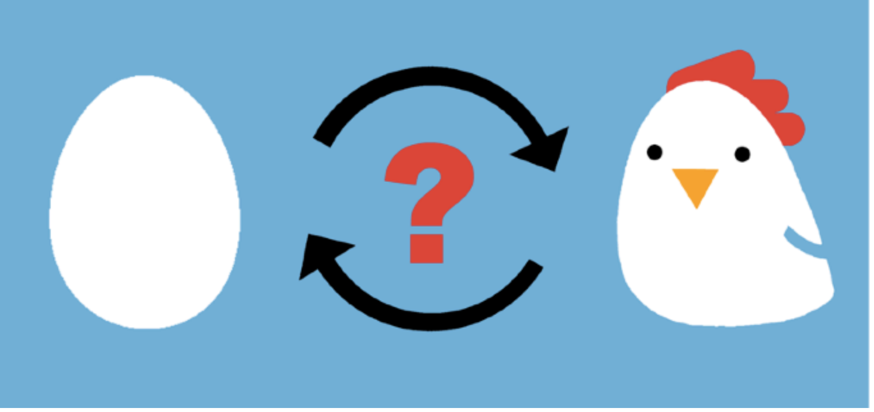


Causality and Machine Learning
(80-816/516)

Classes 5 (Jan 28, 2025)

Multivariate analysis: Goals, techniques, and connections to causal discovery

Instructor:

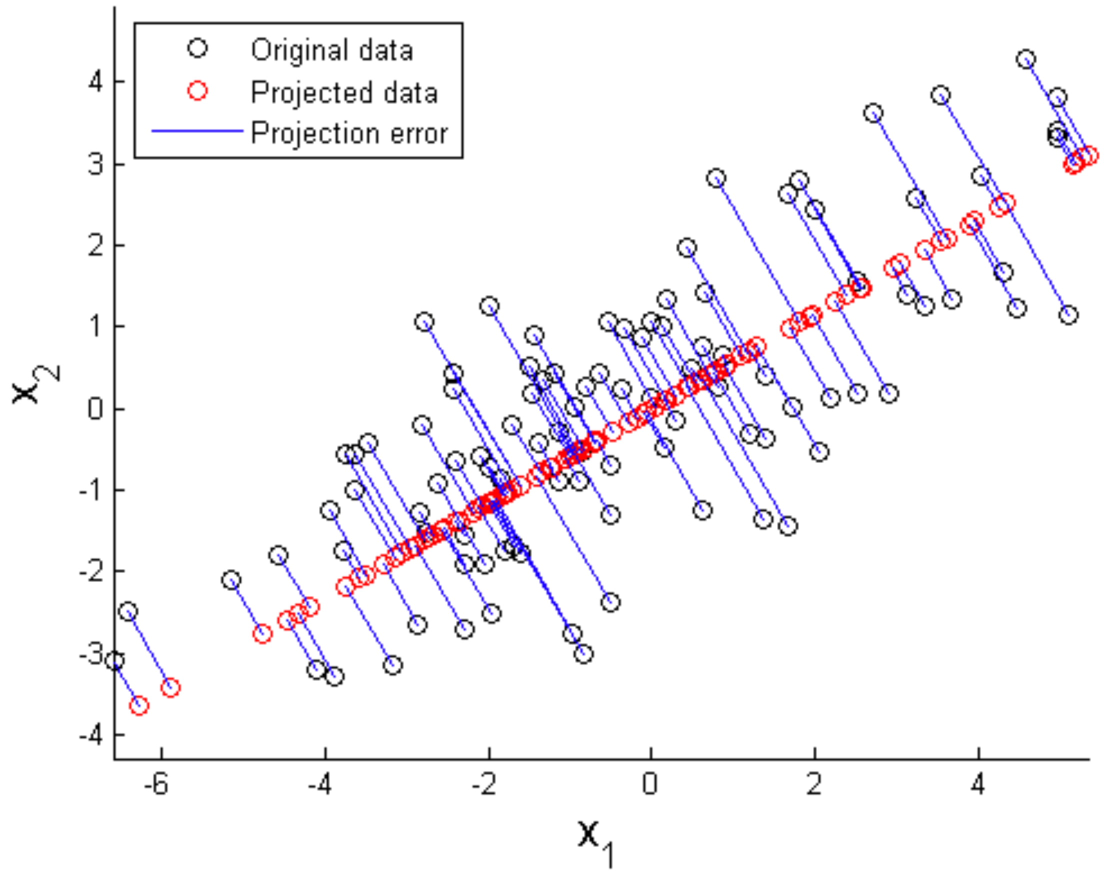
Kun Zhang (kunz1@cmu.edu)

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

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Outline

- Unsupervised learning with multivariate analysis
- Principal component analysis (PCA)
- Factor analysis
  - And probabilistic PCA
- Independent component analysis



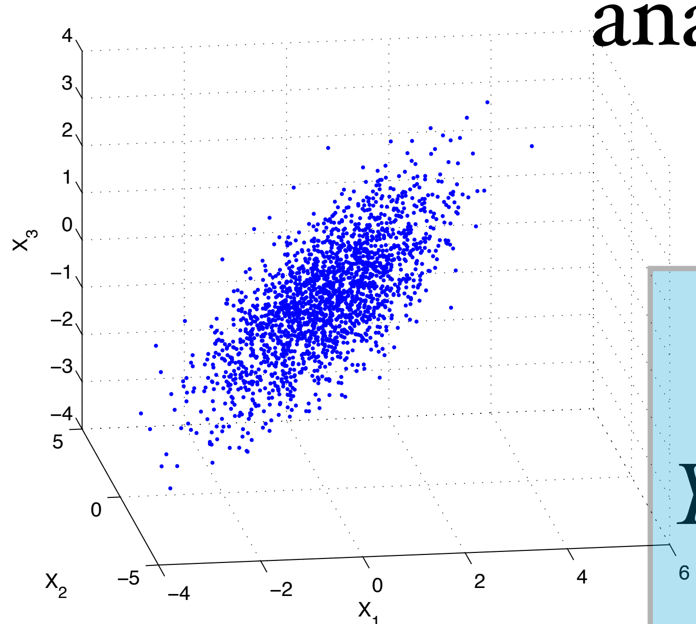
Two Ways of Finding Simpler Data Representations

