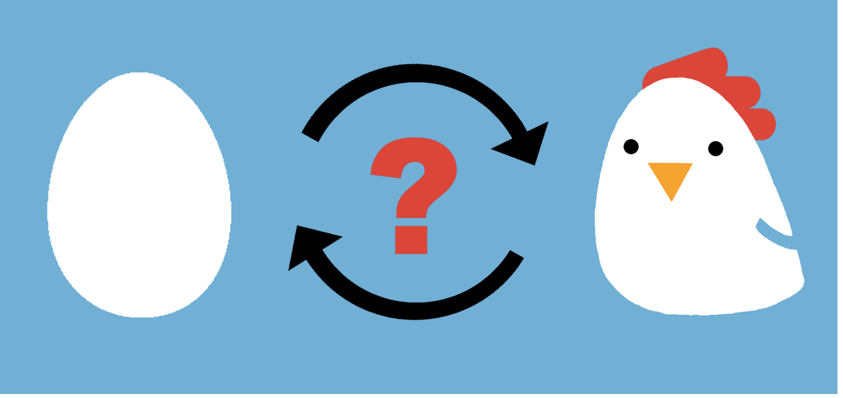


CBMS Conference -- Foundations of Causal Graphical Models and Structure Discovery

Lecture 5

Multivariate Analysis and Traditional constraint- or score-based causal discovery

Instructor: Kun Zhang

Carnegie Mellon University

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Outline

- Basic multivariate analysis & connection to causal analysis
  - PCA, factor analysis, ICA…
- Constraint-based methods for causal discovery
- Basic idea of score-based methods

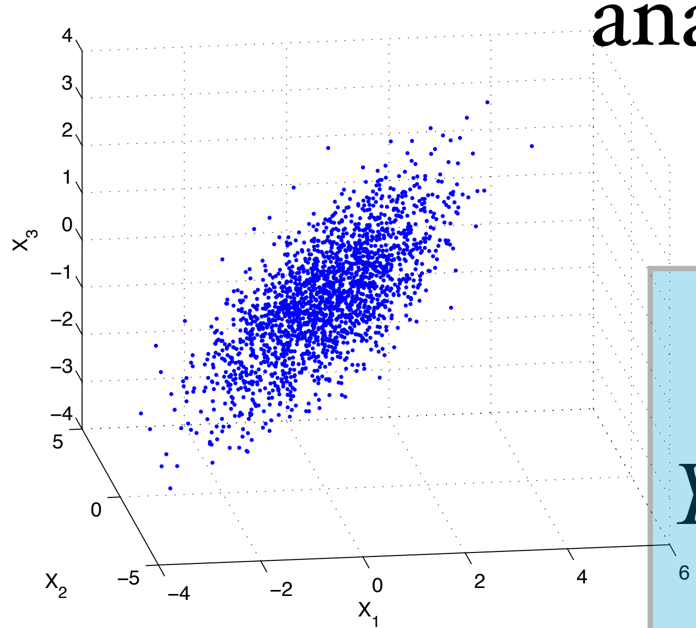
Two Ways of Finding *Simpler* Data Representations

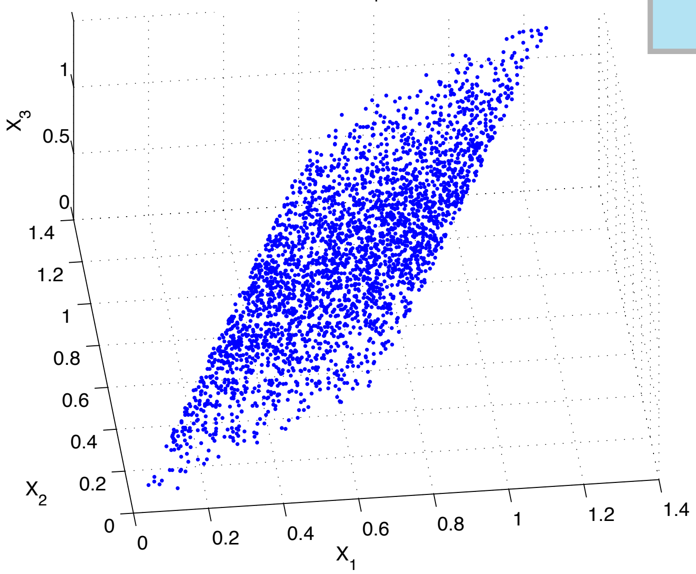
- Fewer “data points” vs. fewer dimensions (#variables)?

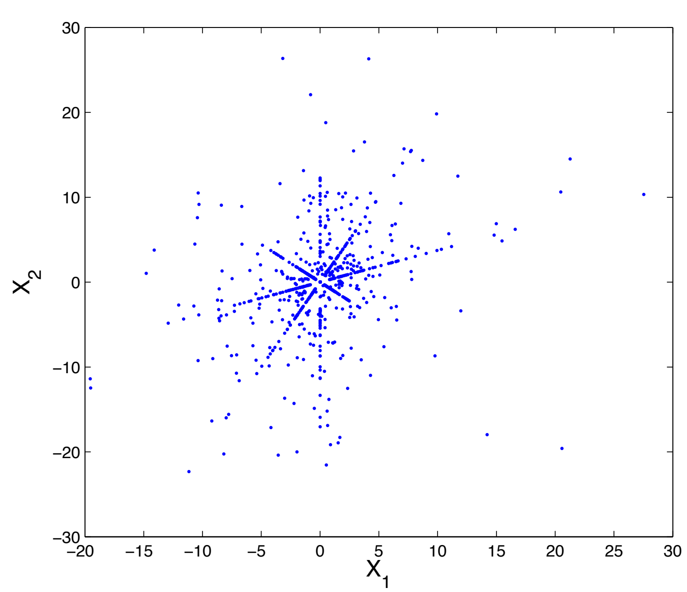
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average at	Attrition pe	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2942	2	3	4	0	1	0	1.5	2	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
9	AUSM246	Australia	Male	647.4586	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
10	AUSM8217	Australia	Male	658.6616	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
11	AUSM8177	Australia	Male	667.5444	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8173	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
13	AUSM8173	Australia	Male	648.7064	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0378	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
15	AUSM8165	Australia	Male	616.55	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
16	AUSM8154	Australia	Male	635.0605	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6959	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
18	AUSF1412	Australia	Female	618.4781	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSF8179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSF8175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8324	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSF8169	Australia	Female	619.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSF8157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSF8155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
25	AUSF1578	Australia	Female	640.6311	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
26	AUSF243	Australia	Female	606.164	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSF8158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	663.6198	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
29	DENM1011	Denmark	Male	651.4847	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
30	DENM1205	Denmark	Male	636.9831	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
31	DENM116_	Denmark	Male	642.9192	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
32	DENM116_	Denmark	Male	646.6609	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
33	DENM116_	Denmark	Male	674.9799	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
34	DENM7_77	Denmark	Male	666.53	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
35	DENM1_58	Denmark	Male	627.4583	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
36	DENM903	Denmark	Male	662.5953	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
37	DENM901	Denmark	Male	672.8408	0	0	1	3	6	0	2.1	NaN	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
38	DENF1559	Denmark	Female	604.4864	0	0	1	3	6	0	2.1	0.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27

INFO | Shape_rawcoordinates | Shape_symmProcCoord. | Factors | Background | Geodistance_extra

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



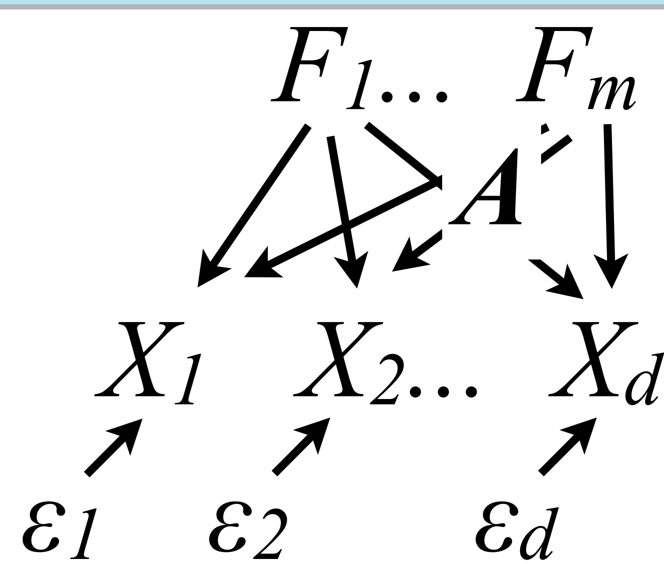




- Regression…

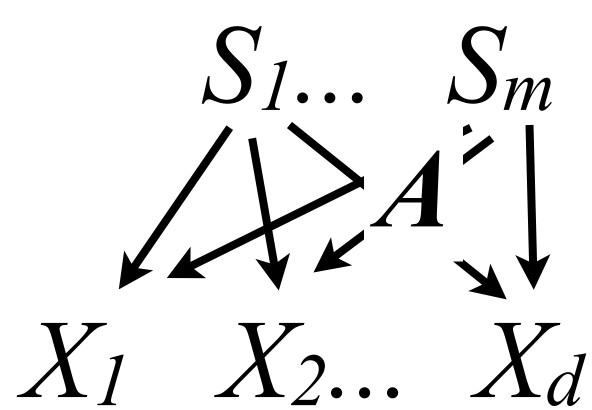
Find a projection of the data: $Y = w^{\mathrm{T}}\mathrm{X}$ with certain properties.

- Principal component analysis



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

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



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

Multiple Regression

- Regress Y on $\boldsymbol{X} = (X_1, X_2)^T$
- $\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + c$
- For simplicity, *assume all variables have zero mean*

$$\text{Minimize} \quad S_E = (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})^\intercal (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})$$

$$\frac{\partial S_E}{\partial \boldsymbol{\alpha}} = 2 \cdot \mathbf{x}^\intercal (\mathbf{y} - \mathbf{x}\boldsymbol{\alpha})$$

$$\text{If } \mathbf{x}^\intercal \mathbf{x} \text{ is invertible, setting } \frac{\partial S_E}{\partial \boldsymbol{\alpha}} = 0$$

$$\Rightarrow \quad \boldsymbol{\alpha} = (\mathbf{x}^\intercal \mathbf{x})^{-1} (\mathbf{x}^\intercal \mathbf{y})$$

$$\begin{array}{cc} & \begin{array}{cc} X_1 & X_2 \end{array} \\ \mathbf{x} = & \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1N} & x_{2N} \end{bmatrix} \end{array}$$

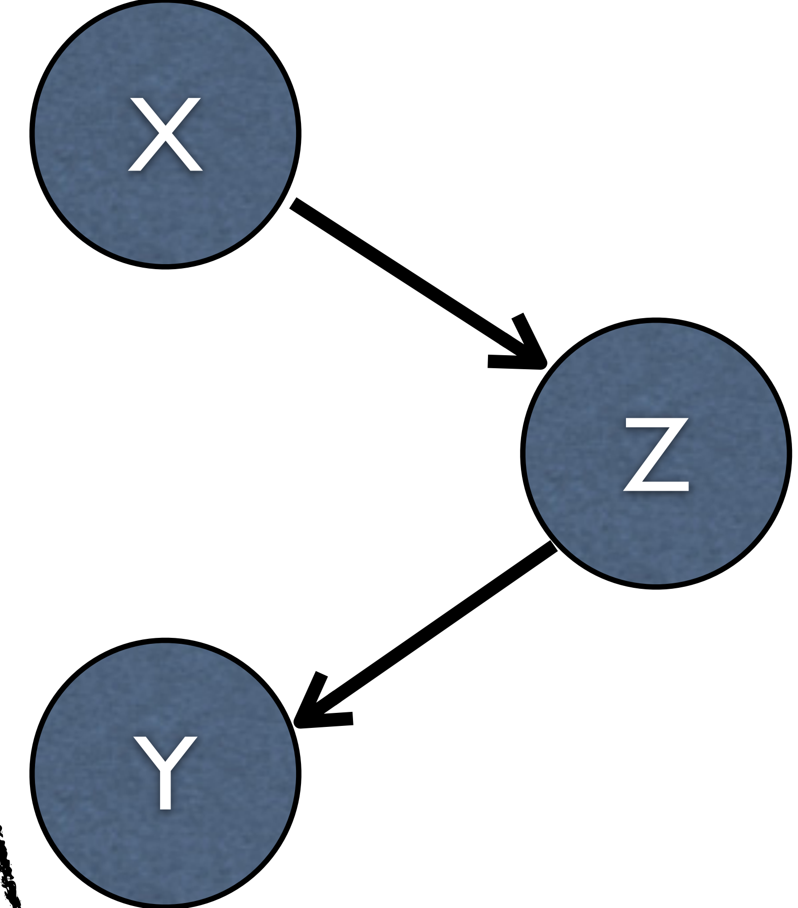
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Simple Regression vs. Multiple Regression

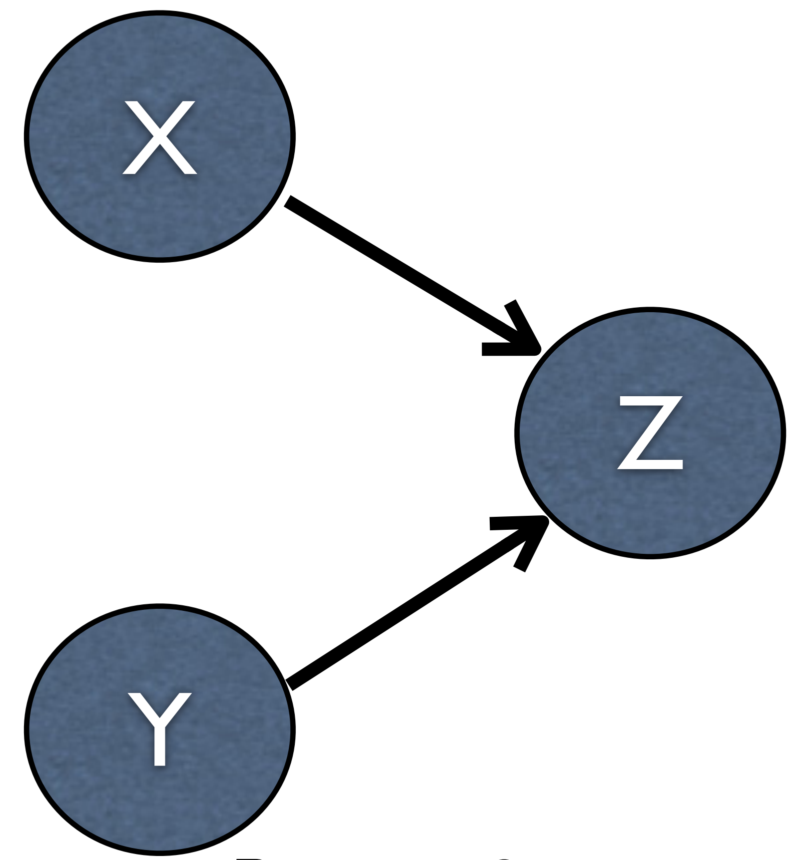
- Let's do simple regression from X to Y: $\hat{y} = \alpha x + c$
- Will α be zero?

Independence vs. conditional independence ;-)
and you can see it from graph!

- Let's do regression from $(X,Z)^T$ to Y: $\hat{y} = \alpha_1 x + \alpha_2 z + c$
- Will the coefficient of x be zero?



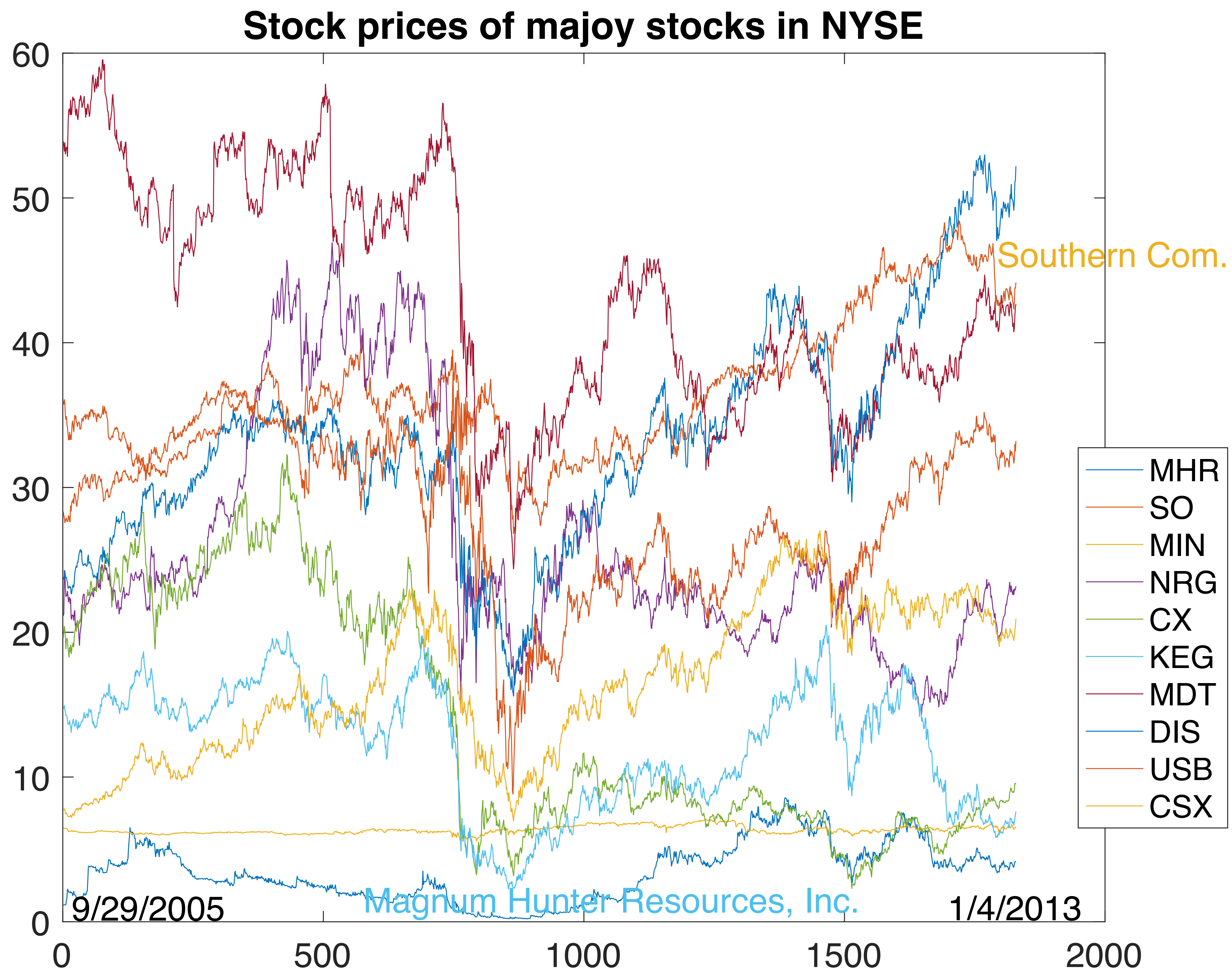
Process 1



Process 2

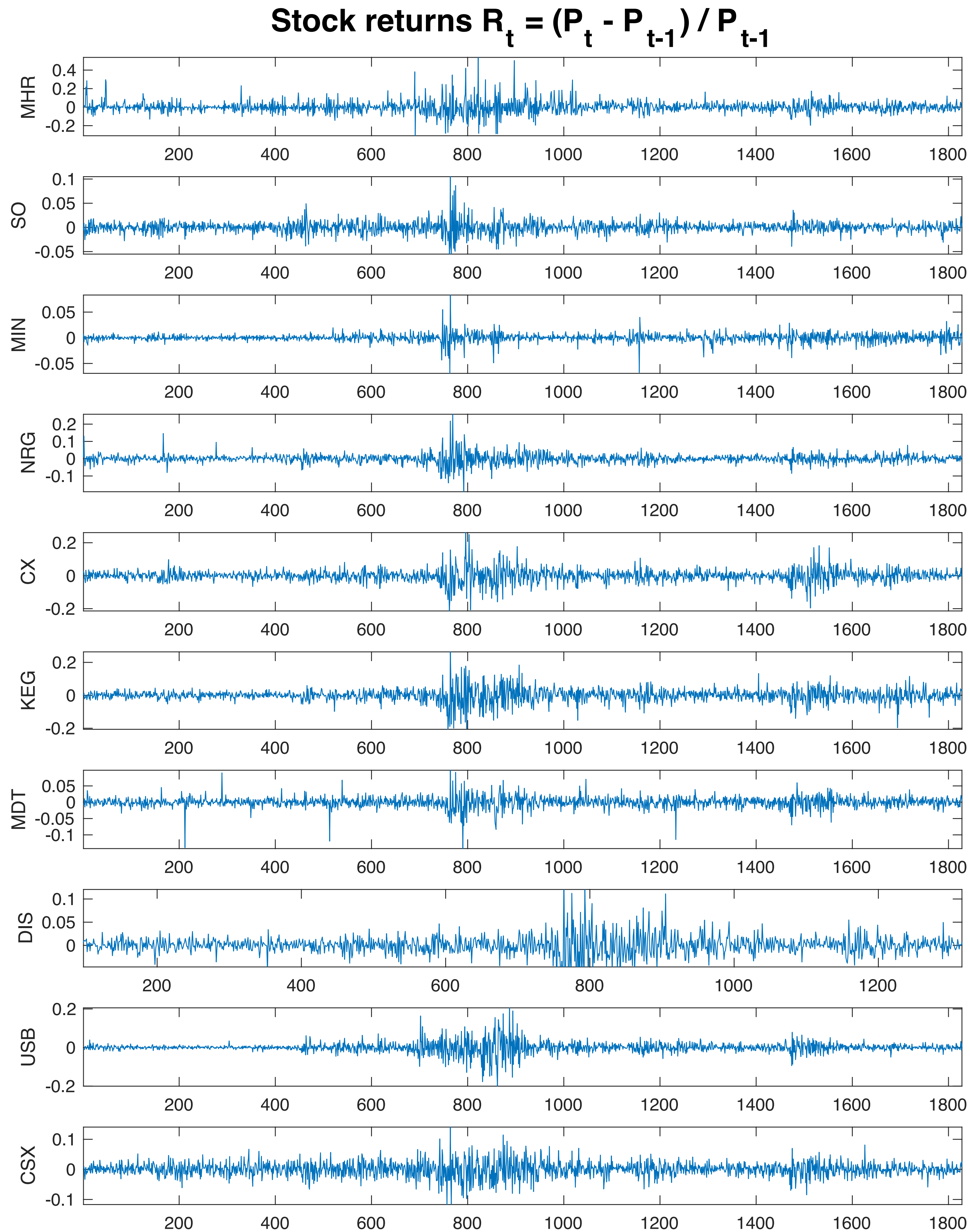
Major Information in the Data?

