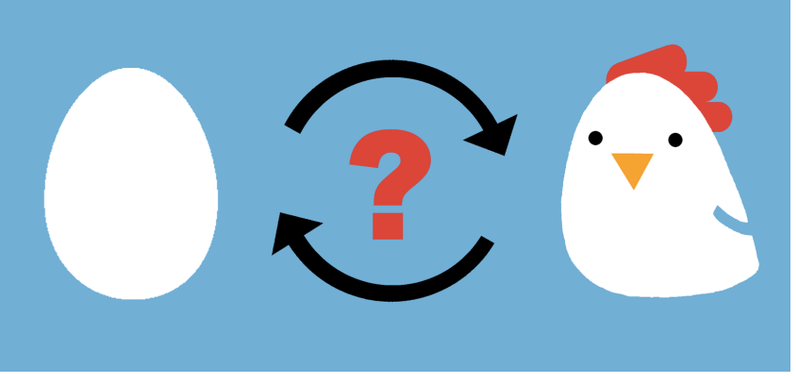


Class 7 (Sept. 22, 2021)

Causality and Machine Learning (*80-816/516*)

Multivariate analysis,

Identification of causal effects &

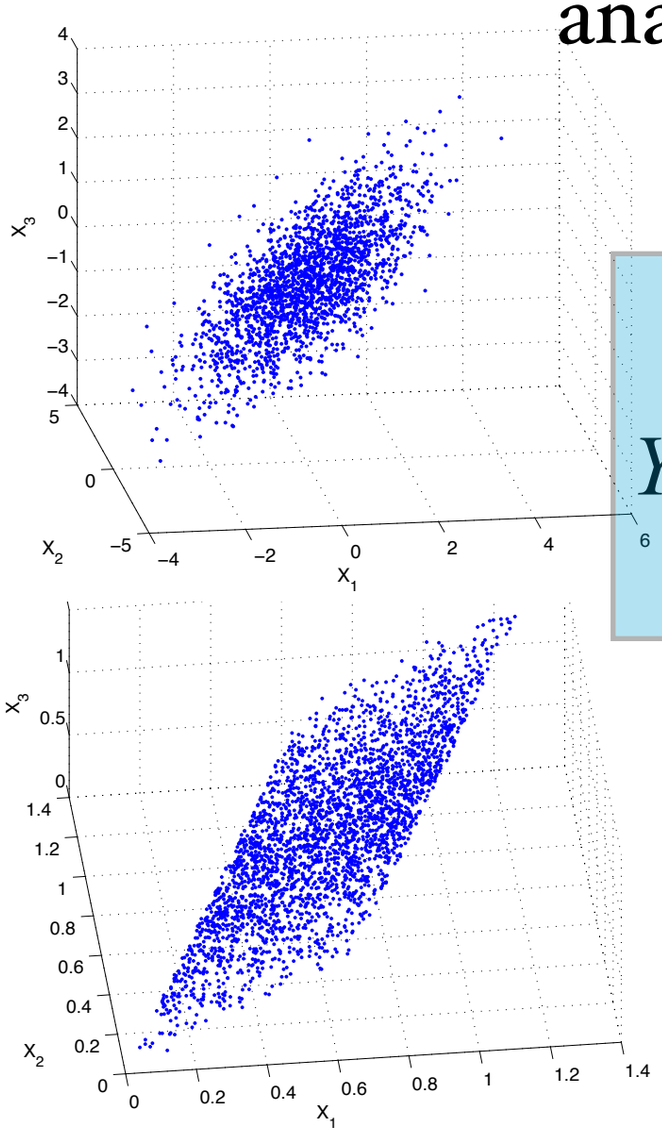
Counterfactual reasoning

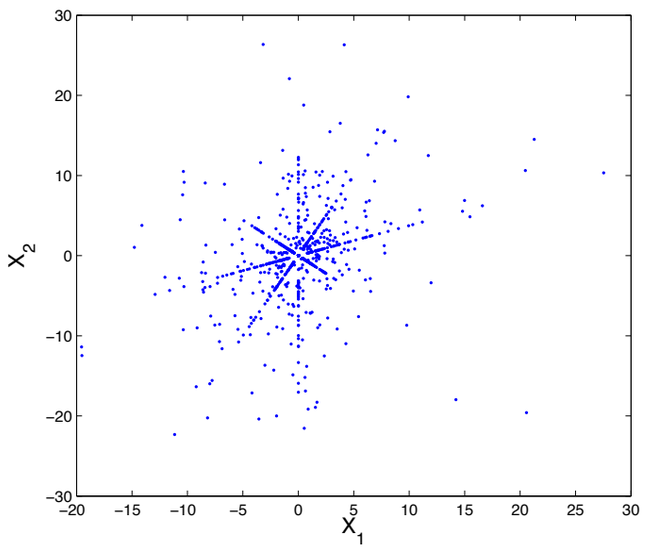
Instructor: Kun Zhang

Zoom link: https://cmu.zoom.us/j/8214572323

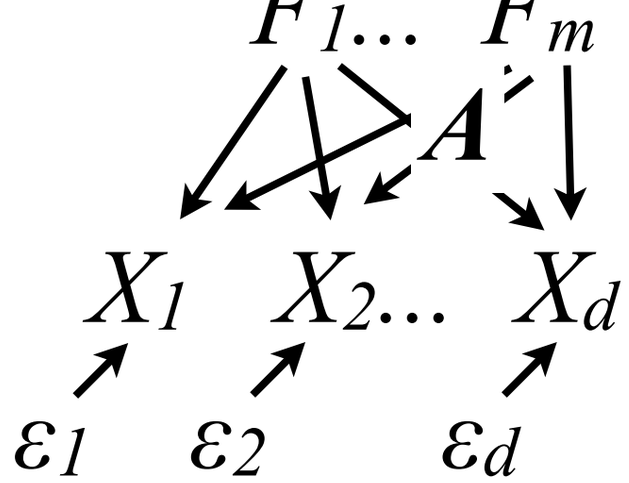
Office Hours: Wednesdays 3 – 4PM (on Zoom or in person);
other times by appointment

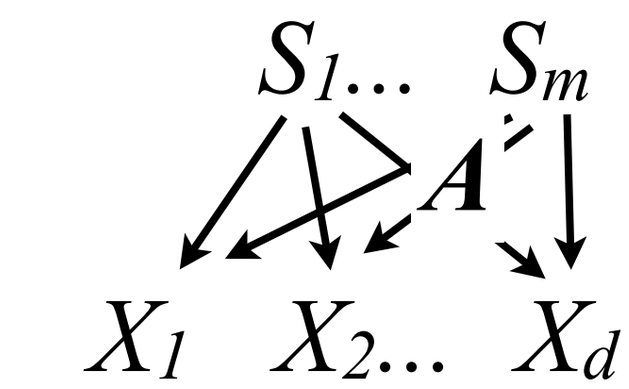
Multivariate analysis (MVA): involves observation and analysis of more than one outcome variable at a time.





Find a projection of the data:
$Y = w^{\mathrm{T}}\mathrm{X}$ with certain properties.



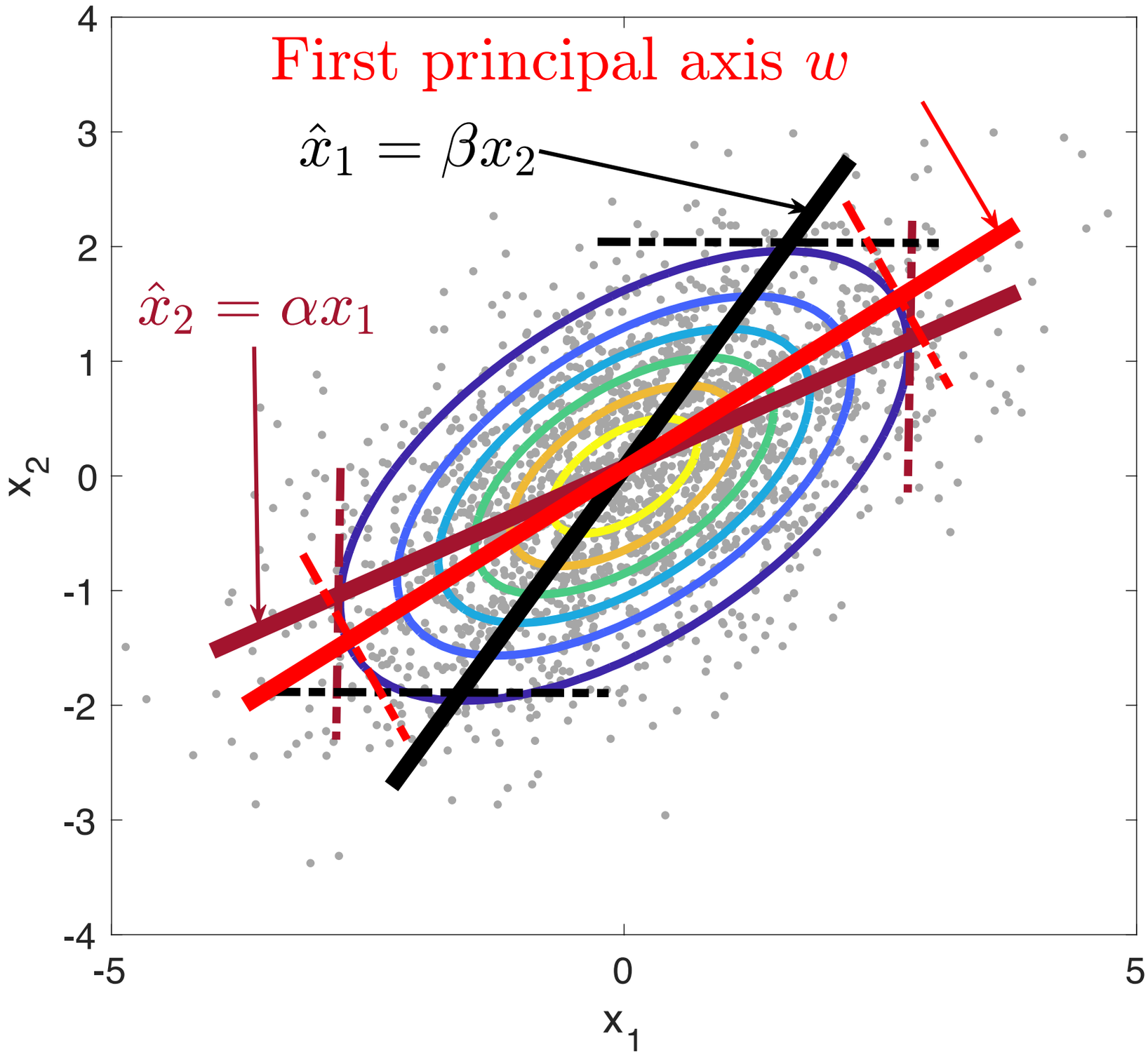


- Regression…
- Principal component analysis
- Factor analysis:
  $\mathbf{X} = \mathbf{A}\cdot F + \varepsilon$

$$\mathbf{X} = [X_1, X_2, \ldots, X_d]^{\mathrm{T}}$$

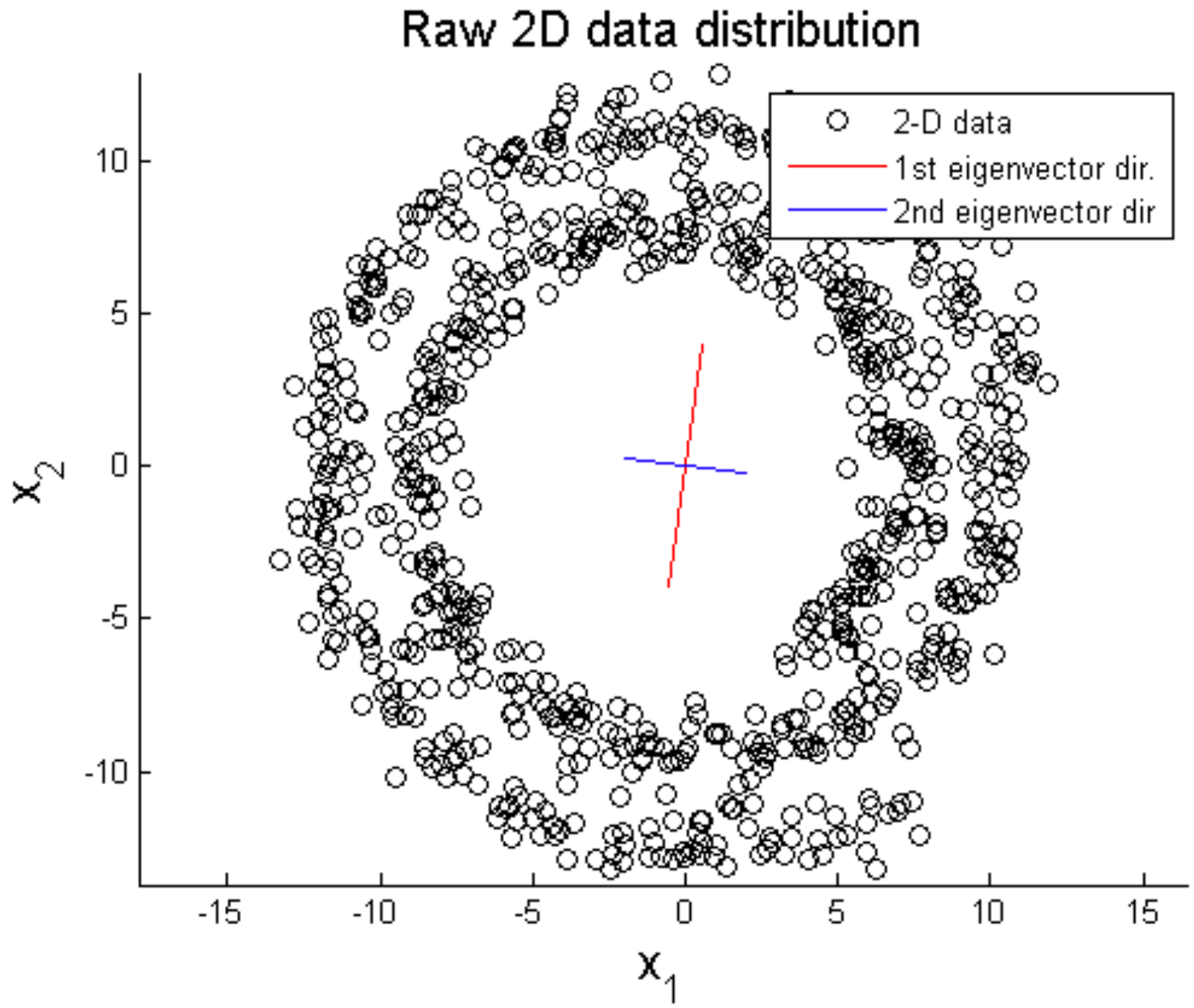
- Independent component analysis:
  $\mathbf{X} = \mathbf{A}\cdot\mathbf{S}$

Principal Axis vs. Regression Line



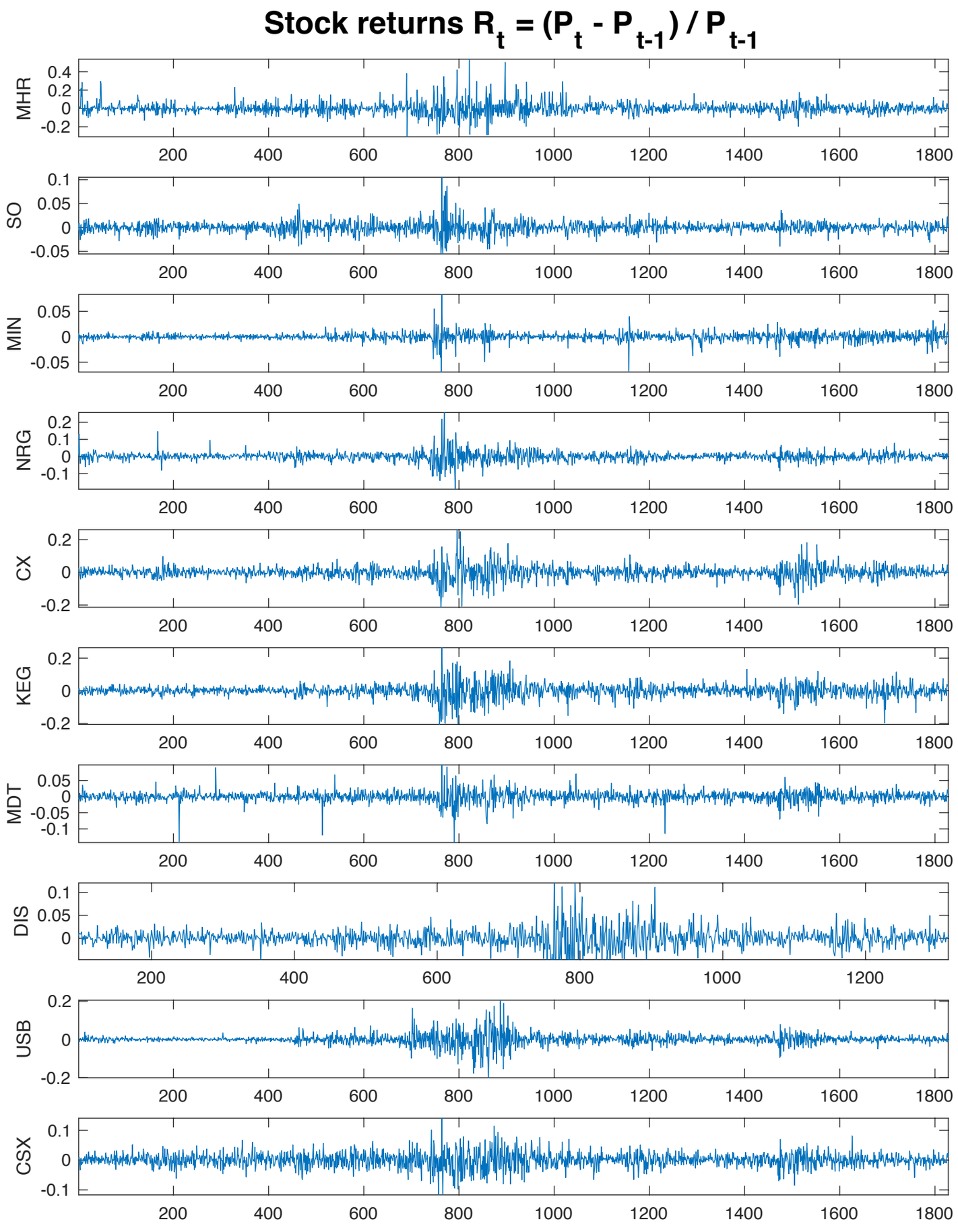
Nonlinear PCA...

