers: Multidimensional Comparison

$d=4, N=2000, 1000$ repetitions

$d=8, N=5000, 100$ repetitions
Unmixing Image Sources with Outliers

Original  Mixed  SWICA  FastICA
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 (d=20)
  – Can be used to solve ISA

• Future work
  – Further acceleration of SWICA
  – What types of sources does SWICA do well on and why?
  – Non-parametric estimation of mutual information
Software

http://www.cs.ualberta.ca/~sergey/SWICA