From Last Meeting

Studying Independent Component Analysis (ICA)

Idea: Find "directions that maximize independence"

Parallel Idea: Find directions that maximize "non-Gaussianity"

References:

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

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

ICA, Last Time (cont.)

"Cocktail party problem":

Have "signals" $\underset{\rightarrow}{s}(t) = \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}$ that are linearly mixed: $\underset{\rightarrow}{x} = A\underset{\rightarrow}{s}$

Show ICAeg1p1d1Ori.ps

Try to "recover signals from mixed versions

Show ICAeg1p1d1Mix.ps

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

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

ICA, Last Time (cont.)

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, Last Time (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

- saw how PCA fails

show ICAeg1p1d1MixPCA.ps

- saw how ICA works

show ICAeg1p1d1MixICA.ps

Fundamental concept

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

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

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

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

Fundamental concept (cont.)

Illustrative examples:

Assess normality with Q–Q plot,

scatterplot of “data quantiles” vs. “theoretical quantiles”

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

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

show QQToyEg1.ps

Fundamental concept (cont.)

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

Assess variability with overlay of simulated data curves

Show EGQQWeibull1.ps

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

Fundamental concept (cont.)

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

a. Uniform marginals

Show HDLSS\HDLSSProjUnif.mpg

- $k = 1$ very poor fit (Unif. “far from” Gaussian)
- $k = 2$ much closer? (Triang. Closer to Gaussian)
- $k = 4$ very close, but still have stat’ly sig’t difference
- $k \geq 6$ all differences could be sampling variation

Fundamental concept (cont.)

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

b. Exponential marginals

Show HDLSS\HDLSSProjExp.mpg

- 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

Show HDLSS\HDLSSProjBim.mpg

- Similar lessons to above

Fundamental concept (cont.)

Summary:

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

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. HDLSS statistics is a strange place

ICA, Algorithm

Summary of Algorithm:

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": $S = \hat{\Sigma}^{1/2} S_S$

4. Explored "nonidentifiability", (a) permutation (b) rescaling

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

ICA, Algorithm (cont.)

An attempt at Functional Data Analysis Scale identification:

(Motivation: care about “energy in columns, not rows”)

Make matrix $W_S{}^t$ “work like a matrix of eigenvectors”

i.e. want col’ns of $W_S{}^t$ (thus rows of W_S) orthonormal

ICA, Algorithm (cont.)

Since FastICA gives ortho'l col'ns, define diagonal matrix

$$D_S = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{pmatrix} \text{ by } D_S = W_S{}^t W_S$$

and define $D_S{}^{-1/2} = \begin{pmatrix} |\lambda_1|^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |\lambda_d|^{-1/2} \end{pmatrix}$

then define the "basis": $B_S{}^t = D_S{}^{-1/2} W_S{}^t$ i.e. $B_s = W_S D_S{}^{-1/2}$

ICA, Algorithm (cont.)

Note that:

- B_S is orthonormal:

$$B_S{}^t B_S = D_S{}^{-1/2} W_S{}^t W_S D_S{}^{-1/2} = D_S{}^{-1/2} D_S D_S{}^{-1/2} = I$$

- the B_S based decomp'n $\vec{\underline{s}}_S(t) = B_S \underrightarrow{z(t)}$ "preserves power":

for each column, $\underrightarrow{\underline{s}_S}{}^t \vec{\underline{s}}_S = \vec{\underline{z}}^t B_S{}^t B_S \vec{\underline{z}} = \vec{\underline{z}}^t \vec{\underline{z}}$

ICA, Algorithm (cont.)

Application 1: in “sphere’d scale”, can proceed as with PCA:

- project data in “interesting directions” to “reveal structure”
- analyze “components of variability” (ANOVA)

Application 2: Can “return to original scale”:

- Basis matrix is $\hat{\Sigma}^{1/2} B_S$
- No longer orthogonal, so no ANOVA decomposition
- Still gives interesting directions????
- Still gives useful “features for discrimination”????

ICA, Toy Examples

More Toy examples:

1. 2 sine waves, original and “mixed”

show ICAeg1p1d2Ori.ps and ICAeg1p1d2Mix.ps (everything on this page is combined in ICAeg1p1d2Combine.pdf)

- Scatterplots show “time series structure”(not “random”)

  show ICAeg1p1d2OriSP.ps and ICAeg1p1d2MixSP.ps

- PCA finds wrong direction

  show ICAeg1p1d2MixPCA.ps and ICAeg1p1d2PCAdecomp.ps

- Sphering is enough to solve this (“orthogonal to PCA”)

  Again show ICAeg1p1d2MixSP.ps

- So ICA is good (note: “flip”, and “constant signal power”)

  show ICAeg1p1d2MixICA.ps and ICAeg1p1d2ICAdecomp.ps

ICA, Toy Examples (cont.)

2. Sine wave and noise

Show ICAeg1p1d4Ori.ps, ICAeg1p1d4OriSP.ps, ICAeg1p1d4Mix.ps and ICAeg1p1d4MixSP.ps
(everything on this page is combined in ICAeg1p1d4Combine.pdf)

- PCA finds “diagonal of parallelogram”

Show ICAeg1p1d4MixPCA.ps and ICAeg1p1d4PCAdecomp.ps

- Sine is all in one, but still “wiggles” (noise still present)

- ICA gets it right (but note noise magnified)

Show ICAeg1p1d4MixICA.ps and ICAeg1p1d4PCAdecomp.ps

ICA, Toy Examples (cont.)