- Projections onto nonlinear manifold instead...
- Easily kernelized

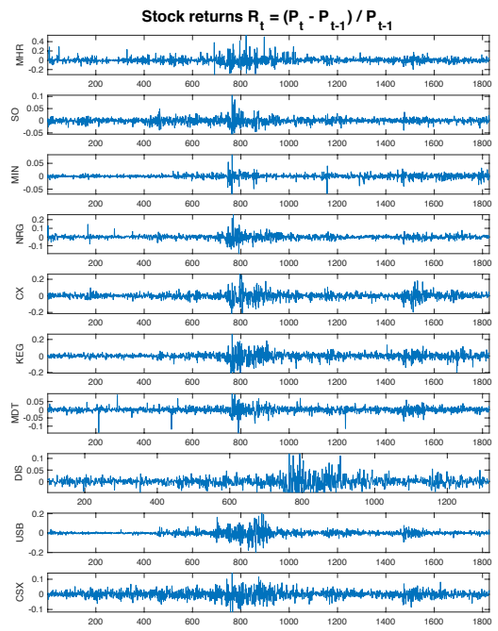


Underlying Factors?

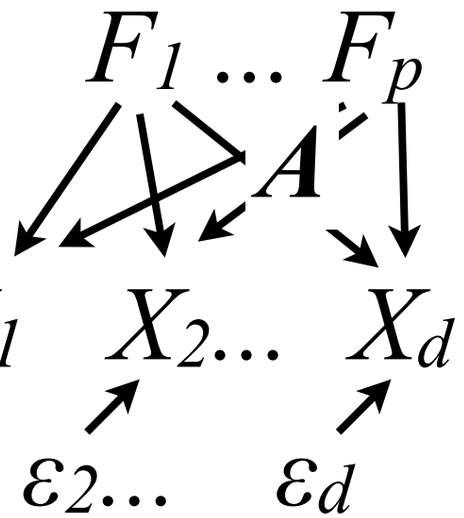
- Major information in the NYSE stock market?



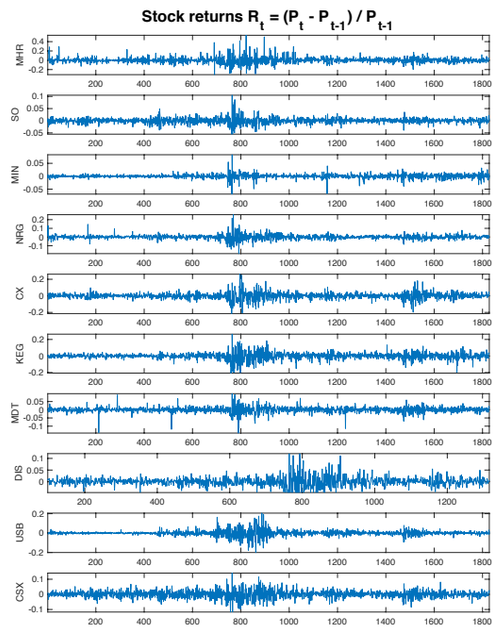
Factor Analysis



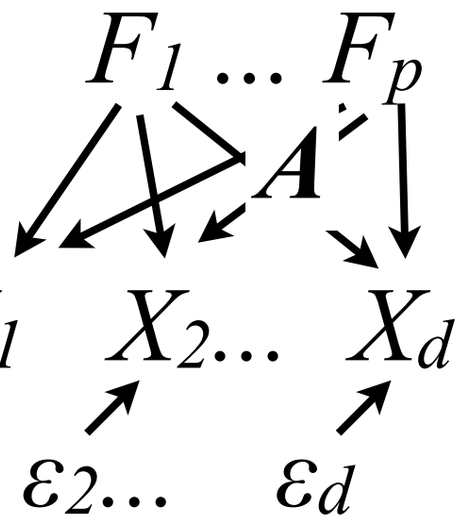
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<d$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
    - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of $\mathbf{A}$ & $\boldsymbol{F}$
- Estimation: MLE



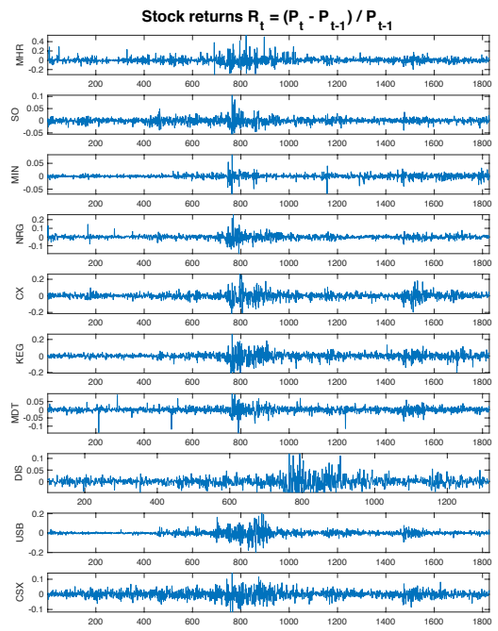
Factor Analysis



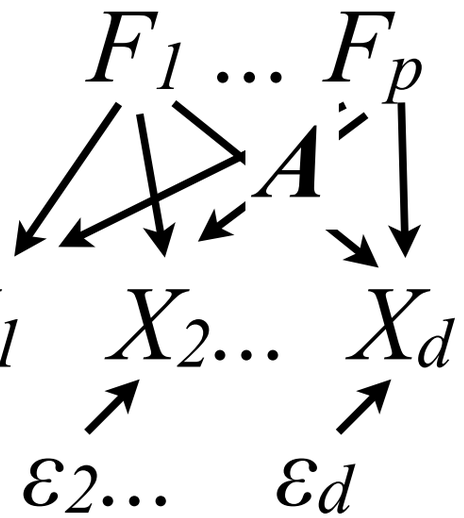
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<d$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
    - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of $\mathbf{A}$ & $\boldsymbol{F}$
- Estimation: MLE
  - $p_X(\mathbf{x})$?
  - Likelihood?



Factor Analysis



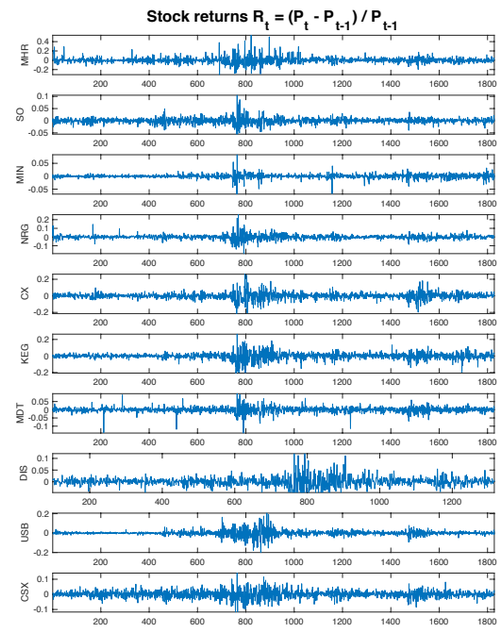
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<d$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
    - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of $\mathbf{A}$ & $\boldsymbol{F}$
- Estimation: MLE
  - $p_X(\mathbf{x})$?
  - Likelihood?

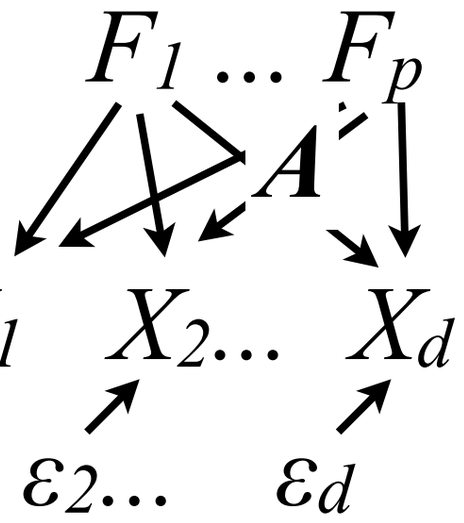


$$\hat{\mathbf{F}} = \mathbf{B}\mathbf{X},$$
where $\mathbf{B} = \mathbf{A}^{\intercal}(\mathbf{A}\mathbf{A}^{\intercal} + \boldsymbol{\Psi})^{-1}$,
because $\begin{bmatrix}\mathbf{X}\\ \mathbf{F}\end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix}\mathbf{A}\mathbf{A}^{\intercal} + \boldsymbol{\Psi} & \mathbf{A}\\ \mathbf{A}^{\intercal} & \mathbf{I}\end{bmatrix}\right)$.

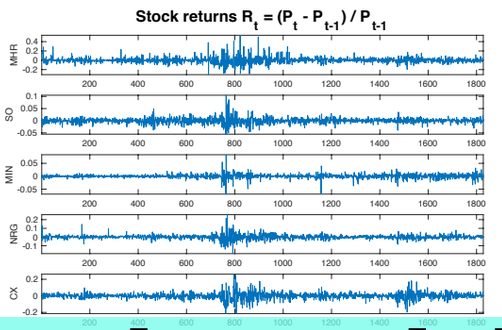
Factor Analysis

- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
    - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of $\mathbf{A}$ (*up to right orthogonal transformation*)
- Estimation: MLE



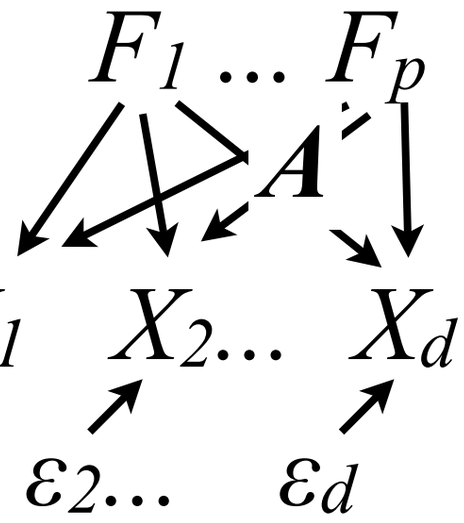


Factor Analysis



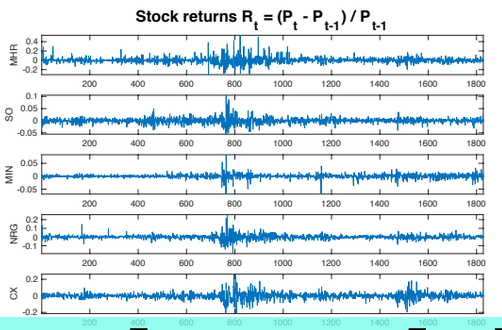
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - E[$\boldsymbol{F}$]=$\mathbf{0}$; Cov[$\boldsymbol{F}$]=I.
  - Cov[$\boldsymbol{\varepsilon}$]= $\boldsymbol{\Psi}$, which is diagonal.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \vdots & \vdots & \vdots & \vdots \\ a_{10,1} & a_{10,2} & a_{10,3} & a_{10,4} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{10} \end{bmatrix}$$



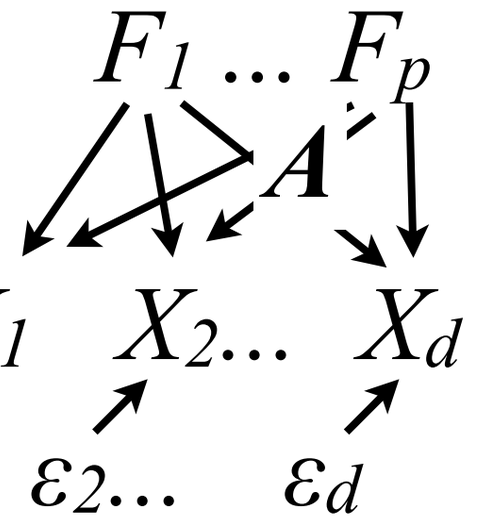
- Partial identifiability of $\mathbf{A}$ (*up to right orthogonal transformation*)
- Estimation: MLE

Factor Analysis



- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
  - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \vdots & \vdots & \vdots & \vdots \\ a_{10,1} & a_{10,2} & a_{10,3} & a_{10,4} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{10} \end{bmatrix}$$

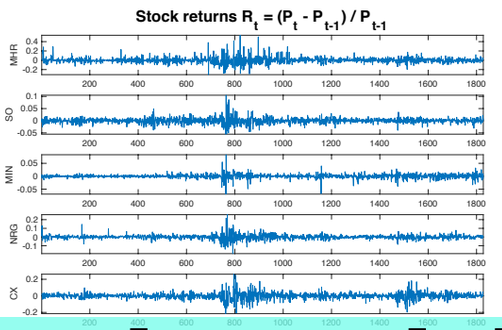


$$\mathbf{A}\mathbf{A}^{\intercal} + \boldsymbol{\Psi} = \mathbf{A}\mathbf{U}\mathbf{U}^{\intercal}\mathbf{A}^{\intercal} + \boldsymbol{\Psi},$$

where $\mathbf{U}$ is an orthogonal matrix.

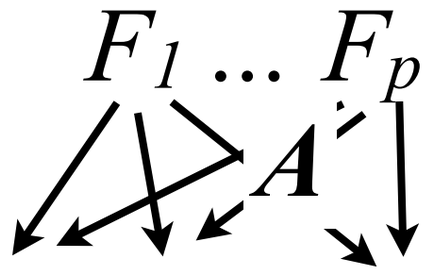
- Partial identifiability of $\mathbf{A}$ (*up to right orthogonal transformation*)
- Estimation: MLE

Factor Analysis



- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - E[$\boldsymbol{F}$]=$\mathbf{0}$; Cov[$\boldsymbol{F}$]=I.
  - Cov[$\boldsymbol{\varepsilon}$]= $\boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of $\mathbf{A}$ (*up to right orthogonal transformation*)
- Estimation: MLE

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \vdots & \vdots & \vdots & \vdots \\ a_{10,1} & a_{10,2} & a_{10,3} & a_{10,4} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{10} \end{bmatrix}$$



$$\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} = \mathbf{A}\mathbf{U}\mathbf{U}^\intercal\mathbf{A}^\intercal + \boldsymbol{\Psi},$$

where $\mathbf{U}$ is an orthogonal matrix.

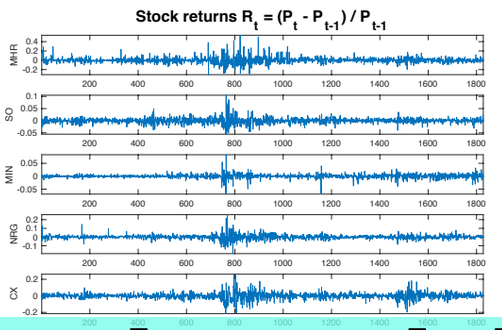
Estimated factors:

$$\hat{\mathbf{F}} = \mathbf{B}\mathbf{X},$$

where $\mathbf{B} = \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi})^{-1}$,

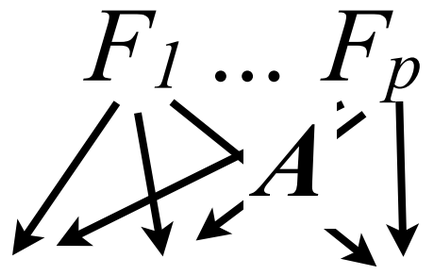
because $\begin{bmatrix} \mathbf{X} \\ \mathbf{F} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} & \mathbf{A} \\ \mathbf{A}^\intercal & \mathbf{I} \end{bmatrix}\right)$

Factor Analysis



- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
  - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \vdots & \vdots & \vdots & \vdots \\ a_{10,1} & a_{10,2} & a_{10,3} & a_{10,4} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{10} \end{bmatrix}$$



- Partial identifiability of $\mathbf{A}$ (*up to right orthogonal transformation*)
  - $p_X(\mathbf{x})$?
  - Likelihood?
- Estimation: MLE

$$\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} = \mathbf{A}\mathbf{U}\mathbf{U}^\intercal\mathbf{A}^\intercal + \boldsymbol{\Psi},$$

where $\mathbf{U}$ is an orthogonal matrix.