- Major information in the NYSE stock market? Better to analyze returns…



Major Information in the Data?

- Major information in the NYSE stock market?

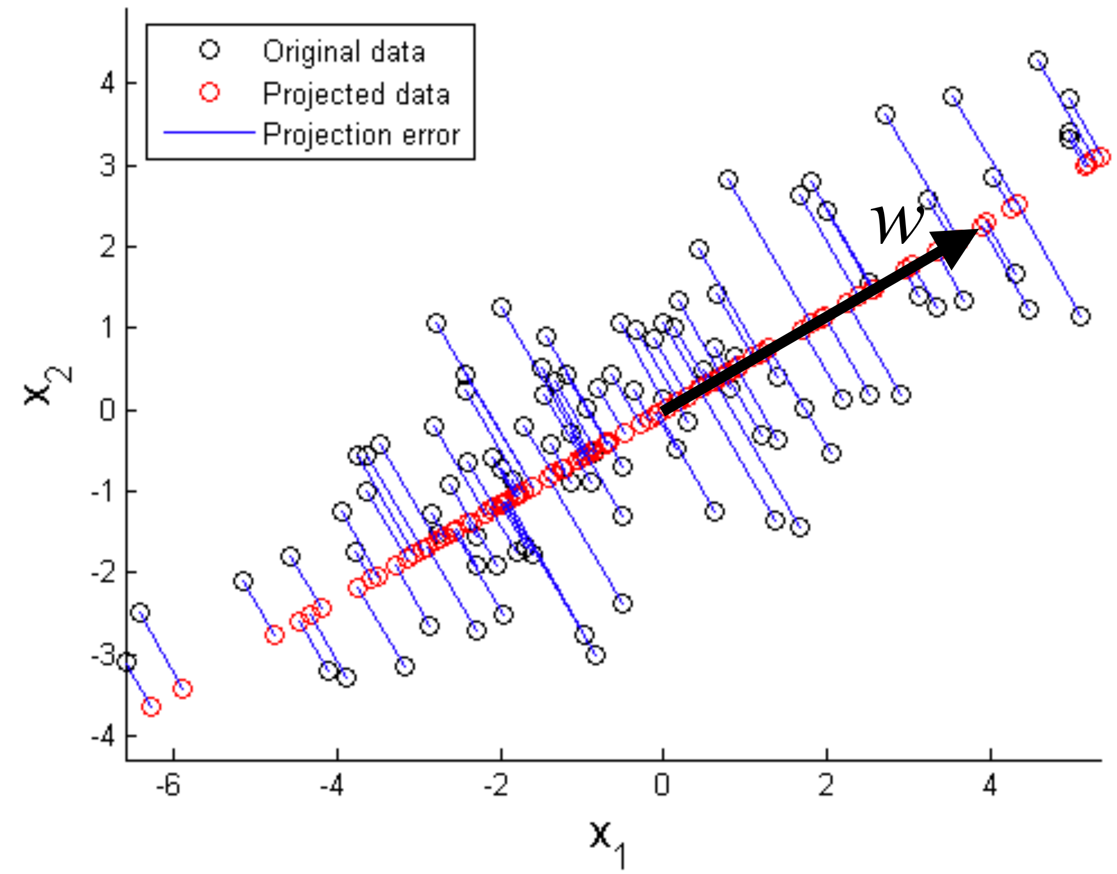


Principal Component Analysis (PCA)

- Find a projection of the data

$$Y = w^{\mathrm{T}}\boldsymbol{X}$$

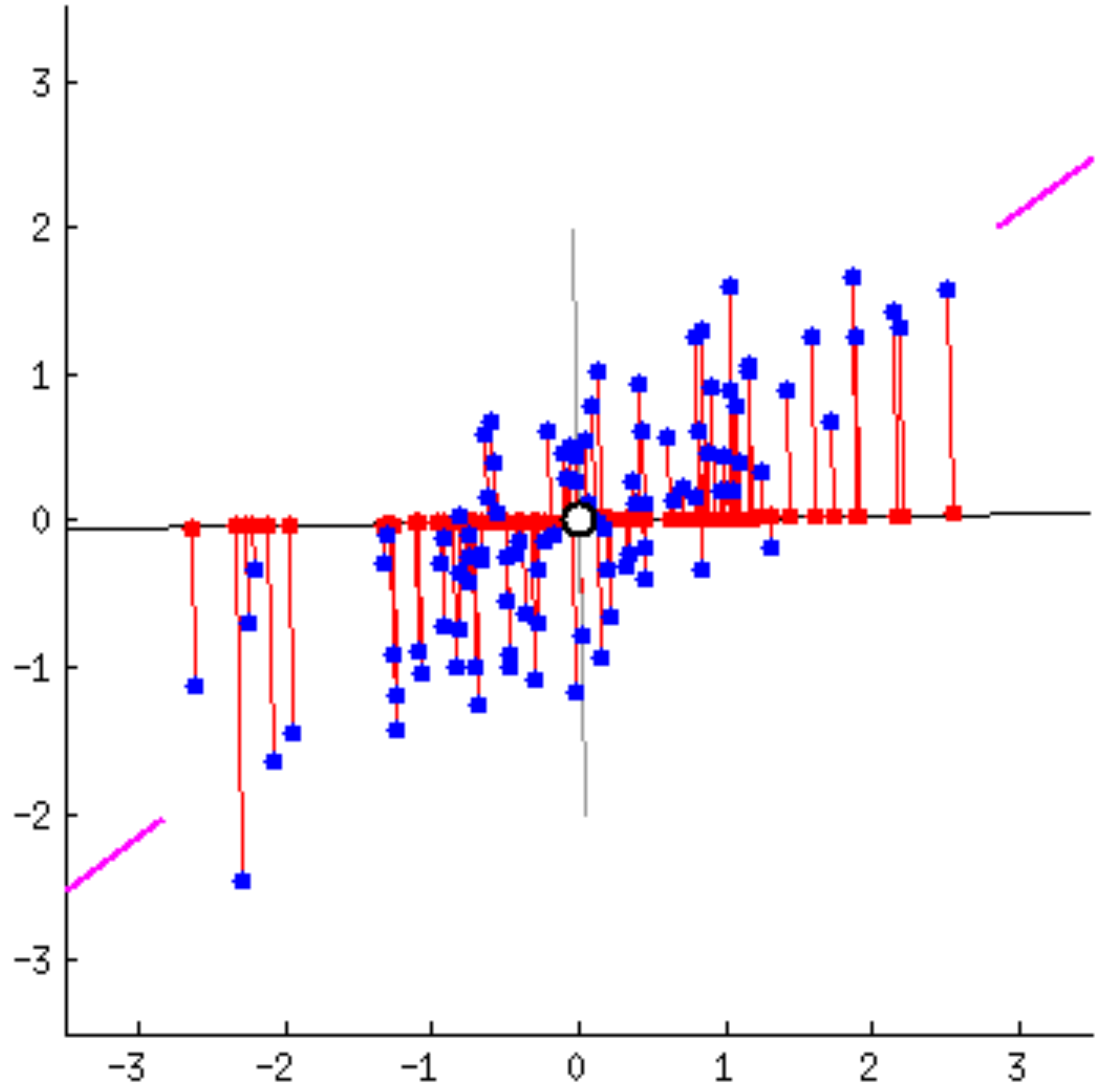
to give the maximum variance (minimal squared reconstruction/ projection error?)



w: principal axis/direction;
$w^{\mathrm{T}}\boldsymbol{X}$: principal component

PCA was invented in 1901 by Karl Pearson, as an analogue of the principal axis theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s. Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT) in signal processing… (https://en.wikipedia.org/wiki/Principal_component_analysis#History)

PCA: Effect of Weight Vector w



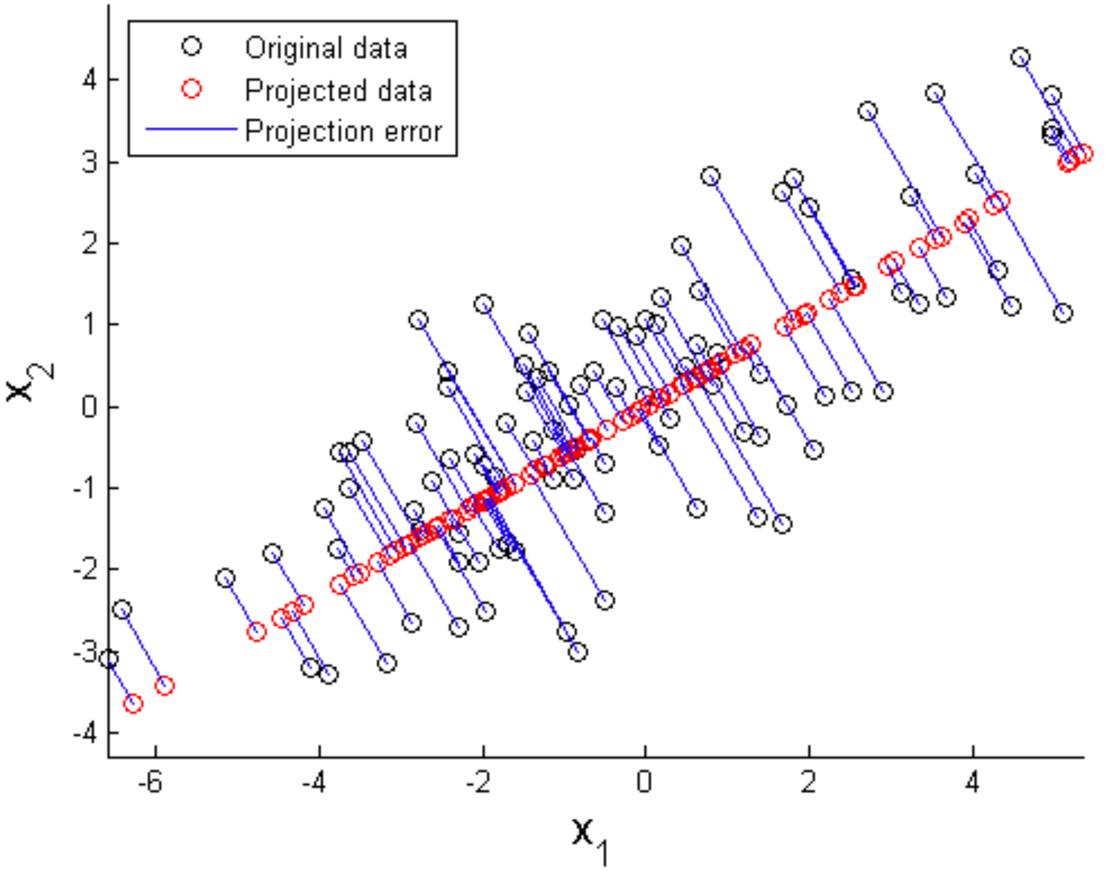
PCA

- Find a projection of the data

$$Y = w^{\mathsf{T}} X$$

to give the maximum variance

- Find next ones if needed…



- Assume $\mathbf{X}$ has a zero mean.

- Maximize the sample variance of Y, which is $\frac{1}{N}\mathbf{Y}^{\intercal}\mathbf{Y} = \frac{1}{N}w^{\intercal}\mathbf{X}\mathbf{X}^{\intercal}w = w^{\intercal}Cw$, where $C = \frac{1}{N}\mathbf{X}\mathbf{X}^{\intercal}$, s.t. $||w||^2 = w^{\intercal}w = 1$.

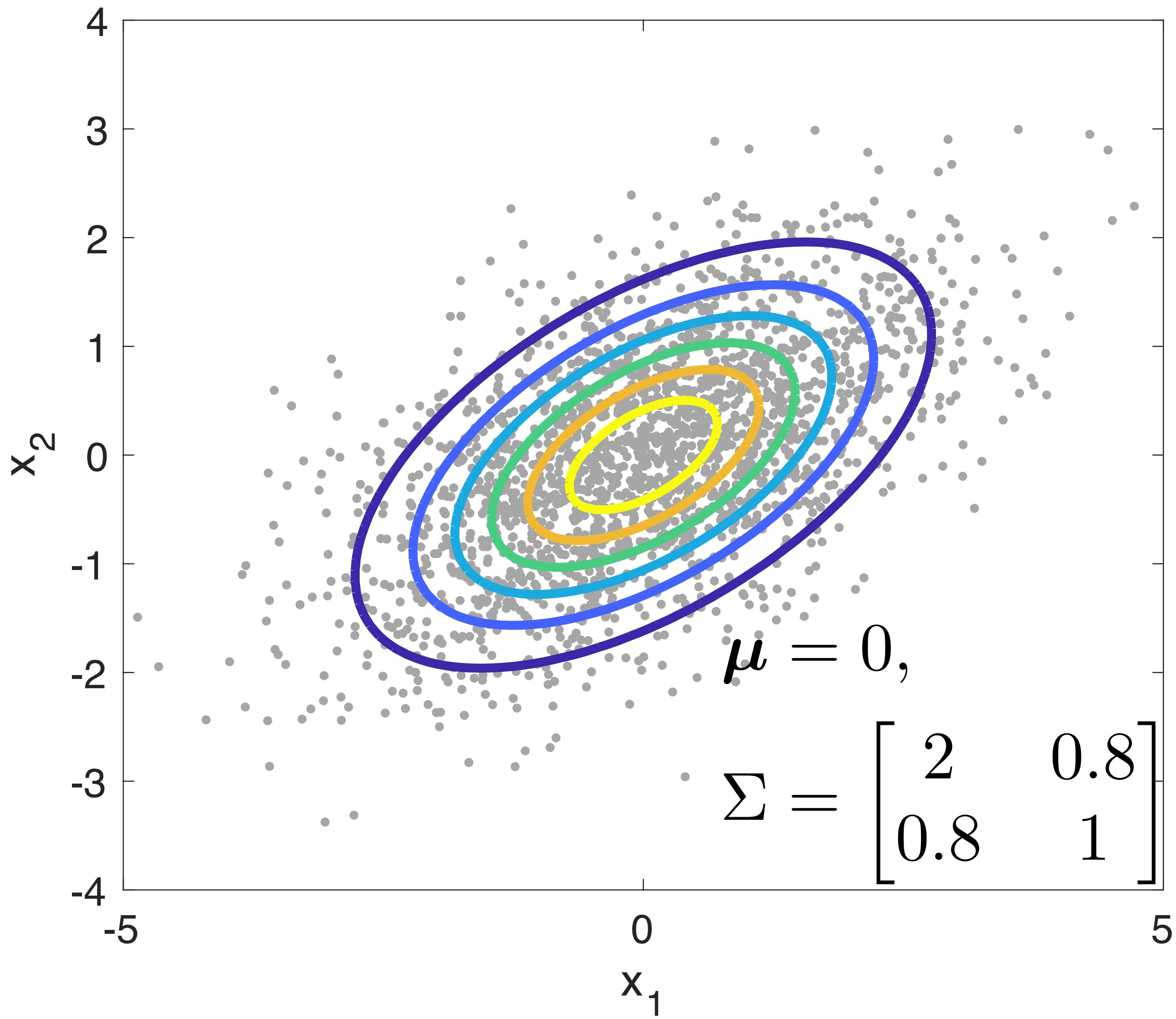
- Let $\mathcal{L} = w^{\intercal}Cw - \lambda w^{\intercal}w$. Setting $\frac{\partial \mathcal{L}}{\partial w} = 0$ gives

$$2Cw - 2\lambda w = 0 \Rightarrow Cw = \lambda w.$$

- So w is an eigenvalue of C and λ is the corresponding eigenvalue.

- The sample variance of Y is then $w^{\intercal}Cw = w^{\intercal} \cdot \lambda w = \lambda w^{\intercal}w = \lambda$. So λ corresponds to the larges eigenvalue.

Principal Axis vs. Regression Line

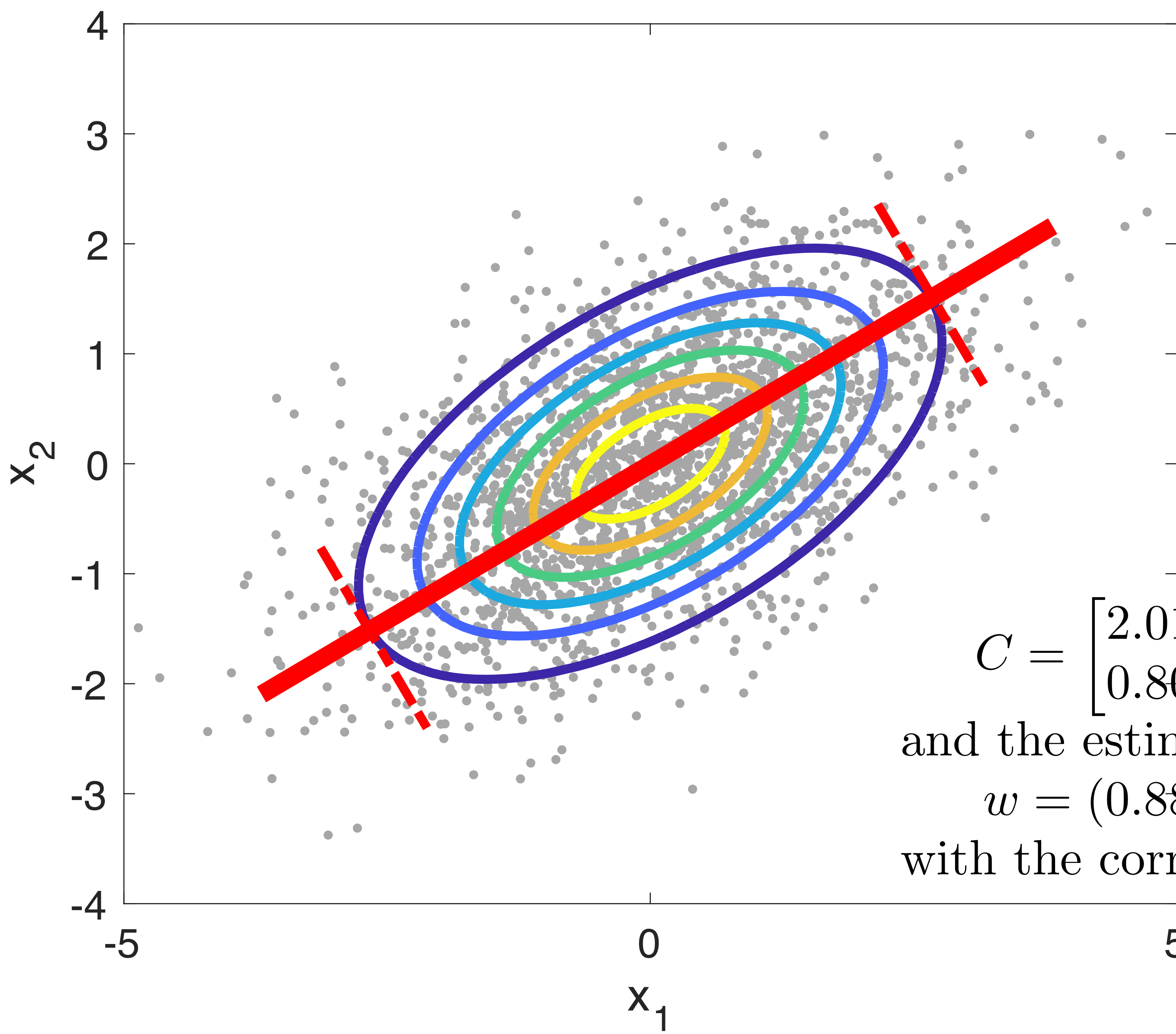


$$\boldsymbol{\mu} = 0,$$

$$\Sigma = \begin{bmatrix} 2 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

