Lecture 6: Latent Variable Models

Applied Multivariate Analysis

Math 570, Fall 2014

Xingye Qiao
Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu
The course deck is courtesy of Sungkyu Jung

Latent Variable Models

- Latent variable models assume the observable random variables $X_1, \ldots, X_p$ are explained by a smaller set of unobservable latent variables (factors, hidden variables, blind sources).
- Examples:
  1. PCA and CCA.
  2. Factor models (Psychology): often referred to as **exploratory** factor analysis.
  3. Structural equation modeling (SEM) (Social Science)
  4. Independent Component Analysis (Engineering, Medical imaging)

Factor models

Consider a simple one-factor model. X_{ji}, the jth random variable in a p-dimensional random vector $\boldsymbol{X}_i$ (corresponding to the ith individual), is modeled as

$$X_{ji} = \lambda_j F_i + U_{ji}, \quad 1 \leq j \leq p, 1 \leq i \leq n.$$

equivalently

$$\boldsymbol{X}_i = \boldsymbol{\lambda} F_i + \boldsymbol{U}_i, \quad 1 \leq i \leq n.$$

- Factor F_i is a univariate random variable defined for each individual i,
- Factor loadings $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ explain the relation between the factor F and the observable variables $\boldsymbol{X}$.
- Errors U_{ji} (independent random variables)

Factor models

Generalization to k-factor model:

$$X_{ji} = \mu_j + \sum_{\ell=1}^{k} \lambda_{j\ell} F_{\ell i} + U_{ji}, \quad 1 \leq j \leq p, 1 \leq i \leq n$$

equivalently

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_{(p\times k)} \boldsymbol{F}_{i(k\times 1)} + \boldsymbol{U}_i, \quad 1 \leq i \leq n.$$

- $\boldsymbol{\mu} \in \mathbb{R}^p$ constant.
- Factor $\boldsymbol{F}_i = (F_{1i}, \ldots, F_{ki})$ is a k-variate random vector with $\mathrm{E}(\boldsymbol{F}_i) = 0, \mathrm{Cov}(\boldsymbol{F}_i) = \mathbb{I}_k$.
- Matrix of factor loadings $\boldsymbol{\Lambda}_{(p\times k)}$ explain the relation between the factors in $\boldsymbol{F}_i$ and the observable variables $\boldsymbol{X}_i$.
- Errors $\boldsymbol{U}_i$ with $\mathrm{E}(\boldsymbol{U}_i) = 0, \mathrm{Cov}(\boldsymbol{U}_i) = \mathrm{diag}(\psi_1, \ldots, \psi_p) = \Psi$.
- $\mathrm{Cov}(\boldsymbol{F}_i, \boldsymbol{U}_i) = \boldsymbol{0}$

Factor models

As a consequence,

$$\text{Cov}(\boldsymbol{X}_i) = \Sigma = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi. \tag{1}$$

- Ψ is diagonal.
- One way to put the problem of factor models is to estimate the covariance matrix in the form of (1).
- Relation to PCA?

PCA and FA

Covariance

- FA: $\text{Cov}(\boldsymbol{X}_i) = \Sigma = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi$.
- PCA: $\text{Cov}(\boldsymbol{X}_i) = \boldsymbol{U}\Lambda\boldsymbol{U}' = \boldsymbol{U}\Lambda^{1/2}(\boldsymbol{U}\Lambda^{1/2})'$

Representation

- FA: $\mathbf{X} - \boldsymbol{\mu}\mathbf{1}_n' = \boldsymbol{\Lambda}_{(p\times k)}\mathbf{F}_{(k\times n)} + \mathbf{U}$. Each row of $\mathbf{F}_{(k\times n)}$ is a hidden factor (variable).
- PCA: $\tilde{\mathbf{X}} \approx \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' = \boldsymbol{U}_{(p\times r)}\mathbf{Z}_{(r\times n)}$. Each row of $\mathbf{Z}_{(r\times n)}$ is a principal component.

Interpretation

- FA: Loading λ's describe the relation between an observable variable X and k hidden factors F_j (X is a linear combination of F_j with λ being the weights).
- PCA: Loading u in the PC direction $\boldsymbol{u}$, describe the relation between a principal component Z and the original variables X_j (Z is a linear combination of X_j with u_j being the weights).

Factor models

$$\text{Cov}(\boldsymbol{X}_i) = \Sigma = \Lambda\Lambda' + \Psi.$$

- Another way of interpreting factor model:
  - The variances of X_j (depreciating subscript i) are decomposed into

$$\text{Var}(X_j) = \text{Var}(\sum_{\ell=1}^{k} \lambda_{j\ell} F_\ell) + \text{Var}(U_j) = \sum_{\ell=1}^{k} \lambda_{j\ell}^2 + \psi_j,$$

  where the first part $h_j^2 = \sum_{\ell=1}^{k} \lambda_{j\ell}^2$ is the portion of the variance of X_j which is explained by the common factors, and is denoted by '*communality*.' The second part ψ_j measures the variation caused by noises unique to the jth variable X_j and is called '*uniqueness*.'

Factor models–Interpretation

Similar to PCA, factors F_ℓ are interpreted using their connection to the original variables.
We see this by an example; the Places Rated Almanac (Boyer and Savageau) rates 329 communities according to nine criteria: Climate and Terrain, Housing, Health Care & Environment, Crime, Transportation, Education, the Arts, Recreation, and Economic ($p = 9, n = 329$).
Suppose a factor model with $k = 3$:

$$\boldsymbol{X} = \Lambda_{(9\times 3)}\boldsymbol{F}_{(3\times 1)} + \boldsymbol{U},$$

with an assumption that the variables are all standardized, i.e.

$$E(X_j) = 0, \mathrm{Var}(X_j) = 1, j = 1, \ldots, p.$$

Factor models–Interpretation

Factor loadings are exactly correlation coefficients! (*because of the add'l assumptions*)

$\text{Corr}(X_j, F_\ell) = \text{Cov}(X_j, F_\ell) = \lambda_{j\ell}.$

Variables with large correlation coefficients contribute more to the factor.

$\Lambda =$

Factor Loadings	1	2	3
Climate	0.286	0.076	0.841
Housing	0.698	0.153	0.084
Health	0.744	-0.41	-0.02
Crime	0.471	0.522	0.135
Transportation	0.681	-0.156	-0.148
Education	0.498	-0.498	-0.253
Arts	0.861	-0.115	0.011
Recreation	0.642	0.322	0.044
Economics	0.298	0.595	-0.533

First factor

Factor 1 Large corr. coef. with Arts, Health, Housing, Recreation, Transportation; Smaller coef. with Crime and Education. The first factor explains everything by varying extent but is primarily a measure of the Arts since about 86% of this variation in this first factor is explained by the Arts.
(Subjective judgement)

$\Lambda =$

Factor Loadings	1	2	3
Climate	0.286	0.076	0.841
Housing	0.698	0.153	0.084
Health	0.744	-0.41	-0.02
Crime	0.471	0.522	0.135
Transportation	0.681	-0.156	-0.148
Education	0.498	-0.498	-0.253
Arts	0.861	-0.115	0.011
Recreation	0.642	0.322	0.044
Economics	0.298	0.595	-0.533

Second and third factor

Factor 2 primarily related to Crime, Education and Economics. The lower the level of the Education, the higher level of Crime but the better the Economy.
Factor 3 a measure of Climate.

$\Lambda =$

Factor Loadings	1	2	3
Climate	0.286	0.076	0.841
Housing	0.698	0.153	0.084
Health	0.744	-0.41	-0.02
Crime	0.471	0.522	0.135
Transportation	0.681	-0.156	-0.148
Education	0.498	-0.498	-0.253
Arts	0.861	-0.115	0.011
Recreation	0.642	0.322	0.044
Economics	0.298	0.595	-0.533

Invariance to scale changes

The interpretation of factor models are invariant to *individual* scale changes (Recall PCA is not invariant to scale change).
Consider $\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{U}$, and let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ with $Y_j = c_j X_j$, so that

$$\boldsymbol{Y} = \text{diag}(\boldsymbol{c})\boldsymbol{\Lambda}\boldsymbol{F} + \text{diag}(\boldsymbol{c})\boldsymbol{U}.$$

- The factors ($\boldsymbol{F}$s in both models) are the same.
- If $\boldsymbol{X}$ had variances 1, $\boldsymbol{\Lambda}$ was the matrix of correlation coefficients:
$$\lambda_{j\ell} = \text{Corr}(X_j, F_\ell).$$
- The new factor loadings $\text{diag}(\boldsymbol{c})\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^Y$ are interpreted via the correlation coefficient between Y_j and F_ℓ,

$$\text{Corr}(Y_j, F_\ell) = \text{Cov}(Y_j, F_\ell)/\sqrt{\text{Var}\, Y_j} = c_j\lambda_{j\ell}/\sqrt{c_j^2} = \lambda_{j\ell}.$$

Estimation: Maximum Likelihood

Assuming MVN for n-sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, write the negative log-likelihood function for parameters $\mu, \boldsymbol{\Lambda}, \Psi$:

$$\ell(\hat{\mu}, \boldsymbol{\Lambda}, \Psi) = \log |\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi| + \text{trace}\{\boldsymbol{S}_0(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi)^{-1}\} + c,$$

where $\hat{\mu} = \bar{\boldsymbol{x}}_i$, and the number of factors k fixed.

- The MLEs of $\boldsymbol{\Lambda}$, Ψ are found by an application of *EM algorithm*.
- (Non-uniqueness of factor loadings) If $\widehat{\boldsymbol{\Lambda}}$ is an MLE, so is $\widehat{\boldsymbol{\Lambda}}\boldsymbol{R}$ for any $k \times k$ orthogonal $\boldsymbol{R}$ (why?).
- Practitioners often take the advantage from the non-uniqueness by choosing *rotated* factors satisfying their taste (discussed later).

Hypotheses test: LRT

Assuming MVN for n-sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, one can test whether
H_0: the assumed factor model $\text{Cov}(\boldsymbol{X}) = \Sigma = \Lambda\Lambda' + \Psi$ is correct
H_1: there is no factor, i.e., Σ is *not* structured.
LR statistic is then (asymptotically for large n)

$$W = -2\log(L_0/L_1) = n\log\left(\frac{|\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}' + \widehat{\Psi}|}{|\boldsymbol{S}_0|}\right) \sim \chi^2_\nu.$$

- Degrees of freedom under H_0 is $q = pk + p - k(k-1)/2$ (rotation); under H_1, $r = p(p+1)/2$. $\nu = r - q$.
- (Bartlett correction) Approximation improved if the rejection region is

$$W^* = (n - 1 - (2p + 4k + 5)/6)\log\left(\frac{|\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}' + \widehat{\Psi}|}{|\boldsymbol{S}_0|}\right) > \chi^2_{1-\alpha,\nu}.$$

Side note: how is the degree of freedom calculated.

- pk parameters in $\mathbf{\Lambda}$
- p diagonal elements in Ψ
- Constraints: to make $\mathbf{\Lambda}$ identifiable, impose constraints so that $\mathbf{\Lambda}'\mathbf{\Lambda}$ is diagonal. If $\mathbf{\Lambda}'\mathbf{\Lambda}$ is not diagonal, can always use a $k \times k$ orthogonal $\boldsymbol{R}$ so that for $\tilde{\mathbf{\Lambda}} := \mathbf{\Lambda}\boldsymbol{R}$, $\tilde{\mathbf{\Lambda}}'\tilde{\mathbf{\Lambda}}$ becomes diagonal. To analyze how many effective constraints are there:
  1. $\mathbf{\Lambda}'\mathbf{\Lambda}$ is symmetric and hence would have had $k(k+1)/2$ free elements if not for the constraints.
  2. $\mathbf{\Lambda}'\mathbf{\Lambda}$ is restricted to be diagonal, and hence we are left with k free element.
  3. This means $k(k-1)/2$ fewer elements.