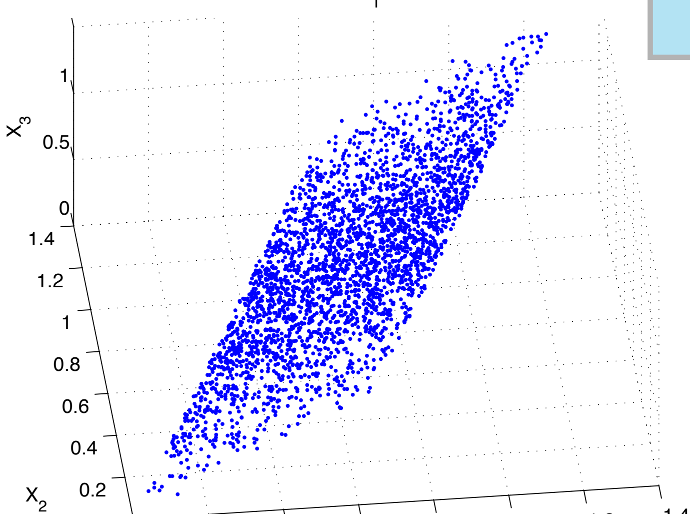
- Fewer "data points" vs. ***fewer dimensions (#variables)***?

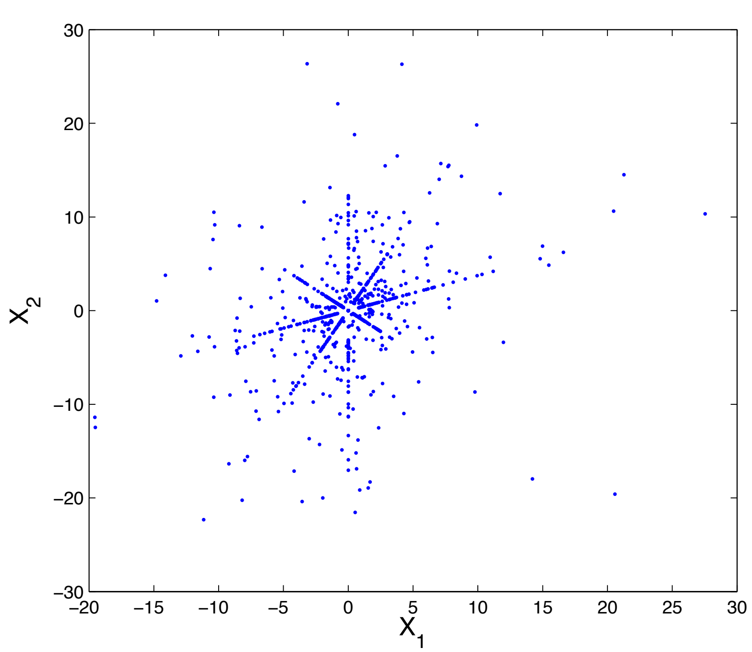
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	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.



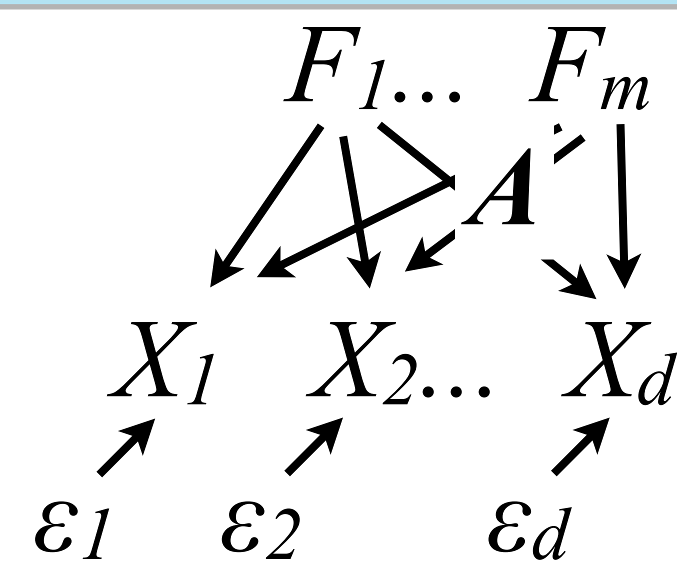




- Regression…

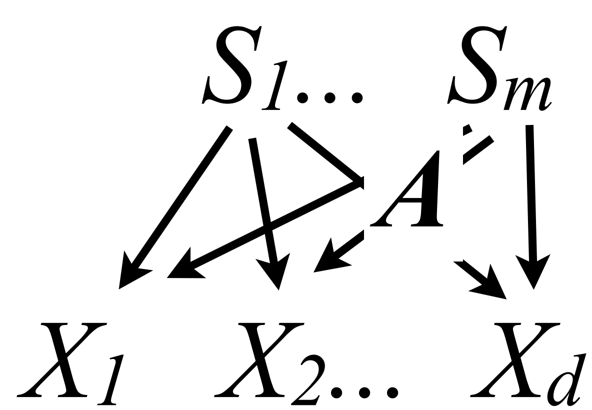
Find a projection of the data: $Y = w^{T}X$ with certain properties.