Principal Axis vs. Regression Line

- First principal component $PC_1 = w^{\mathrm{T}}\boldsymbol{X}$



$$C = \begin{bmatrix} 2.0126 & 0.8013 \\ 0.8013 & 1.0009 \end{bmatrix},$$

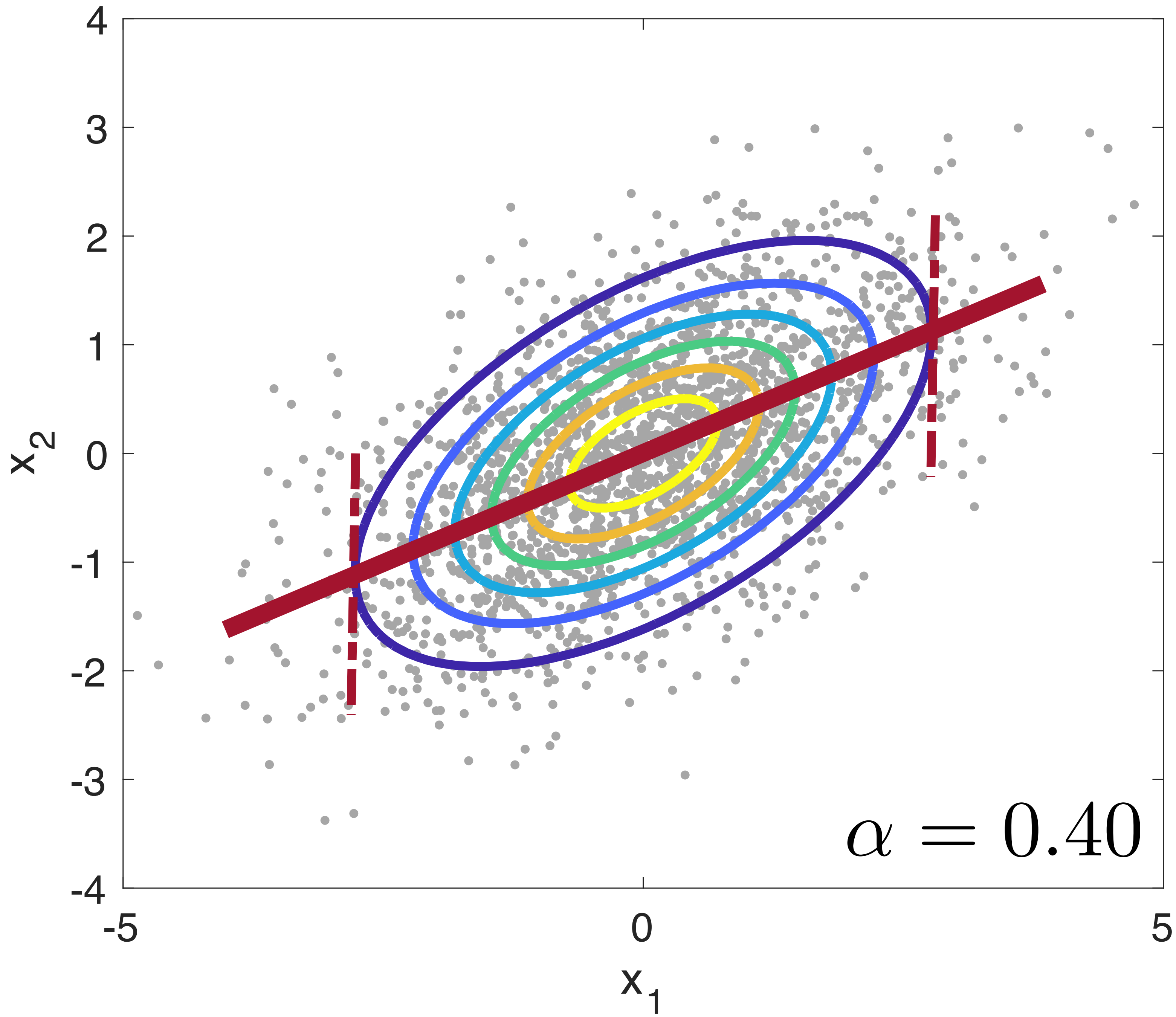
and the estimated weight vector is:

$$w = (0.88, 0.48)^{\intercal},$$

with the corresponding eigenvalue 2.45

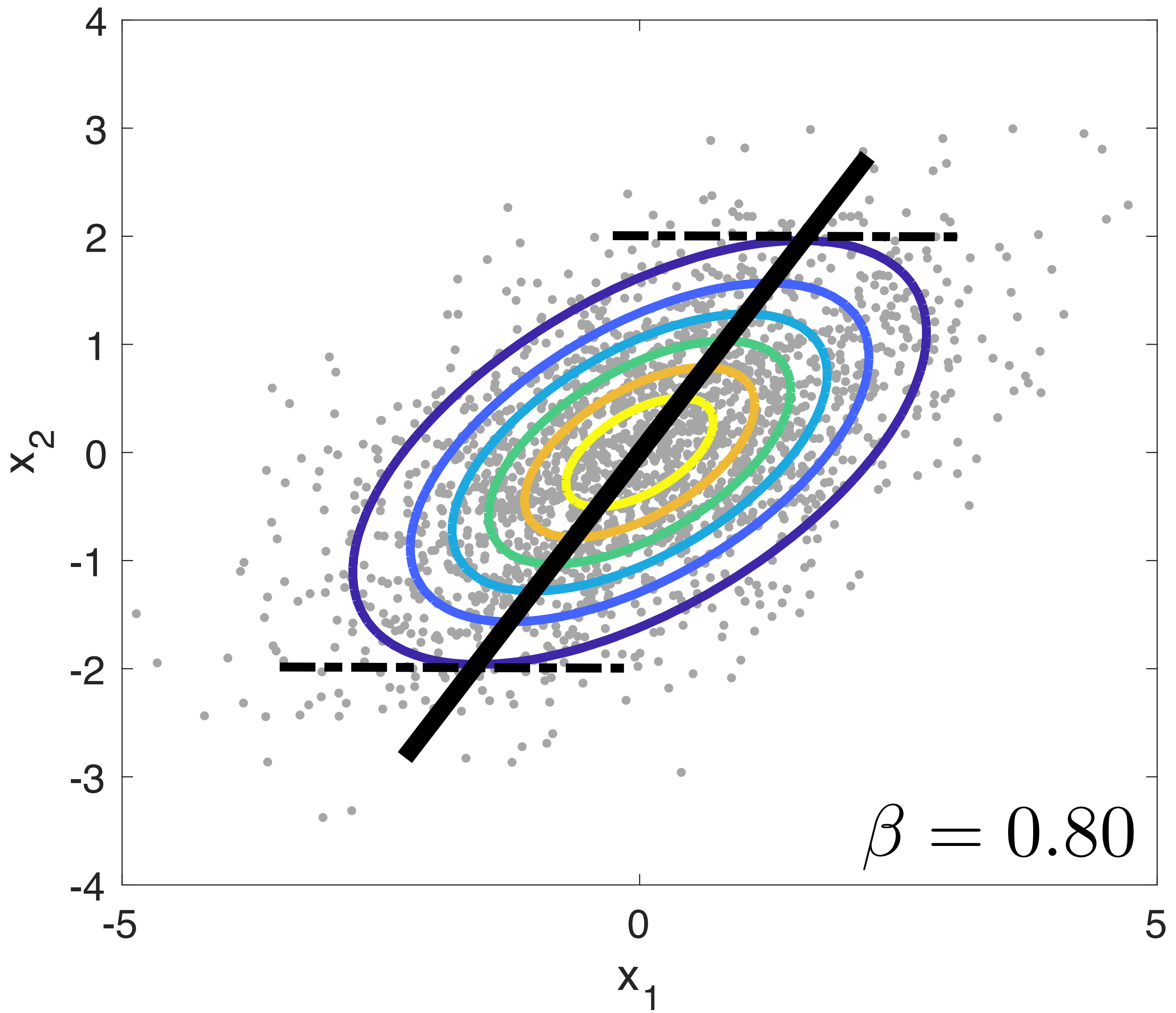
Principal Axis vs. Regression Line

- Regression line from X_1 to X_2: $\hat{x}_2 = \alpha x_1$

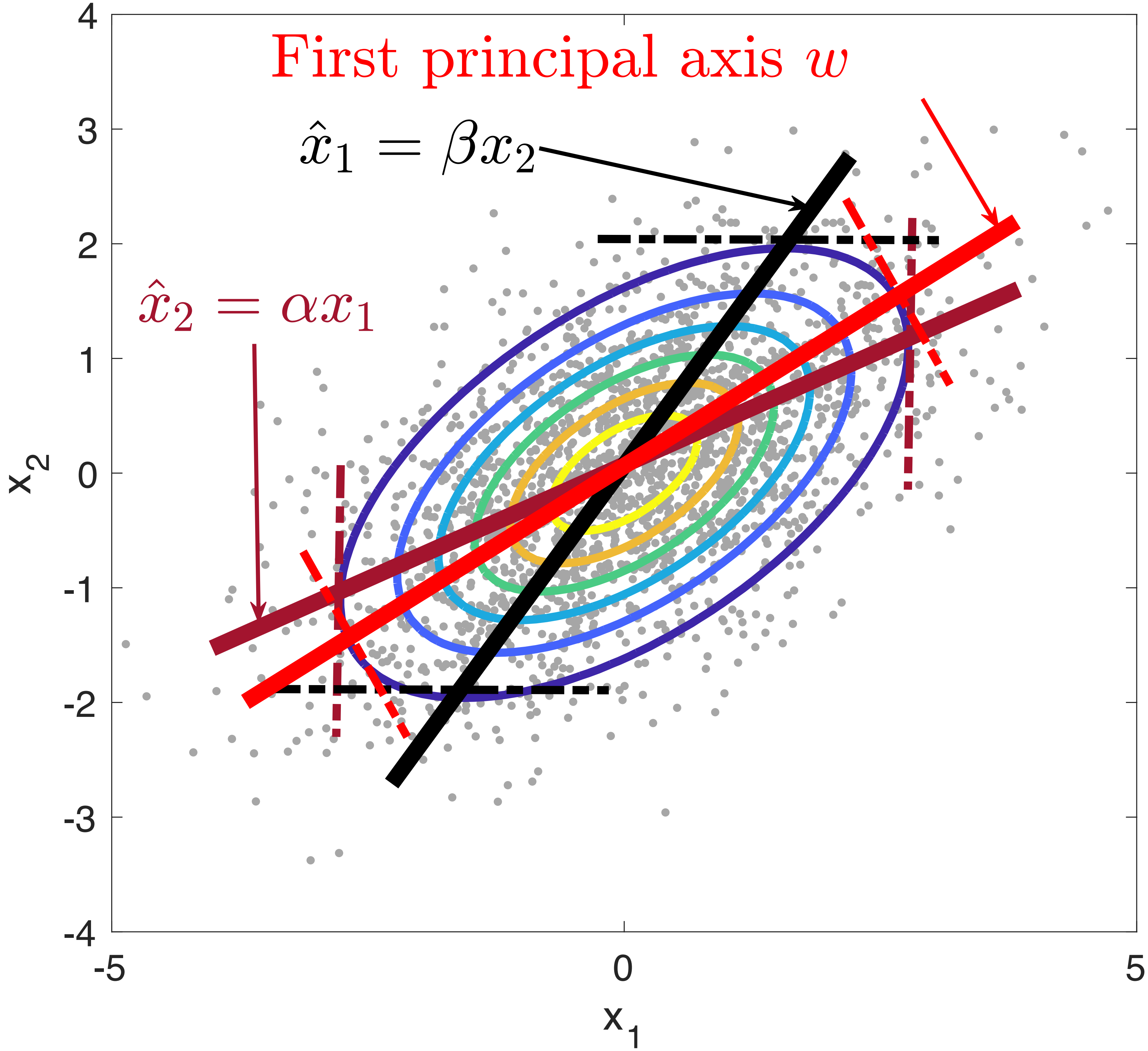


Principal Axis vs. Regression Line

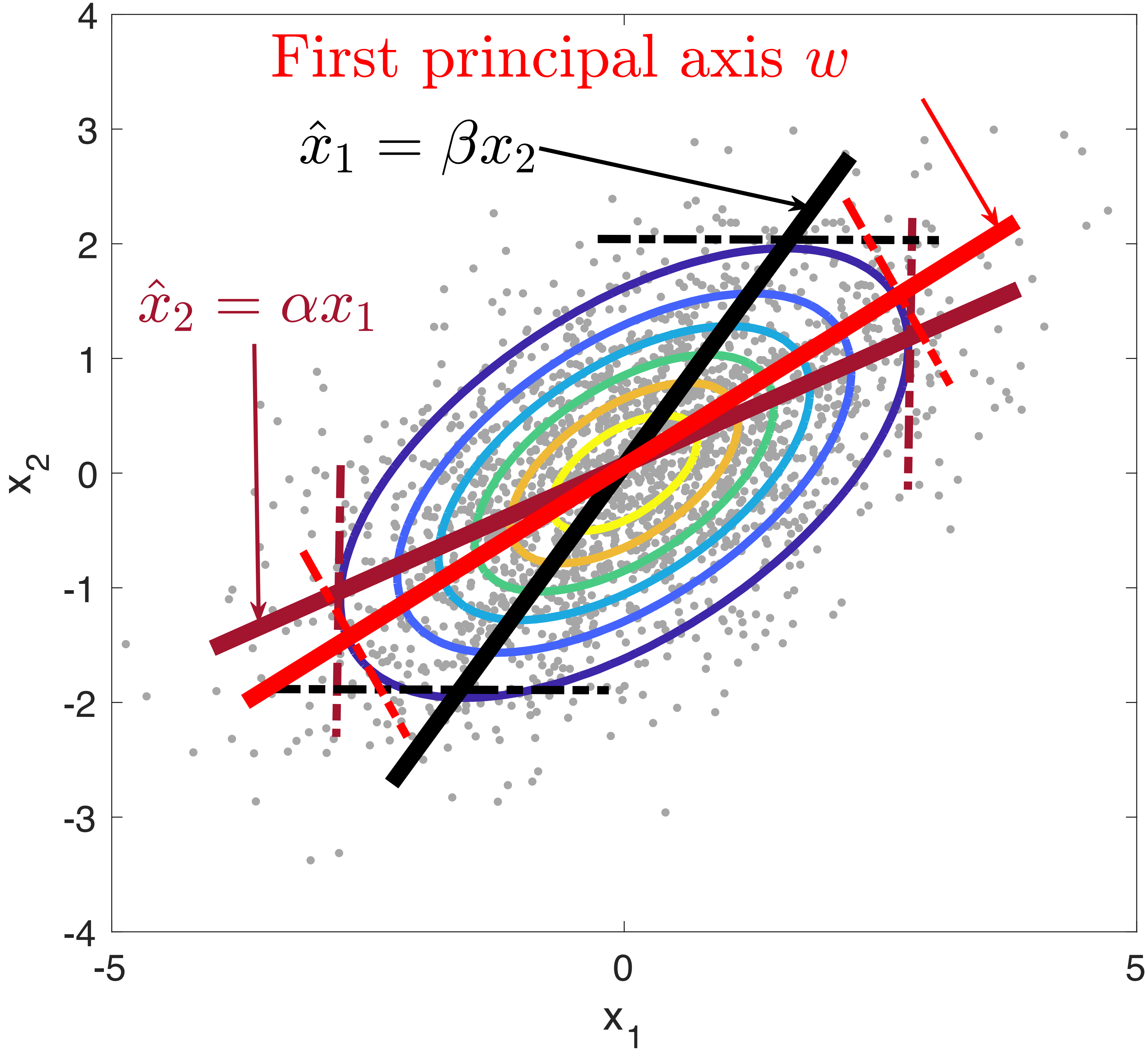
- Regression line from X_2 to X_1: $\hat{x}_1 = \beta x_2$



Principal Axis vs. Regression Line

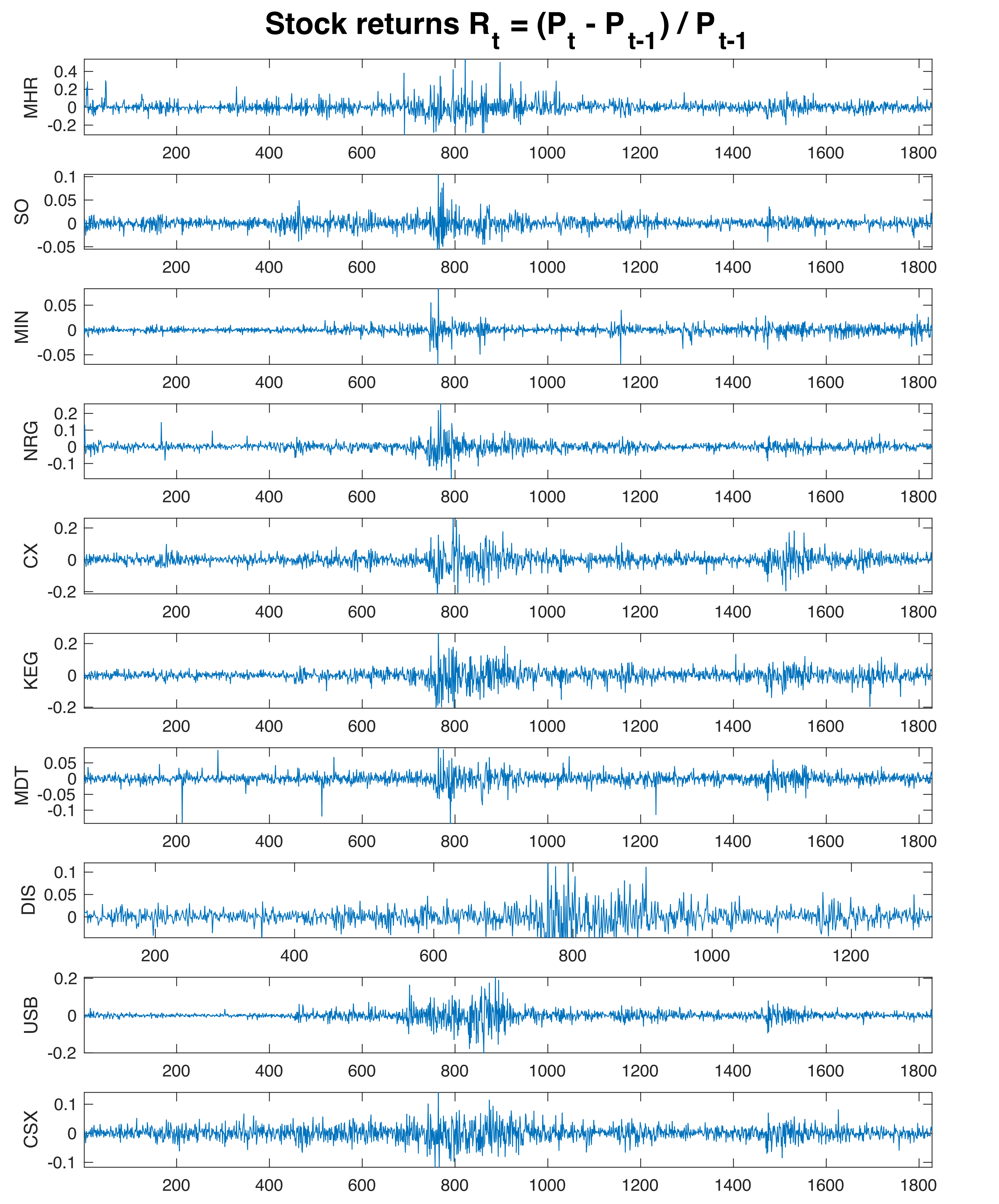


Principal Axis vs. Regression Line

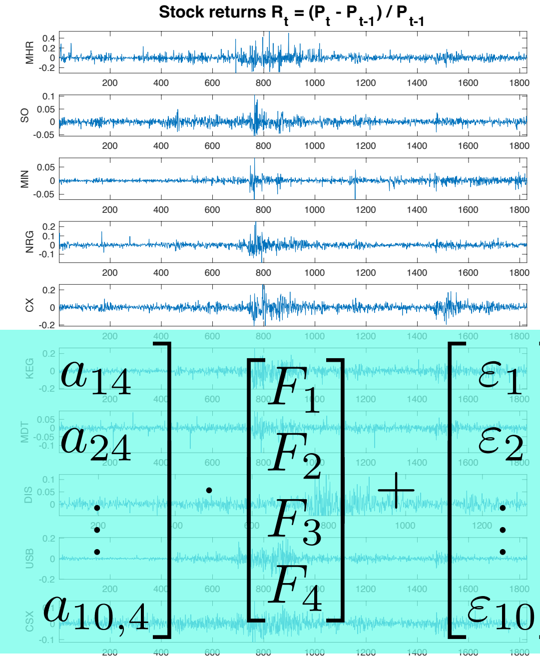


Underlying Factors?

- Major information in the NYSE stock market?



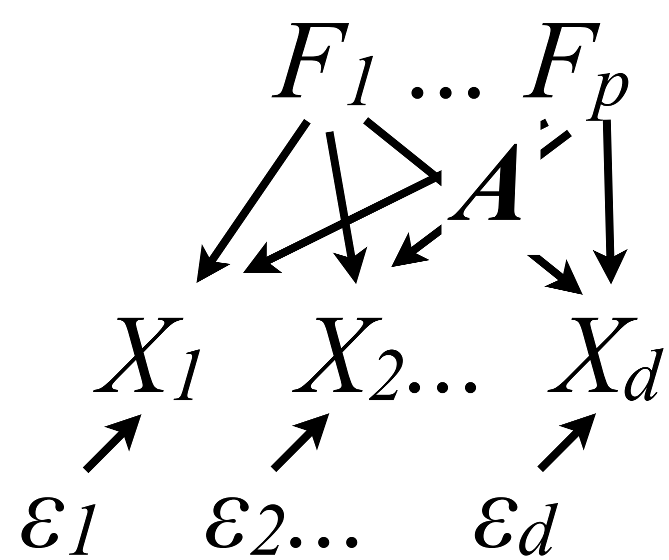
Factor Analysis



- Assume a generating model

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

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



$\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} = \mathbf{A}\mathbf{U}\mathbf{U}^\intercal\mathbf{A}^\intercal + \boldsymbol{\Psi}$,
where $\mathbf{U}$ is an orthogonal matrix.