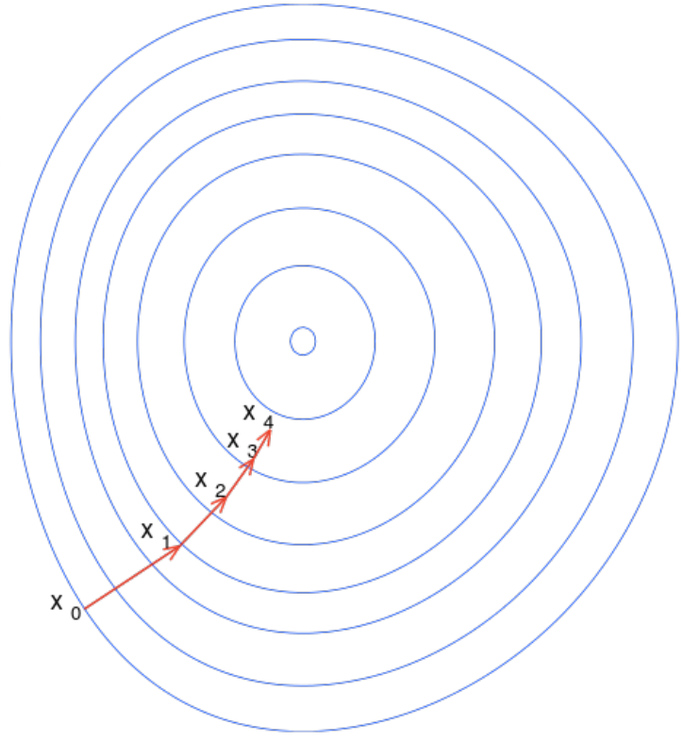
Estimated factors:

$$\hat{\mathbf{F}} = \mathbf{B}\mathbf{X},$$

where $\mathbf{B} = \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi})^{-1}$,

because $\begin{bmatrix} \mathbf{X} \\ \mathbf{F} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} & \mathbf{A} \\ \mathbf{A}^\intercal & \mathbf{I} \end{bmatrix}\right)$

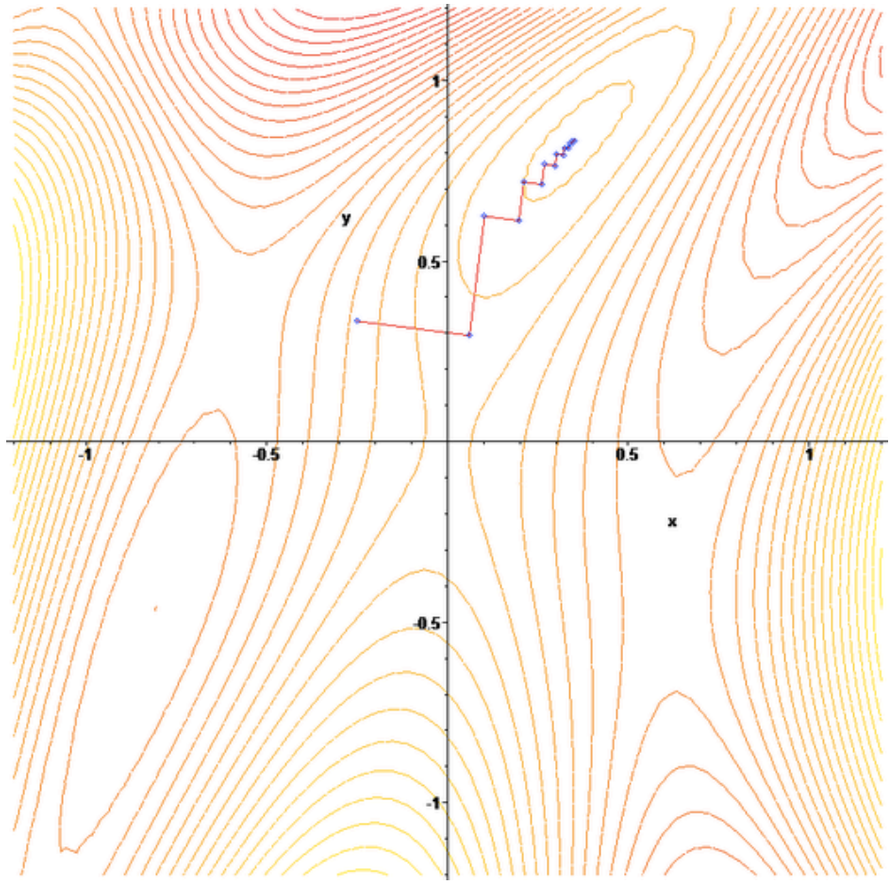
Elementary Optimization: Gradient Method

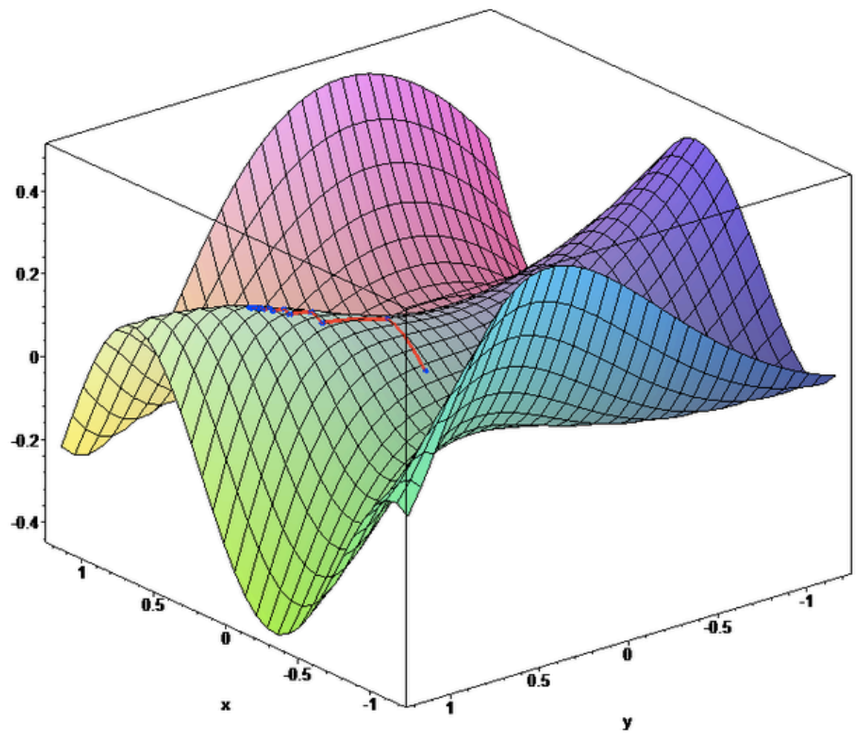


- A general approach to $\min_{x \in \mathbb{R}^n} f(x)$:

$$x^{new} = x^{old} - \gamma \bigtriangledown f(x^{old})$$

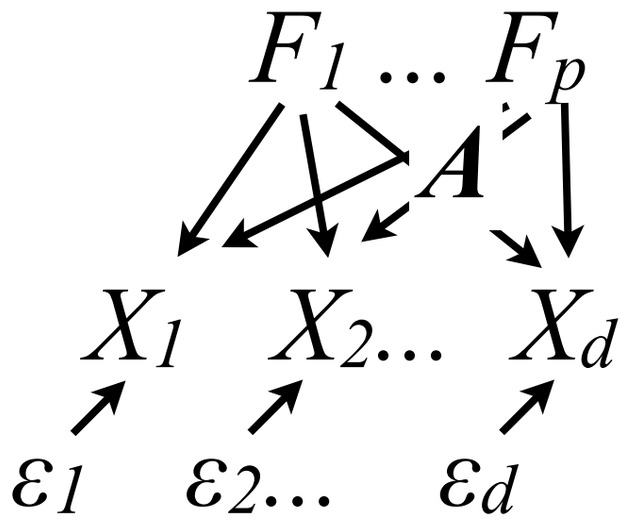
- Example: min $f(x_1, x_2) = \sin\left(\frac{1}{2}x_1^2 - \frac{1}{4}x_2^2 + 3\right)\cos(2x_1 + 1 - e^{x_2})$.

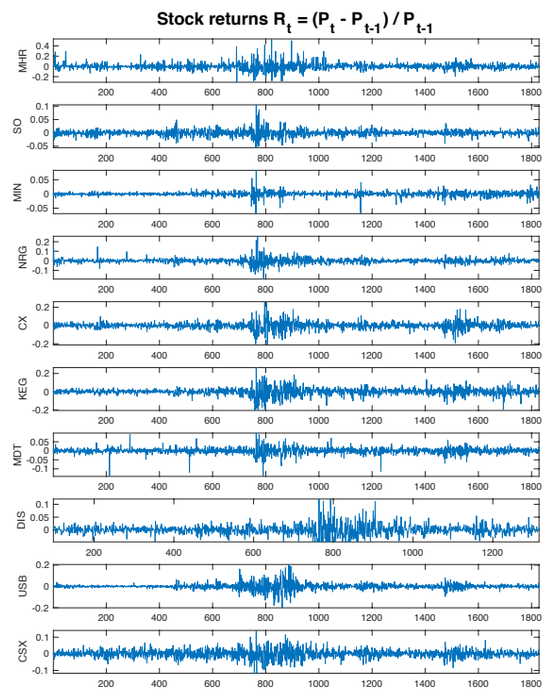




A nice demo is given on https://www.benfrederickson.com/numerical-optimization/

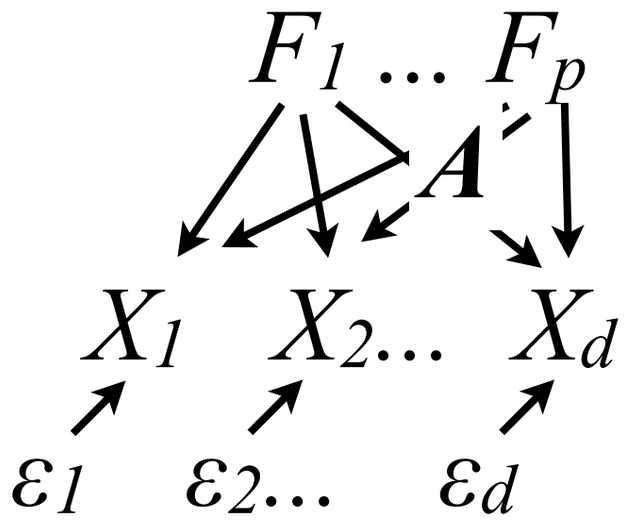
Factor Analysis on the Returns

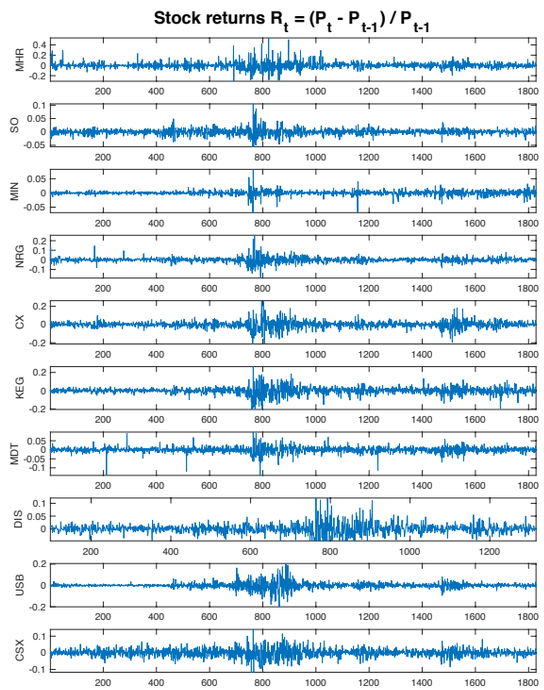




- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
  - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.

Factor Analysis on the Returns





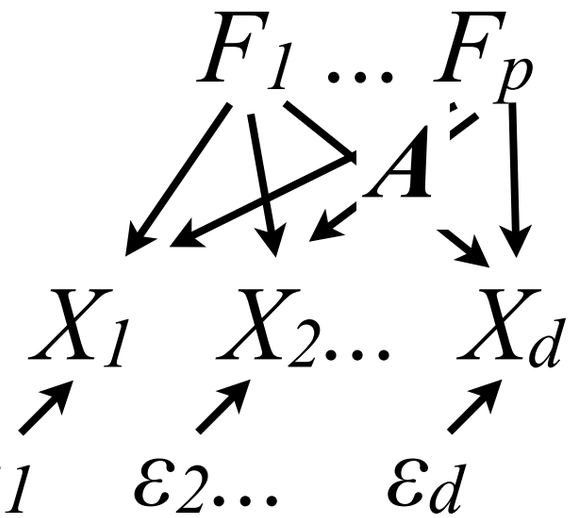
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
  - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.

$\hat{\mathrm{A}} =$

0.3656	0.0003	0.0089	0.1697
0.1175	0.7002	0.1001	0.2019
0.0833	0.1122	0.9837	0.0889
0.3142	0.3506	0.1060	0.6585
0.6793	0.2985	0.1211	0.1736
0.5529	0.2267	0.1164	0.4120
0.3310	0.4828	0.0586	0.1436
0.5881	0.5311	0.0819	0.1465
0.5598	0.3829	0.0210	0.0286
0.5908	0.4224	0.0516	0.1744

Factor Analysis

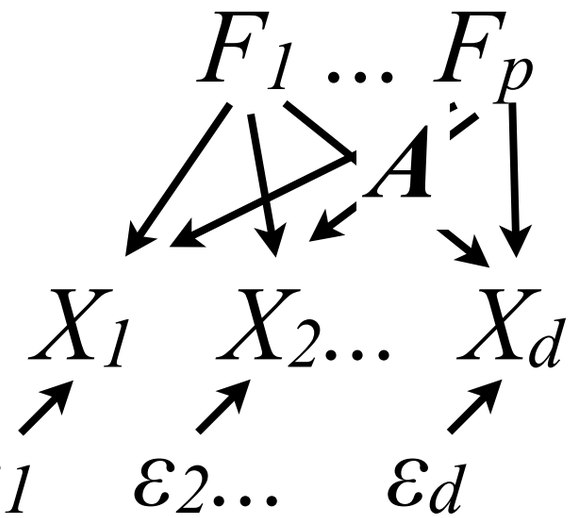
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.
    - $\mathrm{Cov}[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.



- Partial identifiability of A & F
- Estimation: MLE

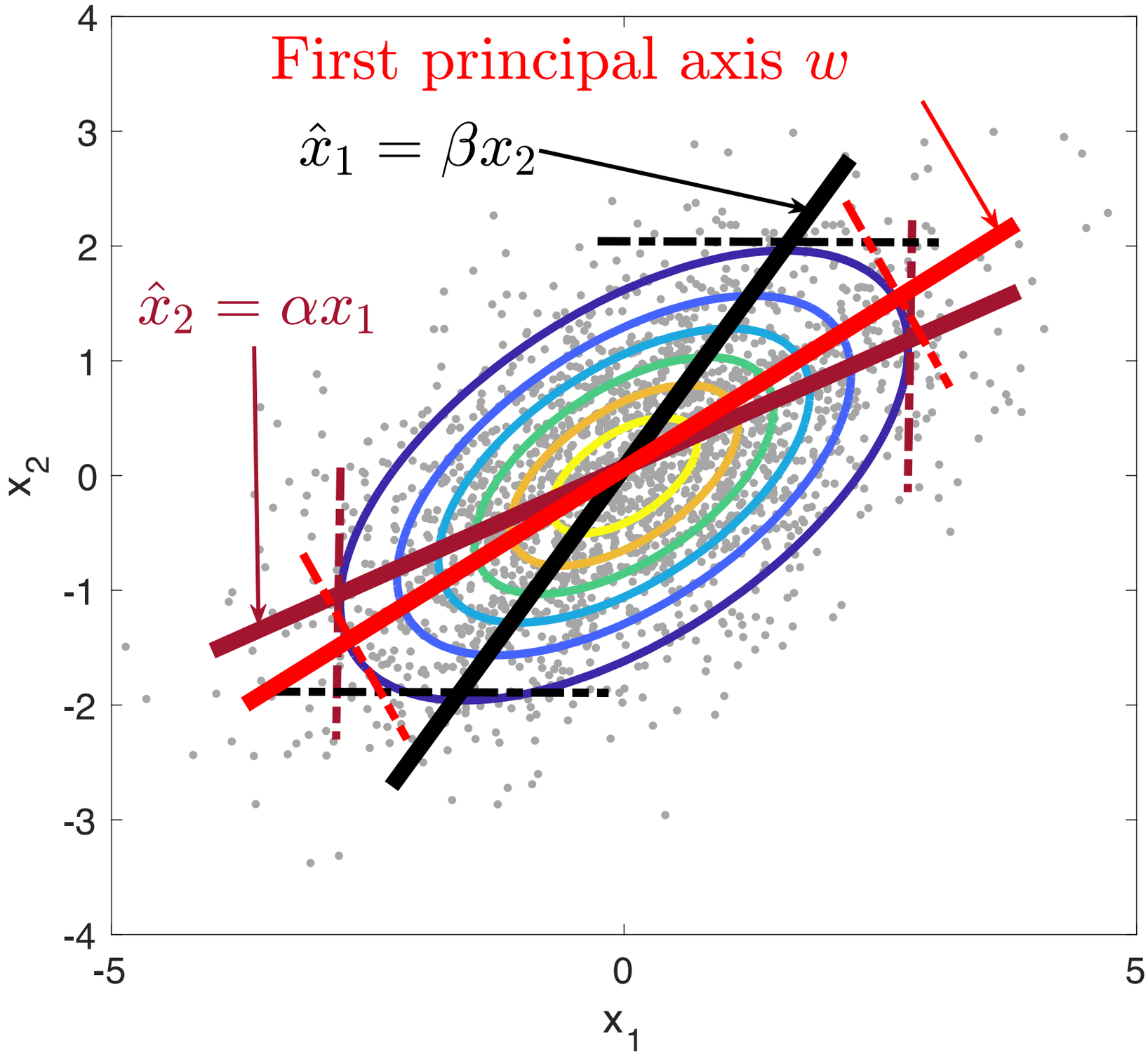
Factor Analysis

- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^T$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
    - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
    - $E[\boldsymbol{F}]=\mathbf{0}$; $Cov[\boldsymbol{F}]=I$.
    - $Cov[\boldsymbol{\varepsilon}]= \boldsymbol{\Psi}$, which is diagonal.
- Partial identifiability of A & F
- Estimation: MLE

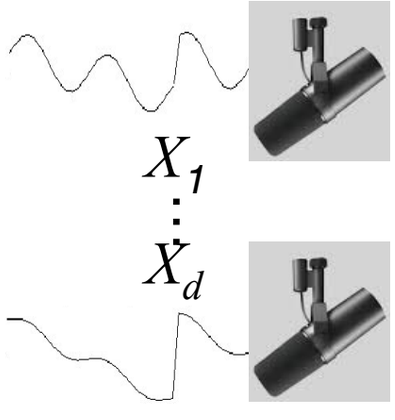


Relationship between FA and PCA?
-What if the noise terms are isotropic?
- What if we add (non)isotropic noise?