- Principal component analysis



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

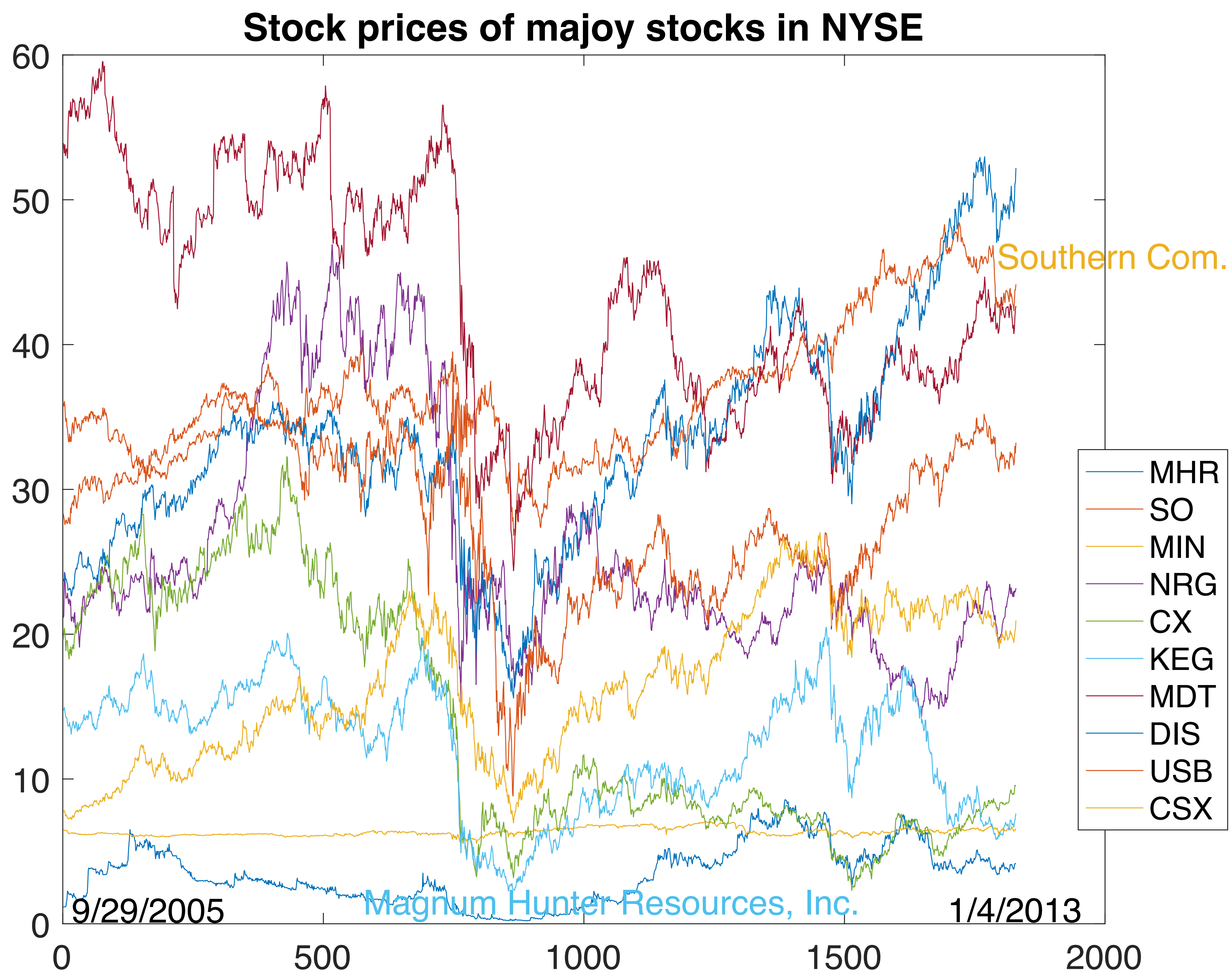
$\mathbf{X} = [X_1, X_2, \ldots, X_d]^{T}$