- Partial identifiability of **A** (*up to right orthogonal transformation*)
- Estimation: MLE, usually EM

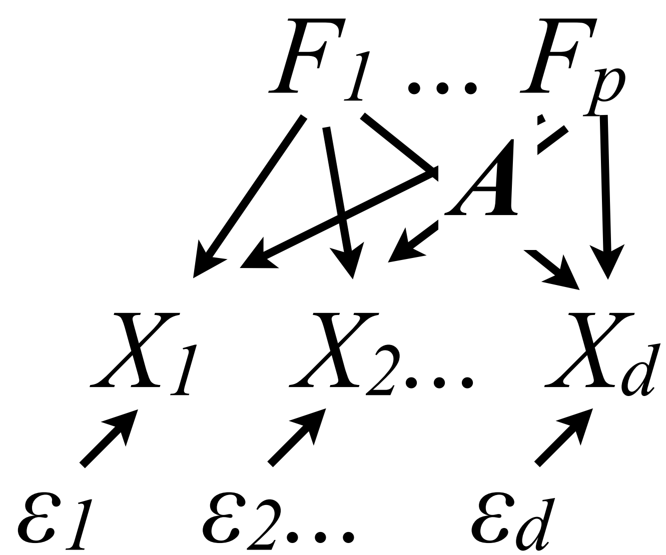
Estimated factors:

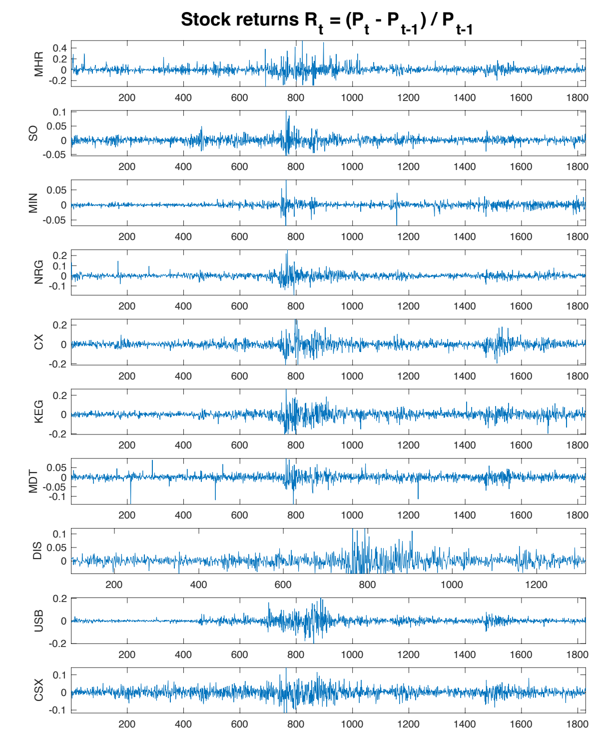
$\hat{\mathbf{F}} = \mathbf{B}\mathbf{X}$,

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

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

Factor Analysis on the Returns





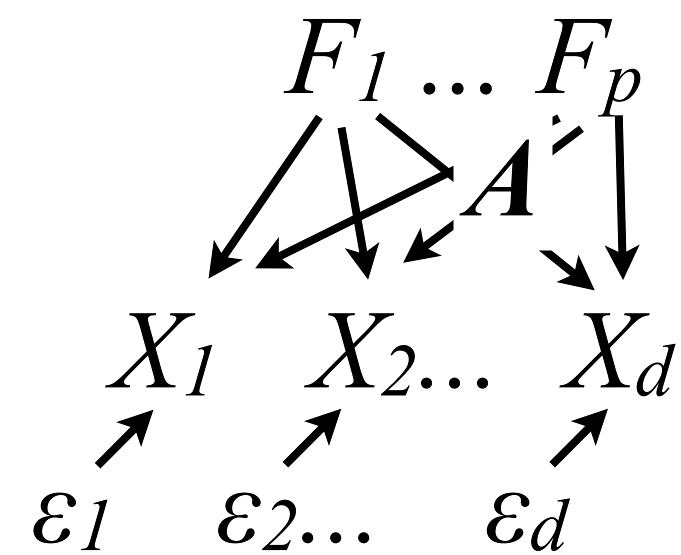
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \dots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \dots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.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

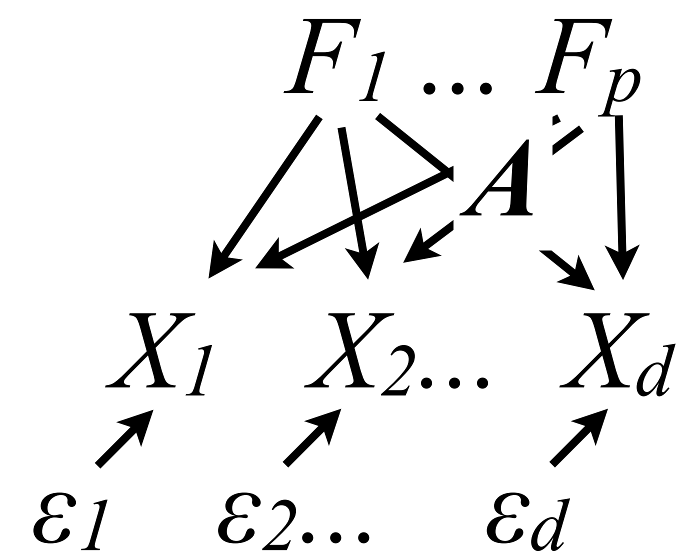
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, ..., X_d]^T$.
  - $\boldsymbol{F} = [F_1, ..., F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $E[\boldsymbol{F}]=\mathbf{0}$; $Cov[\boldsymbol{F}]=I$.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



Relationship between FA and PCA (suppose there is one factor)?
-What if the noise terms are isotropic (Probabilistic PCA)?
- What if we add (non)isotropic noise?

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 d
- Partial identifiability of **A**
- Estimation: MLE

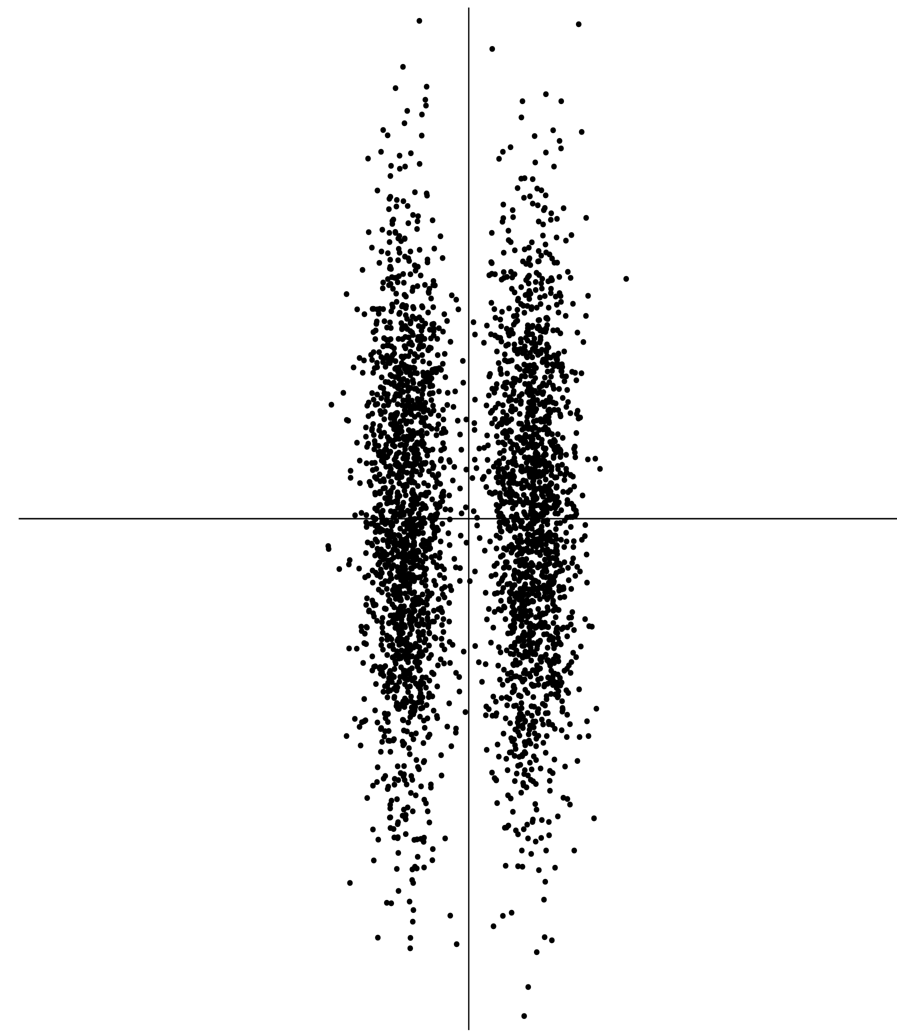


Relationship between FA and PCA:

- What if the noise terms are isotropic?
  - **A** in FA consistent with *w* in PCA.
- What if we add (non)isotropic noise?
  - **A** estimated by FA stays the same; *w* in PCA may change.

Non-Gaussianity is Informative in the Linear Case...

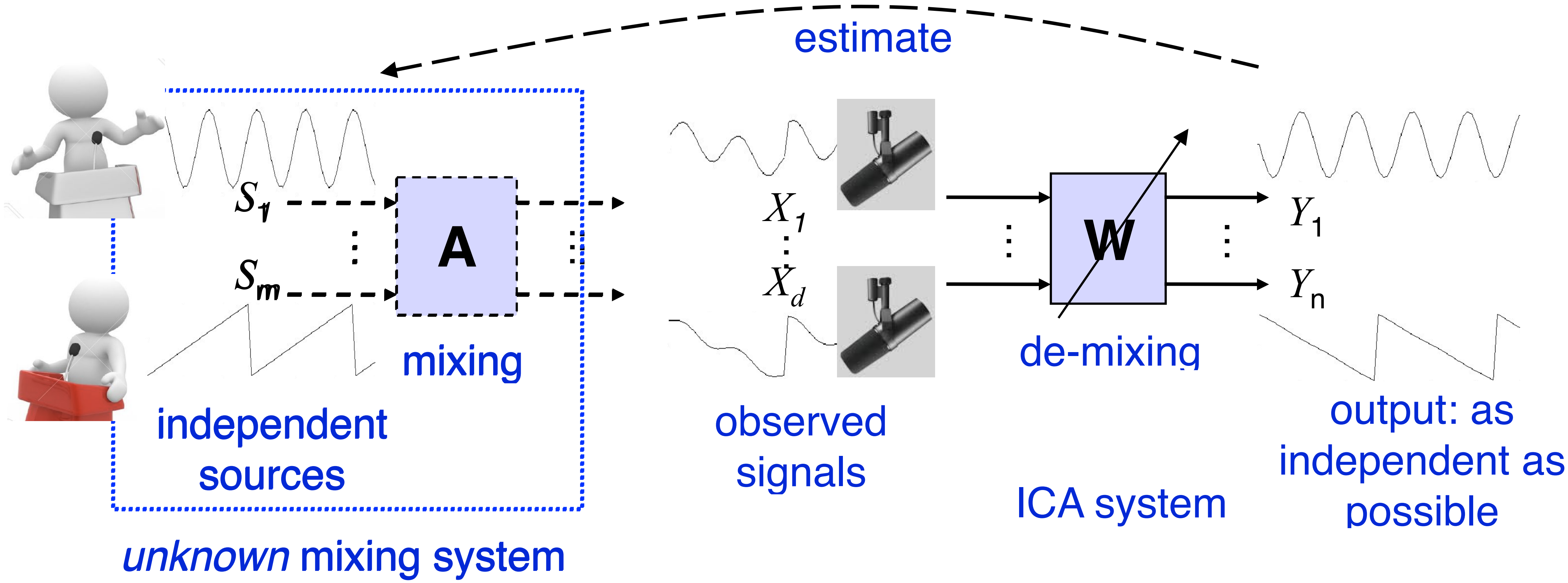
- Smaller entropy, more structural, more interesting
- "Purer" according to the central limit theorem



Which direction is more interesting?

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



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

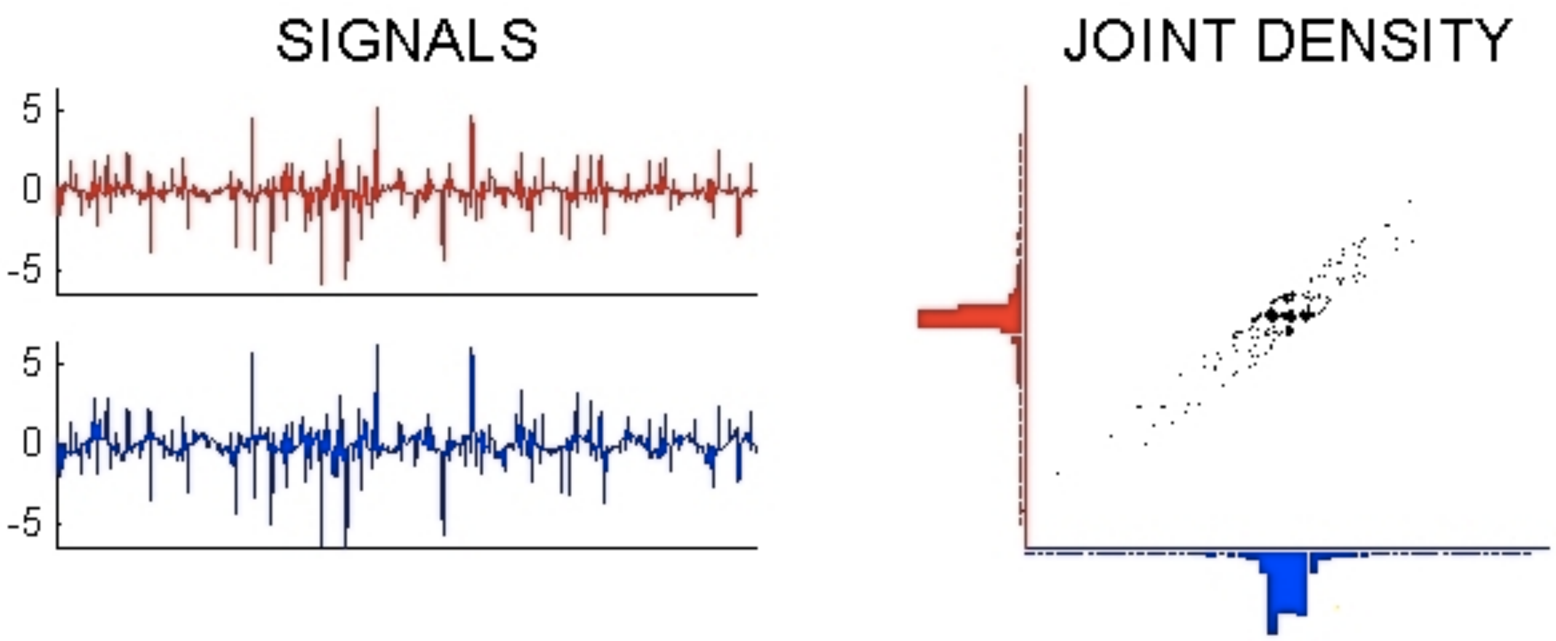
$$\begin{array}{c}X_1\\X_2\end{array}\begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix}\begin{array}{c}s_1\\s_2\end{array}$$

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

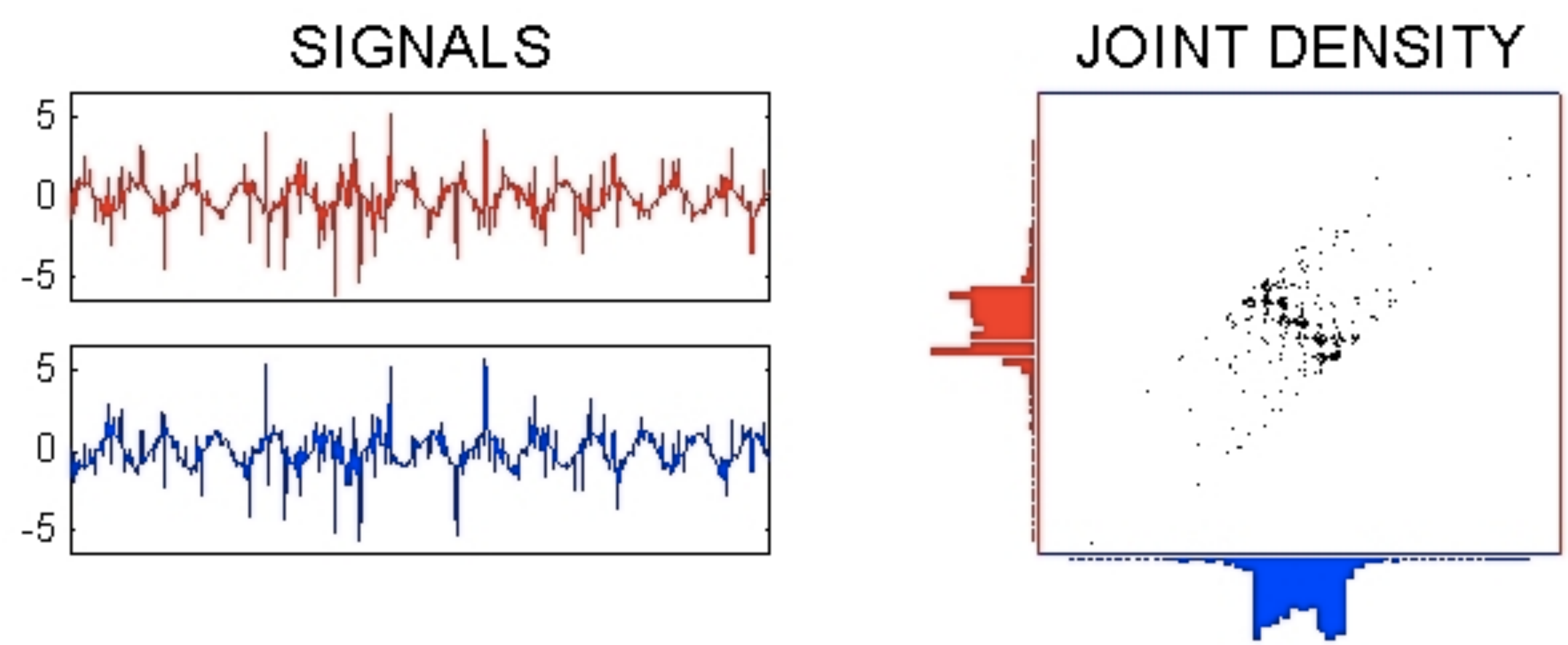
Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