Recap: Principal Axis vs. Regression Line



Independent Component Analysis



observed signals

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

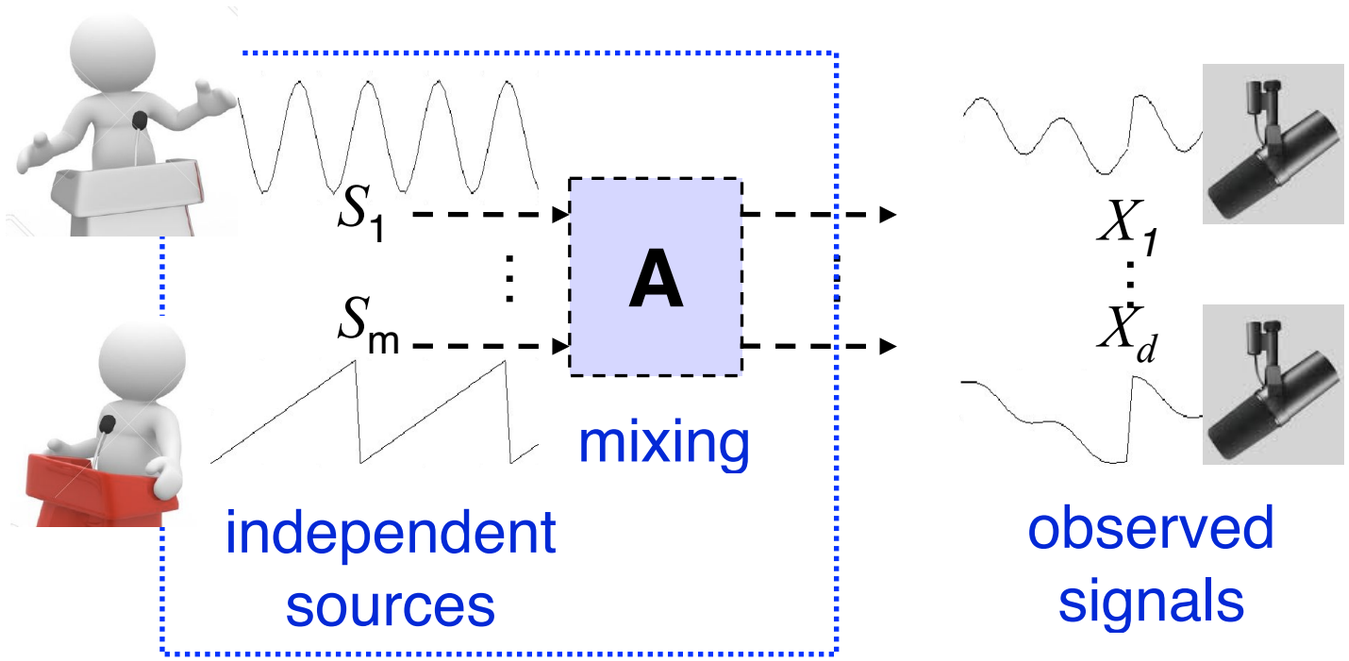
$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



unknown mixing system

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

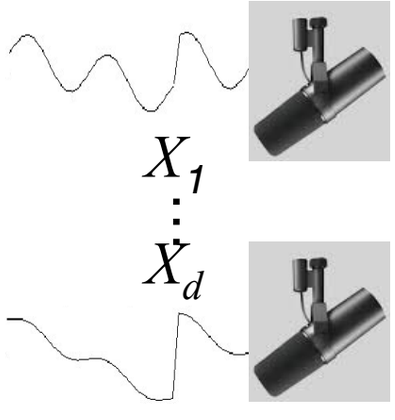
$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



observed signals

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

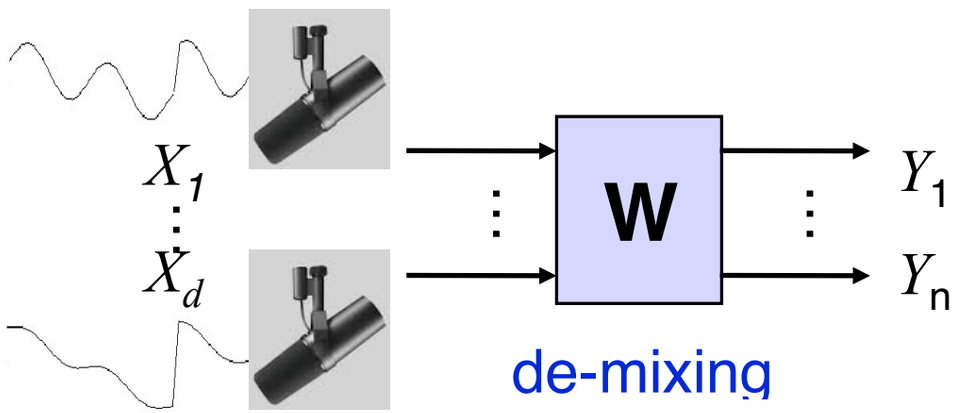
$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



$$\mathbf{X} = \mathbf{A}\cdot\mathbf{S} \qquad \mathbf{Y} = \mathbf{W}\cdot\mathbf{X}$$

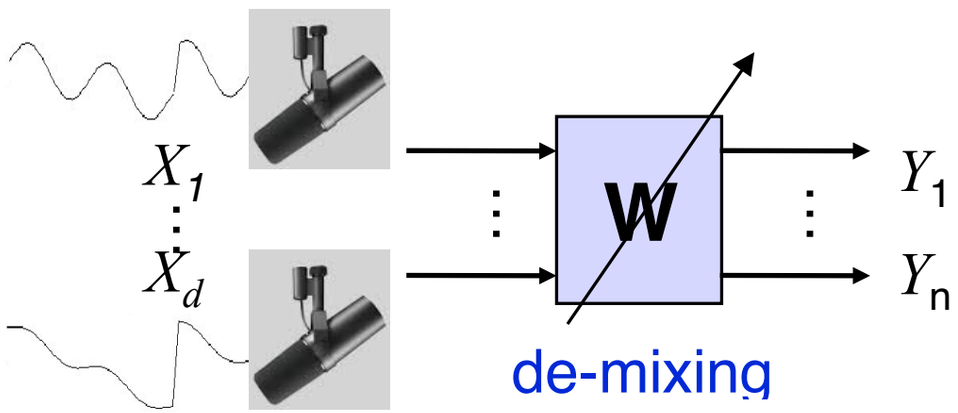
$$\begin{matrix} X_1 \\ X_2 \end{matrix}\begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



$$\mathbf{X} = \mathbf{A}\cdot\mathbf{S} \qquad \mathbf{Y} = \mathbf{W}\cdot\mathbf{X}$$

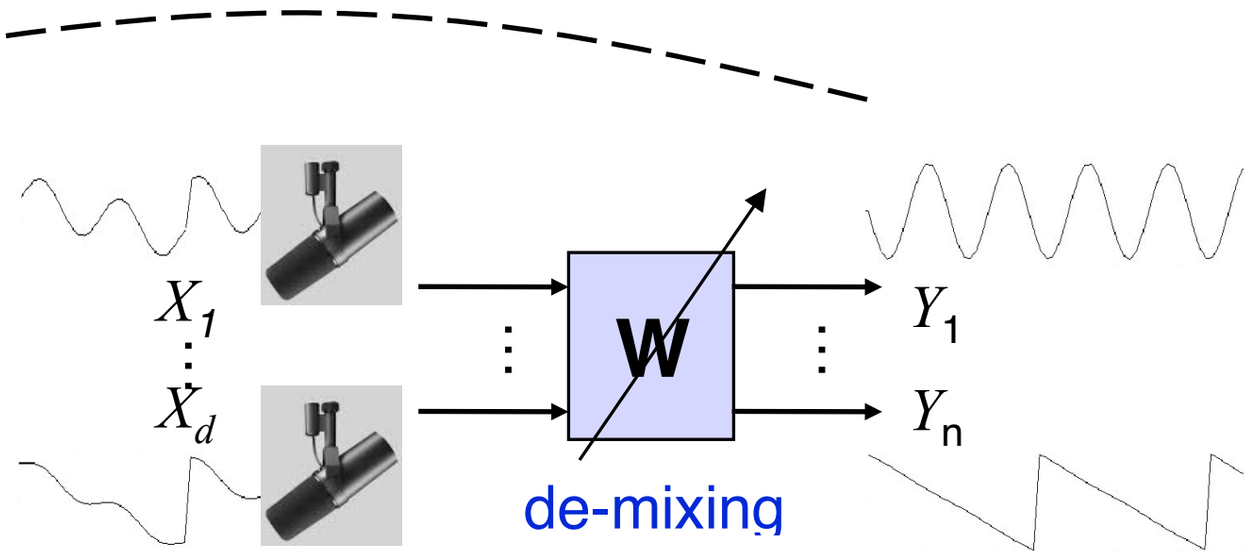
$$\begin{matrix} X_1 \\ X_2 \end{matrix}\begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



$$\mathbf{X} = \mathbf{A}\cdot\mathbf{S} \qquad \mathbf{Y} = \mathbf{W}\cdot\mathbf{X}$$

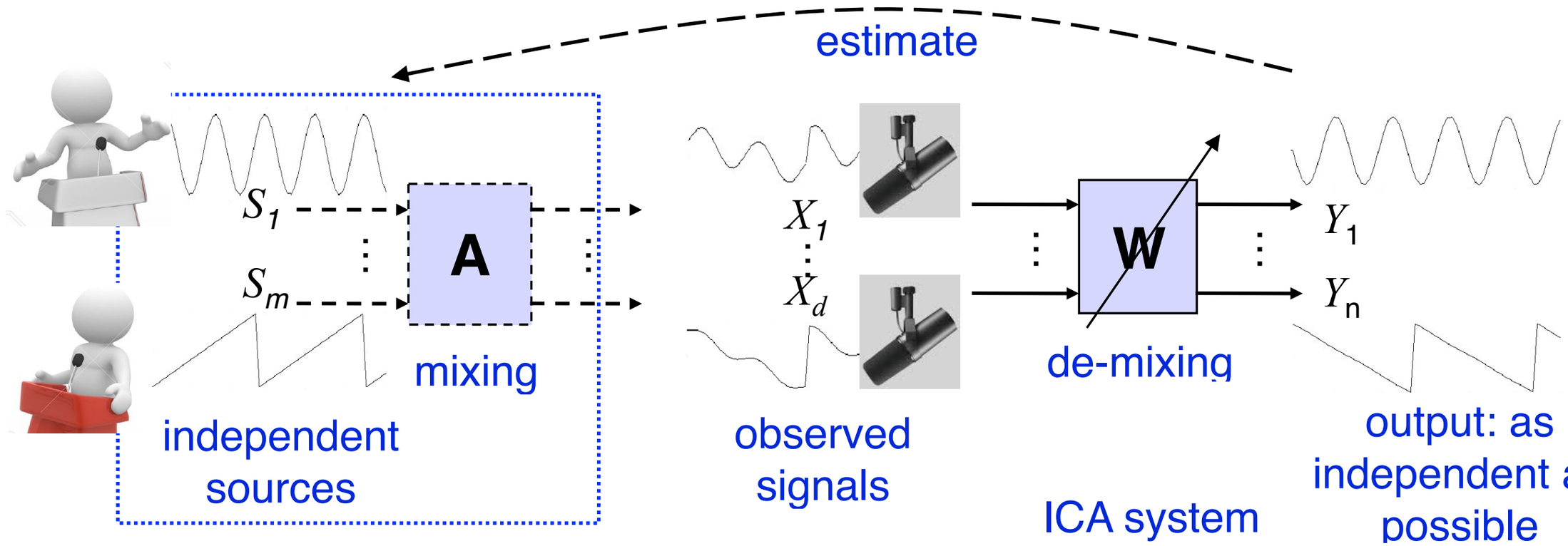
$$\begin{matrix} X_1 \\ X_2 \end{matrix}\begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

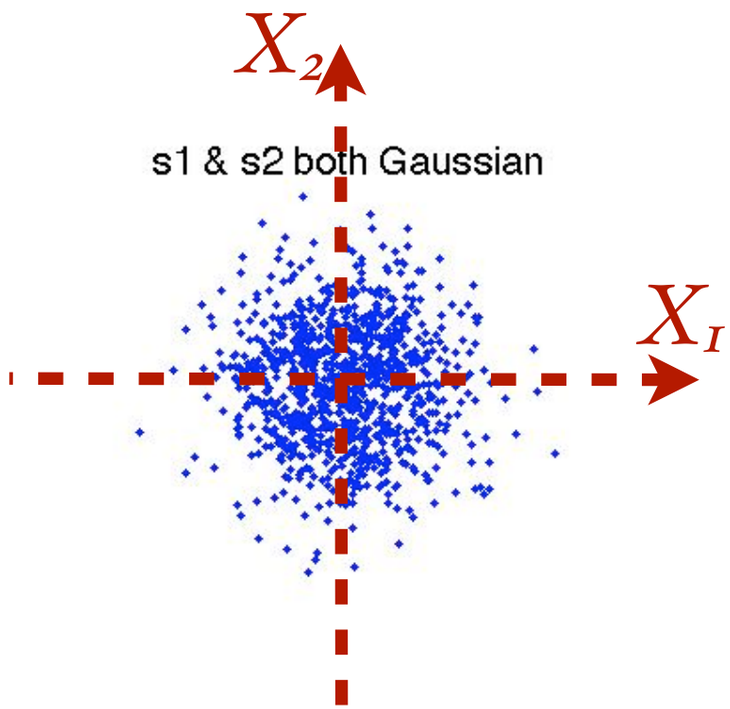
Then A can be estimated up to column scale and permutation indeterminacies

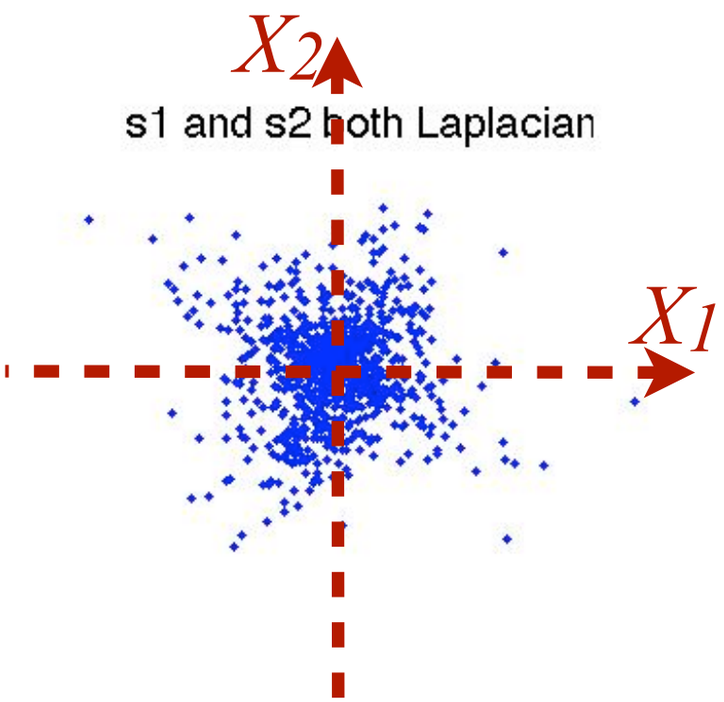
Hyvärinen et al., Independent Component Analysis, 2001

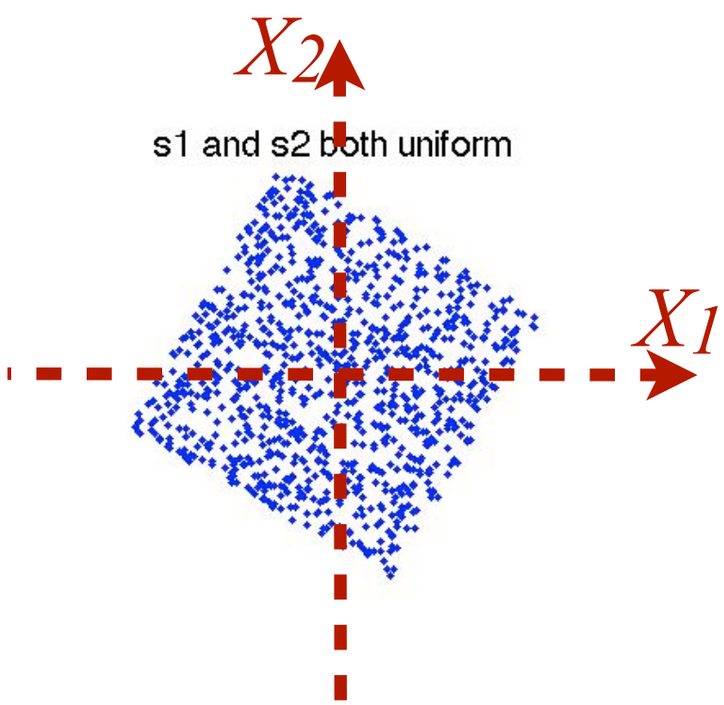
Intuition: Why ICA works?

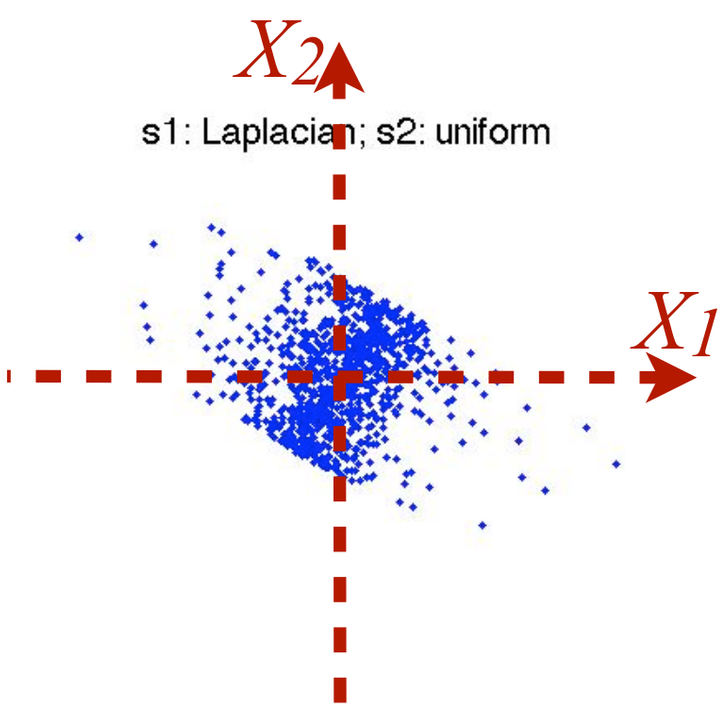
- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent

  - How to achieve the independence?