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

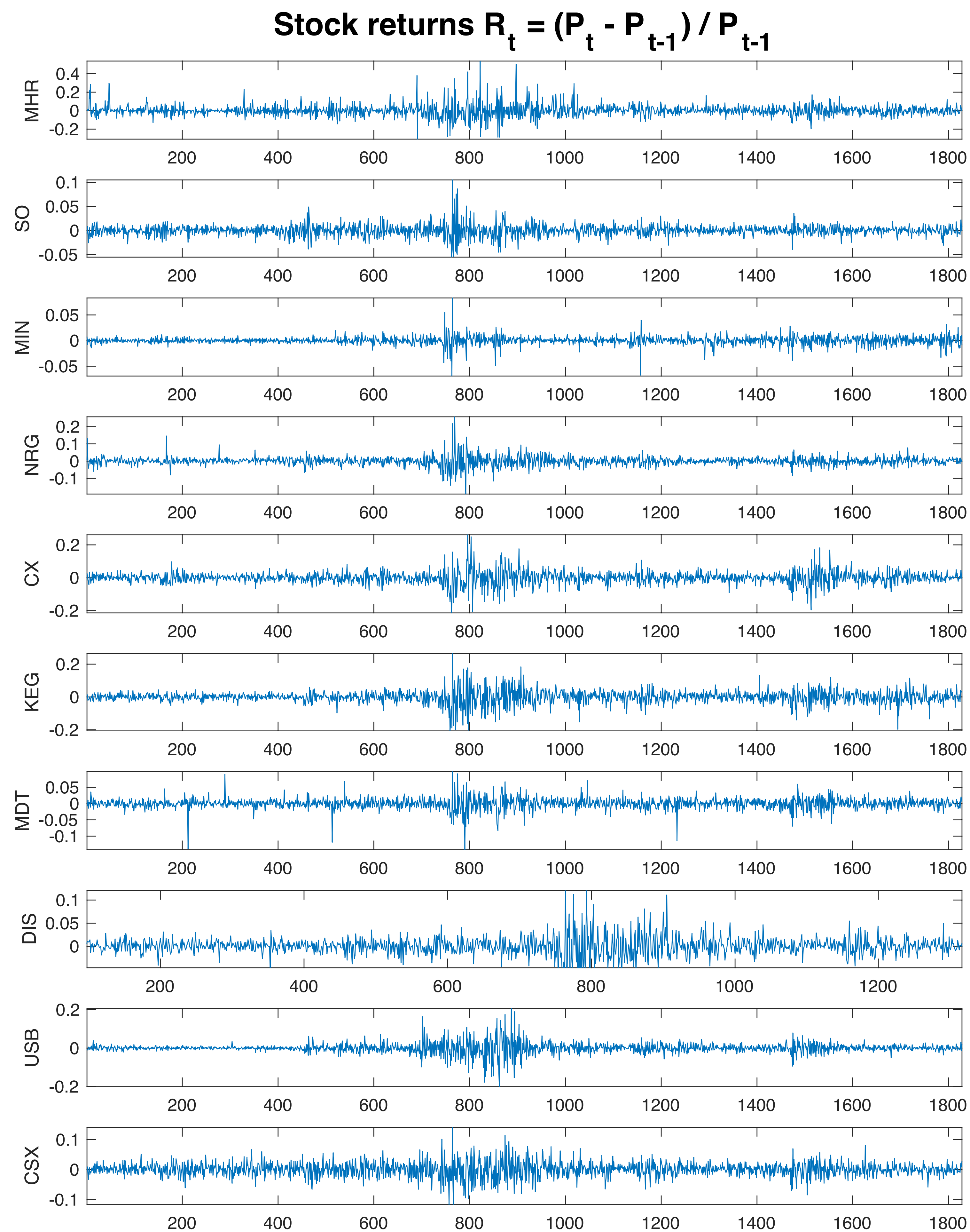
Major Information in the Data?

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



Major Information in the Data?

- Major information in the NYSE stock market?

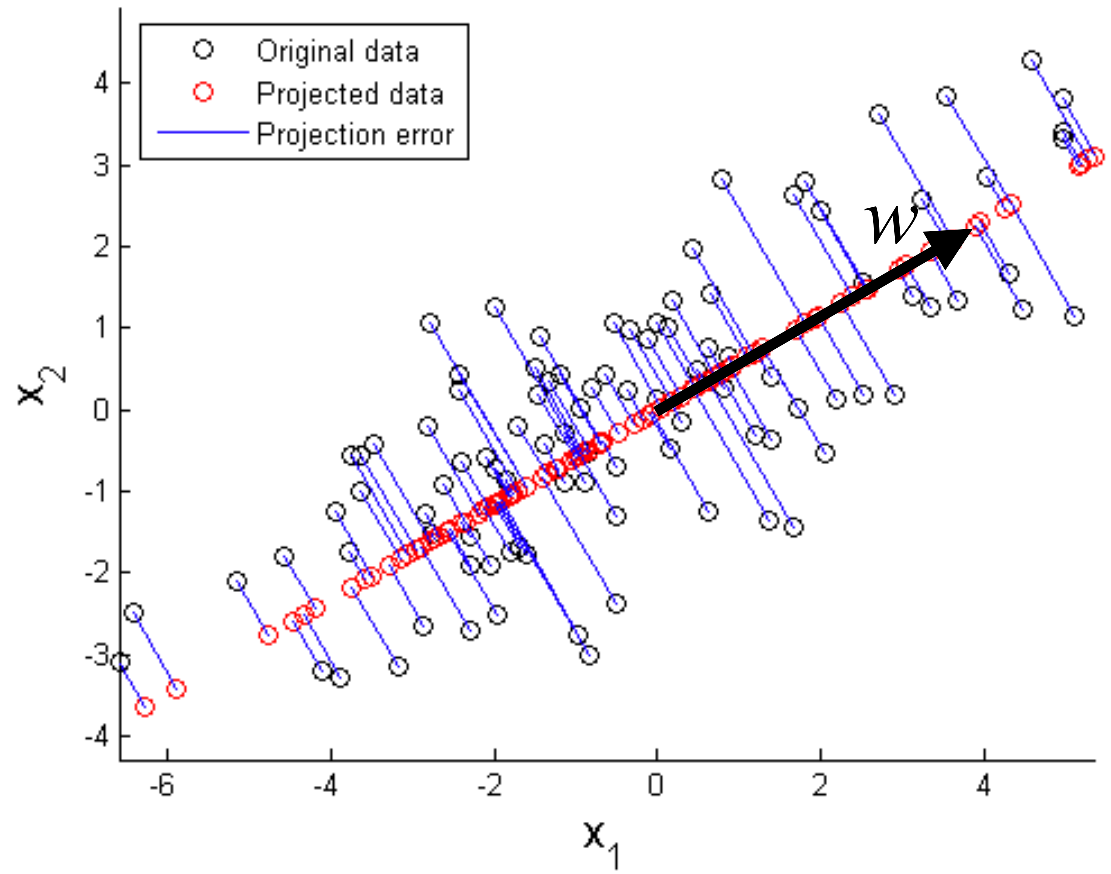


Principal Component Analysis (PCA)