Overall, $pk + p - k(k-1)/2$.

Estimation: Principal Components

Another estimation approach for the structured covariance $\text{Cov}(\boldsymbol{X}) = \Sigma = \Lambda\Lambda' + \Psi$ is to use the PCA.

1. Begin with the (sample) correlation PCA of $\boldsymbol{X}$; Equivalently, eigen-decomposition of

$$\widehat{\text{Corr}}(\boldsymbol{X}) = \boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2} = \sum_{j=1}^{p} l_j \boldsymbol{u}_j \boldsymbol{u}_j'.$$

2. Approximate Λ by $\widehat{\Lambda} = [\sqrt{l_1}\boldsymbol{u}_1, \ldots, \sqrt{l_k}\boldsymbol{u}_k]$.
3. Take the diagonal elements of $\boldsymbol{R} - \widehat{\Lambda}\widehat{\Lambda}'$ as $\hat{\psi}_{jj}$;

$$\widehat{\Psi} = \text{diag}(\boldsymbol{R} - \widehat{\Lambda}\widehat{\Lambda}').$$

Estimation: Principal Components
Number of factors

Heuristics are used when PCA is used to estimate the factor loadings:

1. Use scree plot or cumulative scree plot;
2. or use Kaiser's rule: retain factors 1–k satisfying

$$l_k > \bar{l} = \frac{1}{p} \sum_{i=1}^{p} l_i.$$

When the correlation PCA is used, then

$$\bar{l} = 1.$$

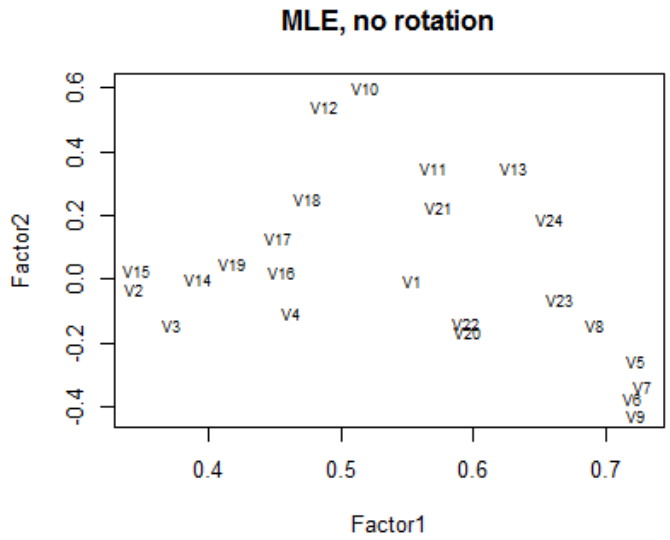
Example: 24 Psychological Tests

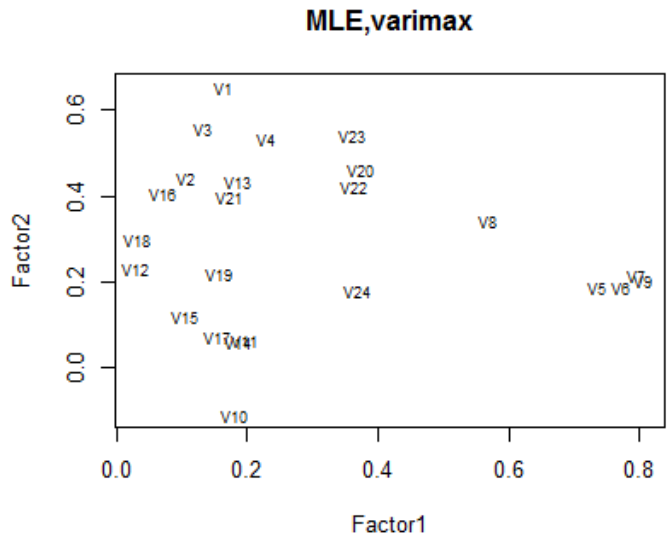
(Section 15.4.4, Izenman) 24 psychological tests administered to 145 students of a grade school near Chicago. ($p = 24, n = 145$)

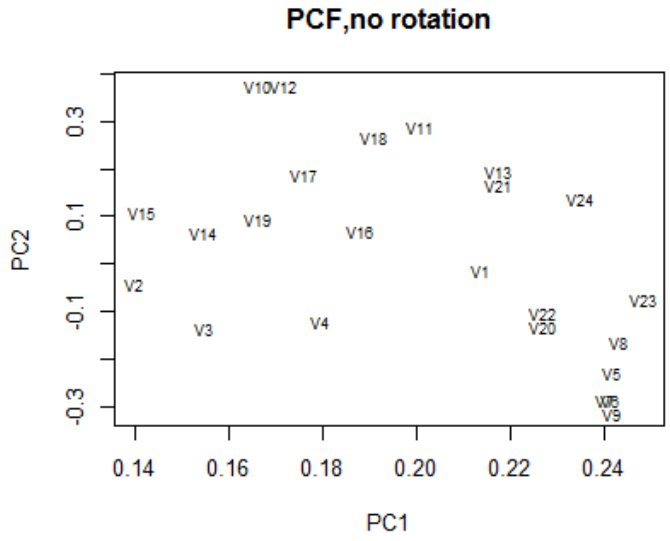
- We build a 5 factor model (recommended by both LRT and Kaiser's rule) and compute estimates for factor loadings.
- Again, estimates are not unique! This is because $\text{Cov}(\boldsymbol{X}) = \Sigma = \Lambda\Lambda' + \Psi = \Lambda \boldsymbol{J}\boldsymbol{J}'\Lambda' + \Psi$. We will see choosing an appropriate $\boldsymbol{J}$ "*improves*" the interpretation.
- Next slide: variable list.

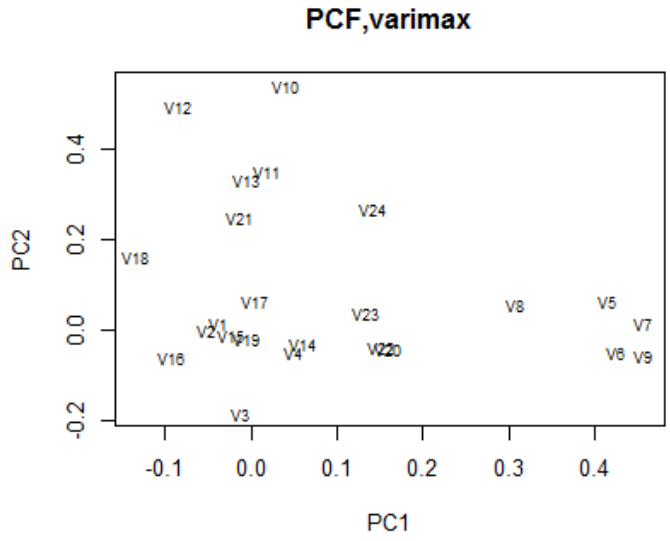
(1) visual perception: spatial relations,
(2) cubes: spatial relations,
(3) paper form board: spatial imagery,
(4) flags: visual-imagery,
(5) general information: scientific and social,
(6) paragraph comprehension,
(7) sentence completion,
(8) word classification,
(9) word meaning,
(10) addition: speed of adding,
(11) code: perceptual speed,
(12) counting dots: arranged in random patterns to be counted by subject,
(13) straight-curved capitals: capital letters to be distinguished,
(14) word recognition,
(15) number recognition,
(16) figure recognition,
(17) object-number: memory,
(18) number-figure: memory,
(19) figure-word: memory,
(20) deduction: logical deduction
(21) numerical puzzles
(22) problem reasoning
(23) series completion
(24) arithmetic problems

See `Factor_24psych.R`









Factor rotation

- In factor rotation, we look for an orthogonal matrix $\boldsymbol{J}_{(k\times k)}$ such that the factor loadings $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{J}$ be more easily interpreted than the original factor loadings Λ.
- Factor loadings can often be easily interpreted if they have simple structure;
  - Each variable has a high loading on a single factor but close-to-zero loadings on other factors.
  - Each factor is constituted by only a few variables with high loadings but close-to-zero loadings of other variables.
- Many proposals – Varimax, Quartimax, Promax, Oblimin rotations.

Varimax rotation

- By (the same) Kaiser (1958). Here the *raw* version is presented.
- Idea is that we want the factor loadings $(\lambda_{1\ell}, \ldots, \lambda_{p\ell})'$ more like $(1, 0, 0, \ldots, 0)'$ but not like $(\frac{1}{3}, -\frac{1}{3}, \ldots, \frac{1}{3})'$;
  $\Rightarrow$ Larger variance for $\lambda_{1\ell}^2, \ldots, \lambda_{p\ell}^2$.
- Let $\lambda_{j\ell}^*$ be the rotated factor loadings $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{J}$. Varimax rotation maximizes the sum of variances of squared loadings (hence, the name); In other words,

$$\boldsymbol{J}^\dagger = \underset{\boldsymbol{J}}{\operatorname{argmax}} \sum_{\ell=1}^{k} \left\{ \widehat{\operatorname{Var}}(\{(\lambda_{1\ell}^*)^2, \ldots, (\lambda_{p\ell}^*)^2\}) \right\}$$

- Varimax rotated factors $\boldsymbol{\Lambda}^\dagger = \boldsymbol{\Lambda}\boldsymbol{J}^\dagger$.

Example: 24 Psychological Tests

Estimated factor loadings (MLE) with Varimax rotation.

Test	S_1	S_2	S_3	S_4	S_5	Unique
1	0.165	**0.655**	0.124	0.181	0.208	0.453
2	0.108	**0.442**	0.087	0.095	0.003	0.777
25	0.134	**0.559**	–0.048	0.111	0.094	0.646
26	0.230	**0.533**	0.089	0.081	0.014	0.648
5	**0.738**	0.189	0.191	0.149	0.056	0.357
6	**0.772**	0.187	0.031	0.248	0.125	0.291
7	**0.798**	0.214	0.143	0.088	0.051	0.286
8	**0.571**	0.343	0.239	0.127	0.044	0.481
9	**0.808**	0.203	0.033	0.219	–0.007	0.257
10	0.181	–0.108	**0.845**	0.180	0.029	0.208
11	0.195	0.066	**0.422**	0.436	0.419	0.413
12	0.030	0.232	**0.694**	0.102	0.131	0.436
13	0.186	0.432	**0.477**	0.077	**0.540**	0.253
14	0.185	0.061	0.044	**0.552**	0.080	0.649
15	0.104	0.122	0.059	**0.509**	–0.002	0.712
16	0.070	0.406	0.056	**0.509**	0.055	0.565
17	0.154	0.072	0.210	**0.595**	–0.026	0.572
18	0.032	0.300	0.322	**0.458**	0.006	0.596
19	0.156	0.221	0.144	0.378	0.046	0.761
20	0.373	**0.462**	0.127	0.293	–0.193	0.509
21	0.172	0.398	0.431	0.238	0.002	0.569
22	0.364	0.423	0.114	0.320	–0.068	0.568
23	0.361	**0.542**	0.249	0.231	–0.113	0.447
24	0.368	0.179	**0.495**	0.321	–0.066	0.480
SS	3.639	2.958	2.450	2.386	0.633	

24 Psychological Tests

- Varimax rotated loadings are easily interpreted.
- $F_1 \sim$ verbal factor
- $F_2 \sim$ deduction of relation factor
- $F_3 \sim$ speed factor
- $F_4 \sim$ memory factor
- $F_5 \sim$ speed factor
- Check the degrees of freedom of LRT is ν, and the test leads that $k = 5$.
- Kaiser's rule for PCA Factor estimation also leads that $k = 5$.

