ORIE 779: Functional Data Analysis

From last meeting

Finished SiZer Background

Started Independent Component Analysis

Independent Component Analysis

Idea: Find "directions that maximize independence"

Motivating Context: Signal Processing

"Blind Source Separation"

References:

Lee, T. W. (1998) *Independent Component Analysis: Theory and Applications*, Kluwer.

Hyvärinen and Oja (1999) *Independent Component Analysis: A Tutorial*, http://www.cis.hut.fi/projects/ica

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*, John Wiley & Sons.

ICA, motivating example

"Cocktail party problem":

- hear several simultaneous conversations

- would like to "separate them"

Model for "conversations": time series:

$$s_1(t) \quad \text{and} \quad s_2(t)$$

Toy Example

ICA, motivating example (cont.)

Mixed version of signals:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$

And also a second mixture (e.g. from a different location):

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$$

Mixed version of above toy example

ICA, motivating example (cont.)

Goal: Recover "signal" $\underset{\rightarrow}{s}(t) = \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}$ from "data" $\underset{\rightarrow}{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$

for *unknown* "mixture matrix" $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, where

$$\underset{\rightarrow}{x} = A\underset{\rightarrow}{s}, \quad \text{for all } t$$

i.e. find "separating weights", W, so that

$$\underset{\rightarrow}{s} = W\underset{\rightarrow}{x}, \quad \text{for all } t$$

Problem: $W = A^{-1}$ would be fine, but A is *unknown*

ICA, motivating example (cont.)

Relation to FDA: recall "data matrix"

$$X = \left(\underset{\rightarrow}{X}_1 \quad \cdots \quad \underset{\rightarrow}{X}_n \right) = \begin{pmatrix} X_{11} & & X_{1n} \\ \vdots & \cdots & \vdots \\ X_{d1} & & X_{dn} \end{pmatrix}$$

Signal Processing: focus on rows (d time series, for $t = 1,...,n$)

Functional Data Analysis: focus on columns (n data vectors)

Note: same 2 different viewpoints as "dual problems" in PCA

ICA, motivating example (cont.)

Scatterplot View (signal processing): plot

- signals & scatterplot $\{(s_1(t), s_2(t)) : t = 1, \ldots, n\}$
- data & scatterplot $\{(x_1(t), x_2(t)) : t = 1, \ldots, n\}$
- scatterplots give hint how ICA works
- affine trans. $\underset{\rightarrow}{x} = A\underset{\rightarrow}{s}$ “stretches indep. signals into dep.”
- “inversion” is key to ICA (even when A is unknown)

ICA, motivating example (cont.)

Scatterplot view of: Why not PCA?

- finds “direction of greatest variability” [PCA - scatterplot]

- which is wrong direction for “signal separation”

[PCA decomposition]

ICA, Algorithm

ICA Step 1:

- "sphere the data" (shown on right in scatterplot view)
- i.e. find linear transf'n to make mean = $\underset{\rightarrow}{0}$, cov = I
- i.e. work with $Z = \hat{\Sigma}^{-1/2}\left(X - \hat{\underset{\sim}{\mu}}\right)$
- requires X of full rank (at least $n \geq d$, i.e. no HDLSS)
  (is this critical????)
- search for "indep." *beyond* linear and quadratic structure

ICA, Algorithm (cont.)

ICA Step 2:

- Find dir'ns that make (sph'd) data as "indep. as possible"

Recall "independence" means:

joint distribution is product of marginals

In cocktail party example [scatterplot]:

Happens only when "support parallel to axes"

Otherwise have "blank areas", but marginals are non-zero

ICA, Algorithm (cont.)

Parallel Idea (and key to algorithm):

Find directions that maximize "non-Gaussianity"

Reason: starting from independent coordinates

"most projections are Gaussian"

(since projection is "linear combo")

Mathematics behind this:

Diaconis and Freedman (1984*) Annals of Statistics*, 12, 793-815.

ICA, Algorithm (cont.)

Worst case for ICA:

- Gaussian
- Then sphered data are independent
- So have “independence” in *all directions*
- Thus can’t find useful directions
- Gaussian distribution is characterized by:

  Independent & spherically symmetric

ICA, Algorithm (cont.)

Criteria for non-Gaussianity / independence:

- kurtosis ($EX^4 - 3(EX^2)^2$, 4th order cumulant)
- negative entropy
- mutual information
- nonparametric maximum likelihood
- "infomax" in neural networks
- $\exists$ interesting connections between these

ICA, Algorithm (cont.)

Matlab Algorithm (optimizing any of above):   "FastICA"

- numerical gradient search method
- can find directions "iteratively"
- or by "simultaneous optimization"
- appears fast, with good defaults
- should we worry about local optima???

Again view raw data, mixed version, ICA decomp.

ICA, Algorithm (cont.)

Notational summary:

1. First sphere data: $Z = \hat{\Sigma}^{-1/2}\left(X - \hat{\mu}\right)$

2. Apply ICA: find W_S to make rows of $S_S = W_S Z$ "indep't"

3. Can transform back to "original data scale": $S = \hat{\Sigma}^{1/2} S_S$

ICA, Algorithm (cont.)