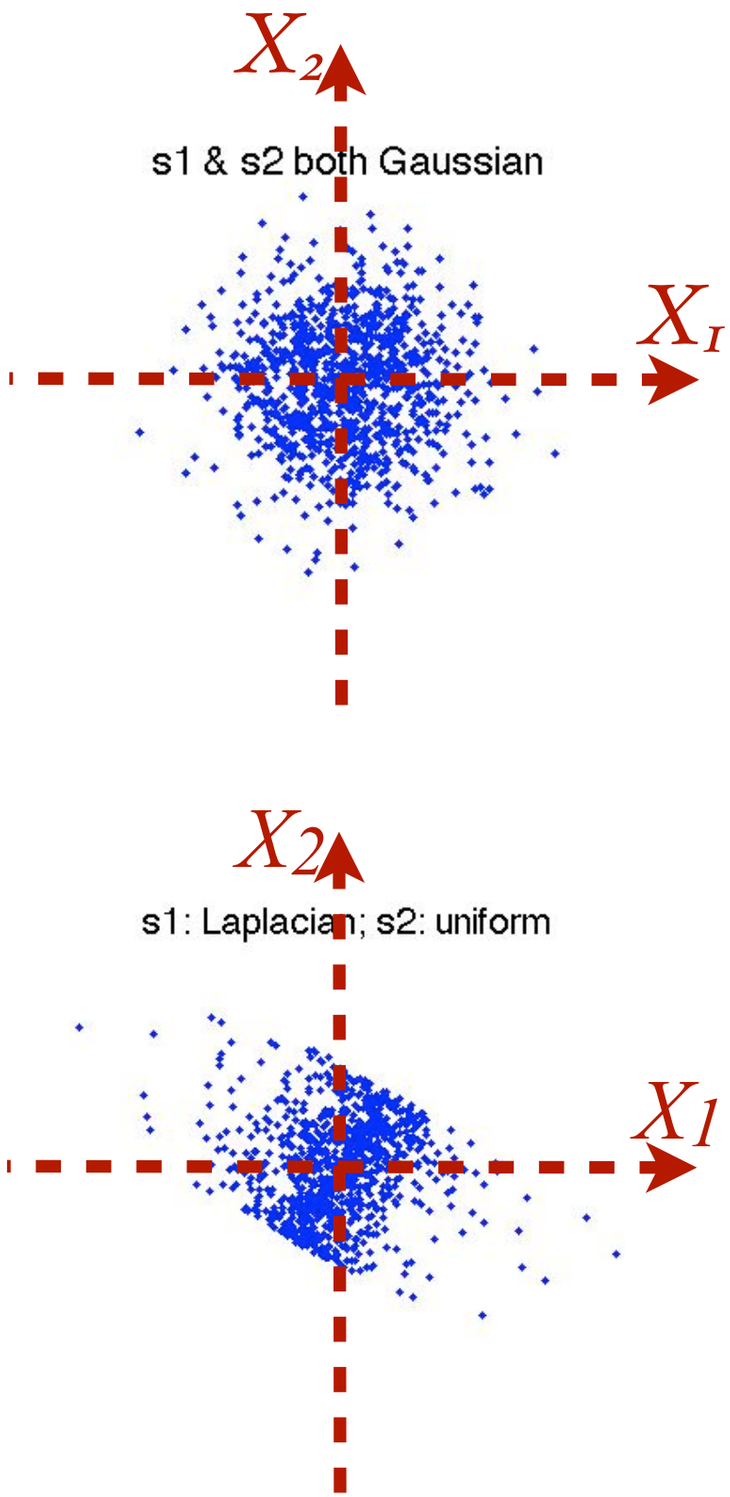


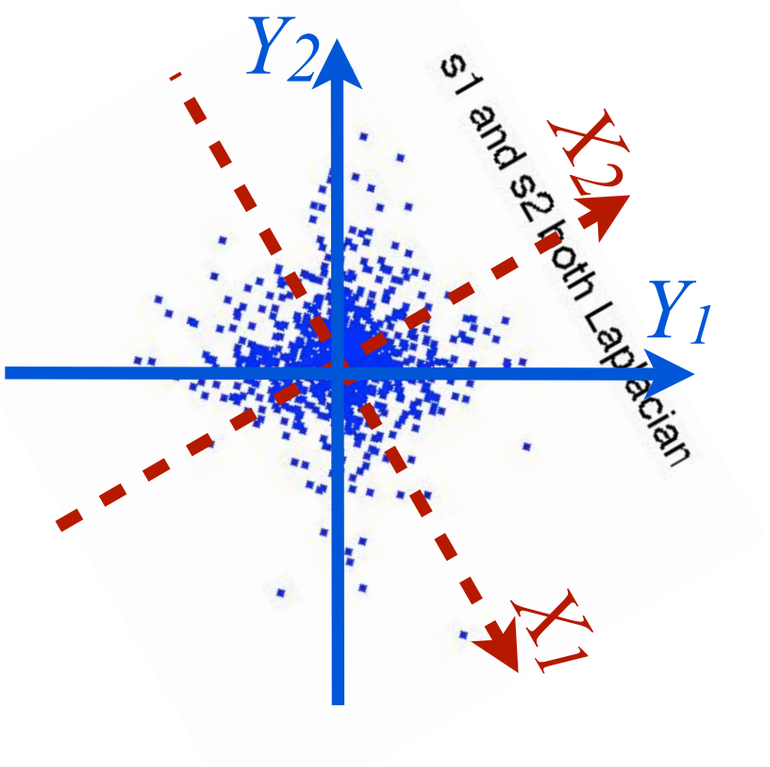


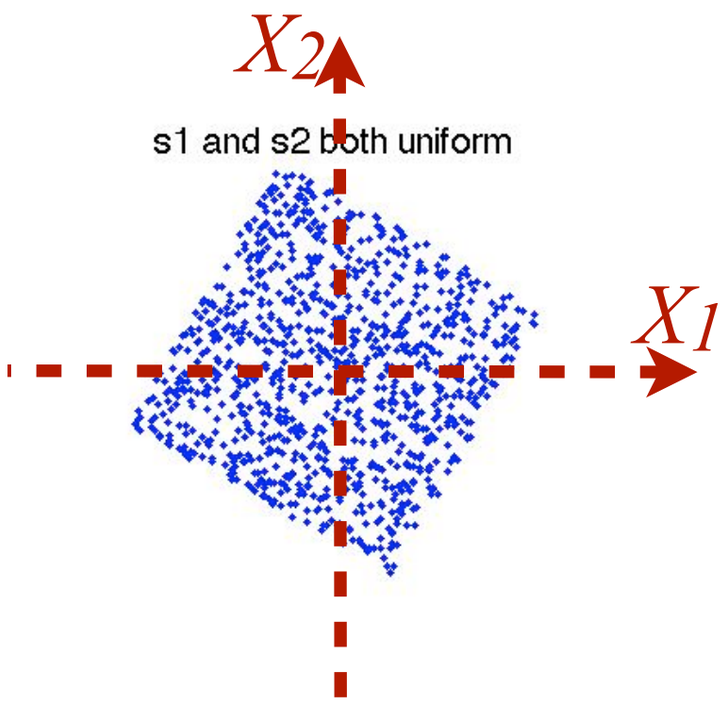
Intuition: Why ICA works?

- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent

  - How to achieve the independence?



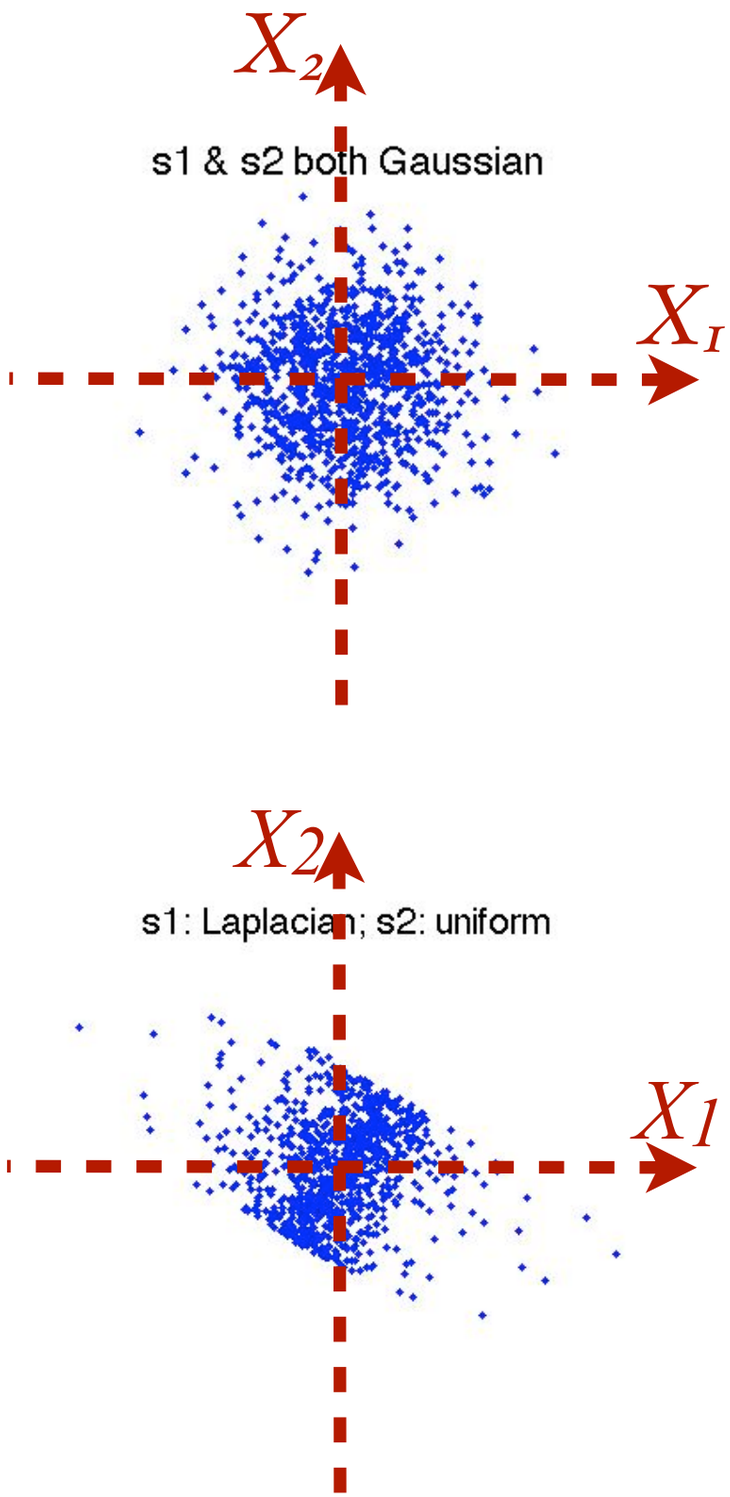


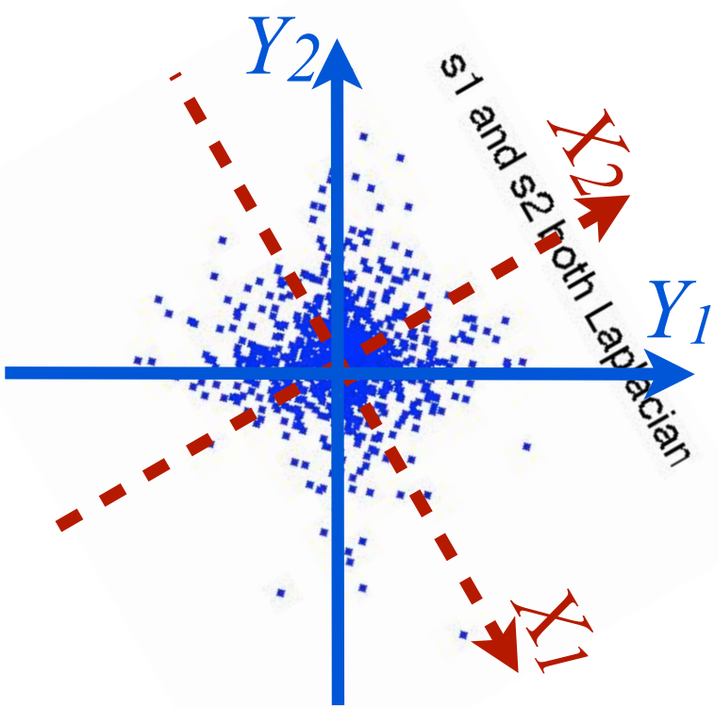


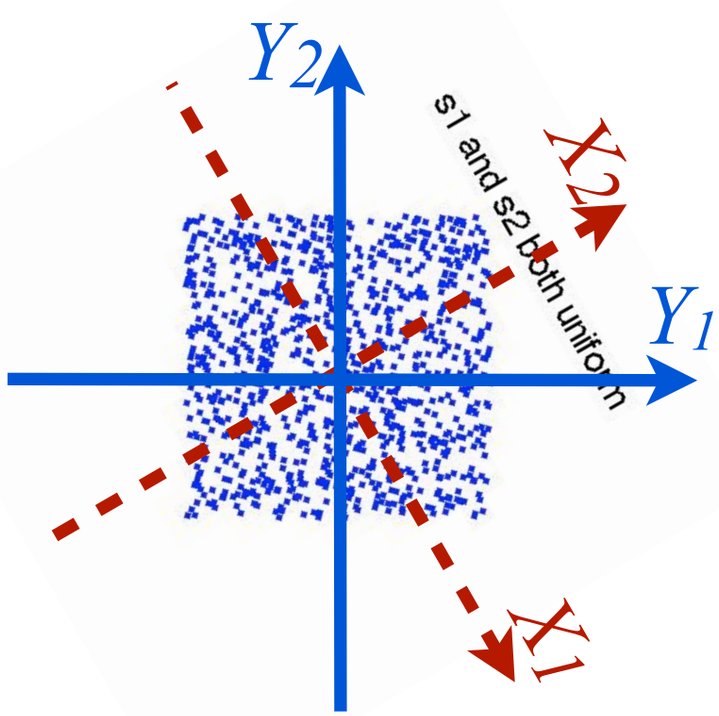
Intuition: Why ICA works?

- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent

  - How to achieve the independence?



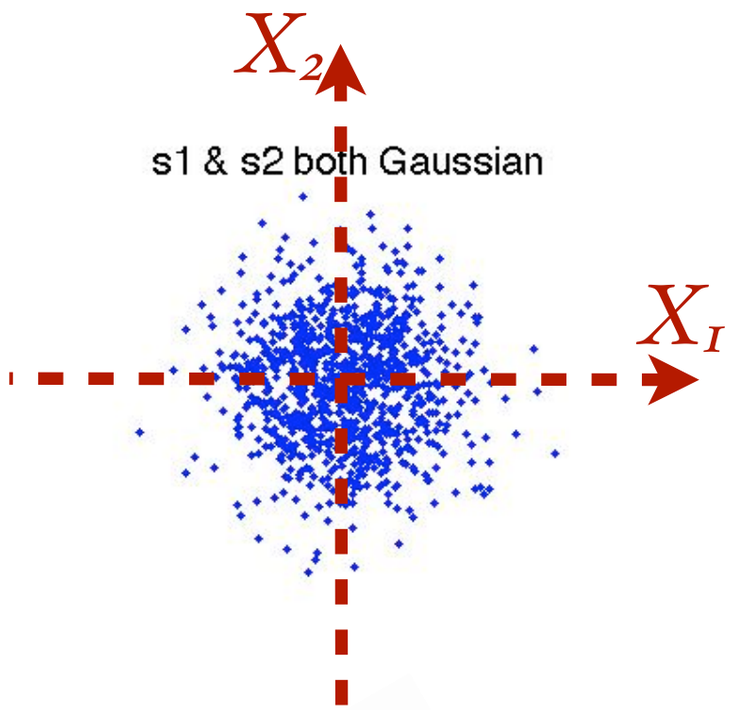


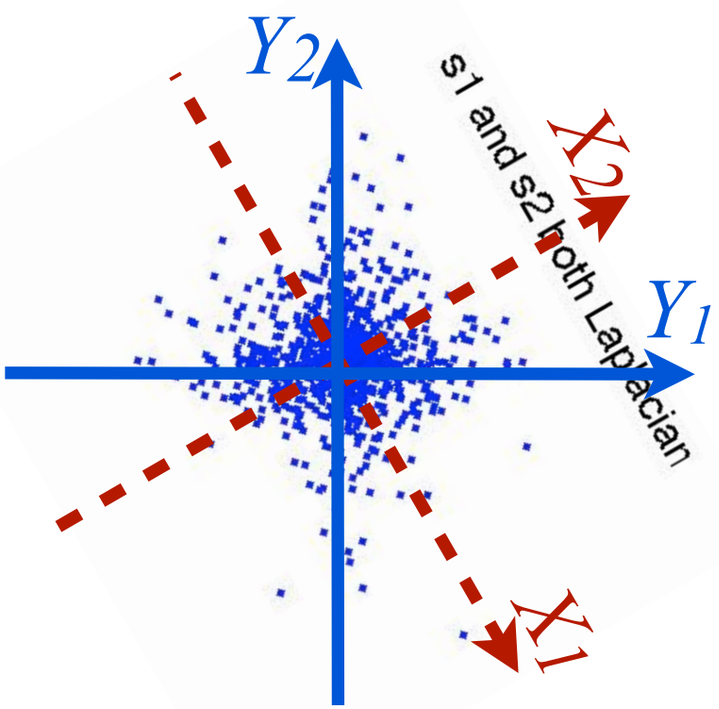


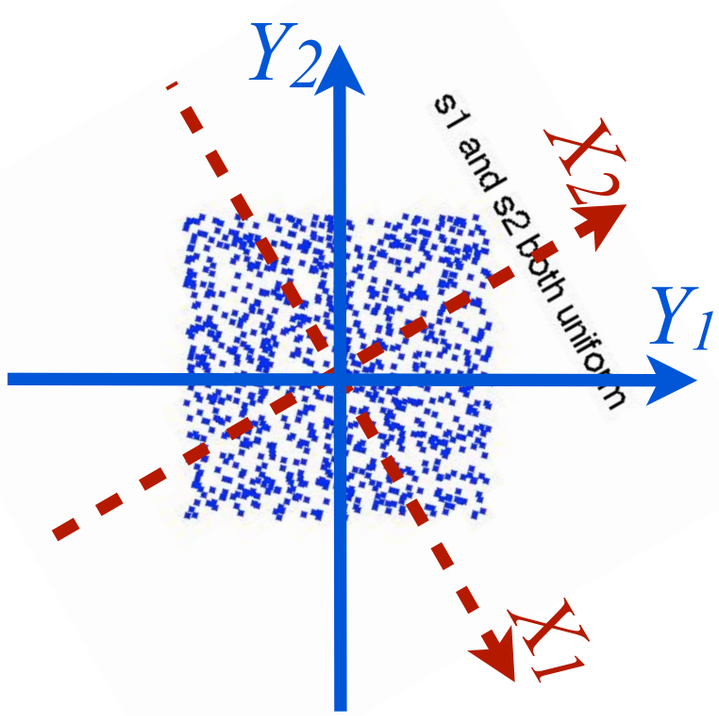
Intuition: Why ICA works?