Factor analysis in R

To perform maximum likelihood estimation of factor models:

```
# number of factors = 3, no ro
fit <- factanal(data, 3, rotation="none")
print(fit, digits=2, cutoff=.3, sort=FALSE)
```

To rotate,

```
fit_varimax <- factanal(data, 3, rotation="varimax")
fit_promax <- factanal(data, 3, rotation="promax")
```

To perform Principal Component estimation of factor models and varimax rotation

```
pc<-prcomp(data,scale=TRUE) #correlation PCA
pc$sdev   # Kaiser's rule
pcfactorloadings <- pc$rotation[,1:5]
rot<-varimax(pcfactorloadings, normalize = FALSE)
```

Factor analysis: Final Remark

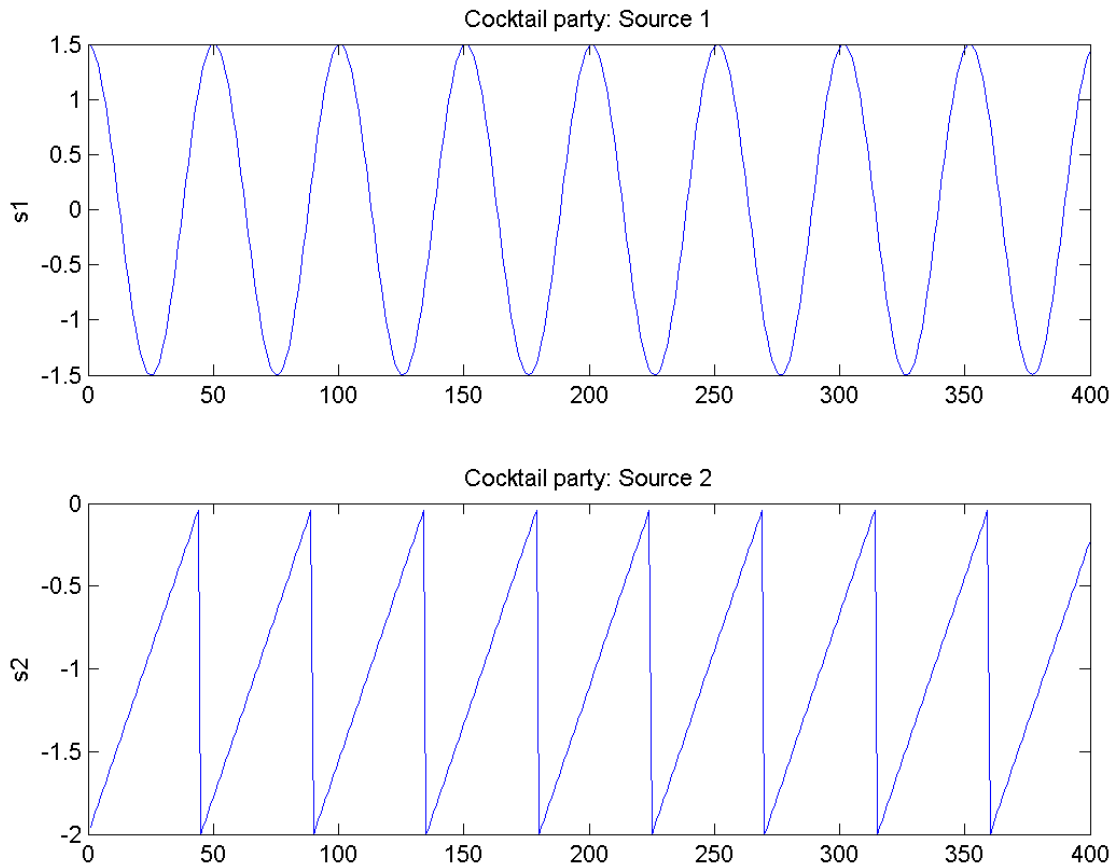
- PCA vs FA
  - Both methods are mostly used in exploratory data analysis.
  - If *uniqueness* is zero, both methods agree.
  - PCA is purely exploratory (no assumption, no model), while FA assumes a specific model.
  - PCA is NOT scale invariant, while FA is so.
  - Computation: PCA–simple vs FA–complex.
- FA is controversial
  - MLE–Normal assumption too strict? numerical problems?
  - MLE–Heywood cases? (referring to the cases where corr. matrix is singular$\rightarrow$ trouble in MLE)
  - Interpretation of factors are based on convenience. One can keep rotating factors until something useful pops out.

Independent Component Analysis (ICA)

- ICA is a multivariate statistical technique that seeks to uncover hidden variables.
- Basic form is a linear dimension reduction; Find *interesting* directions $\boldsymbol{u}_i$, such that its scores ($S_i = \boldsymbol{u}_i' \boldsymbol{X}$) are *independent* to each other.
- Fairly new technique: First appeared in Cardoso 1993.
- Good source of information can be found at `http://research.ics.aalto.fi/ica/`
- Motivating example: Cocktail party problem. Fun examples at `http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi`

Cocktail party problem

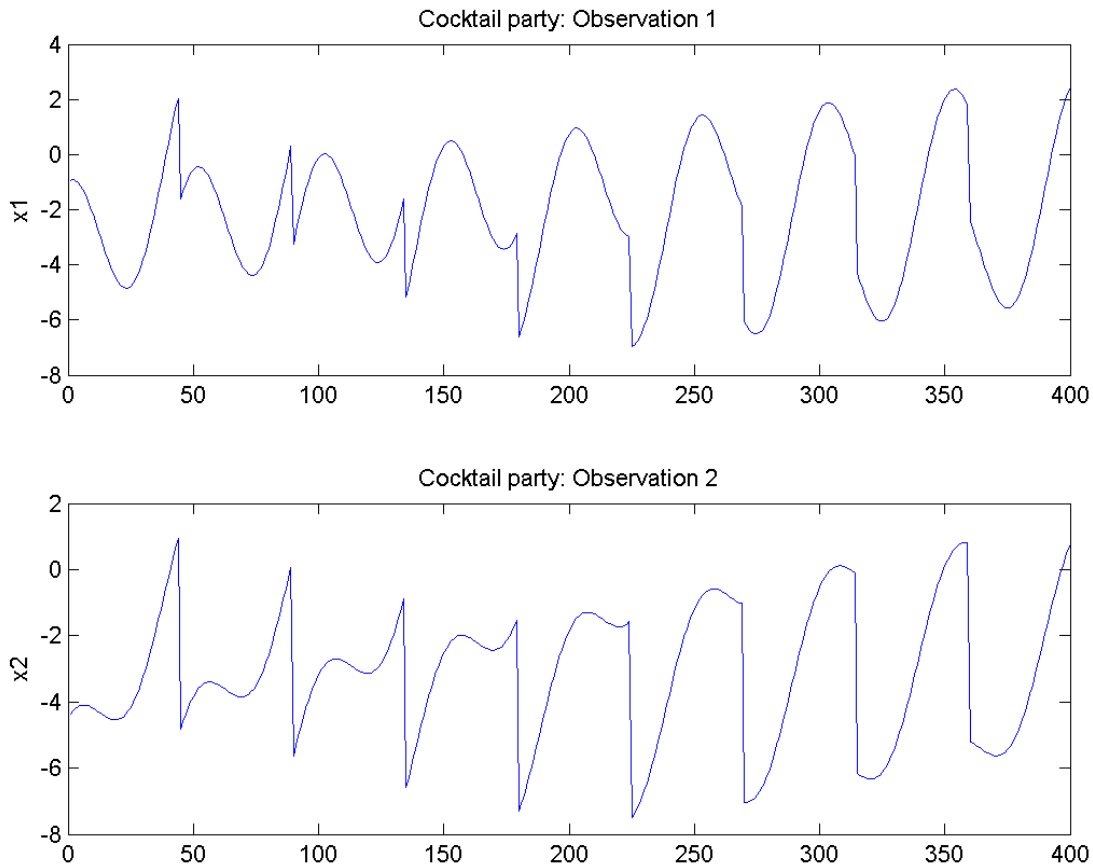
- Hear several simultaneous conversations;
- Wish to separate them.
- Conversations are modeled as time series: $s_1(t)$ and $s_2(t)$.



Cocktail party problem

- What we actually hear is a mixture of both conversations

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t),$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t).$$



Cocktail party problem

- Without knowing the actual source and the mixing matrix $\mathbf{A} = (a_{ij})$, ICA tries to recover the source $\boldsymbol{S} = (S_1, S_2)'$ from data $\boldsymbol{X} = (X_1, X_2)'$;

$$\boldsymbol{X} = (X_1, X_2)' = \mathbf{A}\boldsymbol{S}.$$

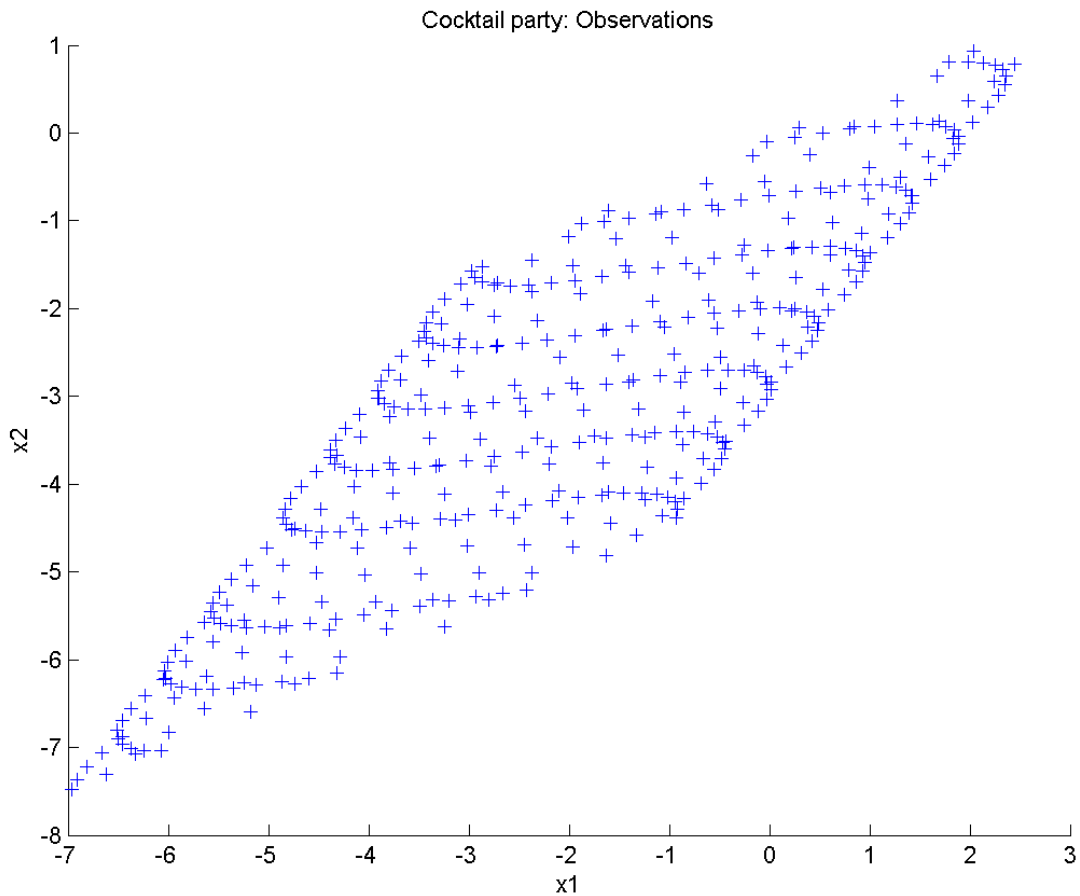
- Since $\boldsymbol{S} = \mathbf{W}\boldsymbol{X}$ for $\mathbf{W} = \mathbf{A}^{-1} = \begin{pmatrix} \boldsymbol{w}_1' \\ \boldsymbol{w}_2' \end{pmatrix}$, the method suggests a linear dimension reduction, with projection vectors $\boldsymbol{w}_i$ and scores S_i:

$$S_i = \boldsymbol{w}_i'\boldsymbol{X}, \ \ i = 1, 2.$$

- In general, the sources are non-Gaussian (with at most one exception). Otherwise, little hope to separate them.