Identifiability problem 1: Generally can't order rows of S_S (& S)

Since for a "permutation matrix" P

(pre-multiplication by P "swaps rows")

(post-multiplication by P "swaps columns")

for each column, $\underset{\rightarrow}{z} = A_S \underset{\rightarrow}{s}_S = A_S P^{-1} P \underset{\rightarrow}{s}_S$ i.e. $P \underset{\rightarrow}{s}_S = P W_S \underset{\rightarrow}{z}$

So PS_S and PW_S are also solutions (i.e. $PS_S = PW_S Z$)

(saw this in "switched order" in Cocktail Party raw, recon'd)

FastICA: appears to order in terms of "how non-Gaussian"

ICA, Algorithm (cont.)

Identifiability problem 2: Can't find scale of elements of $\underrightarrow{s}$

Since for a (full rank) diagonal matrix D

(pre-multiplication by D is scalar mult'n of rows)

(post-multiplication by D is scalar mult'n of columns)

for each col'n, $\underrightarrow{z} = A_S \underrightarrow{s}_S = A_S D^{-1} D \underrightarrow{s}_S$ i.e. $D \underrightarrow{s}_S = D W_S \underrightarrow{z}$

So DS_S and DW_S are also solutions

(also saw this in "inversion" in Cocktail Party raw, recon'd)

ICA, Algorithm (cont.)

Signal Processing Scale identification: (Hyvärinen and Oja)

Choose scale so each signal $s_i(t)$ has “unit average energy”:

$$\sum_t s_i(t)^2$$

(preserves energy along rows of data matrix)

Explains “same scales” in Cocktail Party Example

Again view raw data, ICA decomp.

ICA and non-Gaussianity

For indep., non-Gaussian, stand'zed, r.v.'s: $\underset{\rightarrow}{x} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$,

projections "farther from coordinate axes" are "more Gaussian":

For the dir'n vector $\underset{\rightarrow}{u}_k = \begin{pmatrix} u_{1,k} \\ \vdots \\ u_{d,k} \end{pmatrix}$, where $u_{i,k} = \begin{cases} 1/\sqrt{k} & i = 1,\ldots,k \\ 0 & i = k+1,\ldots,d \end{cases}$

(thus $\left\|\underset{\rightarrow}{u}\right\| = 1$), have $\underset{\rightarrow}{x}^t \underset{\rightarrow}{u} \overset{d}{\approx} N(0,1)$, for large d and k

ICA and non-Gaussianity (cont.)

Illustrative examples:

Assess normality with Q–Q plot,

scatterplot of "data quantiles" vs. "theoretical quantiles"

connect the dots of $\{(q_i, X_{(i)}) : i = 1, \ldots, n\}$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ and $\frac{i - 1/2}{n} = P\{X \leq q_i\}$

ICA and non-Gaussianity (cont.)

Q-Q Plot ("Quantile – Quantile", can also do "Prob. – Prob."):

Assess variability with overlay of simulated data curves [toy e.g.]

E.g. Weibull(1,1) (= Exponential(1)) data ($n = 500$)

- Gaussian dist'n is poor fit (Q-Q curve outside envelope)
- Pareto dist'n is good fit (Q-Q curve inside envelope)
- Weibull dist'n is good fit (Q-Q curve inside envelope)
- Bottom plots are corresponding log scale versions

ICA and non-Gaussianity (cont.)

Illustrative examples ($d = 100 \quad n = 500$):

a. Uniform marginals [graphic]

- $k = 1$ very poor fit (Uniform "far from" Gaussian)
- $k = 2$ much closer? (Triangular closer to Gaussian)
- $k = 4$ very close, but still have stat'ly sig't difference
- $k \geq 6$ all differences could be sampling variation

ICA and non-Gaussianity (cont.)

Illustrative examples ($d = 100 \quad n = 500$):

b. Exponential marginals [graphic]

- still have convergence to Gaussian, but slower

  ("skewness" has stronger impact than "kurtosis")

- now need $n \geq 25$ to see no difference

c. Bimodal marginals [graphic]

- Similar lessons to above

ICA and non-Gaussianity (cont.)

Summary:

For indep., non-Gaussian, stand'zed, r.v.'s: $\underset{\rightarrow}{x} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$,

projections "farther from coordinate axes" are "more Gaussian"

Conclusions:

i. Usually expect "most projections are Gaussian"

ii. Non-Gaussian projections (target of ICA) are "special"

iii. Are most samples really "random"??? (could test???)

iv. High dimensional space is a strange place

ICA Toy Examples

E.g. Two sine waves [combined graphic]

- Scatterplots show “time series structure”(not “random”)
- Since have exactly doubled the frequency
- PCA finds wrong direction
- Sphering is enough to solve this (“orthogonal to PCA”)
- So ICA is good (note: “flip”, and “constant signal power”)
- ICA works even without “honest joint distribution”

ICA, Toy Examples (cont.)

E.g. Sine wave and Gaussian noise [combined graphic]

- PCA finds “diagonal of parallelogram”
- Sine is all in one (since “greatest variability” in that dir’n)
- but still “wiggles” (noise adds to “greatest variation”)
- ICA gets it right
- but magnifies the noise

ICA, Toy Examples (cont.)

E.g. Two realizations of Gaussian noise [combined graphic]

- PCA finds "axis of ellipse" (happens to be "right")
- Note even "realization" of noise is right
- Since that drives PC directions
- ICA is "wrong" (different noise realization)