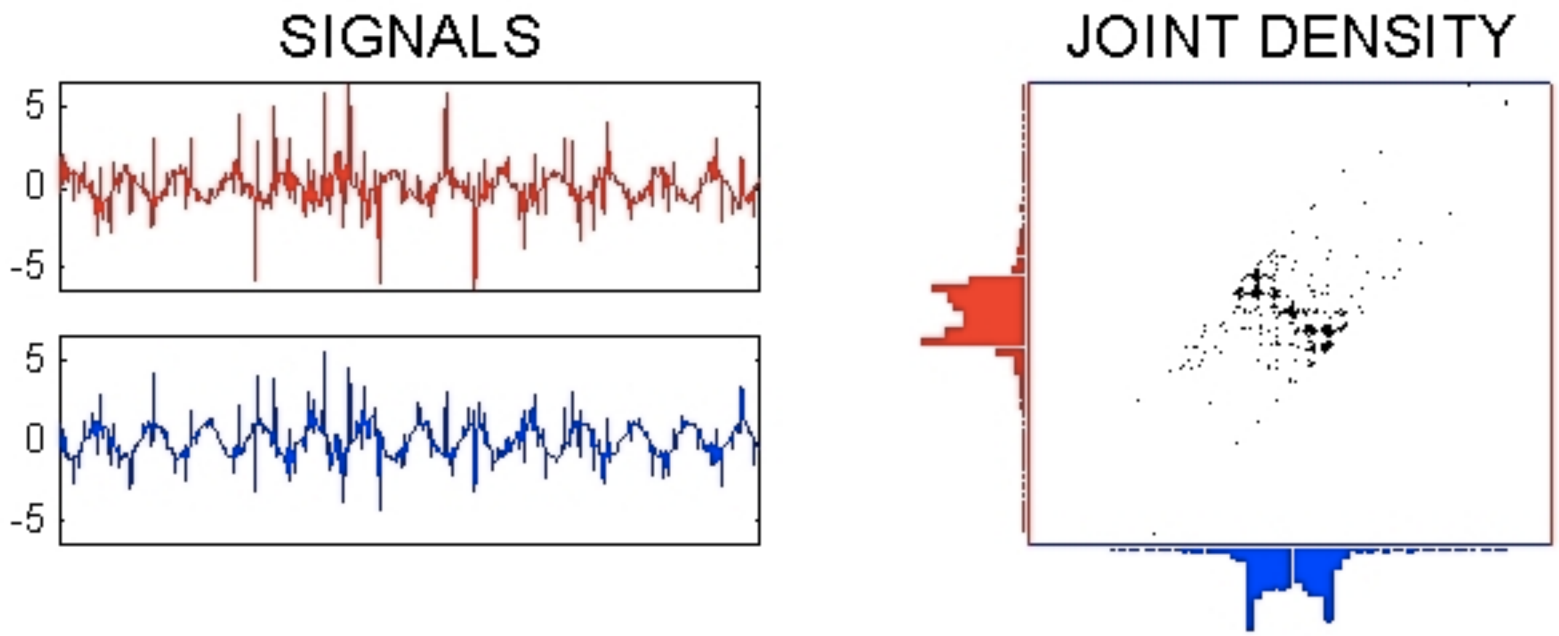
A Demo of the ICA Procedure



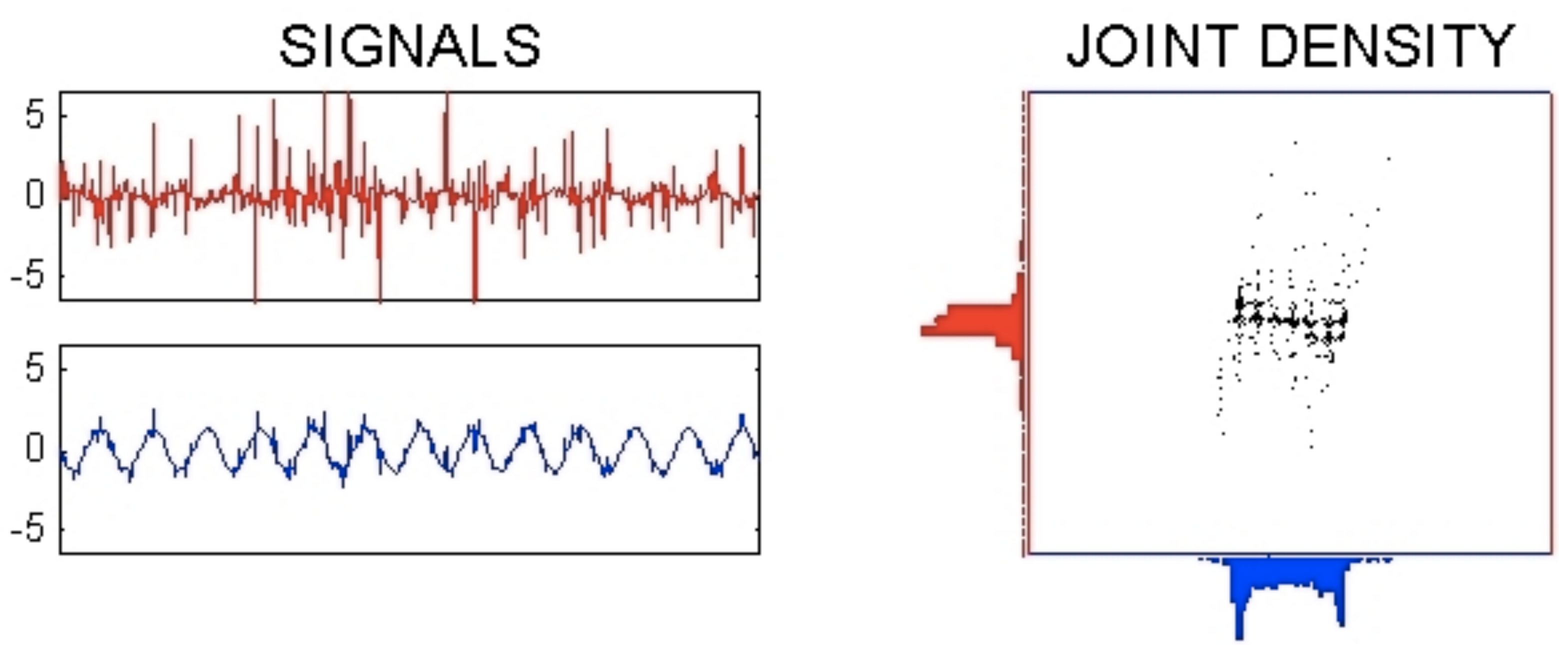
Input signals and density



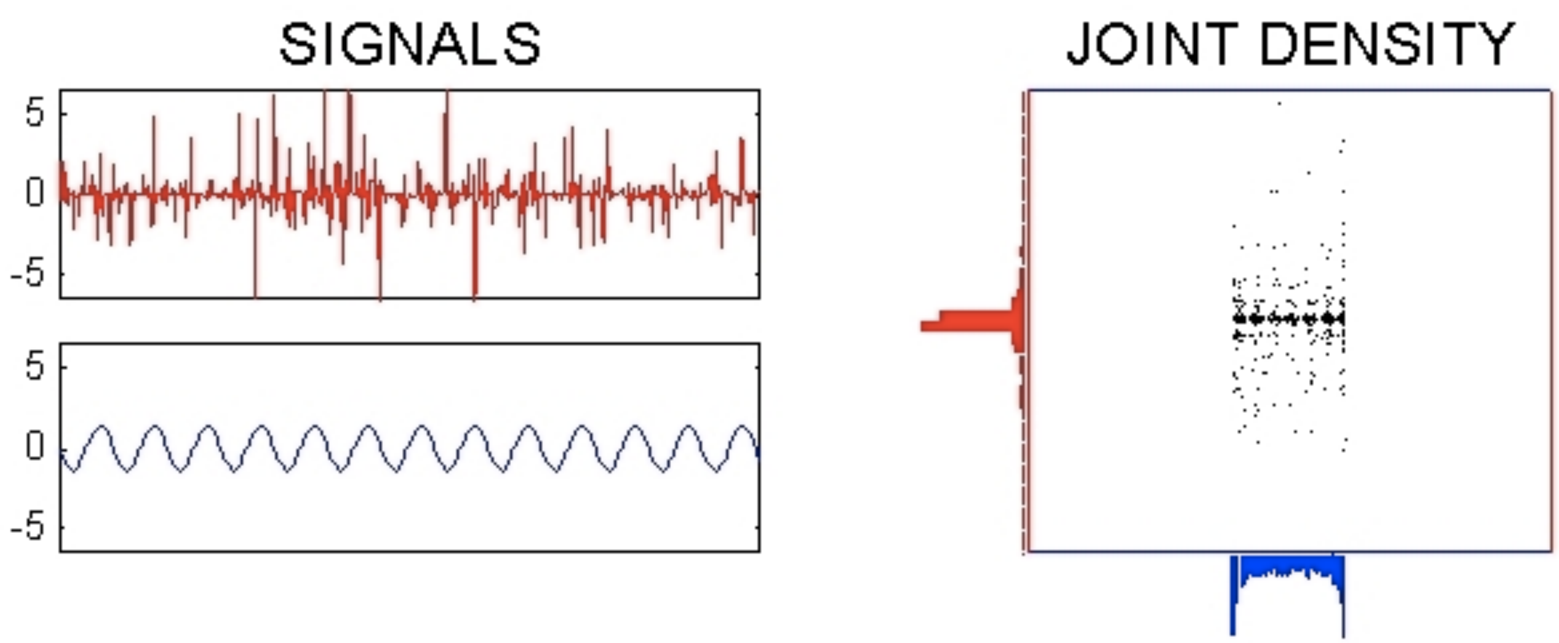
Whitened signals and density



Separated signals after 1 step of FastICA



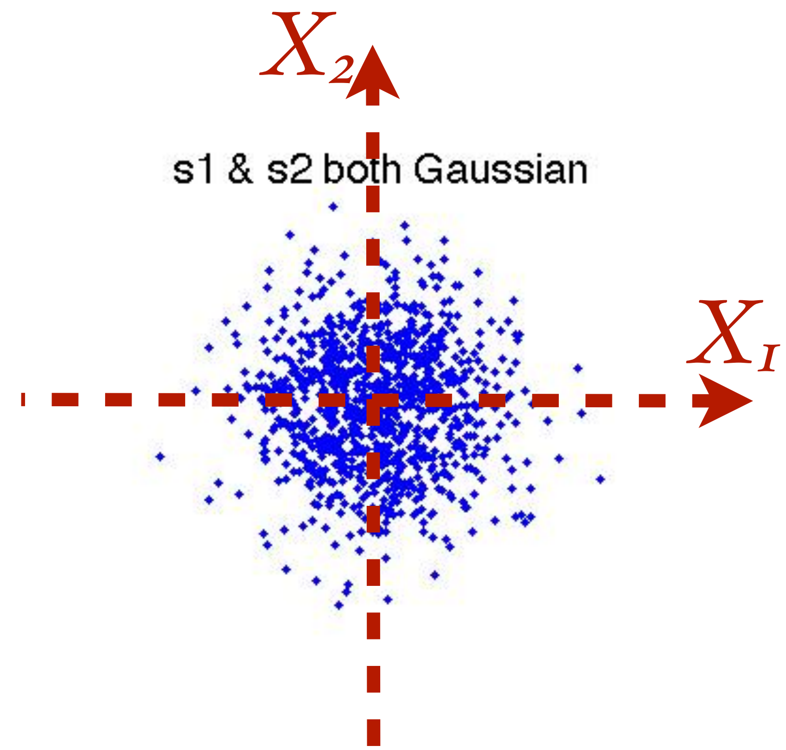
Separated signals after 3 steps of FastICA

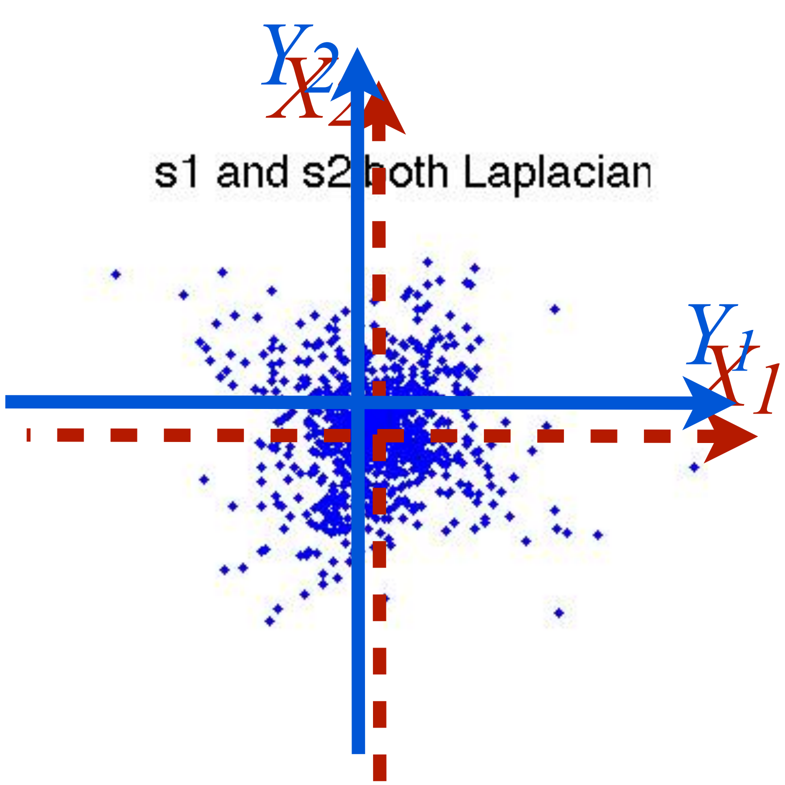


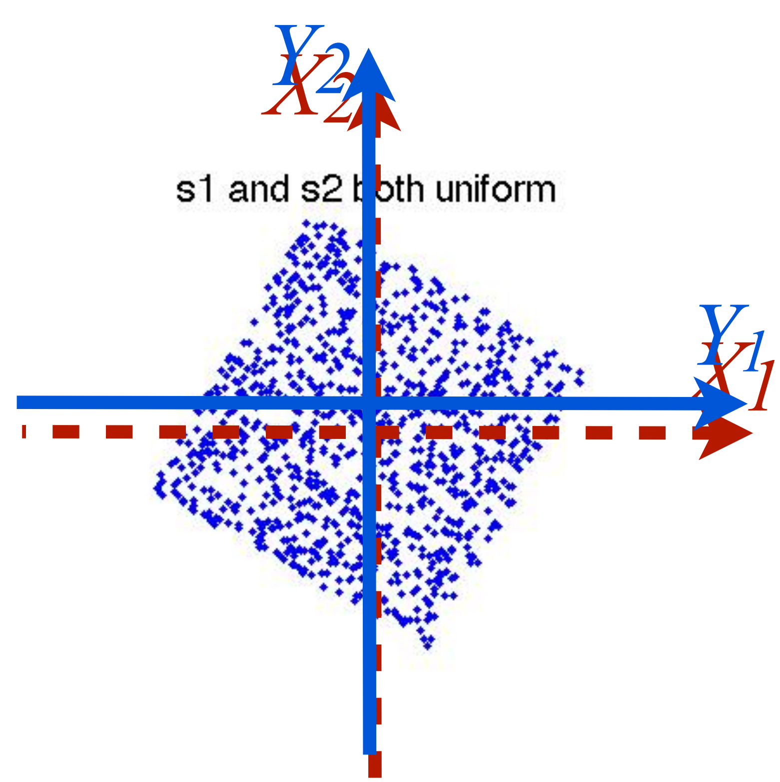
Separated signals after 5 steps of FastICA

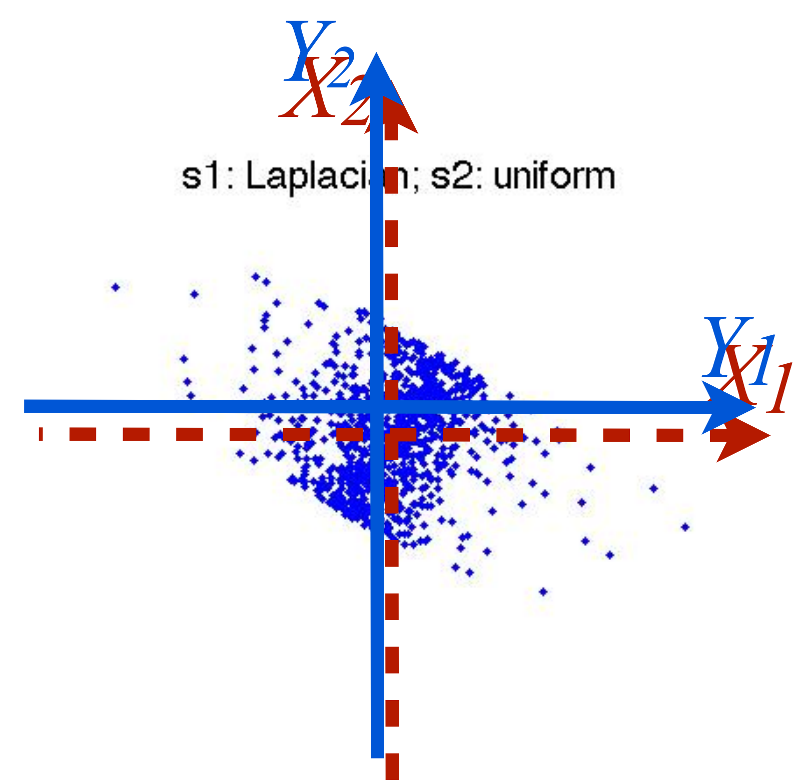
Intuition: Why ICA works?

- (After whitening with $\mathbf{Z}$=$\mathbf{QX}$) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{U} \cdot \mathbf{Z}$ to making Y_i independent
  - How to find $\mathbf{Q}$ such that cov($\mathbf{Z}$) = $\mathbf{I}$?
  - How to find $\mathbf{U}$ to achieve the independence?









Darmois-Skitovich Theorem

Darmois-Skitovitch theorem: Define two random variables, Y_1 and Y_2, as linear combinations of independent random variables S_i, $i = 1, ..., n$:

$$Y_1 = \alpha_1 S_1 + \alpha_2 S_2 + ... + \alpha_n S_n,$$
$$Y_2 = \beta_1 S_1 + \beta_2 S_2 + ... + \beta_n S_n.$$

If Y_1 and Y_2 are statistically independent, then all variables S_j for which $\alpha_j \beta_j \neq 0$ are Gaussian.

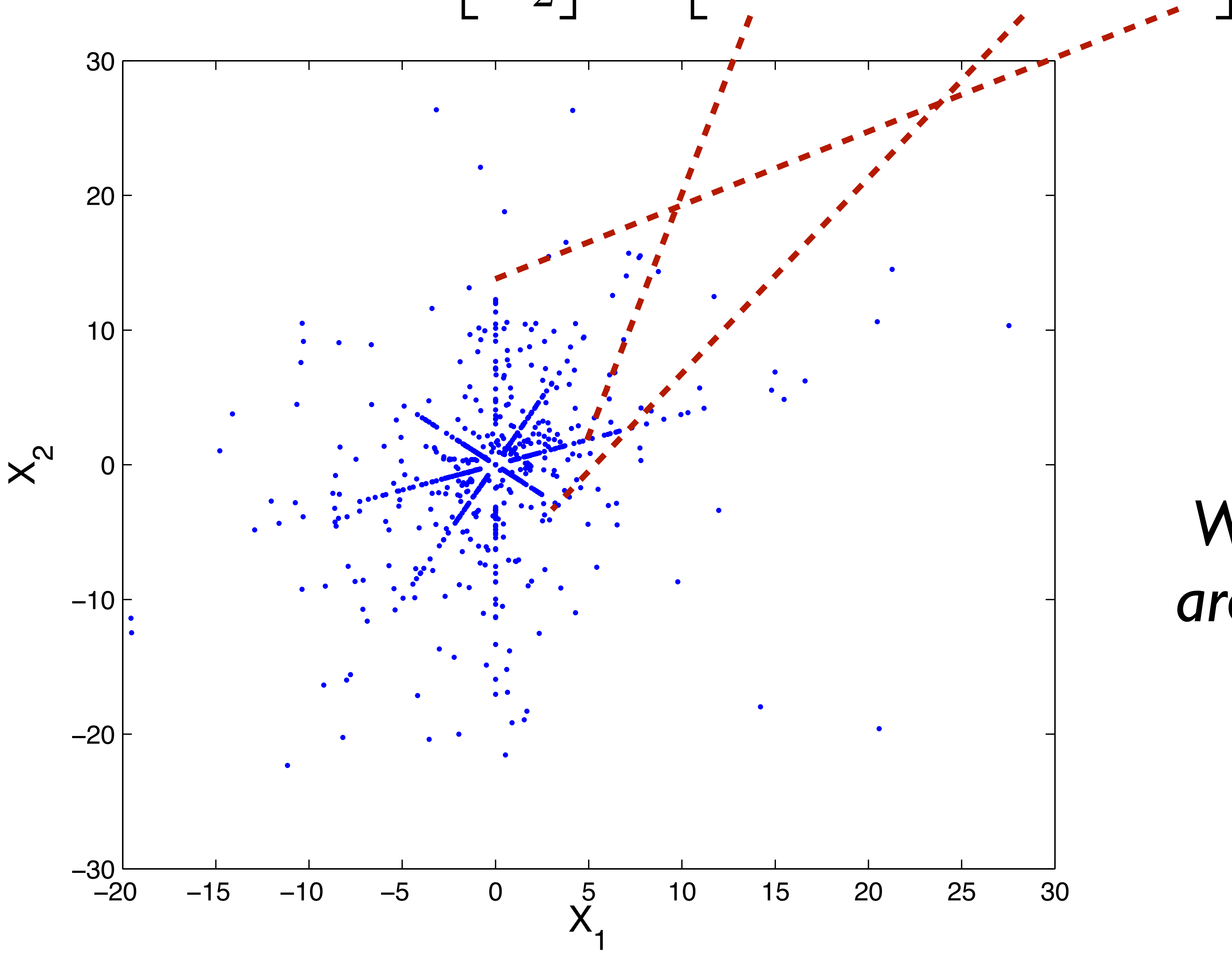
Cool! Can you then see the identifiability of the ICA problem?

Kagan et al., Characterization Problems in Mathematical Statistics. New York: Wiley, 1973

*

Overcomplete ICA: Illustration

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.4 & -0.9 & 0 \\ 0.3 & 0.8 & 0.8 & 1 \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}$$



What if they are Gaussian?

How ICA works? By Maximum Likelihood

- From a maximum likelihood perspective

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

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

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

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

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

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

(Change of variables)

$\log L$

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

- Or by mutual information minimization, or by information maximization…

How ICA works? By Mutual Information Minimization

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

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

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

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

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

- Nonnegative and zero iff Y_i are independent

- $H(X)$ = -E[log $p_X(X)$]: differential entropy--how random the variable is?

Hyvärinen et al., Independent Component Analysis...

How ICA works? Some Interpretation

- Some methods (e.g., FastICA) 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 ⇔ minimizing the entropies

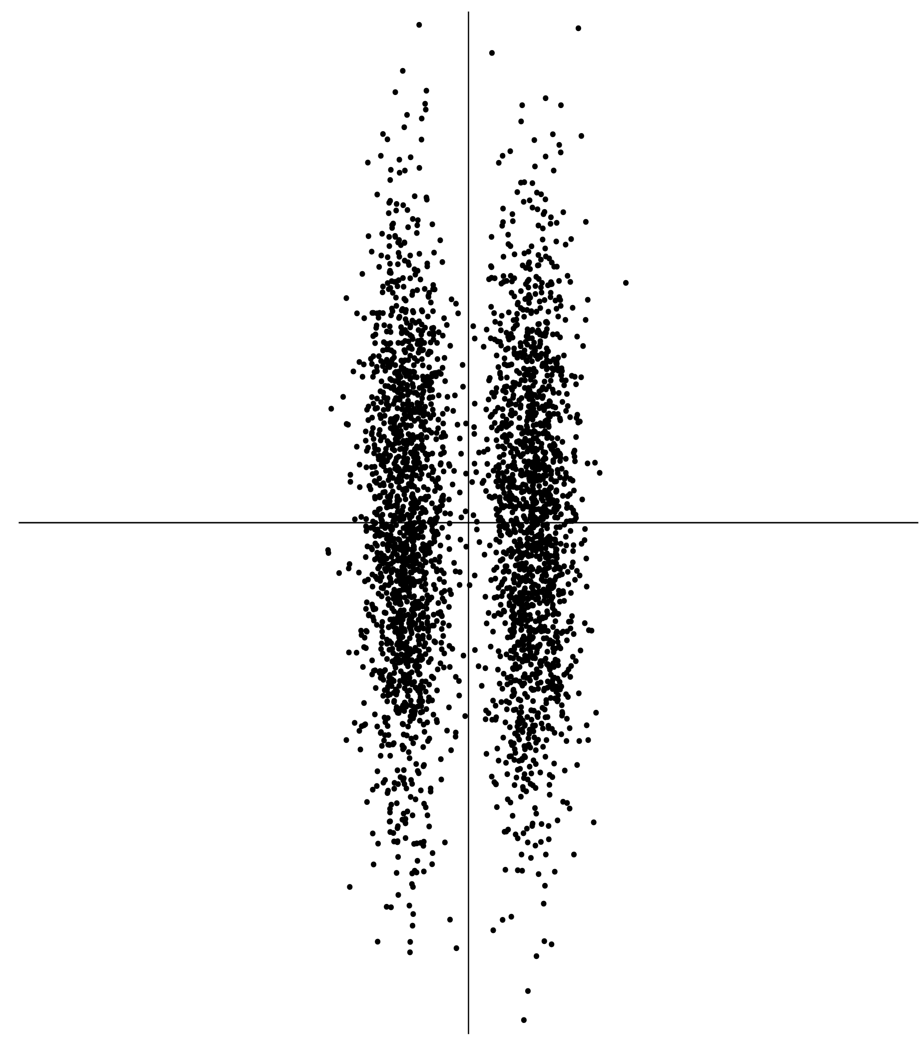
- Given the variance, the Gaussian distribution has the largest entropy (among all continuous distributions)

- Maximizing non-Gaussianity !