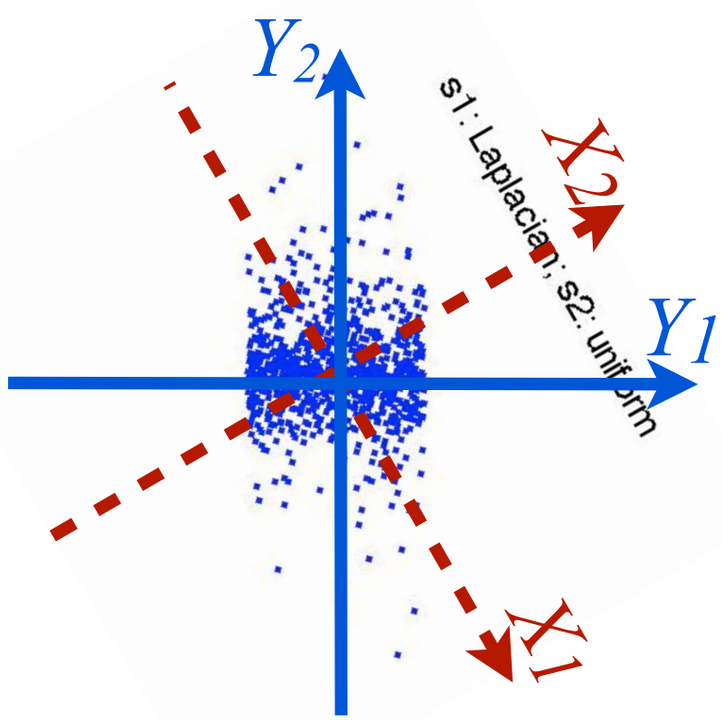
- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent

  - How to achieve the independence?









How ICA works? By Maximum Likelihood

- From a maximum likelihood perspective

$$\mathbf{X} = \mathbf{A}\cdot\mathbf{S}$$

$$\mathbf{Y} = \mathbf{W}\cdot\mathbf{X}$$

$$p_{\mathbf{S}} = \Pi_{i=1}^{d} p_{S_i}$$

$$\Rightarrow p_{\mathbf{X}} = \Pi_{i=1}^{d} p_{S_i}(W_i^{\intercal}\mathbf{X})/|\mathbf{A}|$$

$$\Rightarrow \sum_{t=1}^{n} \log p_{\mathbf{X}}(\,\mathbf{x}_t\,) = \sum_{t=1}^{n}\sum_{i=1}^{d} \log p_{S_i}(W_i^{\intercal}\,\mathbf{x}_t\,) + \boldsymbol{n}\log|\mathbf{W}|$$

($\mathbf{x}_t$: the t-th point of $\mathbf{X}$.)

- To be maximized by the gradient-based method or natural-gradient based method

How ICA works? By Mutual Information Minimization

- Mutual information $I(Y_1,\ldots,Y_d)$ is the Kullback-Leiber divergence from P_Y to $\prod_i P_{Yi}$:

$$I(Y_1,\ldots,Y_d) = \int \ldots \int p_{Y_1,\ldots,Y_d} \log \frac{P_{Y_1,\ldots,Y_d}}{p_{Y_1}\ldots p_{Y_d}} dy_1 \ldots dy_n$$

$$= \int \ldots \int p_{Y_1,\ldots,Y_d} \log P_{Y_1,\ldots,Y_d} dy_1 \ldots dy_d - \int p_{Y_1,\ldots,Y_d} \sum_{i=1}^{d} \log p_{Y_i} dy_i$$

$$= \sum_i H(Y_i) - H(Y)$$

$$= \sum_i H(Y_i) - H(X) - \log |\mathbf{W}| \qquad \text{because } \mathbf{Y} = \mathbf{WX}$$

- Nonnegative and zero iff Y_i are independent

- $H(\cdot)$: differential entropy--how random the variable is?

Hyvärinen et al., Independent Component Analysis...

How ICA works? Some Interpretation

- Some methods (e.g., FastICA, JADE) pre-whiten the data, and then aim to find a rotation, for which $|\mathbf{W}| = 1$

$$I(Y_1, ..., Y_d) = \sum_i H(Y_i) - H(X) - \log |\mathbf{W}| = \sum_i H(Y_i) + \text{const.}$$

- Minimizing $I \Leftrightarrow$ minimizing the entropies

- Given the variance, the Gaussian distribution has the largest entropy (among all continuous distributions)

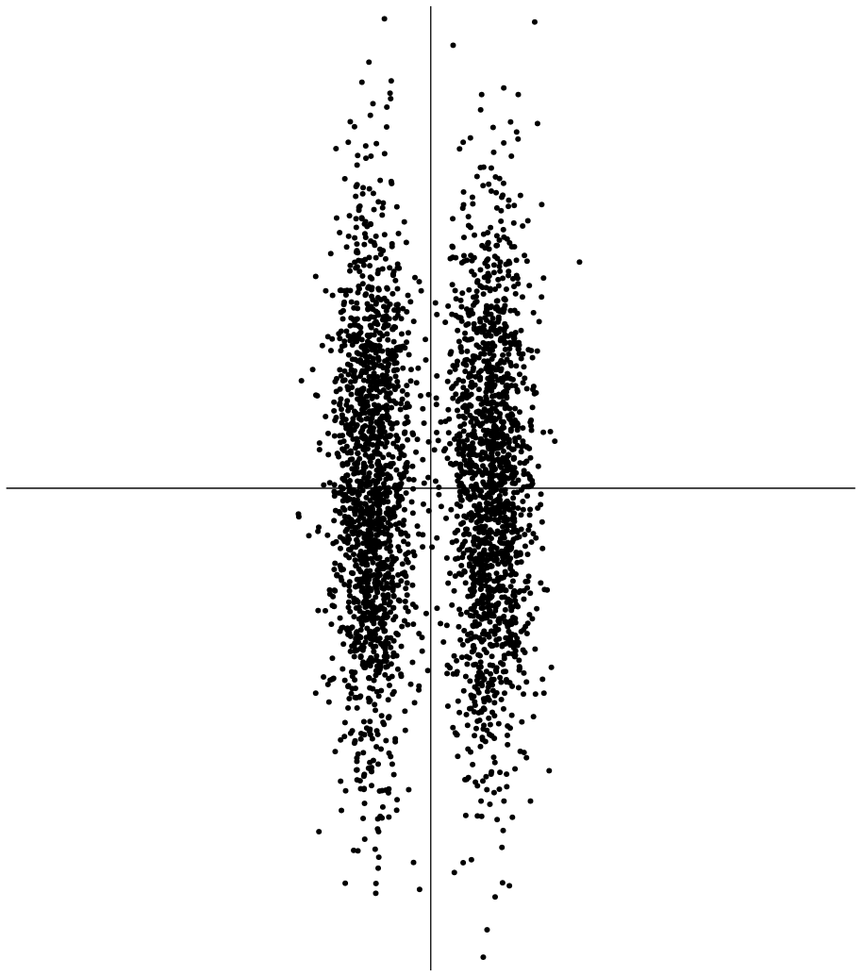
- Maximizing non-Gaussianity !

- FastICA adopts some approximations of negentropy of each output Y_i

Non-Gaussianity is Informative in the Linear Case

- Smaller entropy, more structural, more interesting

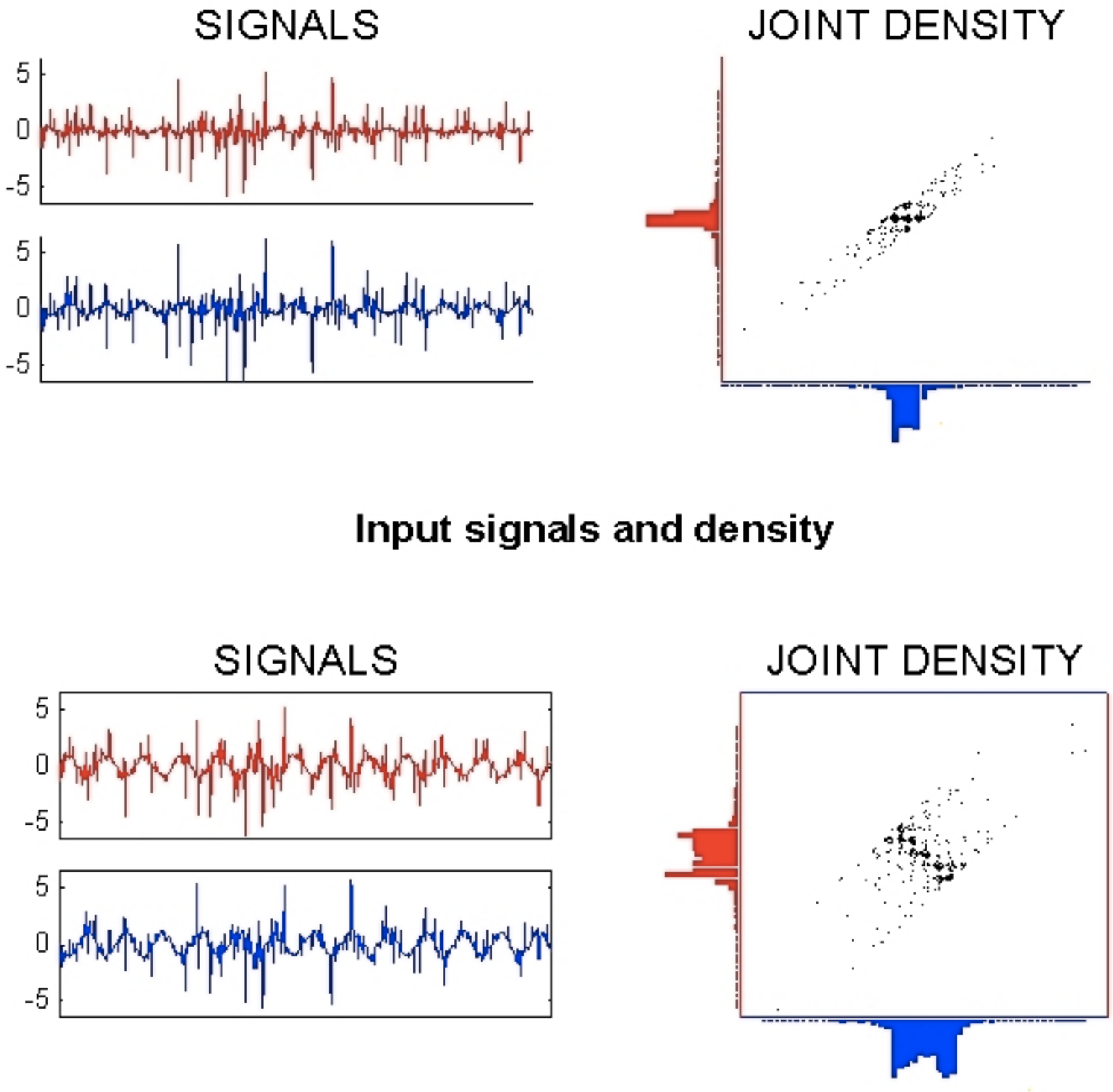
- "Purer" according to the central limit therom

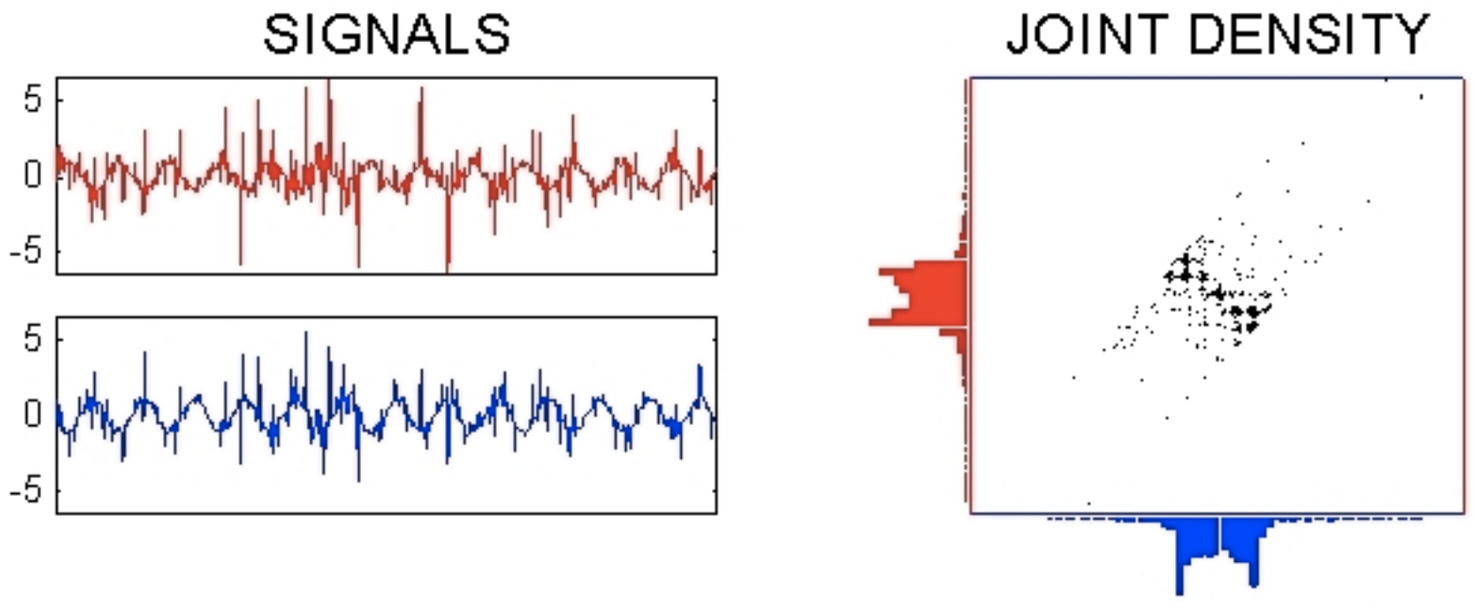


Which direction is more interesting?