3. 2 noise components

Show ICAeg1p1d5Ori.ps, ICAeg1p1d5OriSP.ps, ICAeg1p1d5Mix.ps and ICAeg1p1d5MixSP.ps
(everything on this page is combined in ICAeg1p1d5Combine.pdf)

- PCA finds "axis of ellipse" (happens to be "right")

Show ICAeg1p1d5MixPCA.ps and ICAeg1p1d5PCAdecomp.ps

- Note even "realization" of noise is right

Flip back and forth between ICAeg1p1d5Ori.ps and ICAeg1p1d5PCAdecomp.ps

- ICA is "wrong" (different noise realization)

Show ICAeg1p1d5MixICA.ps and ICAeg1p1d5PCAdecomp.ps

ICA, Toy Examples (cont.)

4. Long parallel points clouds

Show ICAeg1p1d6Ori.ps, ICAeg1p1d6OriSP.ps, ICAeg1p1d6Mix.ps and ICAeg1p1d6MixSP.ps

- PCA finds PC1: “noise” PC2: “signal”

Show ICAeg1p1d6MixPCA.ps and ICAeg1p1d6PCAdecomp.ps

- ICA finds signal in IC1 (most non-Gaussian), noise in IC2

Show ICAeg1p1d6MixICA.ps and ICAeg1p1d6PCAdecomp.ps

ICA, Toy Examples (cont.)

5. 2-d discrimination

show HDLSS\HDLSSod1Raw.ps

- Seek "direction" that separates red and blue projections

- PCA is poor (neither PC1, nor PC2 works)

Show HDLSS\HDLSSod1PCA.ps

- ICA is excellent (since "bimodal" = "most non-Gaussian")

Show HDLSS\HDLSSod1ICA.ps

- No class information used by ICA!

- Thus "useful preprocessing" for discrimination????

- Which is "right", spherical or original scales????

ICA, Toy Examples (cont.)

6. Crossed X Discrimination

show HDLSS\HDLSSxd1Raw.ps

- Common mean (as for Corpora Callosa)

- Want “direction to separate”

- PCA finds good answer

show HDLSS\HDLSSxd1PCA.ps

- So does ICA

show HDLSS\HDLSSxd1ICA.ps

Examples (cont.)

7. Slanted X Discrimination

show HDLSS\HDLSSxd2Raw.ps

- Similar setup and goal to above

- PCA misses (note overlap in projections

show HDLSS\HDLSSxd2PCA.ps

- ICA finds “best direction”

show HDLSS\HDLSSxd2ICA.ps

ICA, CurvDat Examples

Recall PCA for “Parabs”

Show CurvDat\ParabsCurvDat.ps

- Mean captured “parabola” shape
- PC1 is “vertical shift”
- PC2 is “tilt” (hard to see visually)
- Remaining PCs are “Gaussian noise”

ICA, CurvDat Examples (cont.)

Corresponding ICA for “Parabs”

Show ParabsCurvDatICA.ps

- mean and centered data as before
- sphered data has “no structure” (i.e. this structure is “all in covariance”, i.e. have Gaussian point cloud)
- sphered ICs choose “random non-Gaussian” directions
- sphered ICs seem to find outliers
- Original scale versions capture some “vertical shift”
- Non-orthogonality on original scale $\Rightarrow$ hard to interpret

ICA, CurvDat Examples (cont.)

Recall PCA for “Parabs with 2 outliers”

Show CurvDat\Parabs2outCurvDat.ps

- Mean captured “parabola” shape
- PC1 is “vertical shift affected by hi-freq outlier”
- PC2 is “most of high freq.outlier”
- “low freq outlier” and “tilt” are mixed between PC3 & PC4
- hope ICA can “separate these”???

ICA, CurvDat Examples (cont.)

Corresponding ICA for “Parabs with 2 outliers”

Show Parabs2outCurvDatICA.ps

- ICA finds both outliers well (non-Gaussian direction)
- ICA still misses “shift” and “tilt”
- Since these are elliptical point cloud properties, that are ignored through sphering.
- $\exists$ analysis which keeps “both kinds of features”????

ICA, CurvDat Examples (cont.)

Recall PCA for “3 bumps, with 2 independent”

Show CurvDat\Bumps2CurvDat.ps

- Finds both sets of bumps in PC1 and PC2
- Slight mixing of clusters

Corresponding ICA for “3 bumps, with 2 independent”

Show CurvDat\Bumps2CurvDatICA.ps

- Bumps not found (since are “Gaussian” features)
- sphering eliminated bumps

ICA, CurvDat Examples (cont.)

Recall PCA for “Parabs Up and Down” (2 clusters)

Show CurvDat\ParabsUpDnCurvDat.ps

- PC1 finds clusters
- Others find usual structure (vertical shift and tilt)

Corresponding ICA for “Parabs Up and Down”

Show ParabsUpDnCurvDatICA.ps

- Clusters not found????   (seems very “non-Gaussian”)
- sphering killed clusters????
- Problem with numerical search algorithm????

ICA, CurvDat Examples (cont.)

Attempted fix 1:   Change of "nonlinear function"

Show CurvDat\ParabsUpDnCurvDatICAt5.ps

- similar results

- same happened for other choices

Attempted fix 2:   use PCA directions as "starting value"

Show CurvDat\ParabsUpDnCurvDatICAt2.ps

- Gives good solution

- Is this a general problem????

- How generalizable is this solution????

Future plans:

1. Goodness of approximation???
2. Maths for Fisher linear discrimination
3. Polynomial embeddings and SVM discrimination
4. Validation for discrimination (various ways)
5. Internet traffic data?
6. Non elliptical point clouds?