- FastICA adopts some approximations of **neg**entropy of each output Y_i

Non-Gaussianity is Informative in the Linear Case

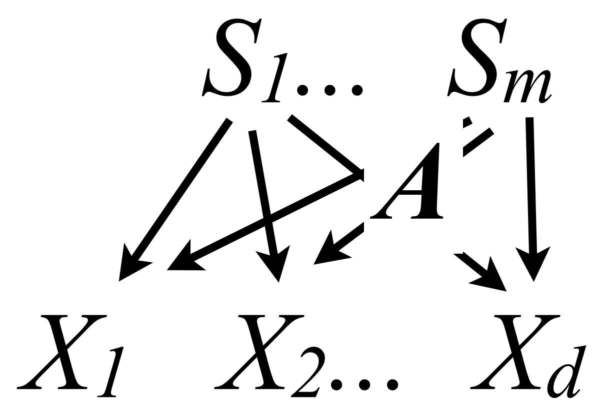
- Smaller entropy, more structural, more interesting
- "Purer" according to the central limit therom



Which direction is more interesting?

Hyvärinen et al., Independent Component Analysis, 2001

Connecting ICA to Causal Analysis



- With identifiability of $\boldsymbol{A}$ (compare it with factor analysis)
- Can we use it for causal analysis?

Outline

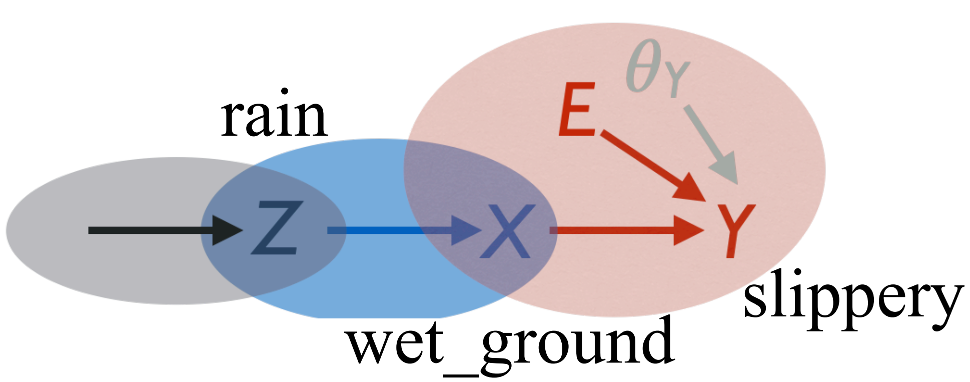
- Basic multivariate analysis & connection to causal analysis
    - PCA, factor analysis, ICA…
- Constraint-based methods for causal discovery
- Basic idea of score-based methods

What Information Helps Find Causality?

- Connection between causal structure and statistical data under *suitable assumptions*

- Note this "irrelevance":

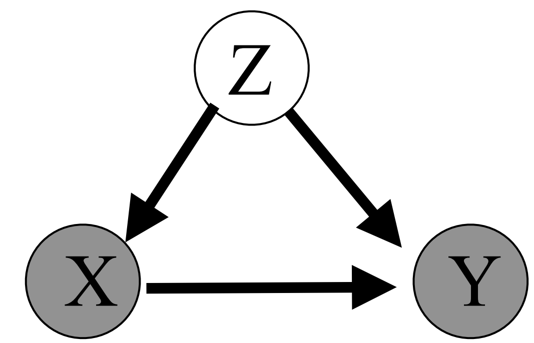
If there is no common cause of X and Y, the generating process for cause X is irrelevant to ("independent" from) that generates effect Y from X



- conditional independence among variables;
- independent noise condition;
- minimal (and independent) changes...

Causal Sufficiency

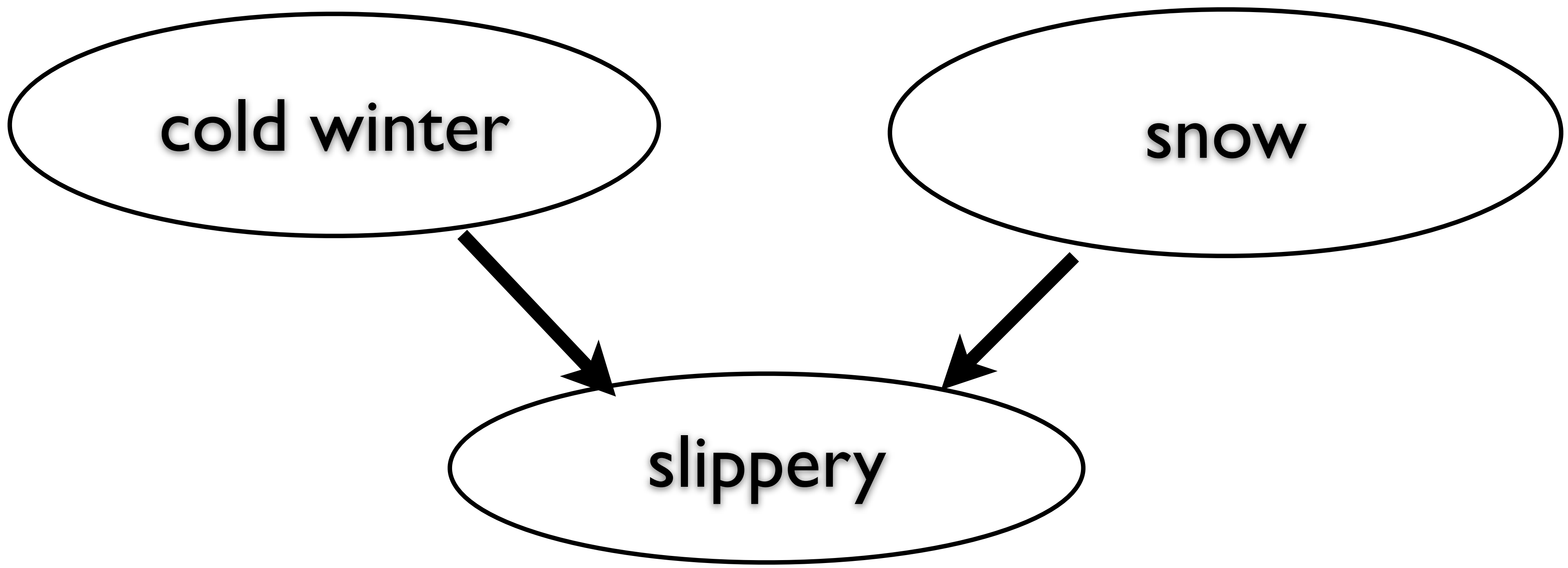
- A set of random variables $\boldsymbol{V}$ is causally sufficient if $\boldsymbol{V}$ contains every common cause (with respect to $\boldsymbol{V}$) of any pair of variables in $\boldsymbol{V}$



- $\boldsymbol{V} = \{X, Y, Z\}$: causally sufficient
- $\boldsymbol{V} = \{X, Y\}$: causally insufficient
- Methods exist in causally **insufficient** cases, e.g., FCI (*Chapter 6 of the SGS book*)

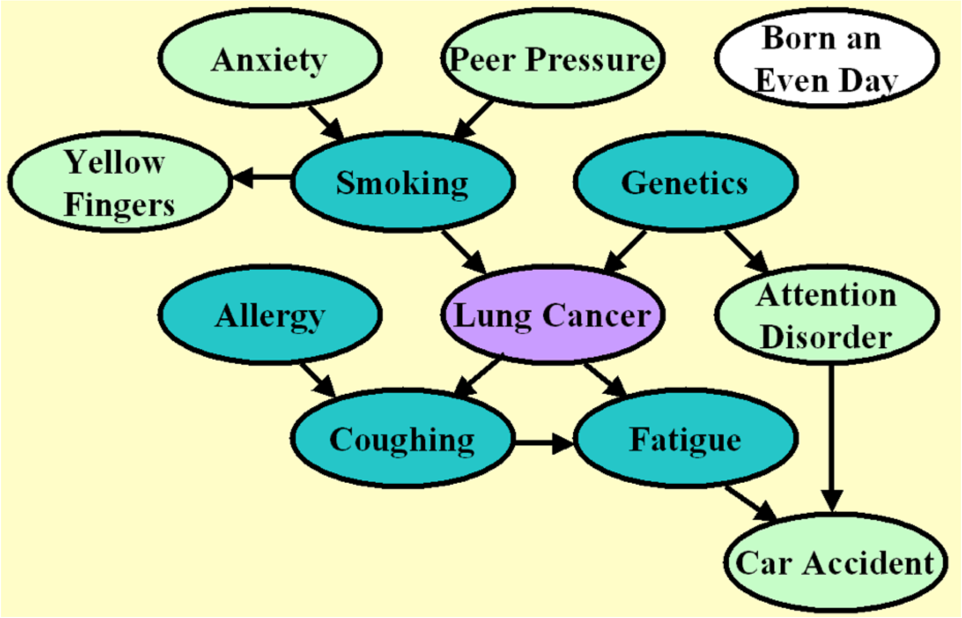
SGS Book, Chapter 5 (for causally sufficient structures); Chapter 6 (without causal sufficiency)

V-Structures



Why so interesting?

We can See CI Relations from DAGs...



- Local Markov condition
- Global Markov condition
  - d-separation implies conditional independence:

$P(\mathbf{V})$, where $\mathbf{V}$ denotes the set of variables, obeys the global Markov condition (or property) according to DAG $\mathcal{G}$ if for any disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, we have

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

Going from CI to Graph?

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

- Contrapositive:
  - Conditional dependence implies d-connection
  - What if variables are conditionally independent?
- Can we recover the property of the underlying graph from CI relations with Markov condition?
  - Arbitrary $P(\mathbf{V})$ would satisfy the global Markov condition according to G^f *in which there is an edge between each pair of variables*: trivial !
  - Under what assumptions can we have CI ⇒ d-separation?

Causal Structure vs. Statistical Independence
(SGS, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

Y → X → Z

Y -- X -- Z **?**

Statistical independence(s)

$Y \perp\!\!\!\perp Z \mid X$

Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

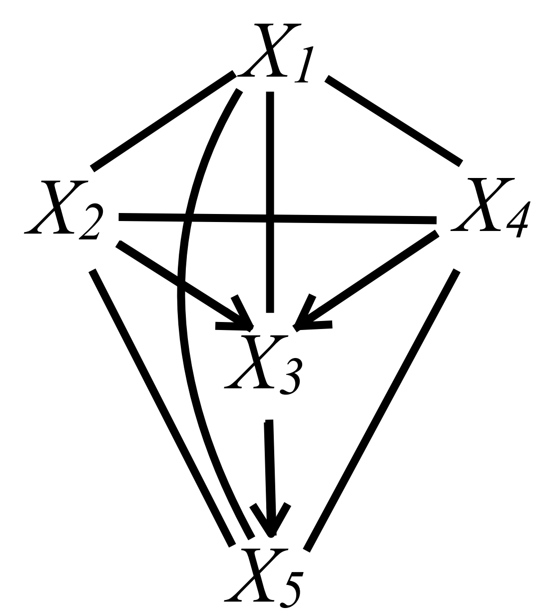
Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

(Typical) Constraint-Based Causal Discovery

- Conditional independence constraints between each variable pair
    - Illustration: the PC algorithm
    - Extensions: the FCI algorithm...

X1	X2	X3	X4	X5
-1.1	1	1.3	0.2	-0.7
2.1	2	3.1	-1.3	-1.6
3.1	4.2	-2.6	0.6	2.1
2.3	-0.6	-3.5	0.8	2.3
1.3	-1.7	0.9	2.4	-1.4
-1.8	0.9	-1.3	0.9	0.7
...	...	...	...	...

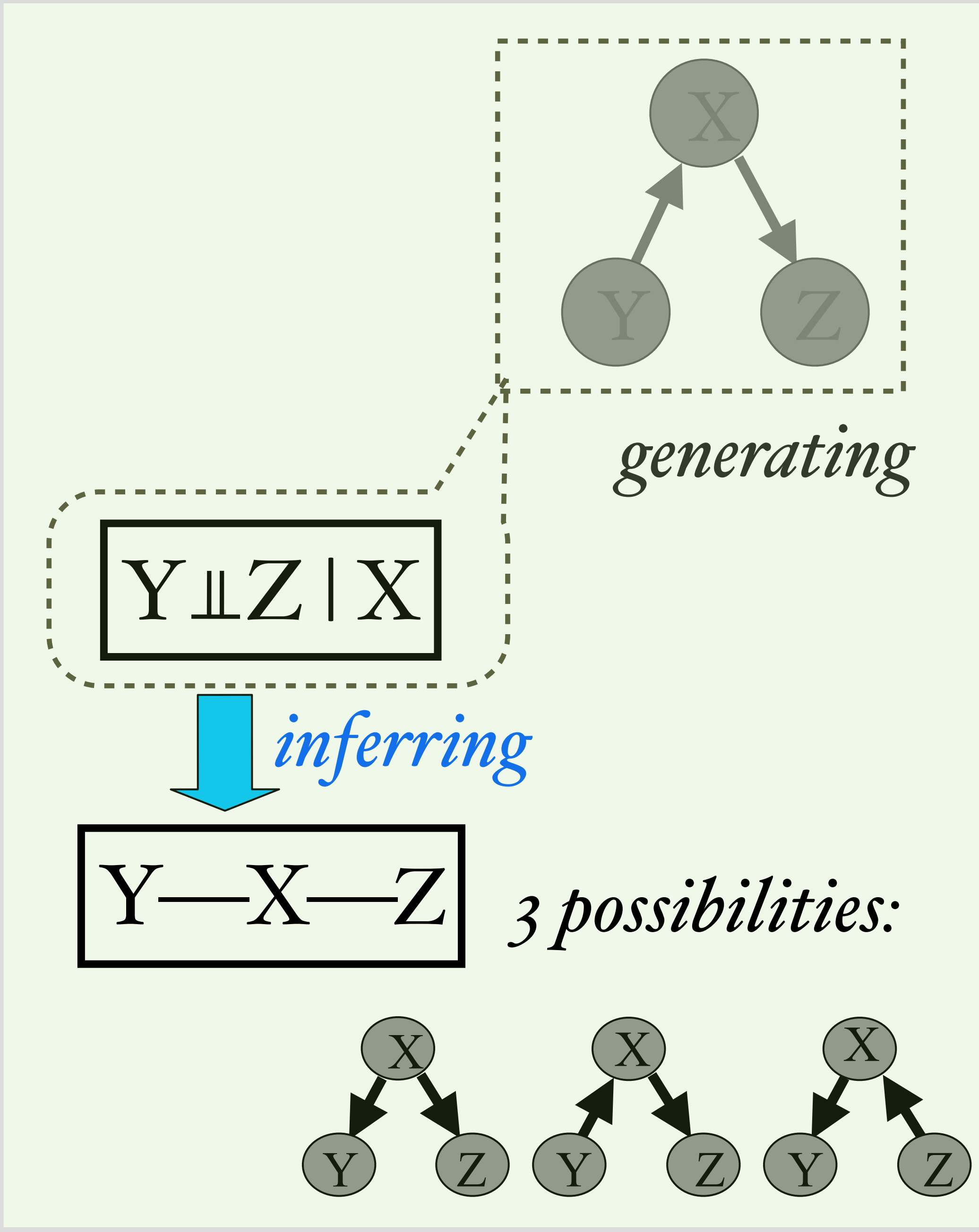
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_2 \perp\!\!\!\perp X_4 \mid X_1$
$X_2 \perp\!\!\!\perp X_5 \mid X_3$
$X_4 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_3 \mid \{X_2, X_4\}$

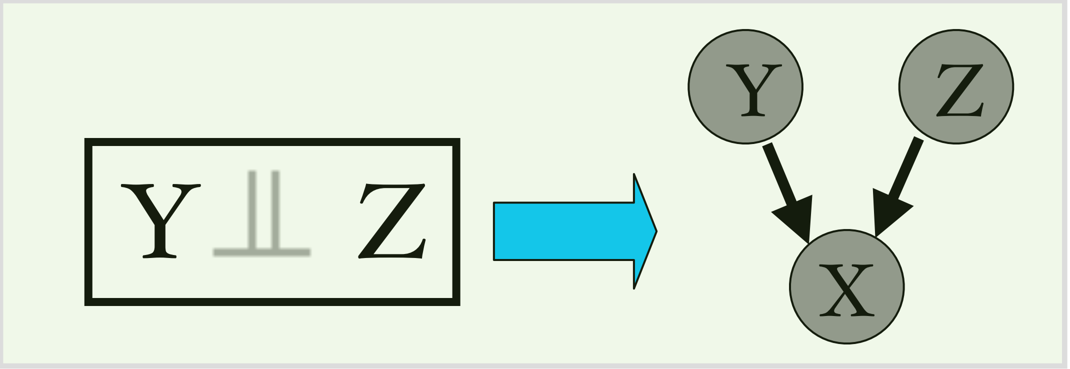


– Spirtes, Glymour, and Scheines. Causation, Prediction, and Search. 1993.

Constraint-Based Causal Discovery

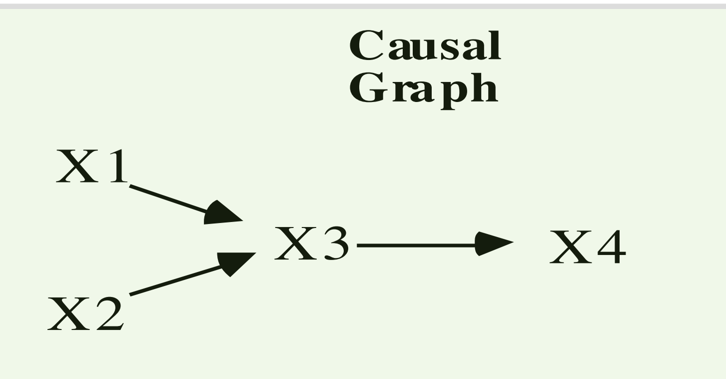
- (Conditional) independence <u>constraints</u> $\Rightarrow$ candidate causal structures
  - Relies on causal Markov condition & faithfulness assumption
  - PC algorithm (Spirtes & Glymour, 1991)
  - *Step 1*: X and Y are adjacent iff they are dependent conditional on every subset of the remaining variables (SGS, 1990)
  - *Step 2*: Orientation propagation
- v-structure
- Markov equivalence class, represented by a pattern
  - same adjacencies; $\rightarrow$ if all agree on orientation; — if disagree





Example I

Step I: finding skeleton



Independcies

$X1 \perp\!\!\!\perp X2$

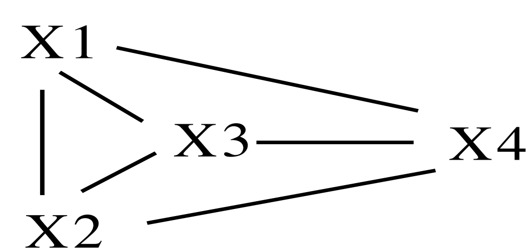
$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Step II: finding v-structure and doing orientation propagation

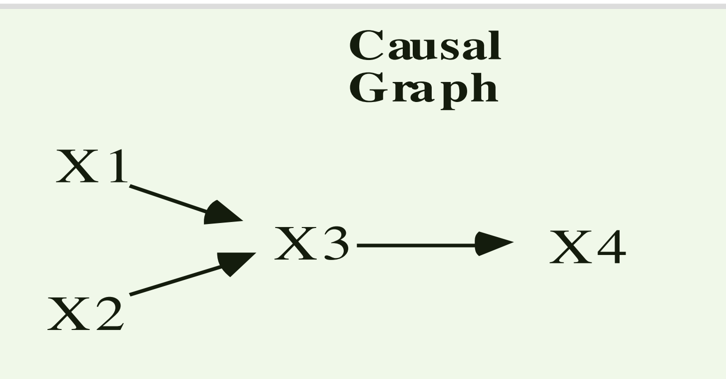
---

Begin with:



Example I

Step I: finding skeleton



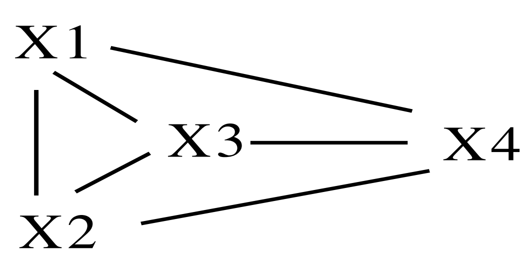
Independcies

$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

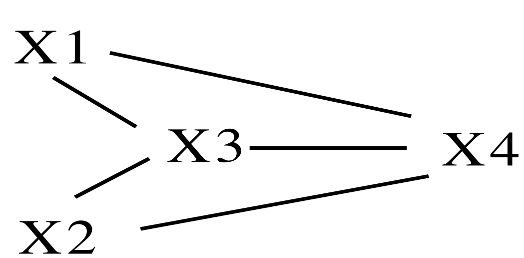
$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Begin with:



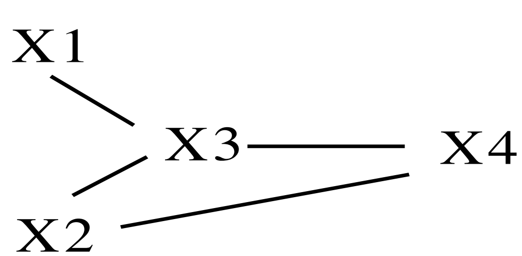
From

$X1 \perp\!\!\!\perp X2$



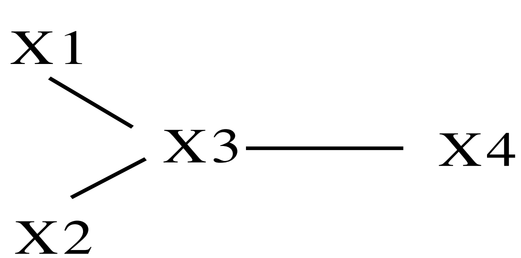
From

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$



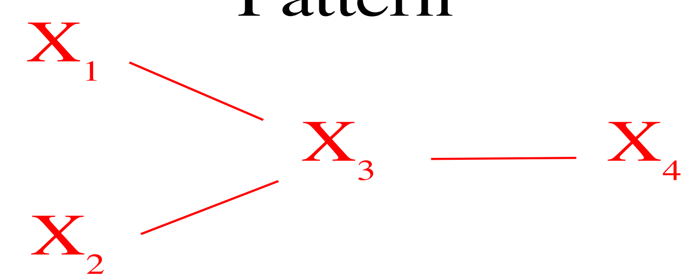
From

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

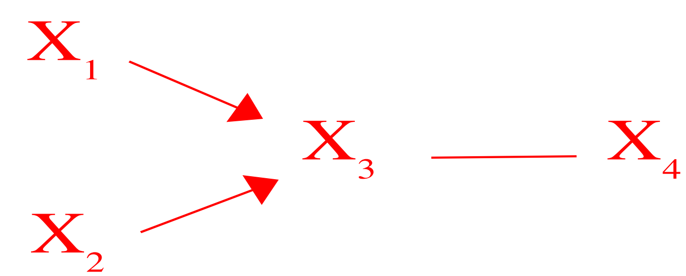


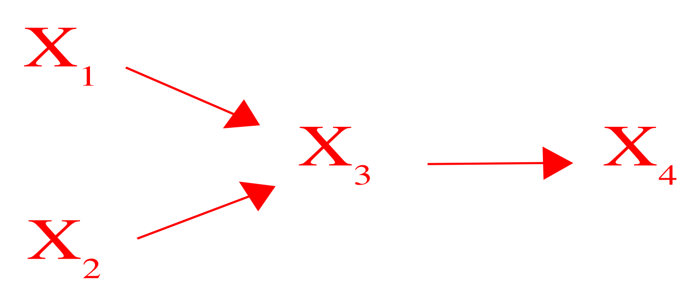
Step II: finding v-structure and doing orientation propagation

Pattern



$X_1 \perp\!\!\!\perp X_2$:





PC Algorithm

Test for (conditional) independence with an increased cardinality of the conditioning set

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

B.)