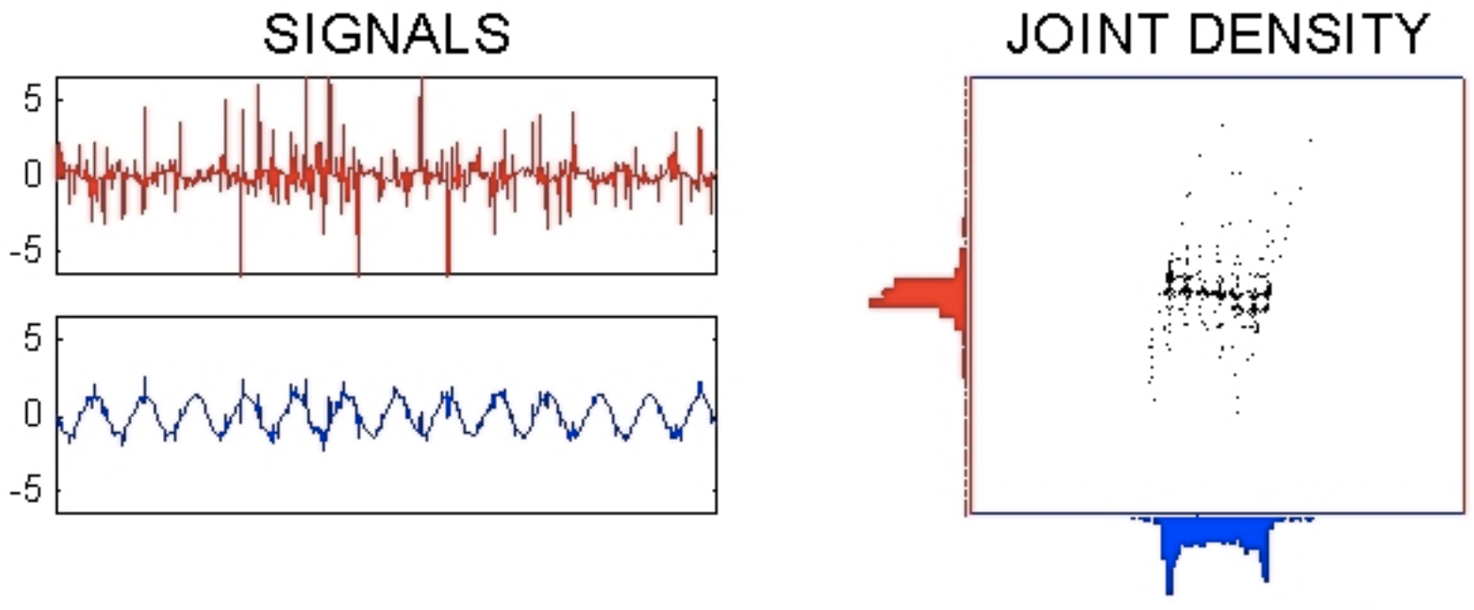
Hyvärinen et al., Independent Component Analysis, 2001

A Demo of the ICA Procedure

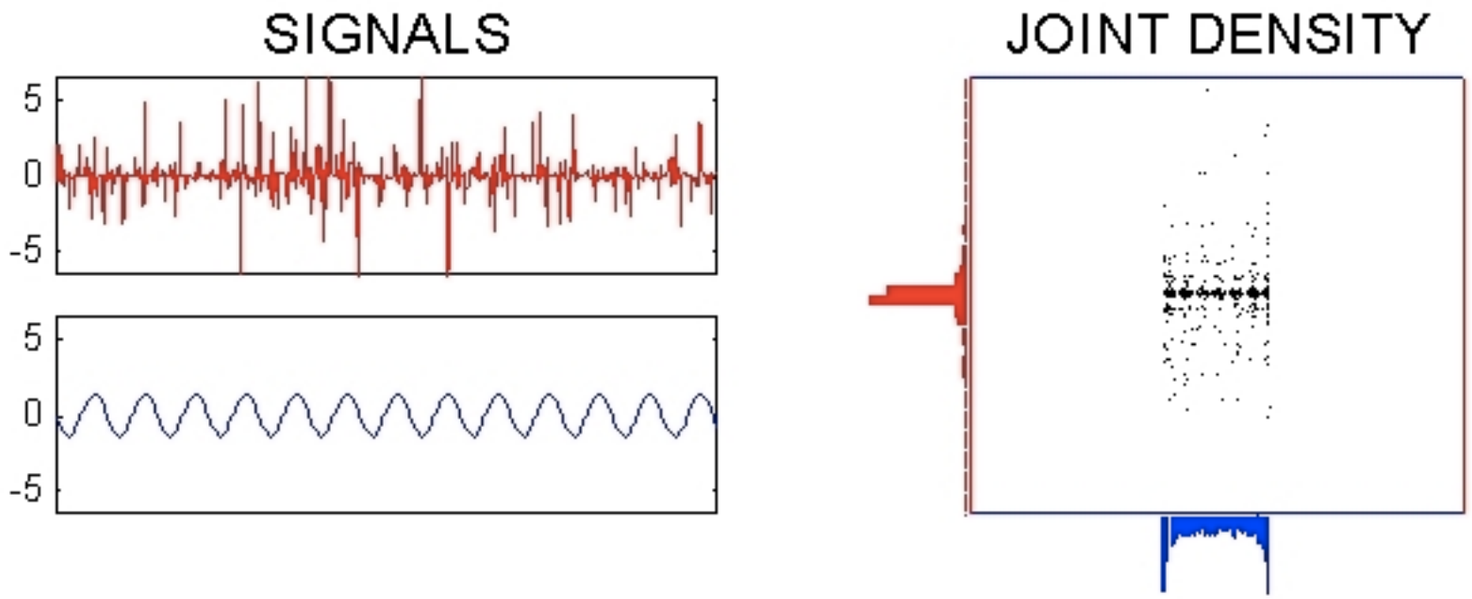




Separated signals after 1 step of FastICA



Separated signals after 3 steps of FastICA



Separated signals after 5 steps of FastICA

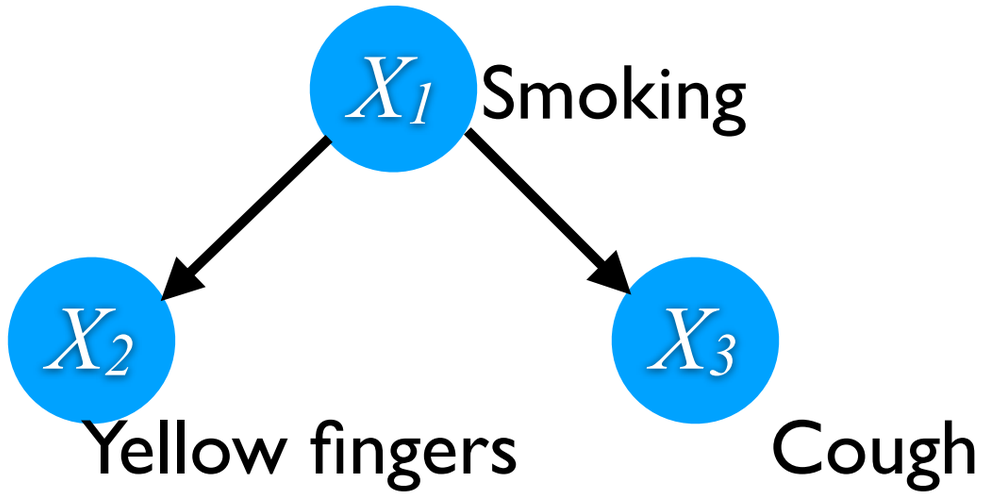
Summary 1: Multivariate Analysis

- Supervised vs. unsupervised learning
- Fewer representative data points vs. fewer representative dimensions
- PCA vs. linear regression
- Factor analysis vs. PCA
- ICA vs. factor analysis
- ICA vs. PCA

Identification of Causal Effects & Counterfactual Inference

- Problem definition
- Backdoor criterion and front door criterion
- Propensity score
- Counterfactual inference

Three Types of Problems in Current AI

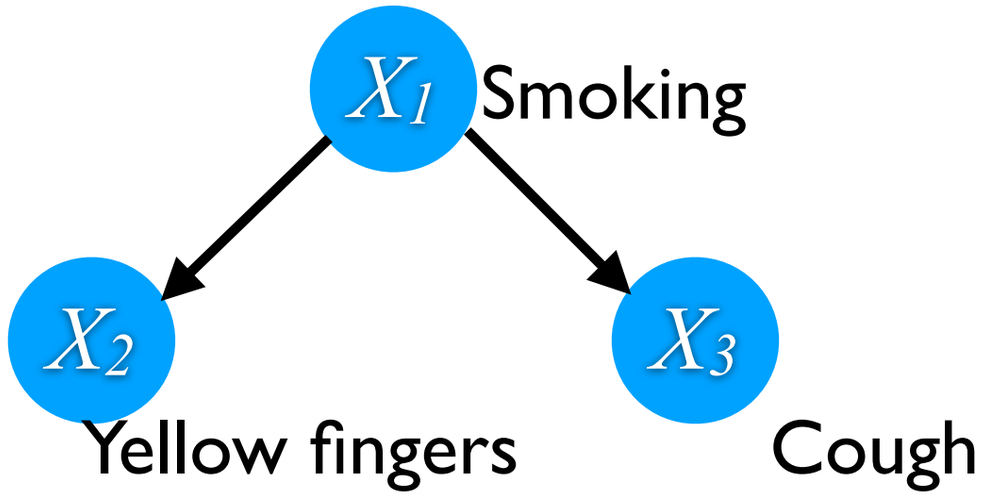


- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?
- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?
- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

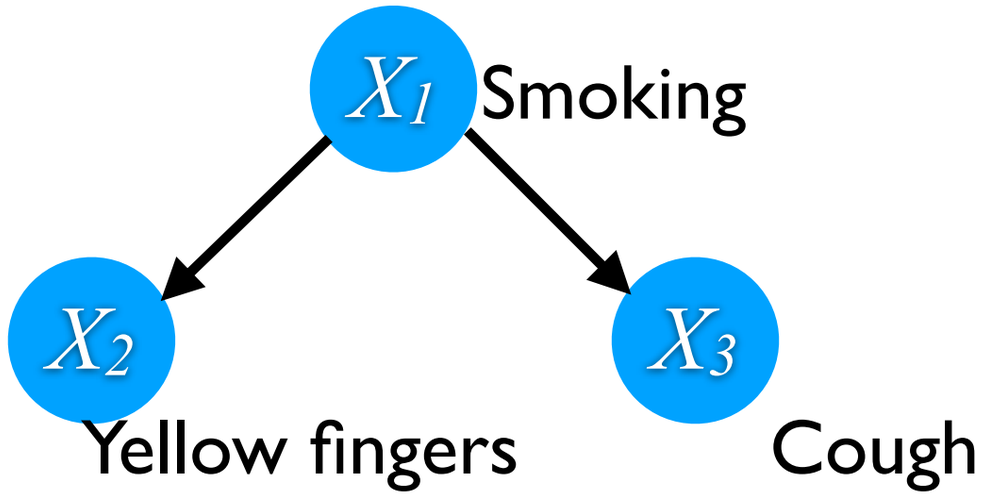
- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

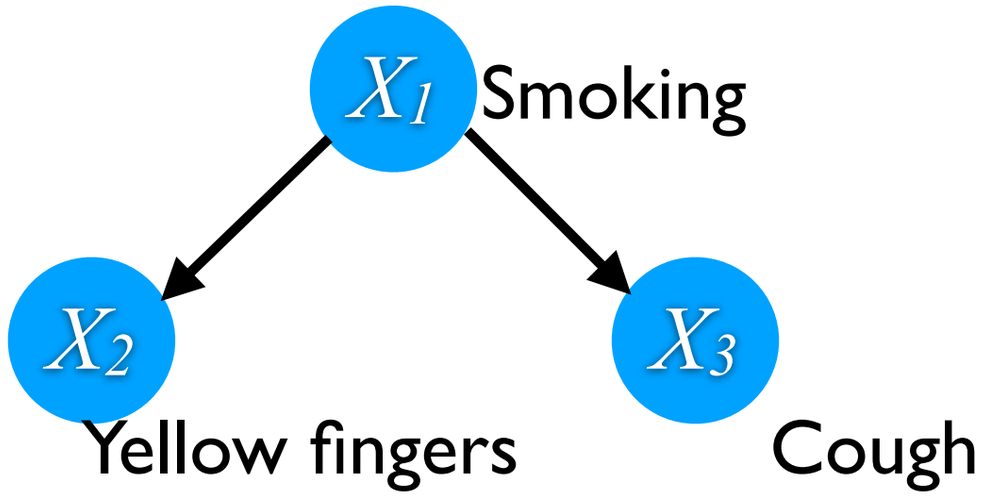
$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

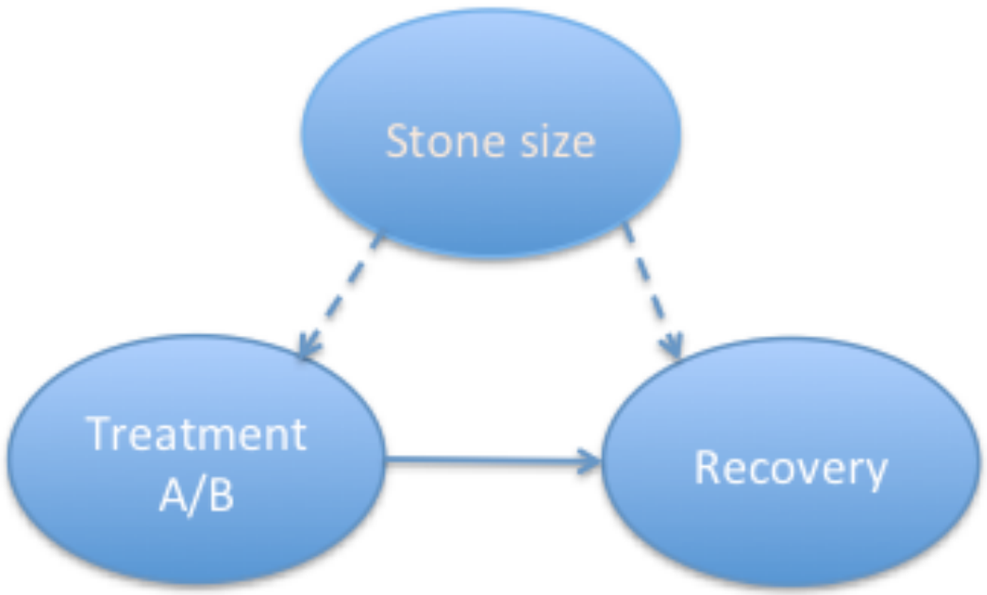
- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{\,X2=1} \mid X2 = 0, X3 = 1)$$

Identification of Causal Effects

P(Recovery | do(Treatment=A)) ?

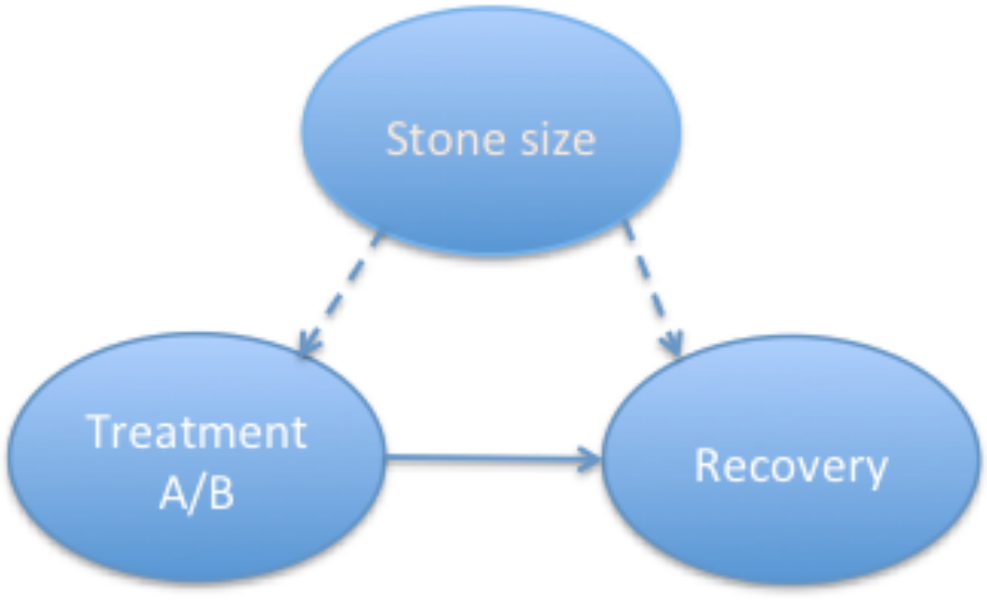
- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



- Usually expensive or impossible to do!

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



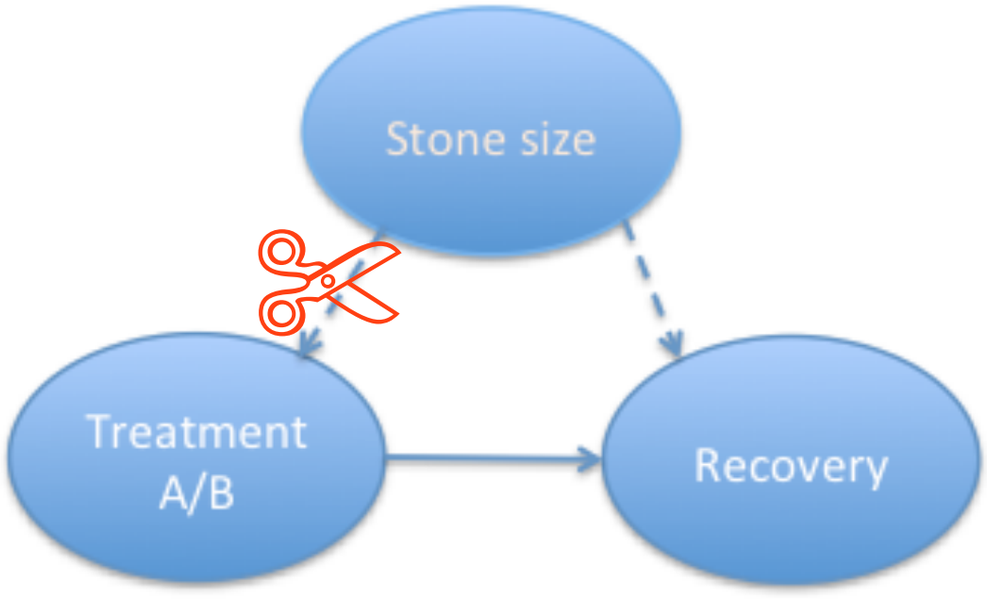
$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

$$P(R\,|\,do(T)) = \sum_S P(R\,|\,T,S)P(S)$$

conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



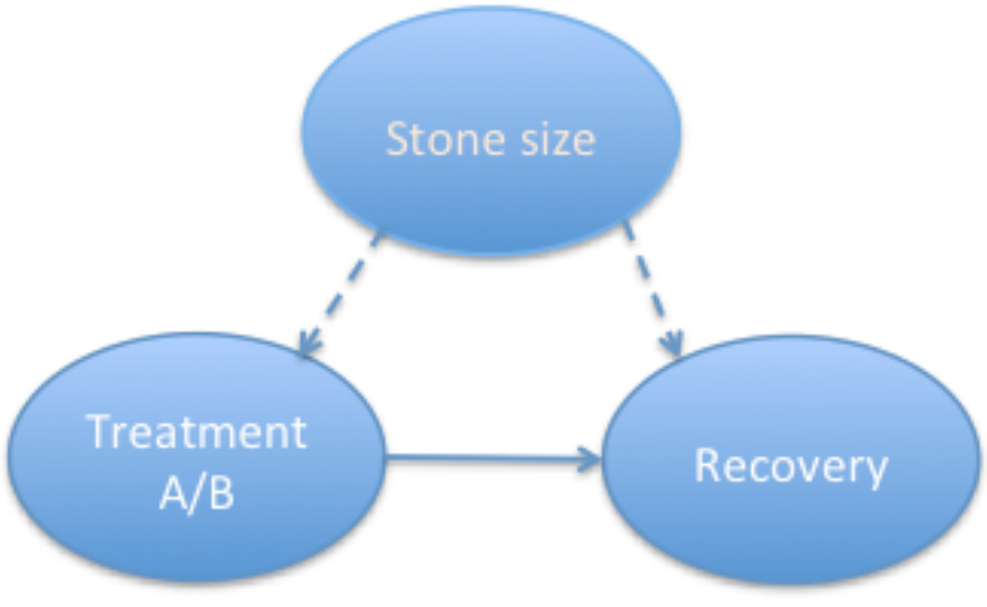
$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