- Find a projection of the data

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

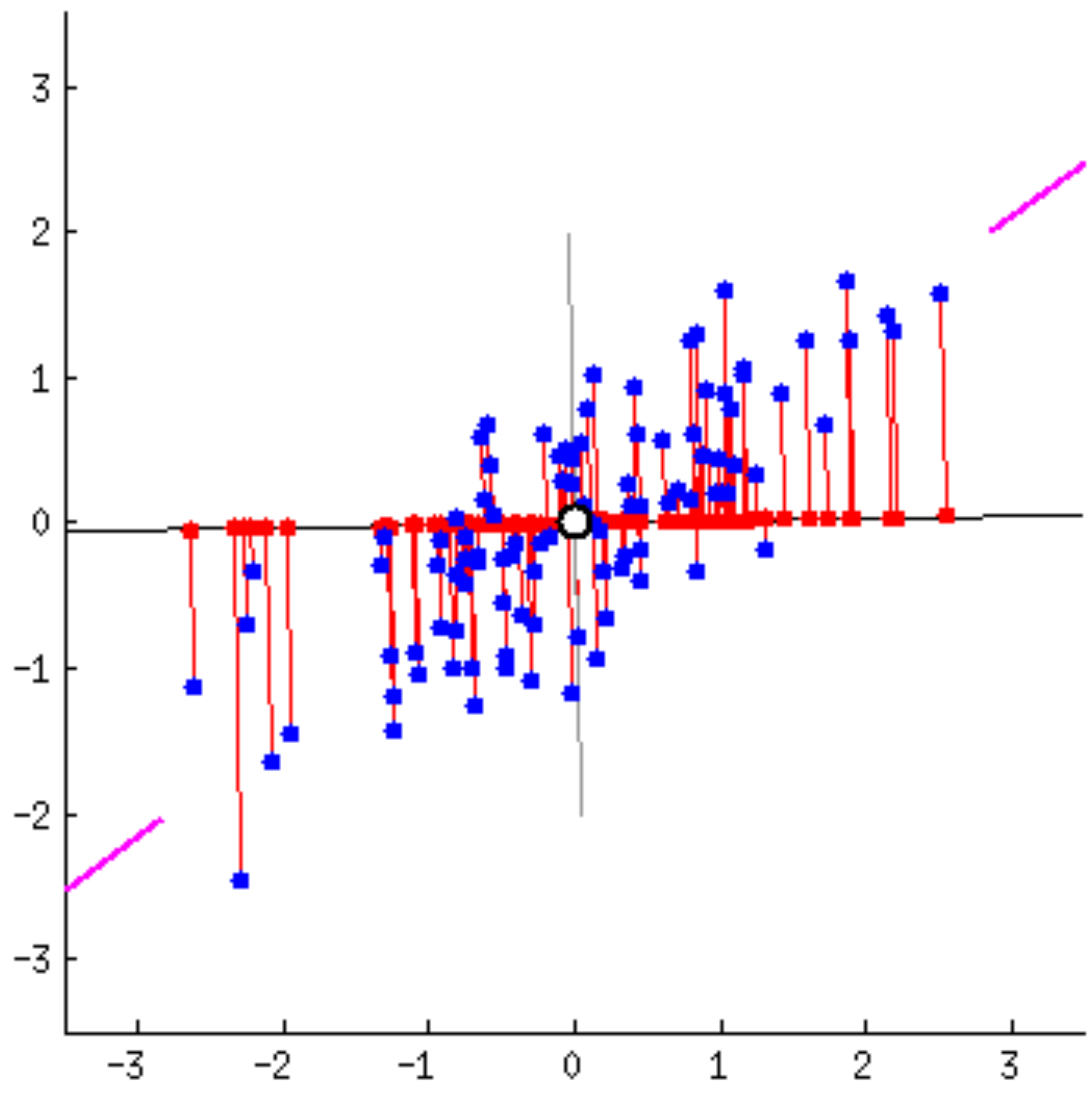
to give the maximum variance (minimal squared reconstruction/ projection error?)