$n = 0$.

repeat

repeat

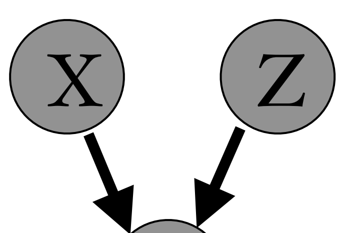
select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ in **Sepset**(X,Y) and **Sepset**(Y,X);

until all ordered pairs of adjacent variables X and Y such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

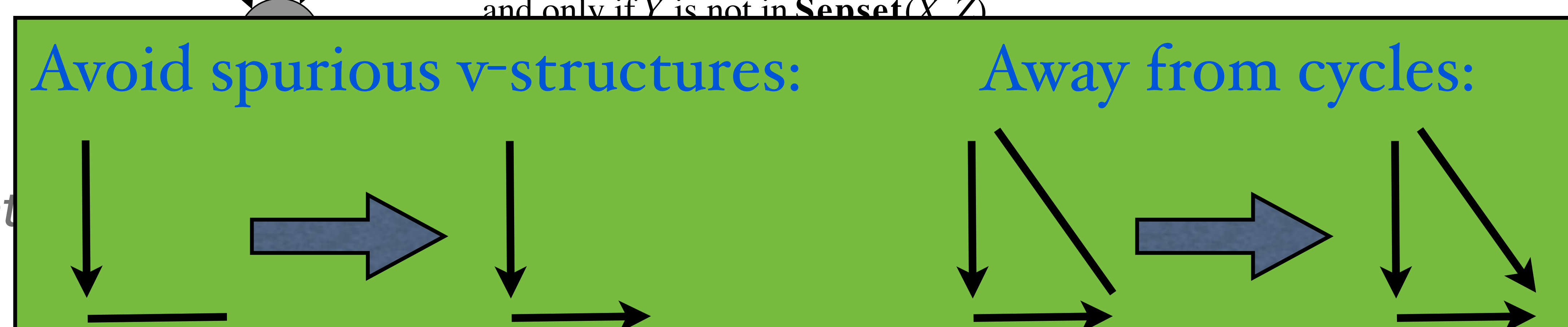
until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

Finding V-structures



C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

Orient



here is no

hen orient

until no more edges can be oriented.

(Independence) Equivalent Classes: Patterns

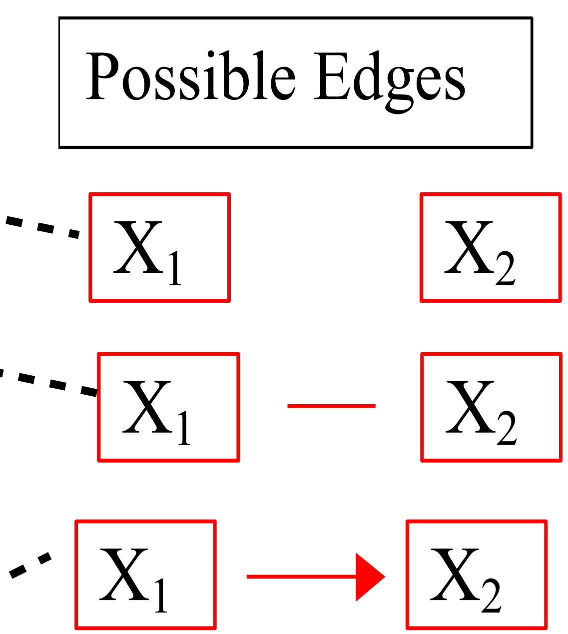
- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)

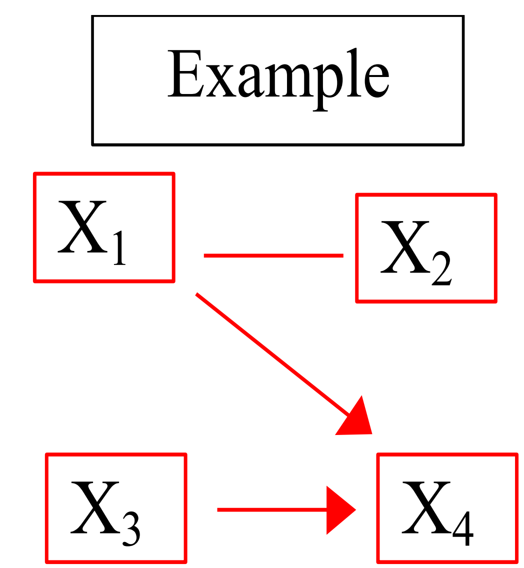
- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

X_1 and X_2 are not adjacent in any member of the equivalent class

$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others

$X_1 \rightarrow X_2$ in every member of the equivalent class



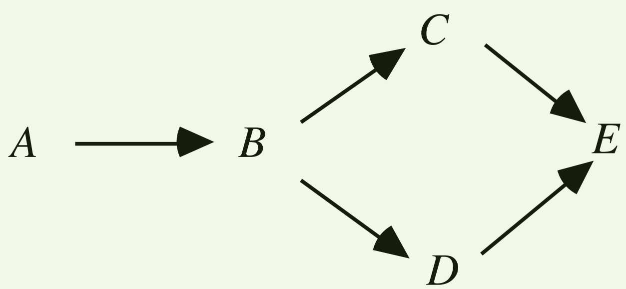


How many DAGs in this class?

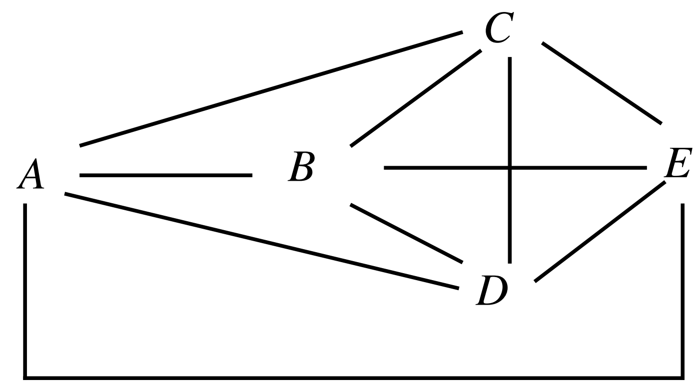
Example II (From SGS Book)

Step I

Step II



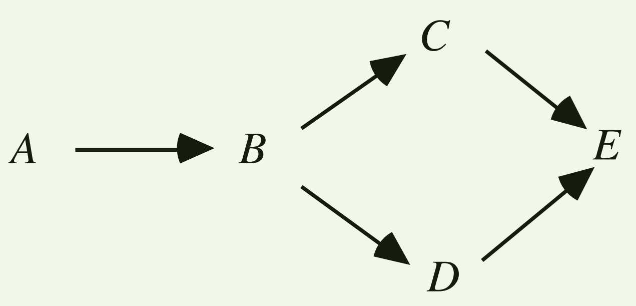
True Graph



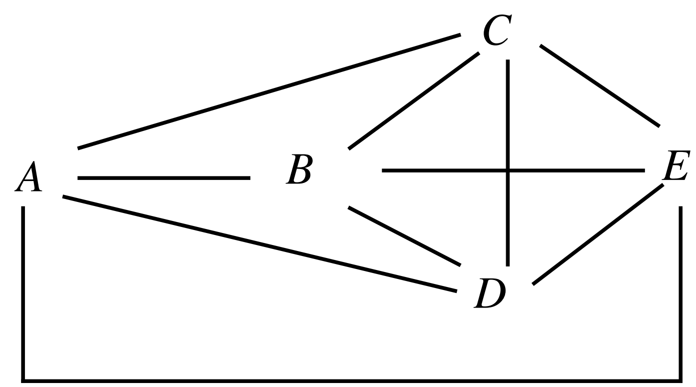
Complete Undirected Graph

Example II (From SGS Book)

Step I



True Graph



Complete Undirected Graph

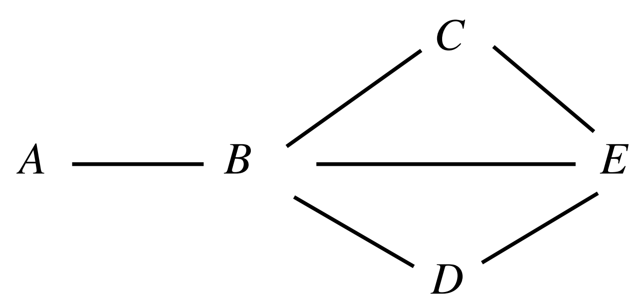
n = 0 No zero order independencies

n = 1 First order independencies

$A \perp\!\!\!\perp C \mid B$ $\quad A \perp\!\!\!\perp D \mid B$

$A \perp\!\!\!\perp E \mid B$ $\quad C \perp\!\!\!\perp D \mid B$

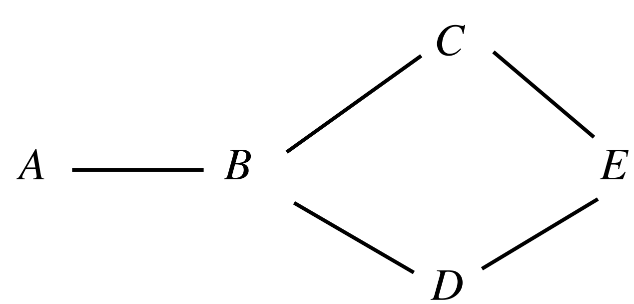
Resulting Adjacencies



n = 2: Second order independencies

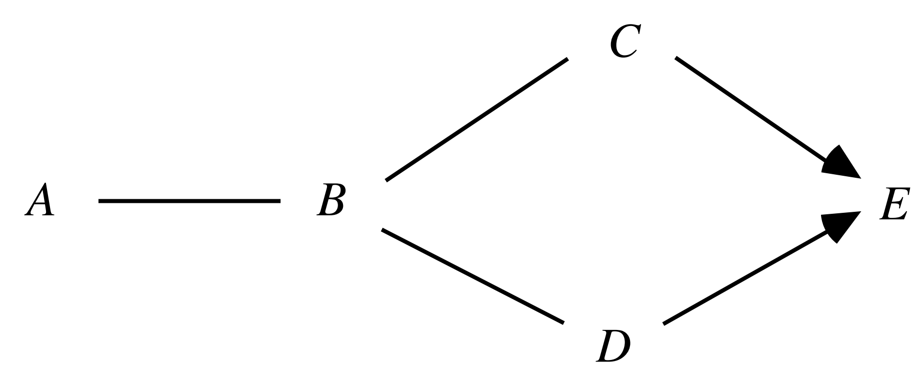
$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies



Step II

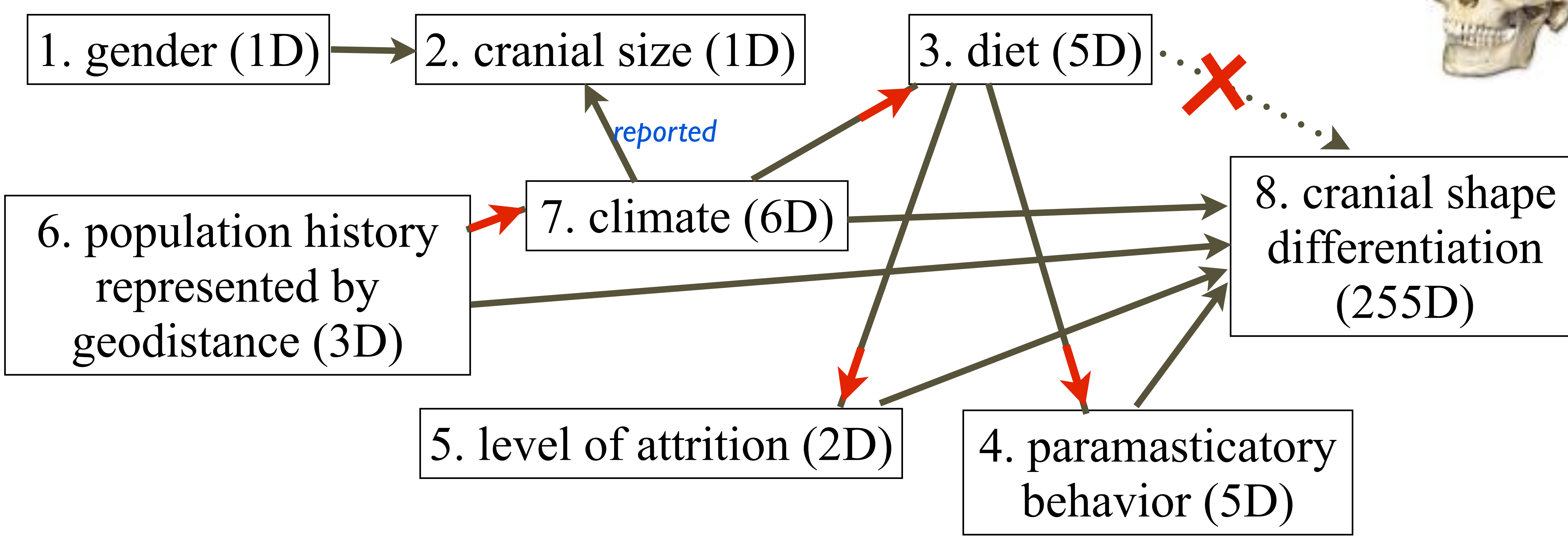
Pattern



Result on the Archeology Data

Thanks to collaborator Marlijn Noback

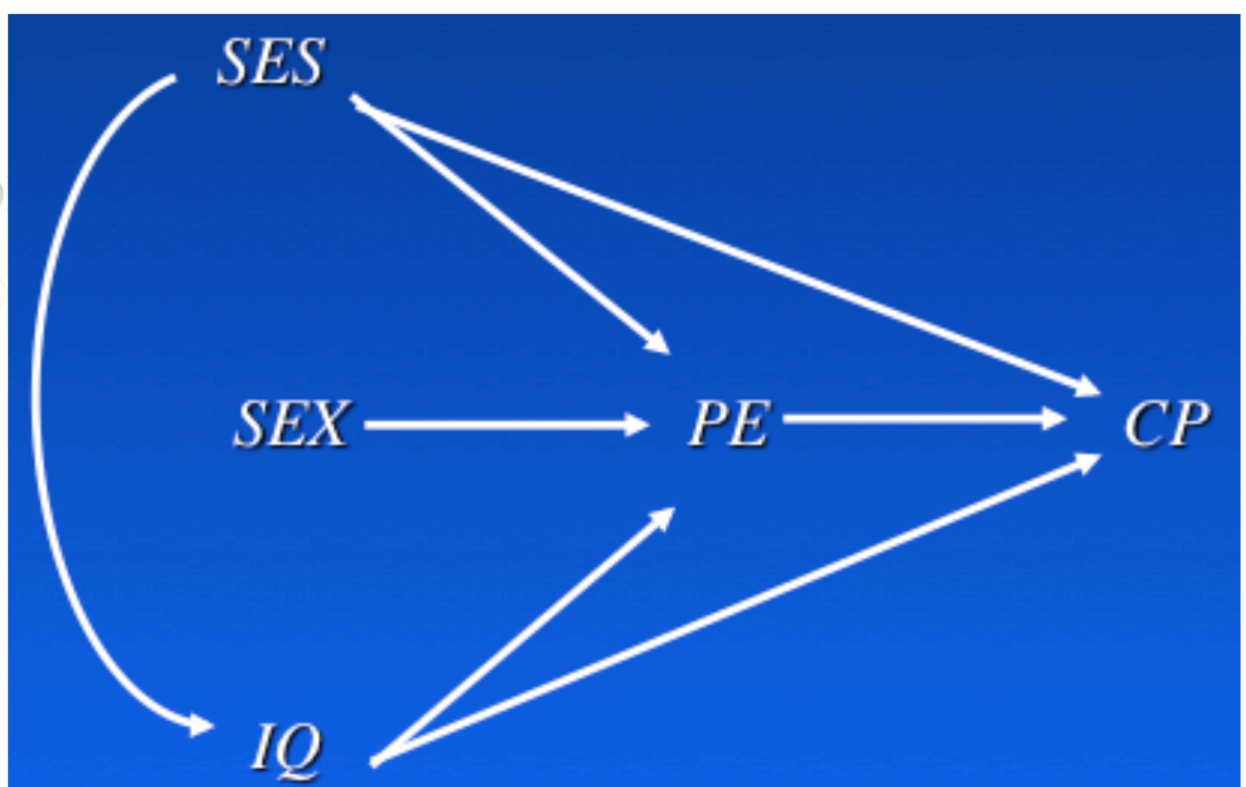
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



Example 2: College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors.

SEX [male = 0, female = 1]
IQ = Intelligence Quotient [lowest = 0, highest = 3]
CP = college plans [yes = 0, no = 1]
PE = parental encouragement [low = 0, high = 1]
SES = socioeconomic status [lowest = 0, highest = 3]



Dealing with Confounders?

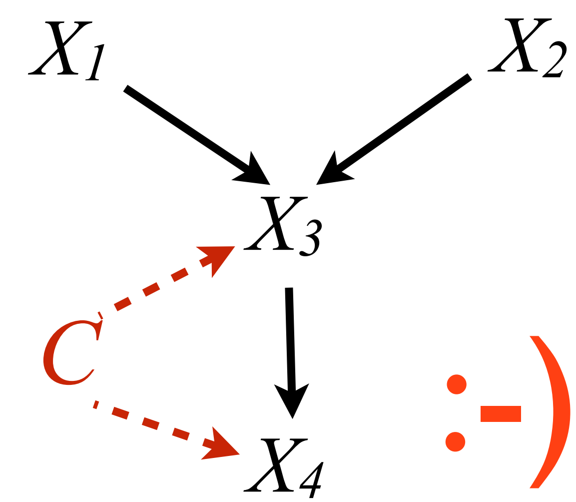
Example I

$X_1 \perp\!\!\!\perp X_2;$

$X_1 \perp\!\!\!\perp X_4 \mid X_3;$

$X_2 \perp\!\!\!\perp X_4 \mid X_3.$

Possible to have confounders behind X_3 and X_4?



E.g., X_1: Raining; X_3: wet ground; X_4: slippery.

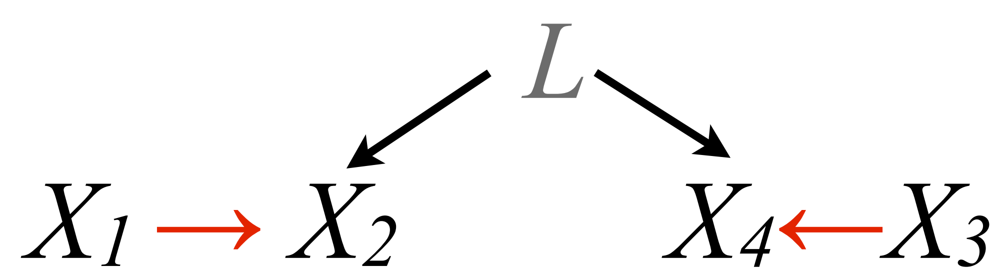
Example II

$X_1 \perp\!\!\!\perp X_3;$

$X_1 \perp\!\!\!\perp X_4;$

$X_2 \perp\!\!\!\perp X_3.$

Are there confounders behind X_2 and X_4?

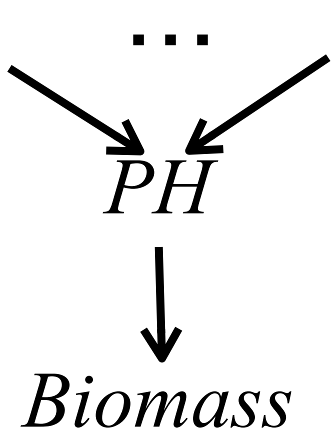


E.g., X_1: I am not sick; X_2: I am in this lecture room; X_4: you are in this lecture room; X_3: you are not sick.