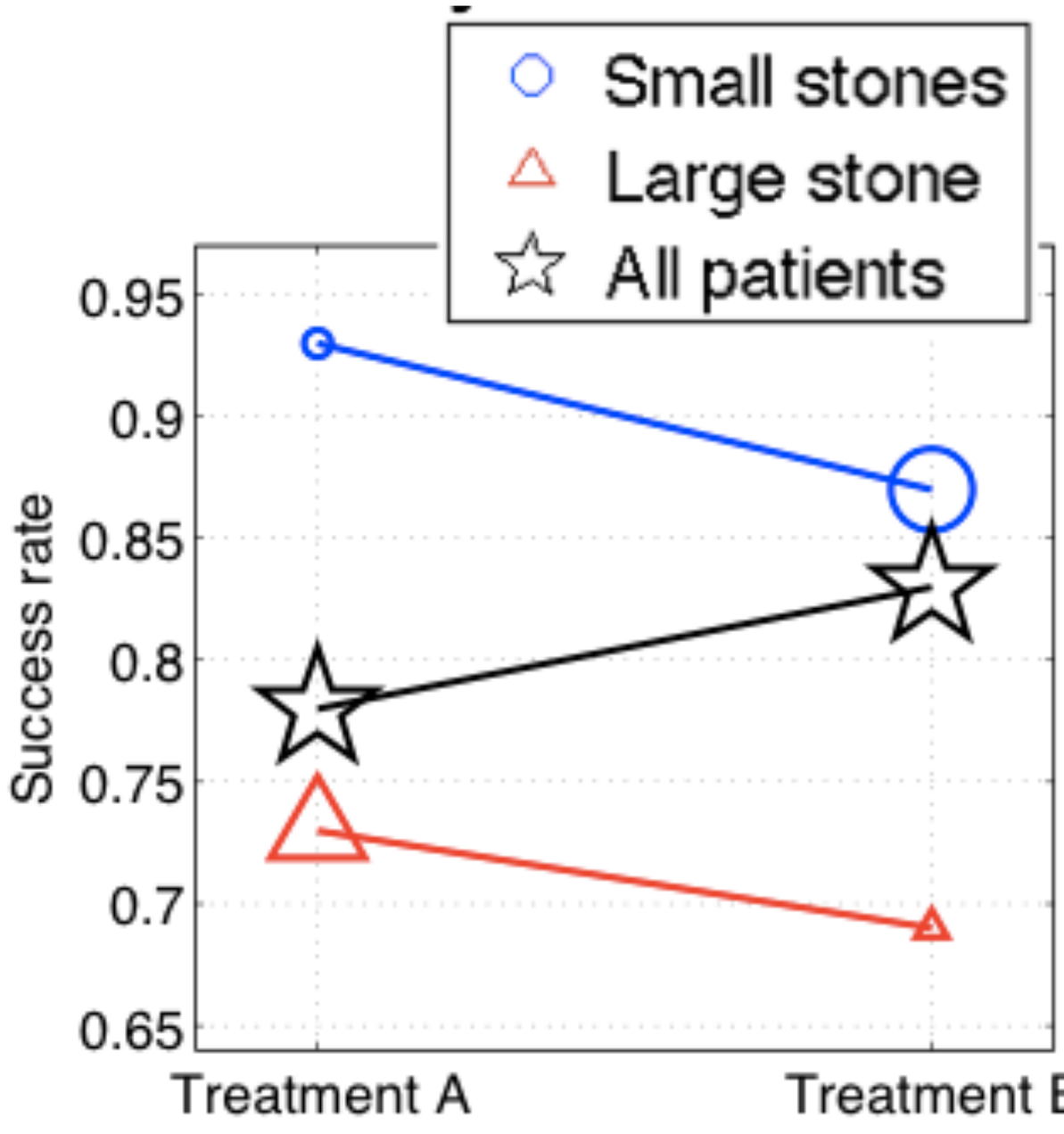
$$P(R\,|\,do(T)) = \sum_S P(R\,|\,T,S)P(S)$$

conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

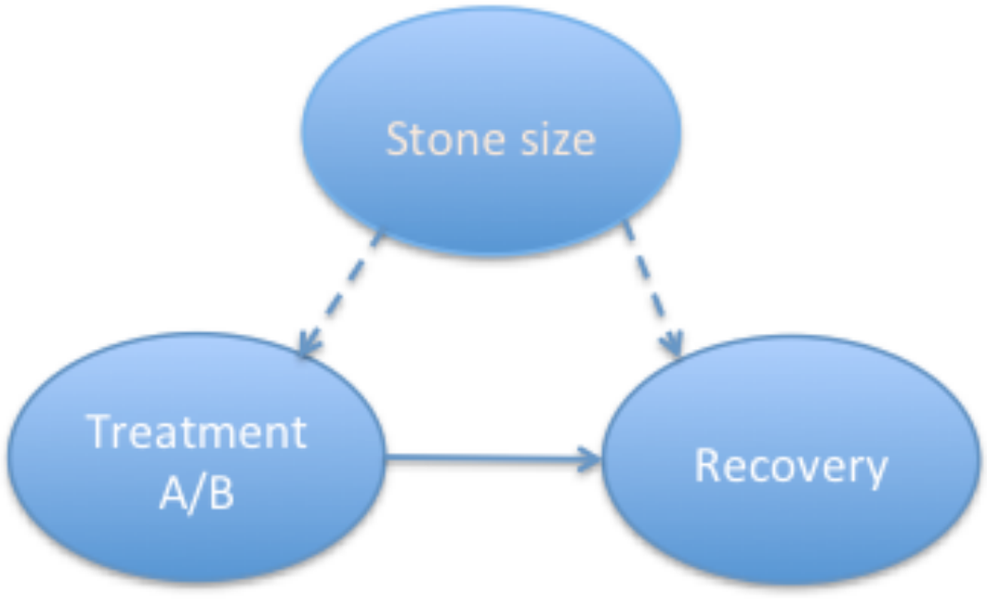


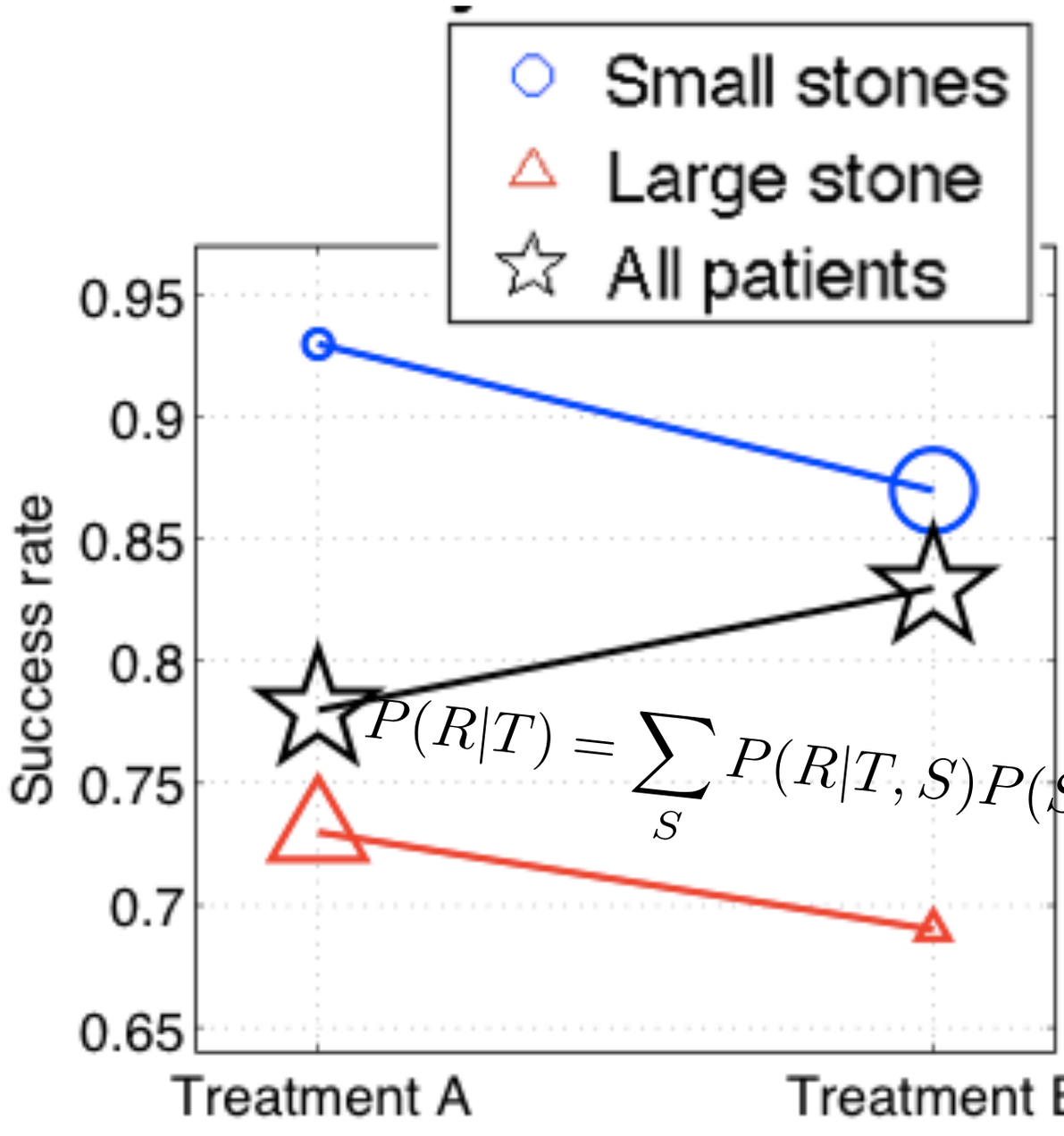


conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

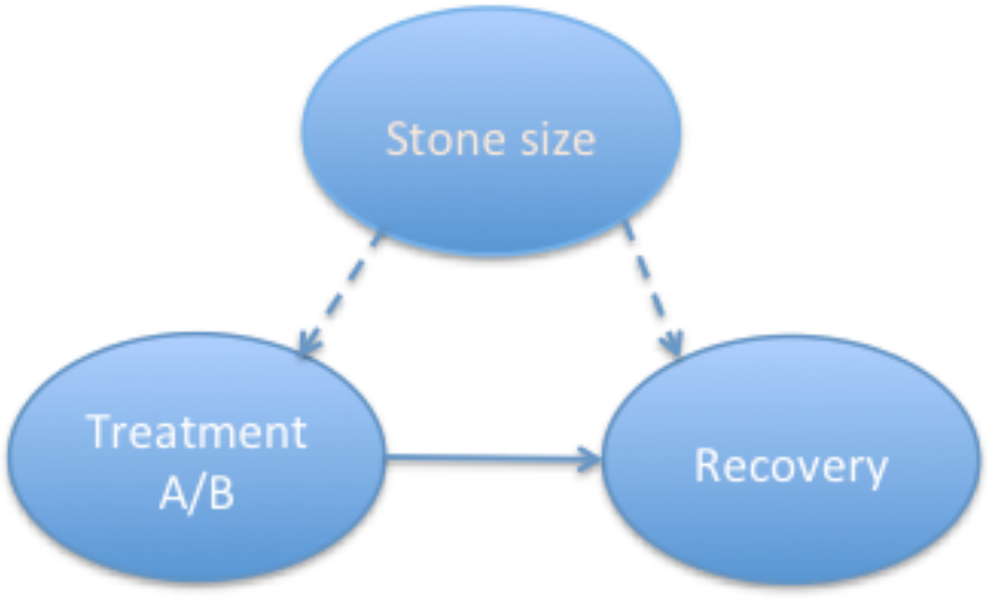




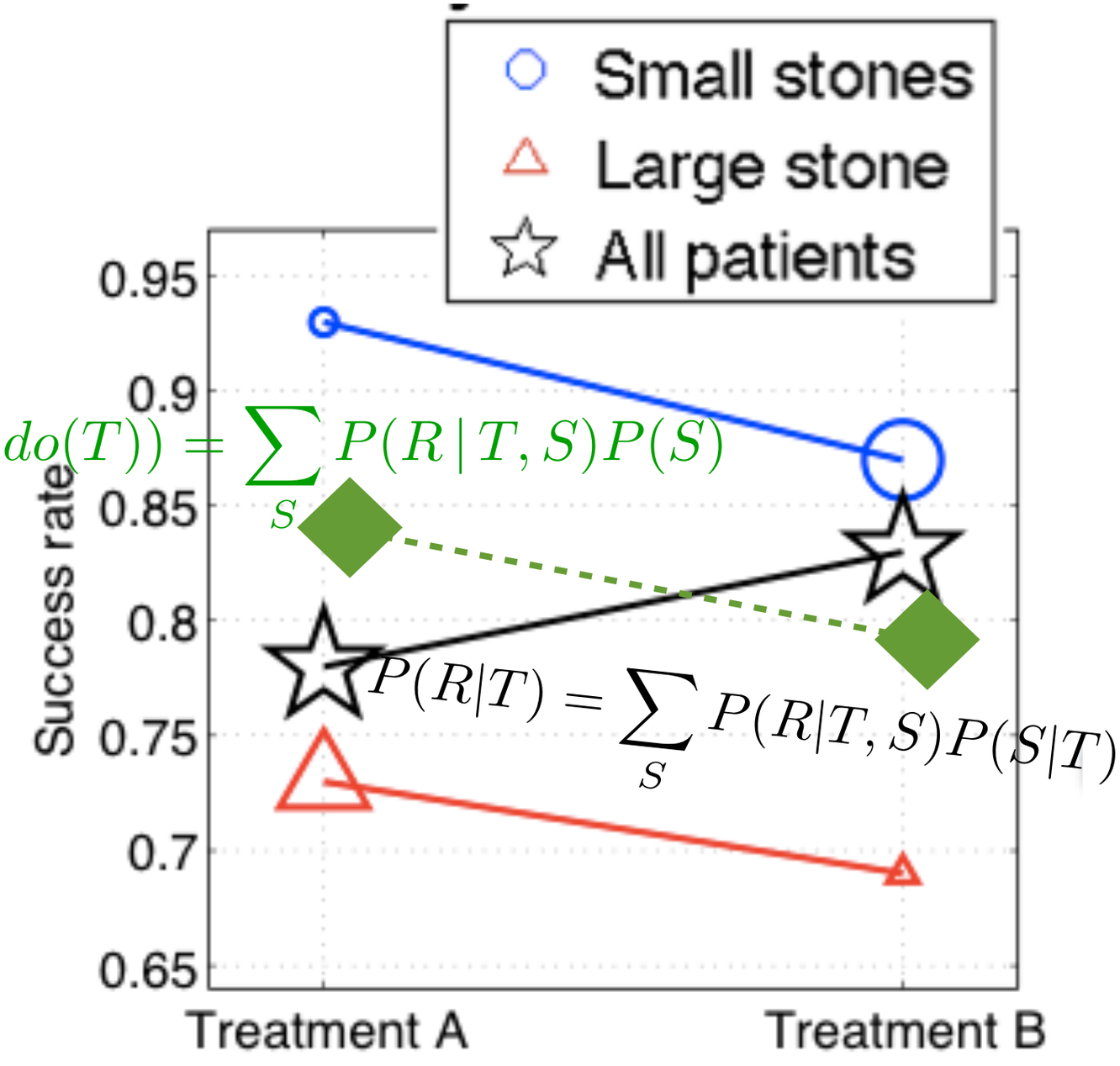
conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

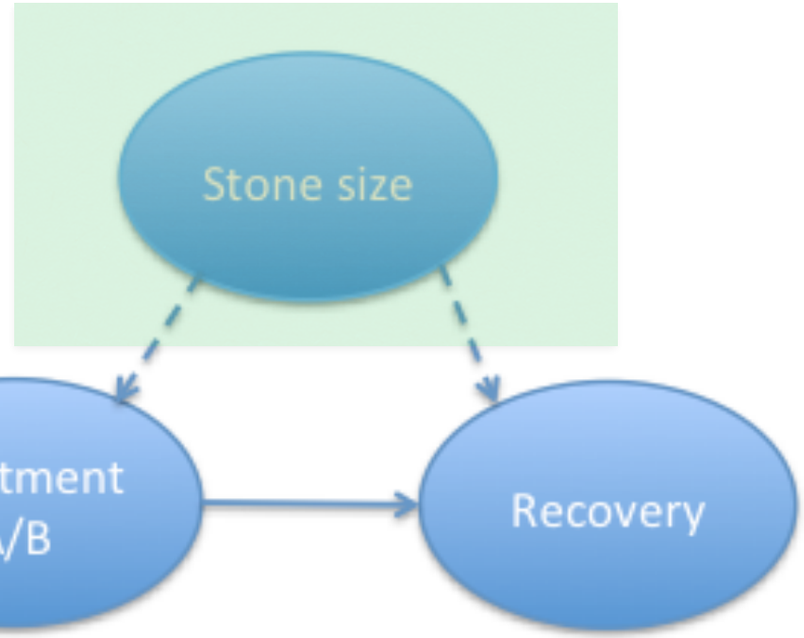


conditioning vs. manipulating

Identifiability of Causal Effects

- Is causal effect, denoted by $P(Y \mid do(X))$, identifiable given complete or partial causal knowledge?

  - Two models with **the same causal structure** and **the same distribution for the observed variables** give the same causal effect?



- How?

- Key issue: Controlling confounding effects

  Examples: Average causal effect (ACE)...