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



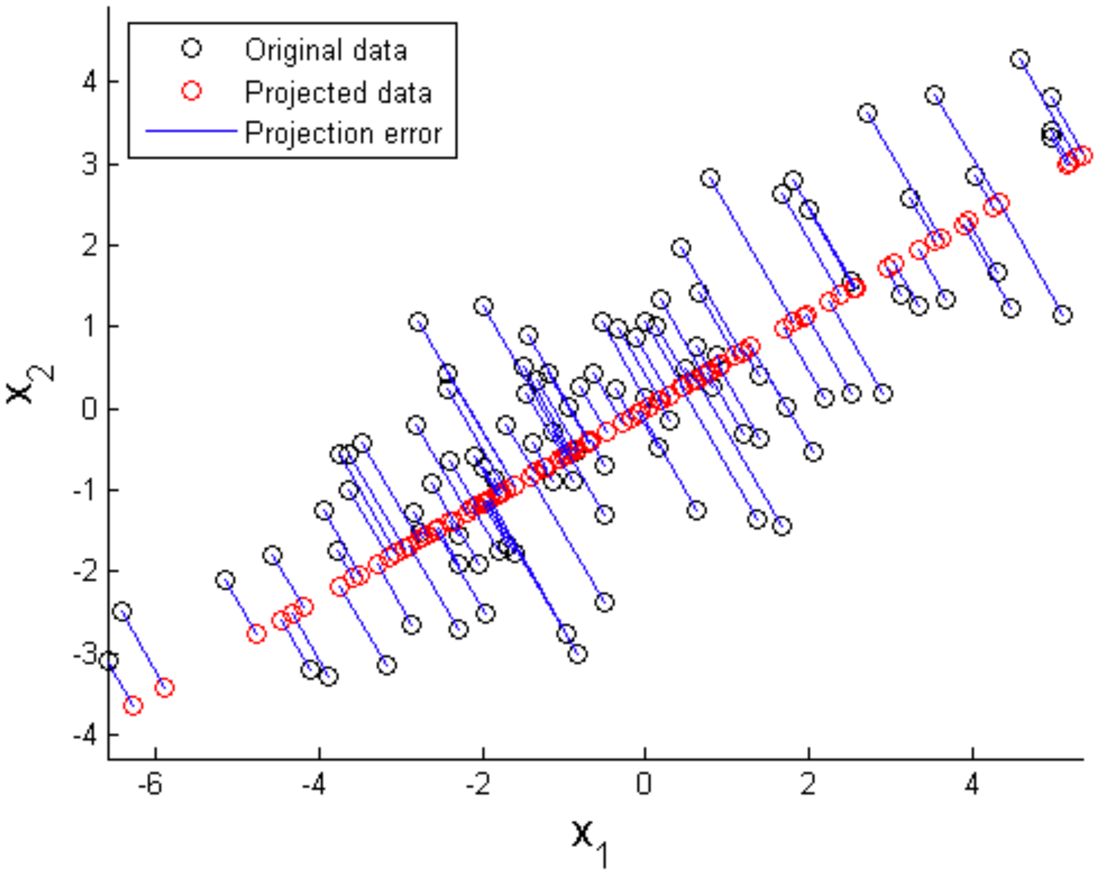
PCA

- 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. $||w||^2 = w^{\intercal}w = 1$.

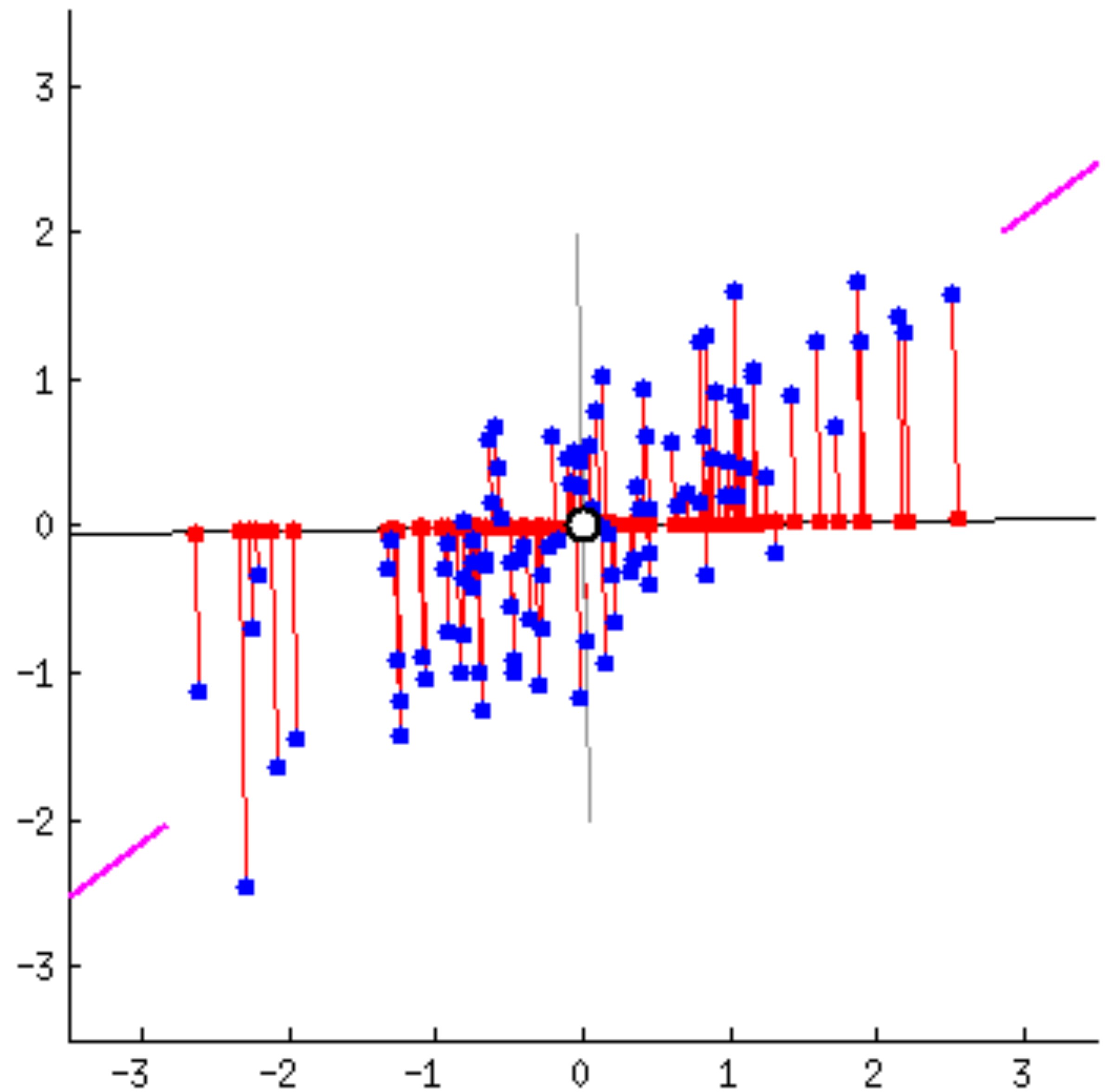
- Let $\mathcal{L} = w^{\intercal}Cw - \lambda w^{\intercal}w$. Setting $\frac{\partial \mathcal{L}}{\partial w} = 0$ gives