(*See the FCI algorithm*)

* I know There Is No Confounder: Example

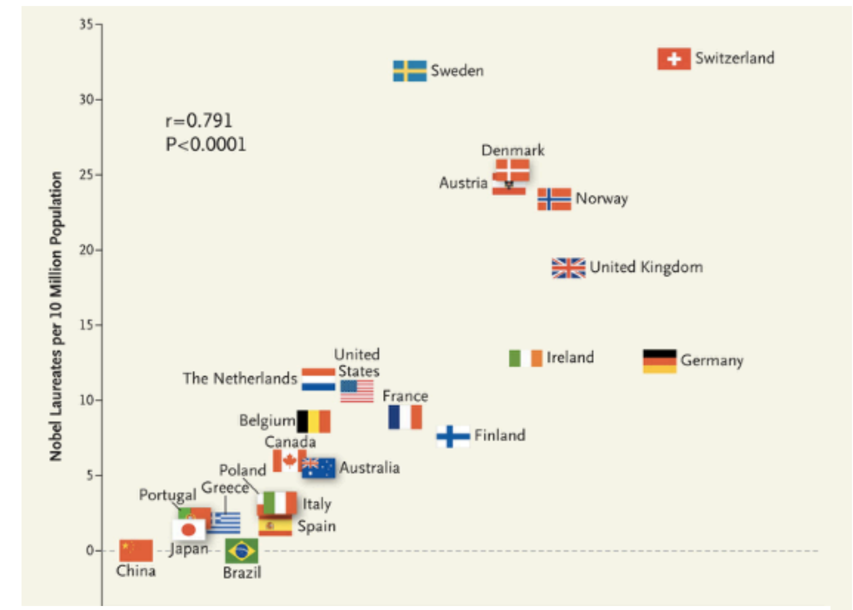
- In the 1970s, the Edison Electric Company in North Carolina was concerned about the effects on plant growth of acid rain produced by emissions from its electric generators.
- The investigators chose samples from the Cape Fear estuary, where the Cape Fear River flows into the Atlantic Ocean.
- obtained 45 samples of Spartina grass up and down the estuary, and measured 13 variables in the samples, including **concentrations of various minerals, acidity (pH), salinity, and the outcome variable, the biomass of each sample**
- The PC algorithm found that among **the measured variables the only *direct* cause of biomass was pH**.
- Y-structure: no confounder!
- Later verified by intervention-based analysis

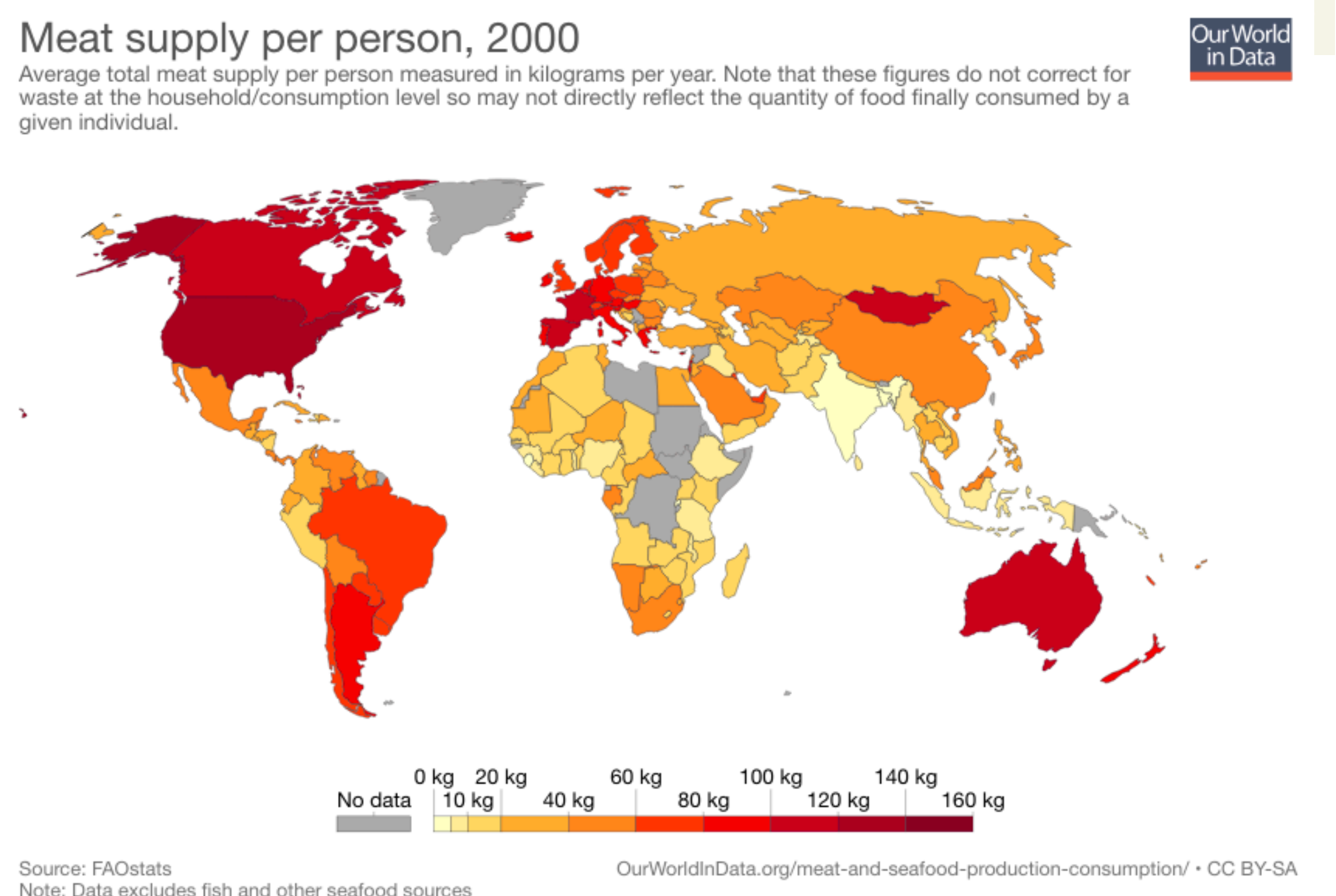


* I Know There must Be Confounders: examples

$$\begin{tikzcd}
 & & L \arrow[dl] \arrow[dr] & & \\
X_1 \arrow[r] & X_2 & & X_4 & X_3 \arrow[l]
\end{tikzcd}$$

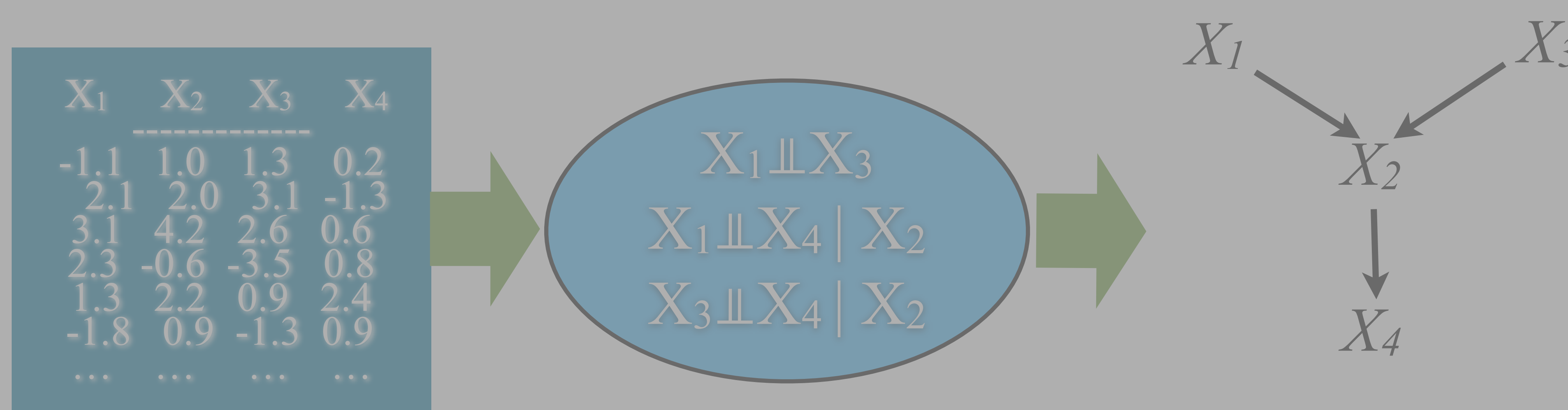
- X_1: I am not sick; X_2: I am in class; X_4: you are in class; X_3: you are not sick
- X_1: European/South American country; X_2: leading in science; X_4: Chocolate consumption; X_3: meat supply per person



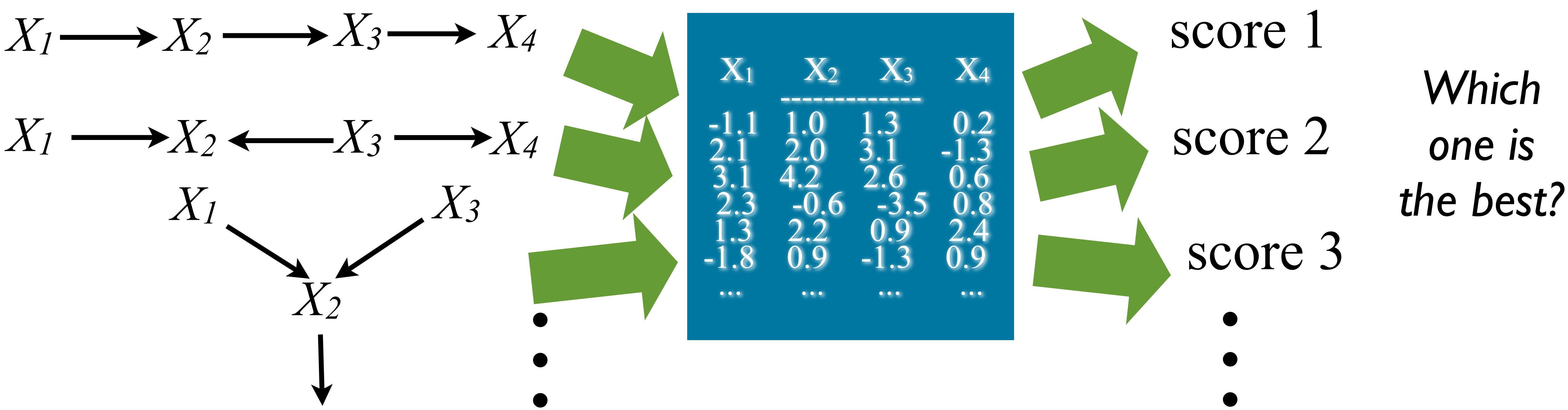


Constraint-Based vs. Score-Based

- Constraint-based methods



- Score-based methods



(Score may be BIC, AIC, etc.)

GES (Greedy Equivalence Search): Score Function

- Assumptions: The score is
  - score equivalent (i.e., assigning the same score to equivalent DAGs)
  - locally consistent: score of a DAG increases (decreases) when adding any edge that eliminates a false (true) independence constraint
  - decomposable: $Score(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} Score(X_i, \mathbf{Pa}_i^{\mathcal{G}})$
- E.g., BIC: $S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D} | \hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$

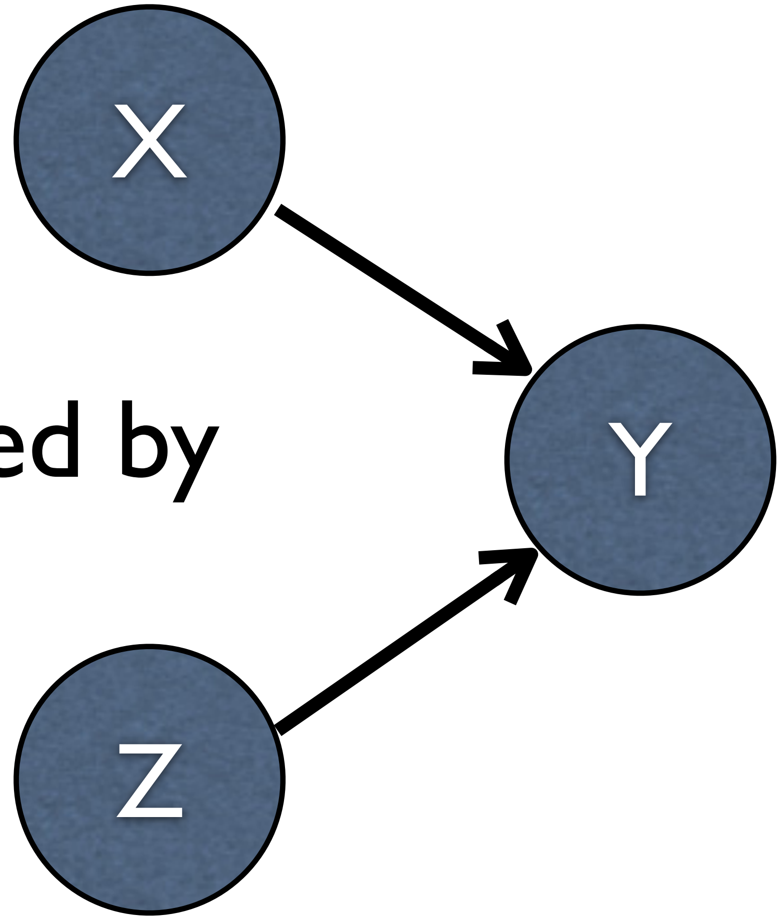
Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002

GES: Search Procedure

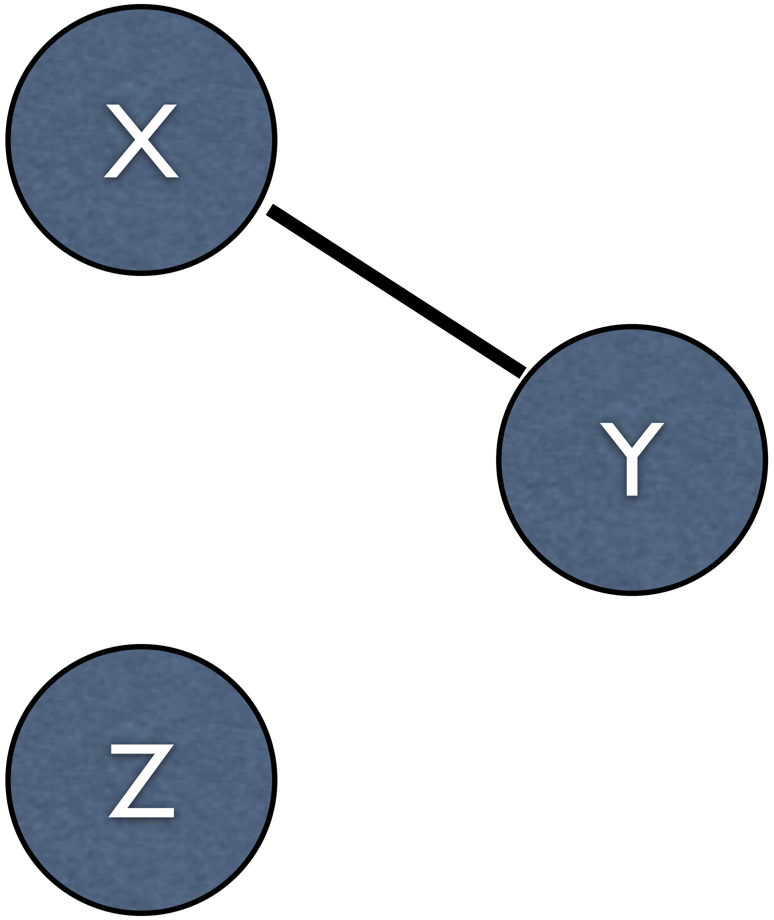
- Performs forward (addition) / backward (deletion) equivalence search through the space of DAG equivalence classes
  - Forward Greedy Search (FGS)
    - Start from some (sparse) pattern (usually the empty graph)
    - Evaluate all possible patterns with one more adjacency that entail strictly fewer CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum
  - Backward Greedy Search (BGS)
    - Start from the output of Stage (1)
    - Evaluate all possible patterns with one fewer adjacency that entail strictly more CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum

GES

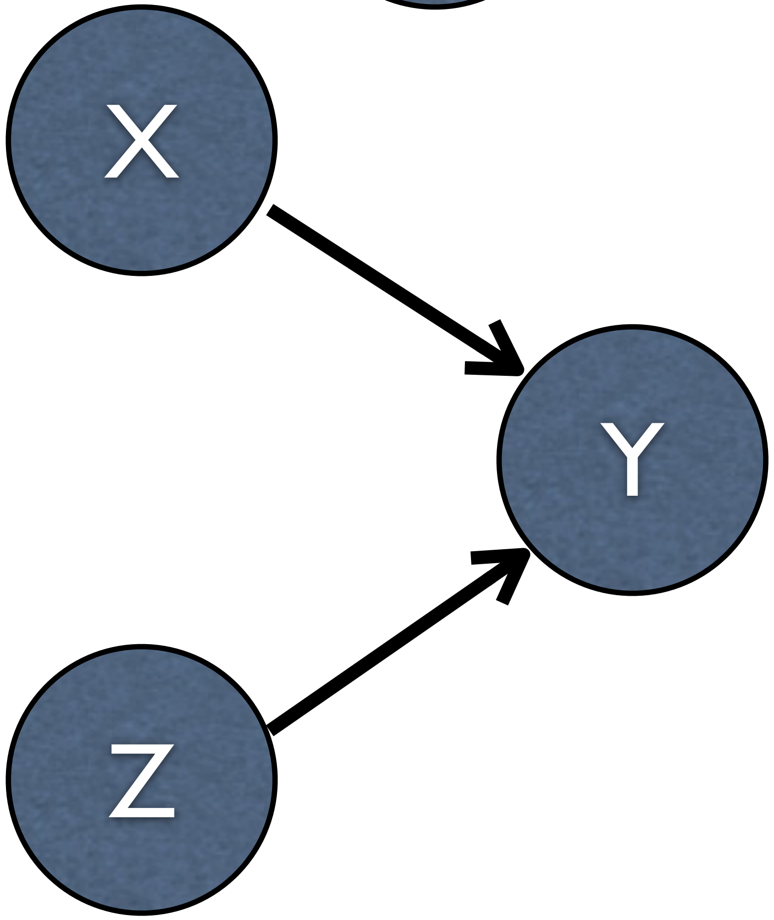
Suppose data were generated by



(1)

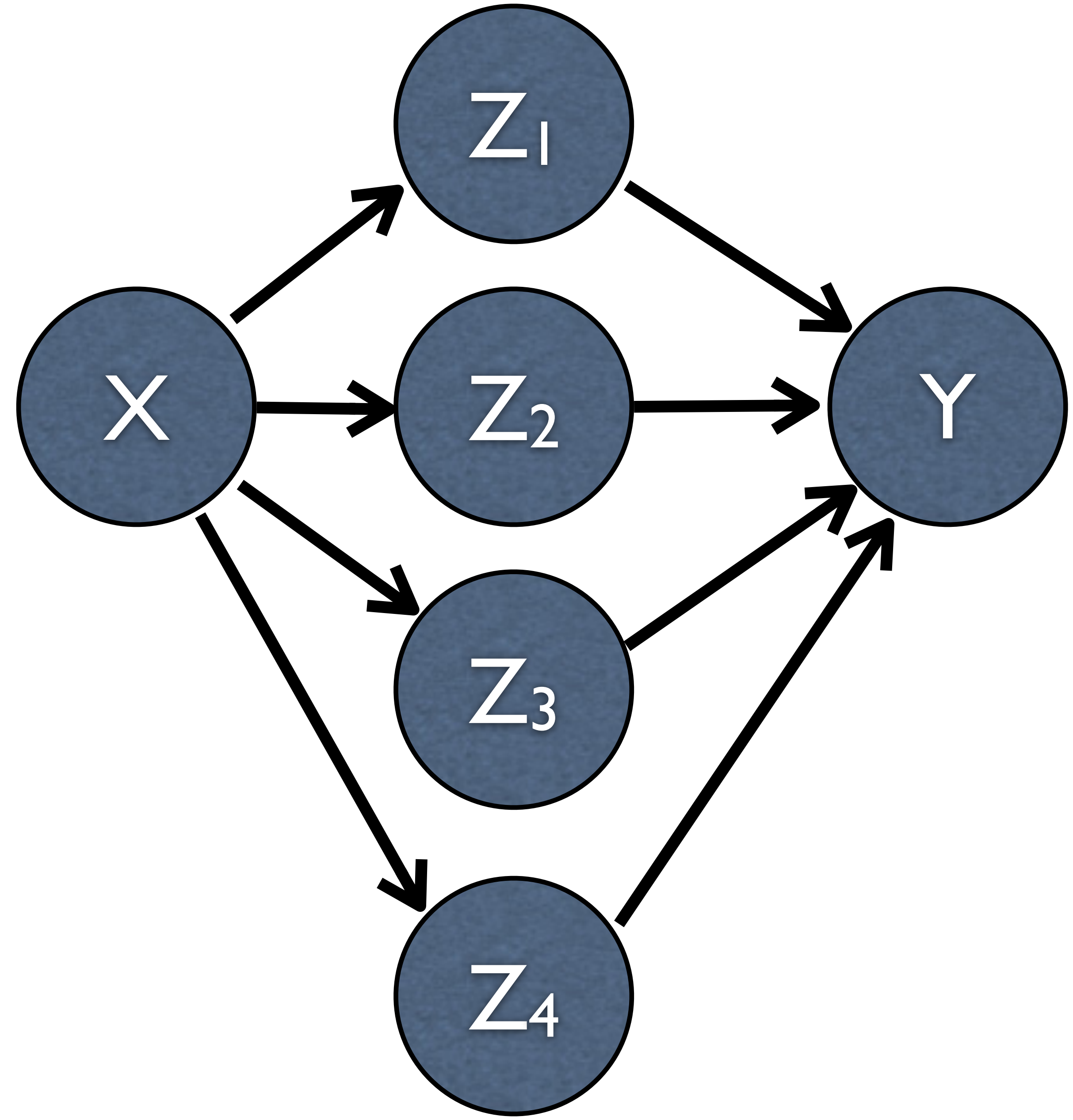


(2)



GES

Suppose data were generated by



Imagine the GES procedure...

Summary: Basic methods for causal discovery

- Basic multivariate analysis: what to discover from dependence among variables?
- Constraint-based methods, especially PC
  - Assumptions
  - Procedure
- Basic idea of GES
- Go beyond equivalence classes?