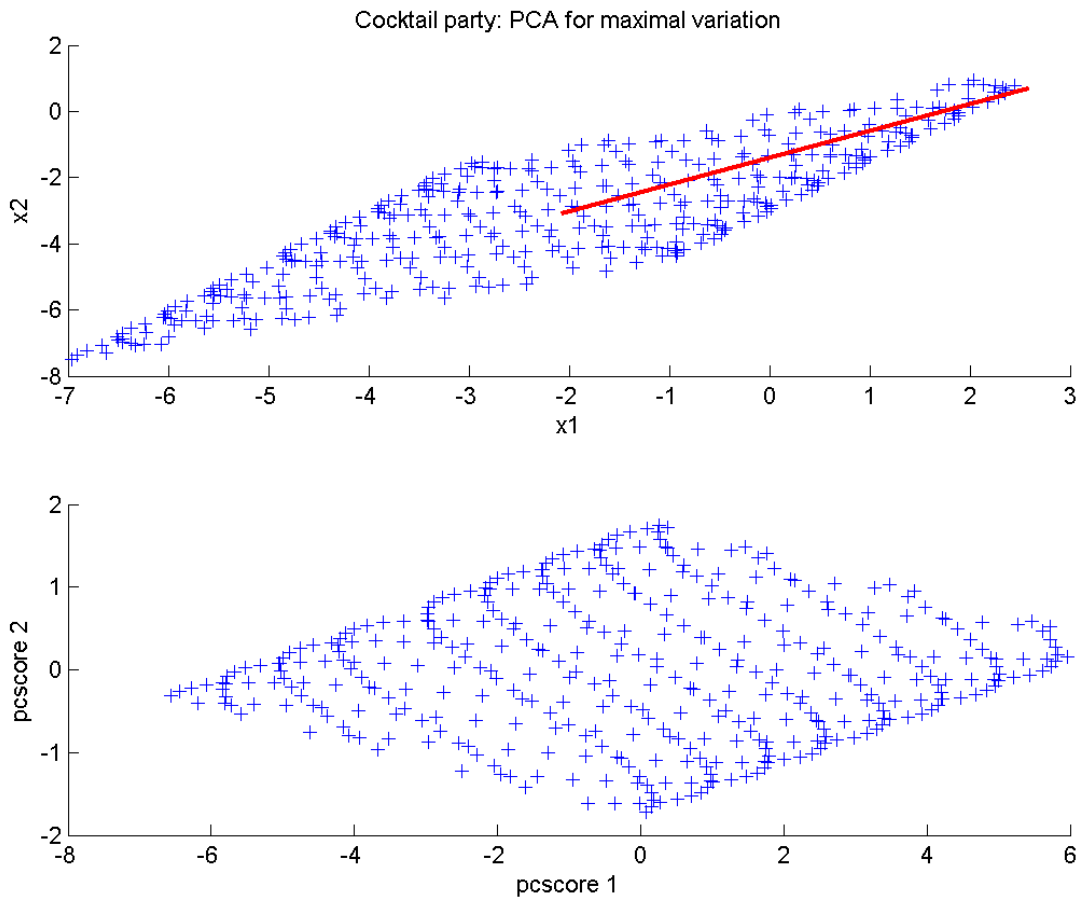
Cocktail party problem

- Ignoring the time label, the data is multivariate ($p = 2$)
- Can PCA help?



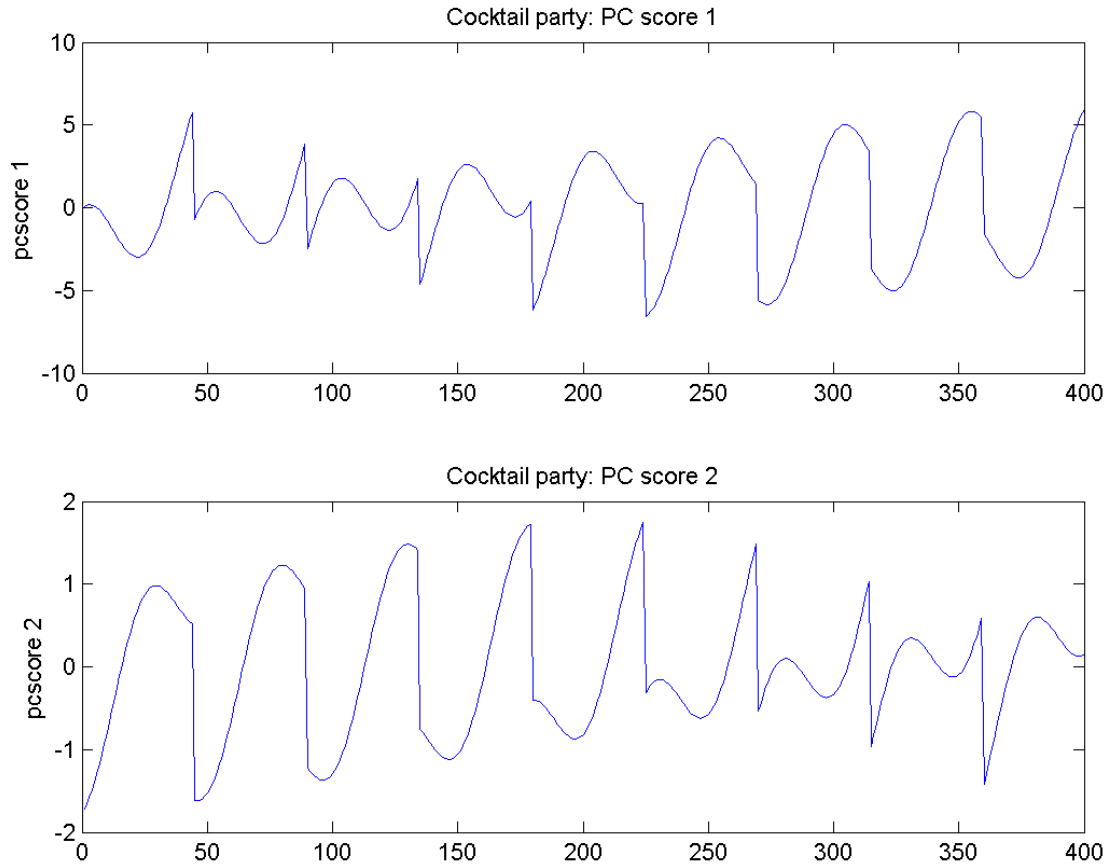
Cocktail party problem–PCA?

- $\boldsymbol{w}_i$ as PC directions is wrong for source separation.
- Since PCA finds the direction of greatest variation.



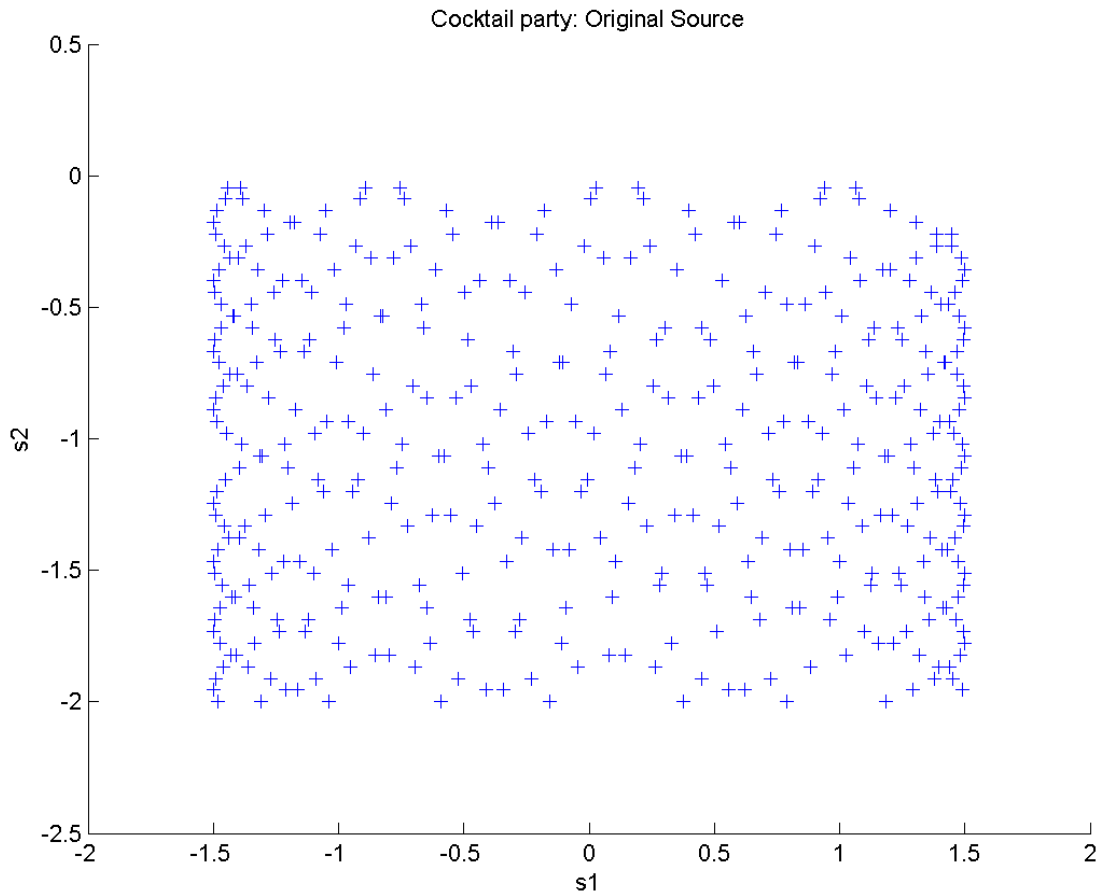
Cocktail party problem–PCA?

- PCA scores are just another mixtures of signals-no good.