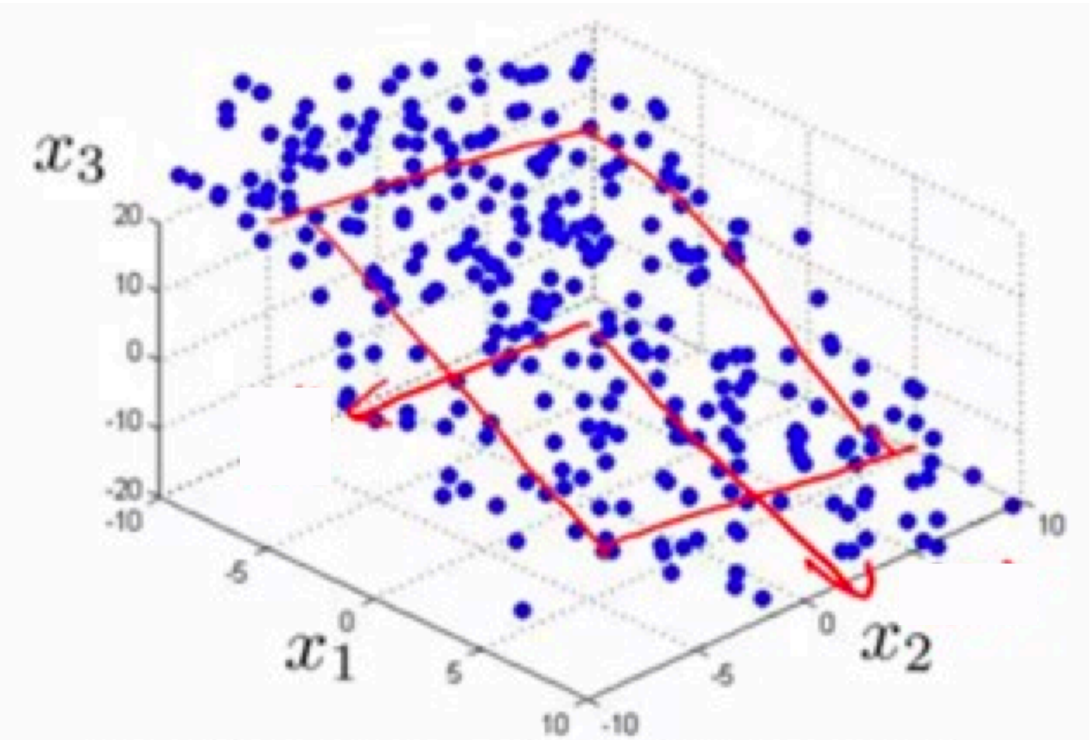
$$2Cw - 2\lambda w = 0 \Rightarrow Cw = \lambda w.$$

- 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.

