From Last Meetings

Studying Covariance vs. Correlation PCA:

From Toy examples:

- It can make a big difference

- Not clear which is “better”

- Issues understood via:
  “how point cloud relates to coordinate axes”

Explore Rescalings (cont.)

E.g. 5: Corpus Callosum Data:

Show CorpColl\CCFrawAlls3.mpg

Recall direct PCA showed interesting population structure:

Show CorpColl\CCFpcaSCs3PC1.mpg, CorpColl\CCFpcaSCs3PC2.mpg, and CorpColl\CCFpcaSCs3PC3.mpg

Expect difference with "correlation PCA"? Parallel coordinates:

Show CorpColl\CCFParCorAlls3.ps

- Coordinate wise variances very different
- So expect large difference

Explore Rescalings (cont.)

Correlation PCA:

Show CorpColl\CCFpcaSCs3PC1Corr.mpg, CorpColl\CCFpcaSCs3PC2Corr.mpg, CorpColl\CCFpcaSCs3PC3Corr.mpg,

- found only “pixel effect directions”
- since these “have been magnified” (see Par. Coord’s)
- similar effect to Fisher Linear Disc.

Show CorpColl\CCFfldSCs3mag.mpg

- Correlation PCA clearly inferior here

Explore Rescalings (cont.)

Summary:

- no apparent “general solution”
- depends on context
- sometimes “unit free” aspect is dominant, use Corr.
  (or other adjustments)
- other times Corr. PCA gives “useless distortion”

Independent Component Analysis

Idea: Find "directions that maximize independence"

Motivating Context: Signal Processing

In particular: "Blind Source Separation"

References:

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

Lee, T. W. (1998) book

ICA, motivating example

"Cocktail party problem":

- hear several simultaneous conversations

- would like to "separate them"

Model for "conversations": time series:

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

show ICAeg1p1d1Ori.ps

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):

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

Show ICAeg1p1d1Mix.ps

ICA, motivating example (cont.)

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

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

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

Goal is to 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.)

“Solutions” for Cocktail Party example:

Approach 1: PCA:

Show ICAeg1p1d1PCAdecomp.ps

“Direction of Greatest Variability” doesn’t solve this problem

Approach 2: ICA:

Show ICAeg1p1d1ICAdecomp.ps

“Independent Component” directions do

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 $t = 1,\ldots,n$)

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

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

ICA, motivating example (cont.)

Scatterplot View: plot

- signals $\{(s_1(t), s_2(t)) : t = 1, \ldots, n\}$

Show ICAeg1p1d1Ori.ps and ICAeg1p1d1OriSP.ps

- data $\{(x_1(t), x_2(t)) : t = 1, \ldots, n\}$

Show ICAeg1p1d1Mix.ps and ICAeg1p1d1MixSP.ps

- 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.)

Why not PCA?

- finds "direction of greatest variability"

show ICAeg1p1d1MixPCA.ps

- which is wrong direction for "signal separation"

show ICAeg1p1d1PCAdecomp.ps

ICA, Algorithm

ICA Step 1:

- "sphere the data"
- i.e. find linear transf'n to make mean = $\underrightarrow{0}$, cov = I
- i.e. work with $Z = \hat{\Sigma}^{-1/2}(X - \hat{\mu})$
- 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

again show ICAeg1p1d1OriSP.ps and ICAeg1p1d1MixSP.ps

ICA, Algorithm (cont.)

ICA Step 2:

- Find dir'ns that make (sph'd) data as "indep. as possible"
- Worst case: Gaussian – sph'd data is independent

Interesting "converse application" of C.L.T.:

- For S_1 and S_2 independent (& non-Gaussian)
- $X_1 = uS_1 + (1-u)S_2$ is "more Gaussian" for $u \approx \frac{1}{2}$
- so independence comes from "least Gaussian directions"

ICA, Algorithm (cont.)

Criteria for non-Gaussianity / independence:

- kurtosis ($EX^4 - 3\left(EY^2\right)^2$, 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

show ICAeg1p1d1ICAdecomp.ps and again show ICAeg1p1d1MixICA.ps

ICA, Algorithm (cont.)

Notational summary:

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

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, $\underset{\rightarrow}{z} = A_S \underset{\rightarrow}{s}_S = A_S P P \underset{\rightarrow}{s}_S$ i.e. $P \underset{\rightarrow}{s}_S = P W_S \underset{\rightarrow}{z}$

So $P S_S$ and $P W_S$ are also solutions (i.e. $P S_S = P W_S Z$)

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

ICA, Algorithm (cont.)

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

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, $\underset{\rightarrow}{z} = A_S \underset{\rightarrow}{s}_S = A_S D^{-1} D \underset{\rightarrow}{s}_S$ i.e. $D \underset{\rightarrow}{s}_S = D W_S \underset{\rightarrow}{z}$

So DS_S and DW_S are also solutions

ICA, Algorithm (cont.)

Signal Processing Scale identification: (Hyvärinen and Oja)

Choose scale to give each signal $s_i(t)$ “unit total energy”:

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

(preserves energy along rows of data matrix)

Explains “same scales” in Cocktail Party Example

Again show ICAeg1p1d1ICAdecomp.ps