Cocktail party problem

- Scatters of original source (which is unknown in practice)
- Understood as uniform distribution on a rectangle
- So marginal distributions are independent

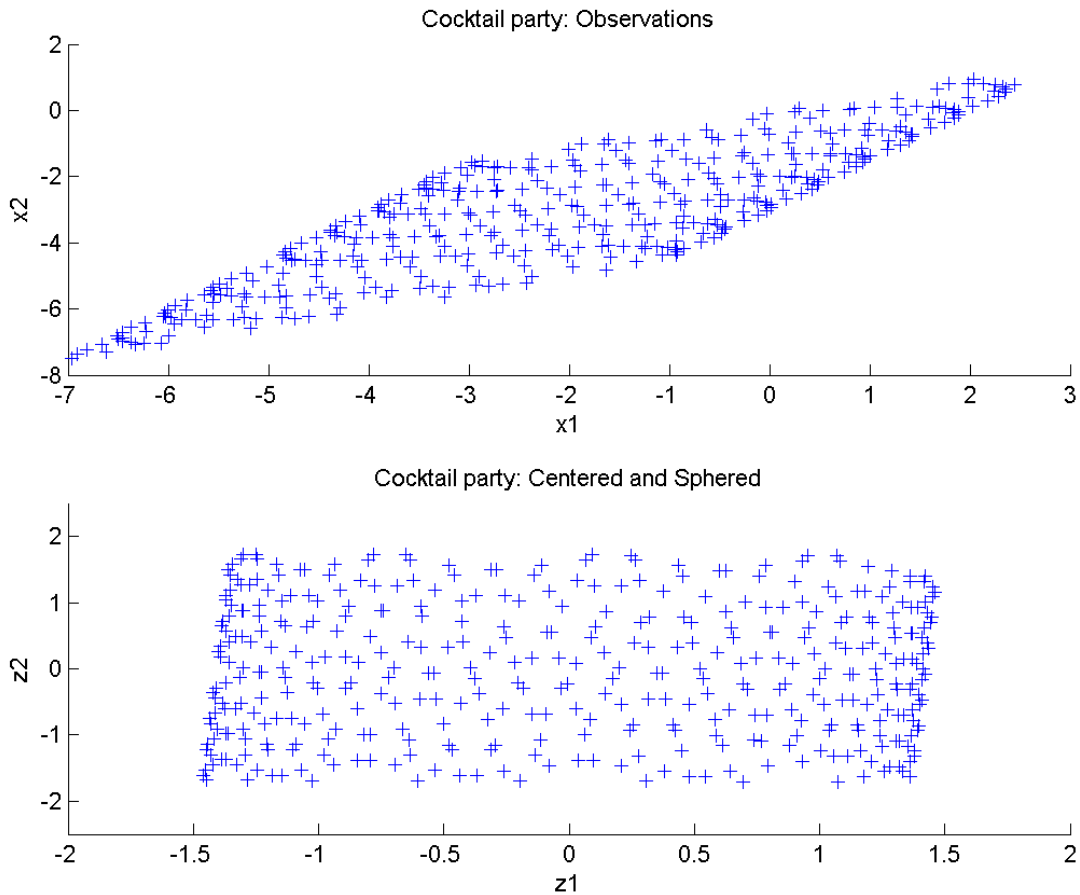


Cocktail party problem–Trial&Error

- Centering and sphering (aka whitening) the observations

$$\boldsymbol{z} = \widehat{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \hat{\mu}).$$

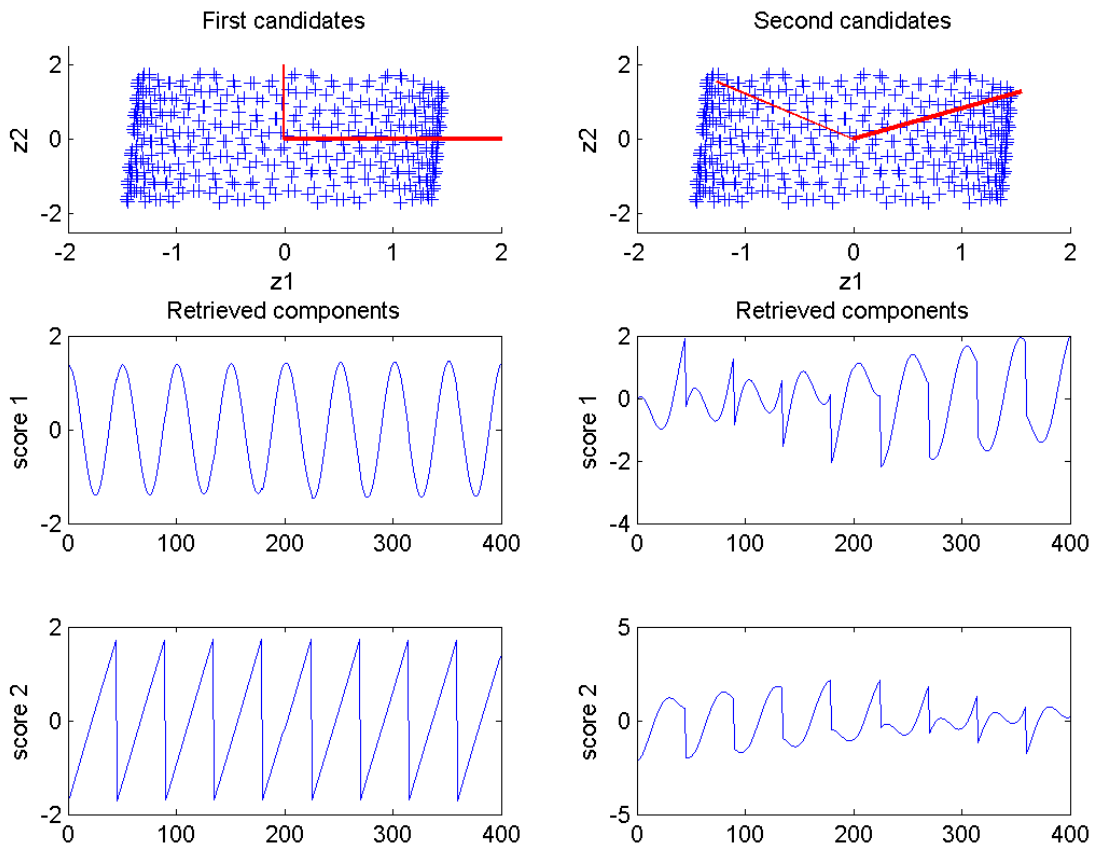
- Find directions in the transformed ($\boldsymbol{z}$-) space



- First candidate $\boldsymbol{U} = [\boldsymbol{u}_1 \boldsymbol{u}_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ surprisingly well-separates two signals.
- Back to original $\boldsymbol{x}$-space,

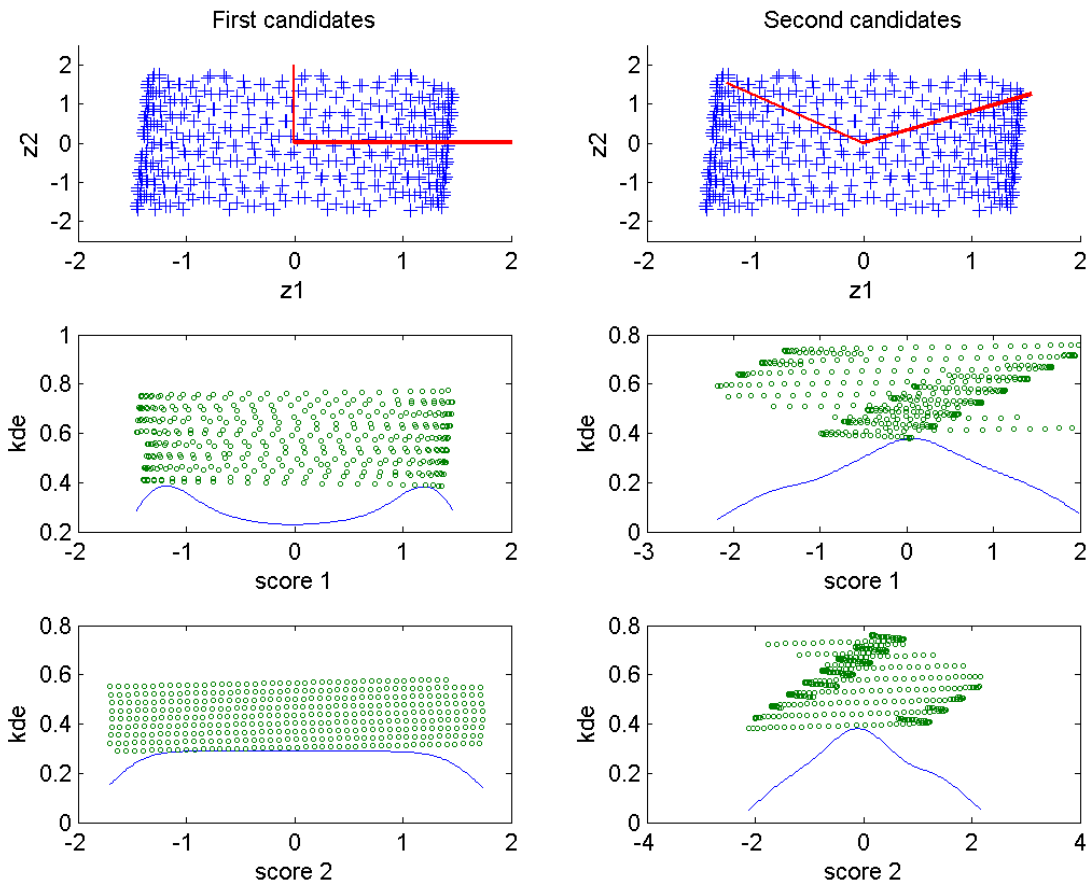
$$s_i = \boldsymbol{u}_i' \boldsymbol{z} = \boldsymbol{u}_i' \widehat{\Sigma}^{-\frac{1}{2}} (\boldsymbol{x} - \hat{\mu}) = \widehat{\boldsymbol{w}}_i' \boldsymbol{x} + \boldsymbol{c},$$

where $\widehat{\boldsymbol{w}}_i = \widehat{\Sigma}^{-\frac{1}{2}} \boldsymbol{u}_i = i$th column of $\widehat{\Sigma}^{-\frac{1}{2}}$.



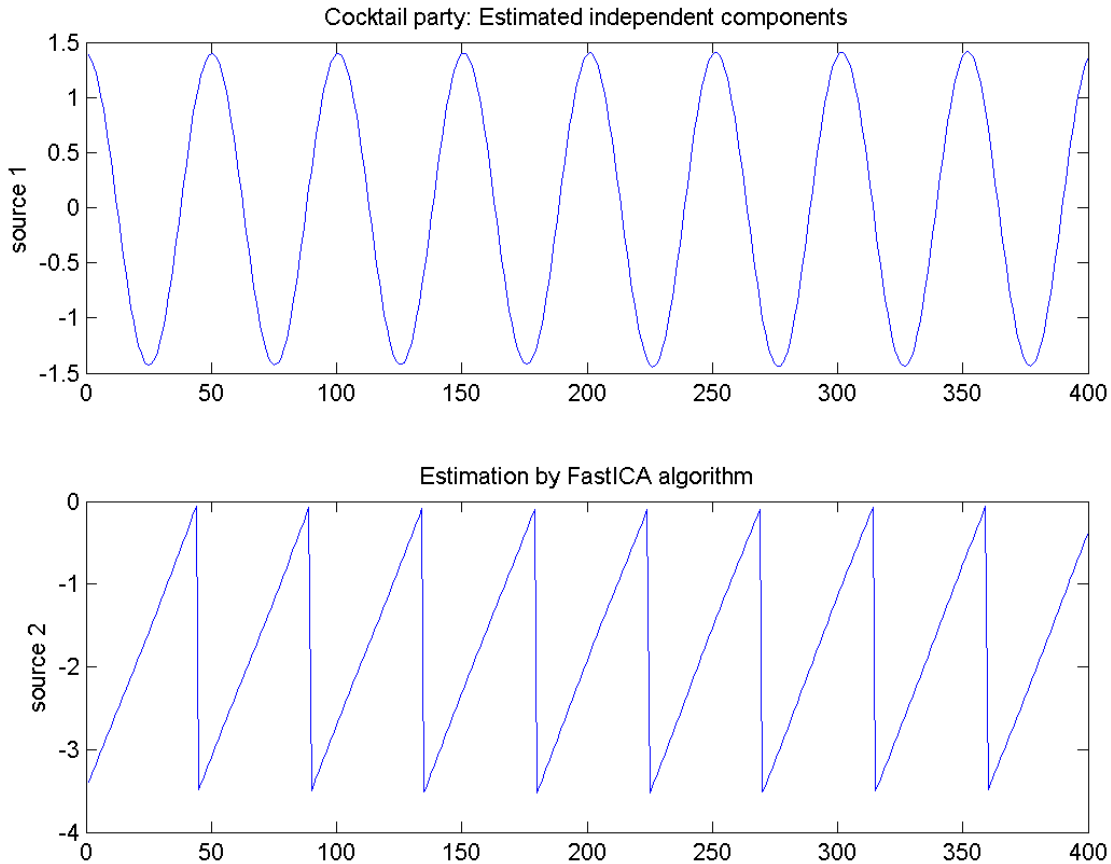
Cocktail party problem–Trial&Error

- Non-Gaussianity is the key
  1st cand. (left) less Gaussian than 2nd (right)
- Systematic way of separating sources?
  (Measuring independence through nongaussianity)



Cocktail party problem–ICA

- Solution by `FastICA` algorithm (to be discussed)



Independent Component Analysis

- The cocktail party is an example of ICA models.
- General ICA model– Observation is a mixed source plus error:

$$\boldsymbol{X} = f(\boldsymbol{S}) + \boldsymbol{e},$$

where the m sources $S_1, \ldots, S_m$ are standardized and independent.

- Special case: Noiseless linear mixing ICA model

$$f(\boldsymbol{S}) = \mathbf{A}\boldsymbol{S}, \ \mathrm{Var}(\boldsymbol{e}) = 0.$$

- If the number of observations (dimension of $\boldsymbol{X}$) equals the number of sources (dimension of $\boldsymbol{S}$, m above), then there exists an unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ such that

$$\boldsymbol{X} = \mathbf{A}\boldsymbol{S} \Leftrightarrow \boldsymbol{S} = \mathbf{W}\boldsymbol{X},$$

so that the sources are exactly recovered from $\boldsymbol{X}$.

- ICA finds an estimate $\widehat{\mathbf{W}}$ of $\mathbf{W}$ so that the components of $\boldsymbol{Y} = \widehat{\mathbf{W}}\boldsymbol{X}$ are as **independent** (and as non-Gaussian) as possible.

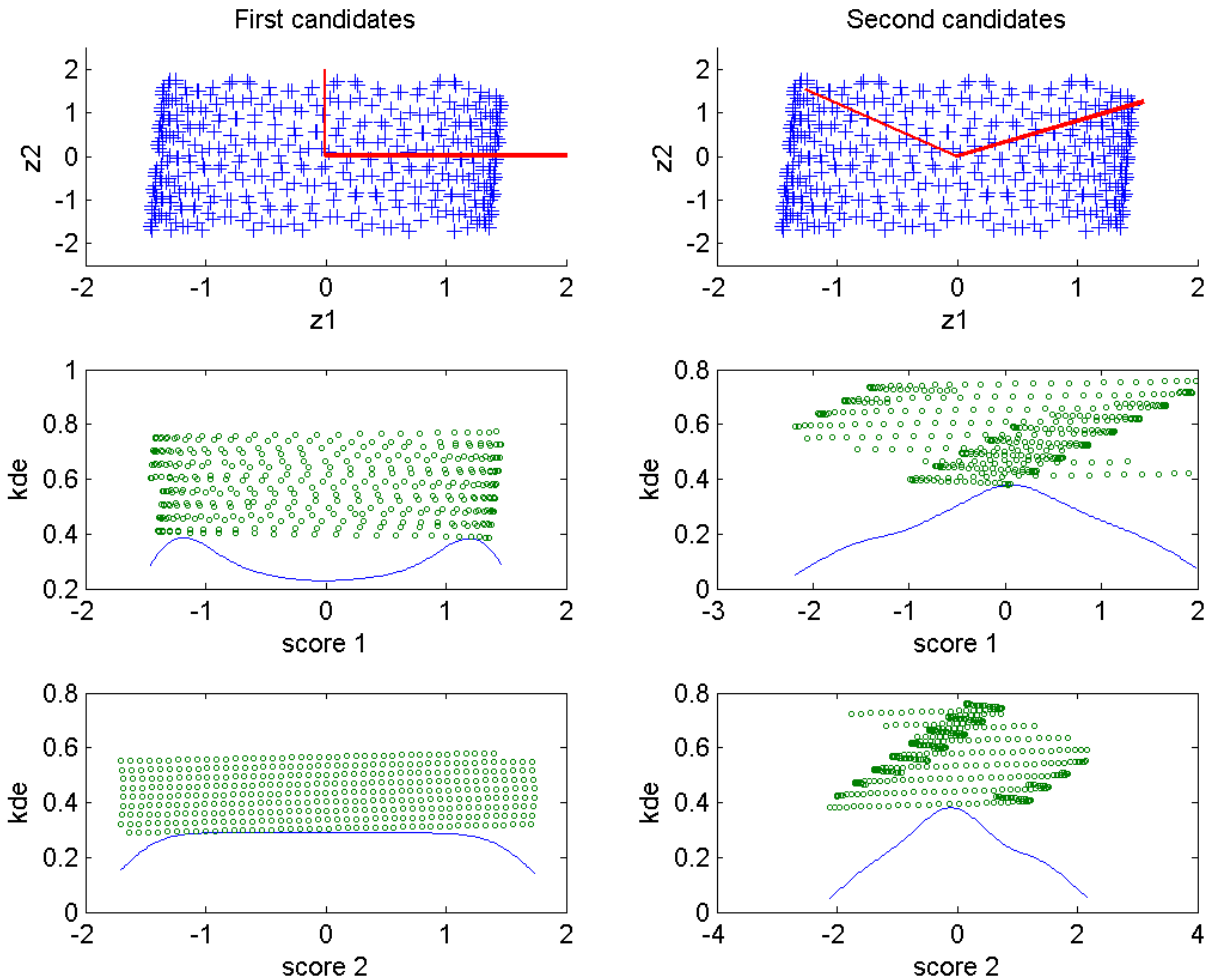
On the next few slides:

- Practical viewpoint: Show that Gaussian is bad for practical purpose.
- Heuristic: data suggests negatively larger kurtosis $\rightarrow$ less normal
- Theory: relations between Negentropy, Gaussian, Independence and Mutual Information.

ICA and MVN

- Worst case for ICA: $\boldsymbol{X} \sim N_p(\mu, \Sigma)$.
- After centered+sphered (whitened), $\tilde{\boldsymbol{X}} = \Sigma^{-1/2}(\boldsymbol{X} - \mu) \sim N_p(0, \mathbb{I}_p)$, independence in every orthogonal directions.
  - In particular, assume that $\text{Cov}(\boldsymbol{S}) = \mathbb{I}$ too. If $\mathbf{A}$ satisfies $\tilde{\boldsymbol{X}} = \mathbf{A}\boldsymbol{S}$, then $\mathbb{I} = \text{Cov}(\tilde{\boldsymbol{X}}) = \mathbf{A}\mathbb{I}\mathbf{A}'$, and hence $\mathbf{A}$ is orthogonal. Then note that $\boldsymbol{S}$ is Gaussian too and $\boldsymbol{S}$ and $\tilde{\boldsymbol{X}}$ share the same density function. In other words, we cannot find (strictly) more interesting variables $\boldsymbol{S}$ than what we already have ($\tilde{\boldsymbol{X}}$) through rotation ($\mathbf{A}$).
- No meaningful directions
- From practical viewpoint, Gaussian data makes it impossible to find better structure.

Heuristic: Back to trial-and-error of Cocktail Party example



kurtosis: measures "peakness". Normal distribution has kurtosis $= 0$. t distribution has kurtosis > 0. Uniform distribution has negative kurtosis.

- Left: uniform on rectangle (independent marginals), negatively larger kurtosis
- Right: uniform on slanted parallelogram (dependent marginals), kurtosis $\sim$ that of Normal
- Kurtosis has been used as a measure of non-Gaussianity (see Projection Pursuit, Section 7.4 Izenman)

Entropy

- *Negentropy* as another measure of departure from normality
- Entropy of a random variable X (or pdf p),

$$\mathcal{H}(X) = -\int p(x) \log p(x) dx = \mathrm{E}_X(\log(p(X)^{-1})),$$

  is a measure of uncertainty about outcomes of X.
- Simple intuitive example: Bernoulli X with $P(X = 1) = u$.
  $\mathcal{H}(X) = u \log(1/u) + (1 - u) \log[1/(1 - u)] \approx 0$ low uncertainty.
  $\mathcal{H}(X)$ large, large uncertainty (and more information the distribution contains).
- The idea here is that the less likely an event is, the more information it provides when it occurs. Conversely, if an event is very likely to occur, then it does not tell much once it actually occurs.

```
> u=0.1
> u*log(1/u)+(1-u)*log(1/(1-u))
[1] 0.325083
> p=0.2
> n=10
> x=0:n
> pmf = vapply(x,dbinom,1,size=n,prob=p)
> sum(pmf*log(1/pmf))
[1] 1.621929
> n=100
> x=0:n
> pmf = vapply(x,dbinom,1,size=n,prob=p)
> sum(pmf*log(1/pmf))
[1] 2.803287
> p=0.5
> pmf = vapply(x,dbinom,1,size=n,prob=p)
> sum(pmf*log(1/pmf))
[1] 3.028368
```

Negentropy measures non-Gaussianity

- Rao (1965) from Izenman,

  Among all random variables having equal variance, the largest value of $\mathcal{H}(Y)$ occurs when Y has a normal distribution.

- Negentropy (Negative Entropy): For $Z \sim N(0,1)$, and $Y \sim (0,1)$

$$\mathcal{J}(Y) = \mathcal{H}(Z) - \mathcal{H}(Y) \geq 0$$

  The equality holds iff $Y \sim N(0,1)$.

  Negentropy $\mathcal{J}(Y) \uparrow \Leftrightarrow$ very NON-Gaussian

Kullback-Leibler divergence

A "distance" between two density functions:

$$KL(p\|q) := \int p(\boldsymbol{y}) \log\left(\frac{p(\boldsymbol{y})}{q(\boldsymbol{y})}\right) d\boldsymbol{y}$$
$$= -\mathcal{H}(\boldsymbol{Y}) - \int p(\boldsymbol{y}) \log(q(\boldsymbol{y})) d\boldsymbol{y}$$

- $KL(p\|q) \geq 0$
- $KL(p\|q) = 0$ iff $p = q$

Measuring Independence

- From the definition:
  $Y_1, \ldots, Y_p$ mutually independent if

$$f(y_1, \ldots, y_p) = \prod_{i=1}^{p} f_i(y_i).$$

- A measure of independence is *Mutual Information* (between r.vs)

$$\mathcal{MI}(\boldsymbol{Y}) = KL(f \| \prod_{i=1}^{p} f_i),$$

  difference between the true joint density of $\boldsymbol{Y}$ and the joint density of $\boldsymbol{Y}$ should Y_j are independent (which is the product of marginal densities).
- "How much difference is the joint density f from its independence version?"

Measuring Independence

- For observations $\boldsymbol{X}$ and unmixing $\mathbf{W}$, retrieved sources $\boldsymbol{S} = \mathbf{W}\boldsymbol{X}$ are more independent if $\mathcal{MI}(\boldsymbol{S})$ is smaller.
- If both $\boldsymbol{X}_{(m\times 1)}$ and $\boldsymbol{S}_{(m\times 1)}$ have zero mean and identity covariance matrix (an assumption we can afford), then

$$\begin{aligned}\mathcal{MI}(\boldsymbol{S}) &= KL(f \| \prod_{i=1}^{p} f_i) \\ &= -\mathcal{H}(\boldsymbol{S}) - \int f(\boldsymbol{s}) \log(\prod_{i=1}^{p} f_i(\boldsymbol{s}))d\boldsymbol{s} \\ &= \sum_{j=1}^{m} \mathcal{H}(S_j) - \mathcal{H}(\boldsymbol{S}),\end{aligned}$$

where $\mathcal{H}(\boldsymbol{S}) = \log|\det(\mathbf{W})| + \mathcal{H}(\boldsymbol{X}) = \mathcal{H}(\boldsymbol{X})$, leading to

$$\mathcal{MI}(\boldsymbol{S}) = \sum_{j=1}^{m} \mathcal{H}(S_j) - \mathcal{H}(\boldsymbol{X}) = (m\mathcal{H}(Z) - \mathcal{H}(\boldsymbol{X})) - \boxed{\textstyle\sum_{i=1}^{n} \mathcal{J}(S_i)}.$$

Independence and Non-Gaussianity

Facts

1. Retrieved sources $\boldsymbol{S} = \mathbf{W}\boldsymbol{X}$ are more independent if $\mathcal{MI}(\boldsymbol{S})$ is smaller.
2. $\mathcal{MI}(\boldsymbol{S}) \downarrow$, negentropy of S_i, $\sum_{j=1}^{m} \mathcal{J}(S_j) \uparrow$
3. The larger the negentropy of S_i is, the more non-Gaussian S_i, $\sum_{j=1}^{m} \mathcal{J}(S_j)$, is.

Hence

Find $\mathbf{W}$ to have S_i most independent $\Leftrightarrow$ To minimize $\mathcal{MI}(\boldsymbol{S}) \Leftrightarrow$ To maximize $\sum_{j=1}^{m} \mathcal{J}(S_j) \Leftrightarrow$ To make each S_i as non-Gaussian as possible.

ICA

- Back to ICA
- noiseless linear mixing model $\boldsymbol{X} = \mathbf{A}\boldsymbol{S}$
- Given $\boldsymbol{X}$, find unmixing matrix $\mathbf{W}$ so that retrieved sources $\boldsymbol{S} = \mathbf{W}\boldsymbol{X}$ are as independent as possible.
- Measure of independence? kurtosis or negentropy.
- Both are parameters of distributions.
- In practice, find optimal $\widehat{\mathbf{W}}$ that maximizes sum of *estimates* of negentropy based on data.

FastICA algorithm by Hyvärinen et al. (2001)

- By Oja and his colleagues (`http://research.ics.aalto.fi/ica/`)
- `FastICA` numerically maximizes

$$(E(G(\mathbf{W}\boldsymbol{X}) - E(G(\boldsymbol{Z})))^2 = \text{ Approx. of } \sum_{j=1}^{m} \mathcal{J}(S_j)$$

with respect to $\mathbf{W}$.
G various non-quadratic functions,
$\boldsymbol{Z} \sim N_m(0, \mathbb{I})$.

Package FastICA

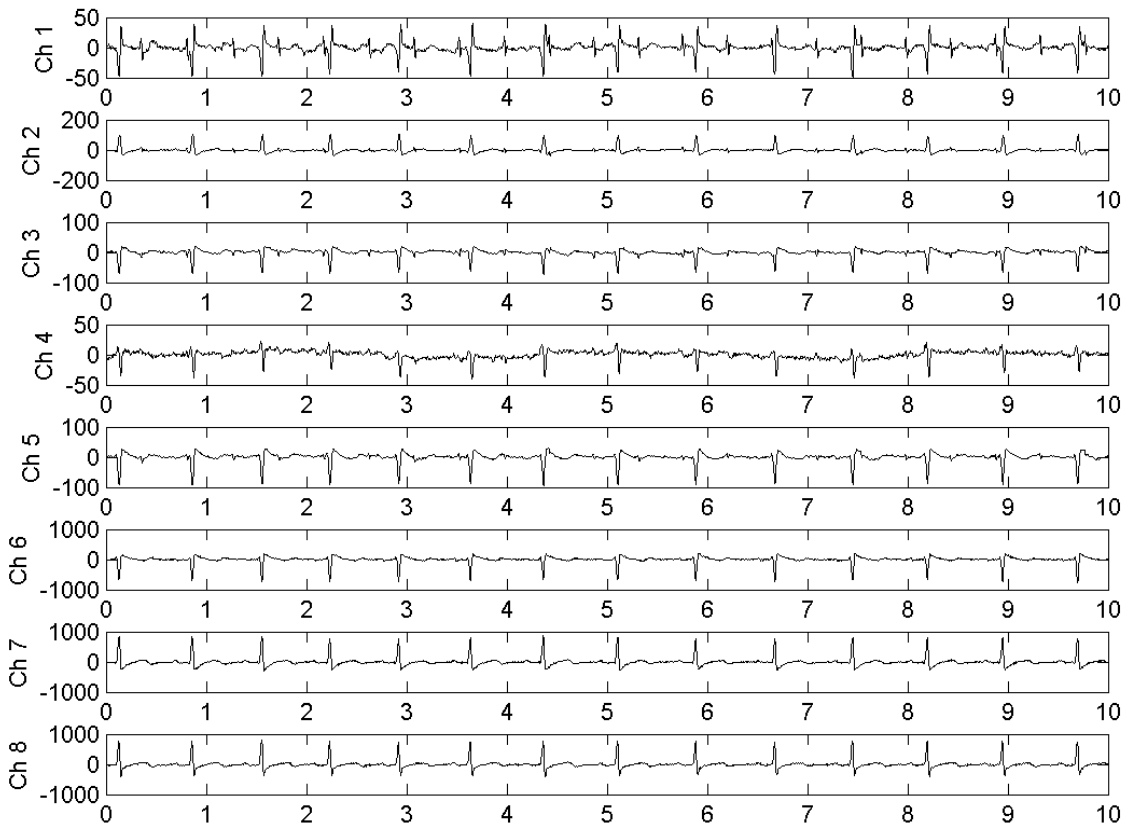
Matlab package: `http://research.ics.aalto.fi/ica/fastica/code/dlcode.shtml`
R package:
`http://cran.r-project.org/web/packages/fastICA/`
Other languages:
`http://research.ics.aalto.fi/ica/fastica/`

Cutaneous potential recordings of a pregnant woman

- Section 15.3.2 Izenman
- L. De Lathauwer, B. De Moor, J. Vandewalle, "Fetal Electrocardiogram Extraction by Blind Source Subspace Separation", IEEE Trans. Biomedical Engineering, Vol. 47, No. 5, May 2000. search for the title at `http://homes.esat.kuleuven.be/~smc/daisy/daisydata.html`
- Monitoring fetal heart activity of a pregnant women to assess health of the fetus.
- Multichannel electrocardiogram (ECG) is used to maternal and fetal electrical activity.
- Challenge: Maternal ECG signal is stronger than fetal, contaminated by respiration.
- Goal: Separate fetal heart activity from mixed signal.

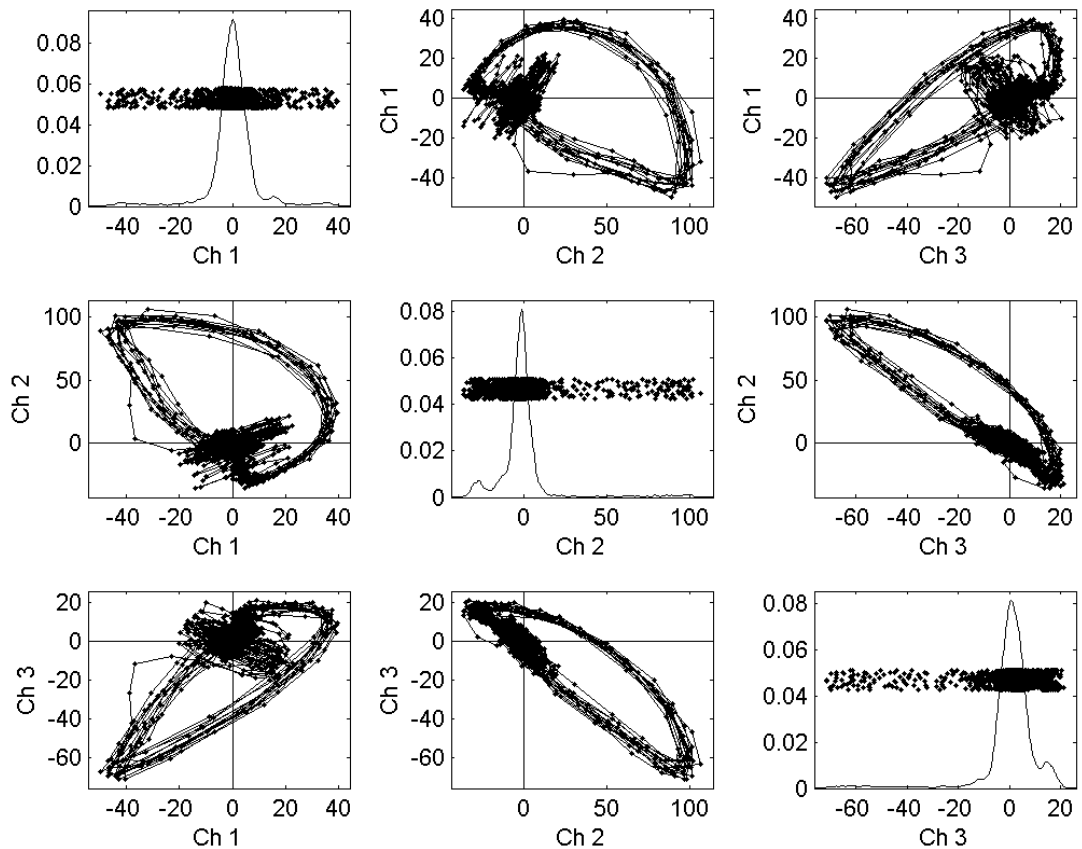
Cutaneous recordings of pregnant woman

- 8-channel recordings of ECG over time (10 secs)
- First five are measured near fetus
- Last three are on the mother's chest



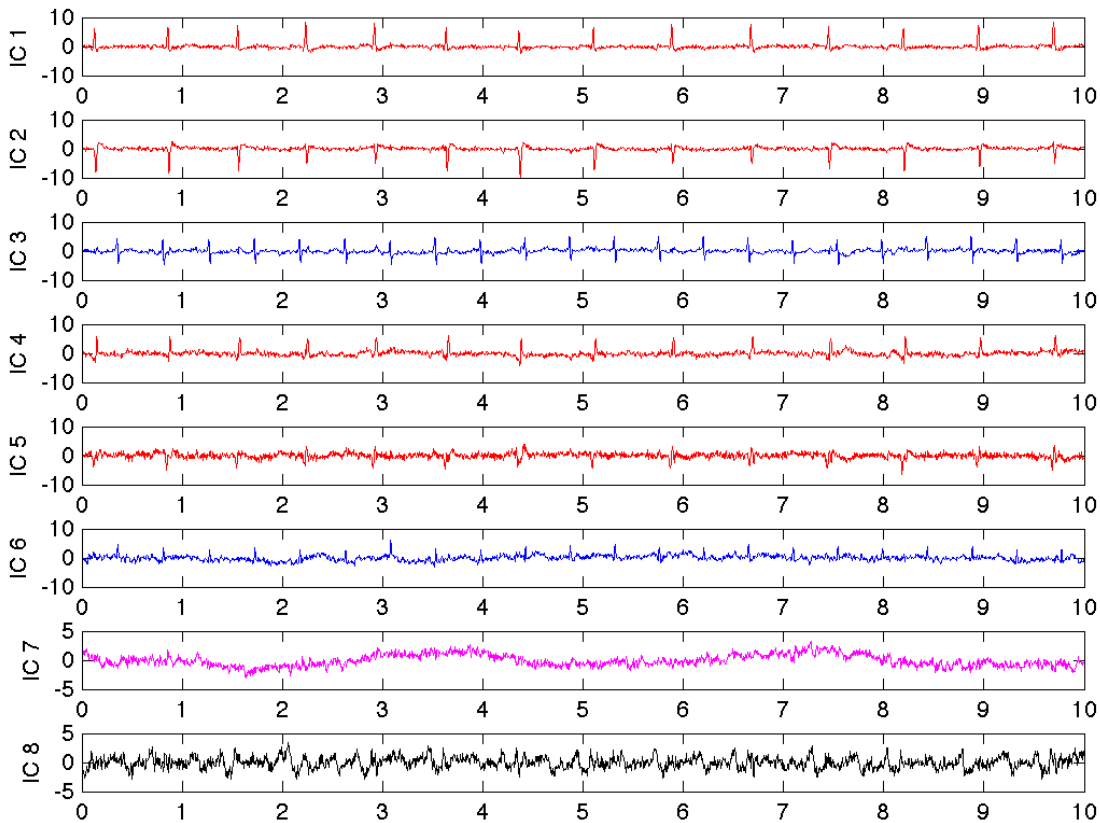
Cutaneous recordings of pregnant woman

- Marginal and joint distributions of the input $\boldsymbol{X}$ are severely non-Gaussian



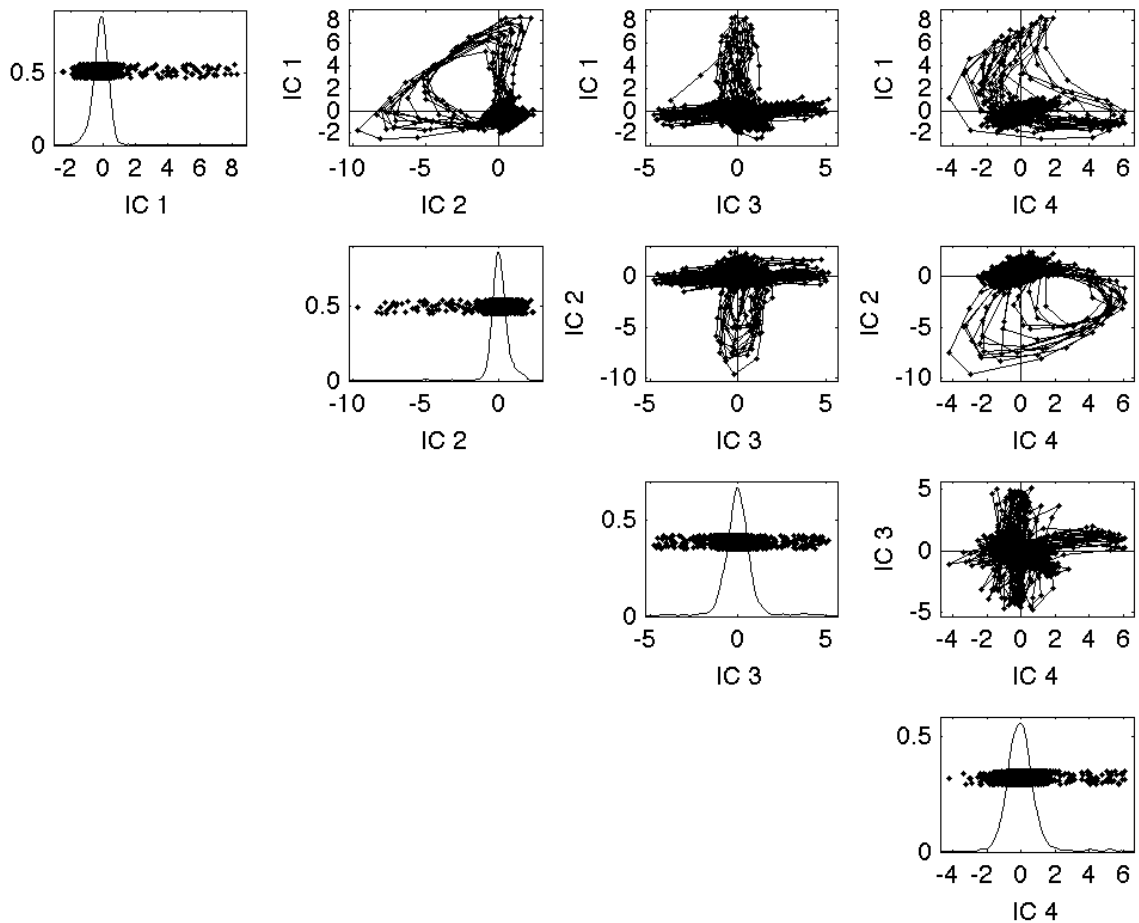
Cutaneous recordings of pregnant woman

- Result from a "FastICA" algorithm
- cardiac rhythms of the mother, cardiac rhythms of the fetus,
- respiration component, sensor noise



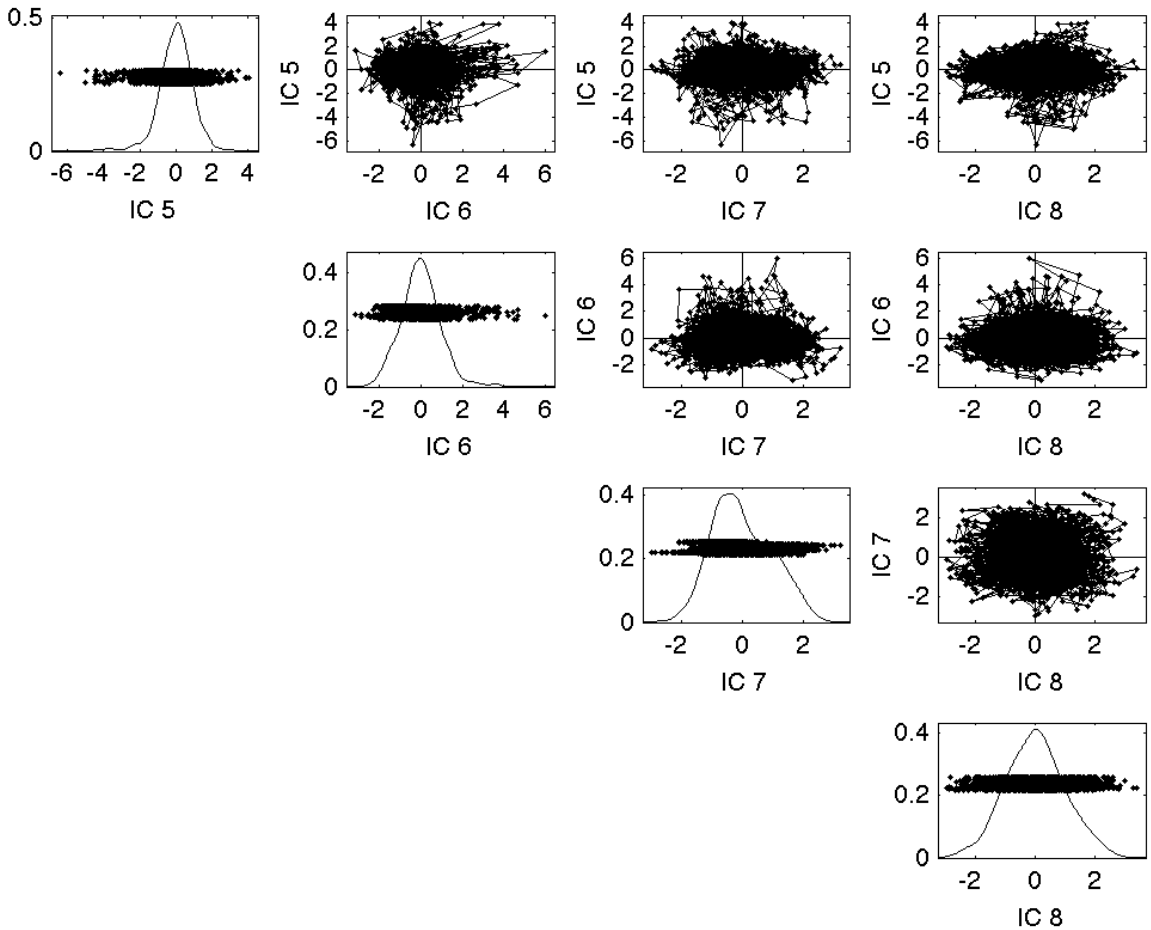
Potential problems

- Every run results in different component (p.556 Izenman)
- Algorithm susceptible to different initial values
- Are the components really independent?



Potential problems

- Are the components really independent? *As ind. as possible*
- Identifiability issues (similar to those of factor analysis)
- The order of ICs given by non-Gaussianity



Few remarks on ICA

- Linear Mixing ICA is a linear dimension reduction method.
- Finds k directions $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$ so that the scores (of projections of observations) $S_i = \boldsymbol{w}_i' \boldsymbol{X}$ are most "interesting" (where the interestingness is independence or non-gaussianity)
- Many different ways to solve. `FastICA` most popular
- Usually applied to analyze time-series data: Autocorrelation of sources?

Projection pursuit (PP)

- PP (Friedman and Tukey) is another linear dimension reduction approach that seeks to find two or three projection directions to achieve non-Gaussianity.
- The goal is non-Gaussianity (as the authors thought that would be the most revealing feature.) However, the resulting model is very similar to that of ICA (which shots for independence.)
- Some projection indices that measures non-Gaussianity have been proposed and are maximized/minimized. Seems to favor kurtosis.
- Read Izenman Section 7.4; HS Section 19.2; ESL Section 14.7.3
  (ESL = The Elements of Statistical Learning by Hastie Tibshirani and Freidman.)