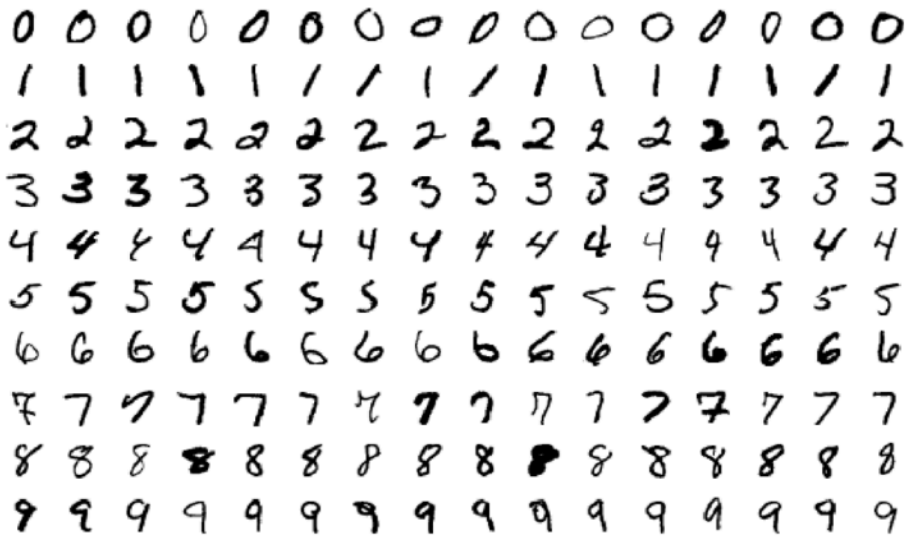
PCA: Effect of Weight Vector *w*

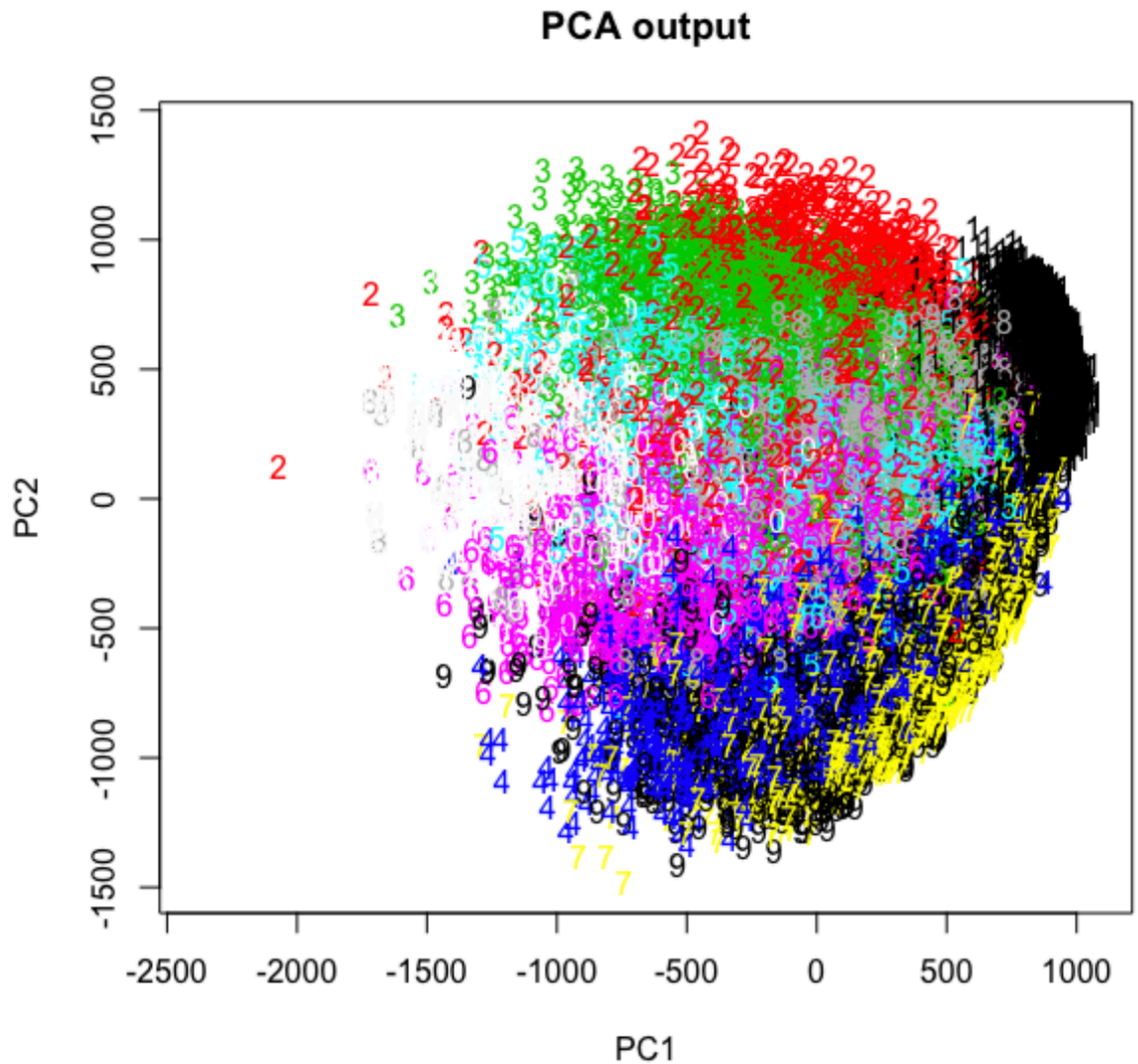




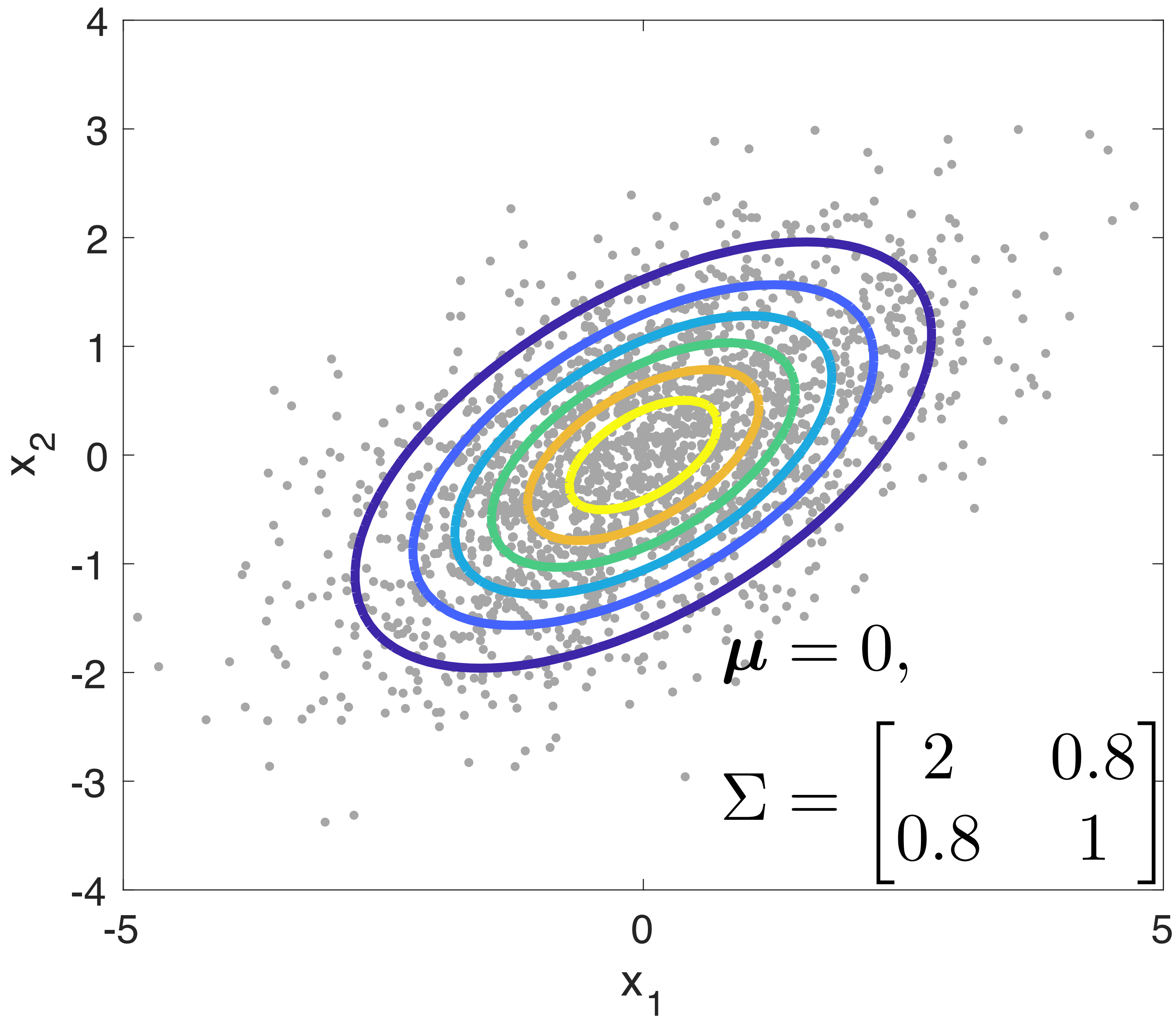
Reduce data from 3D to 2D

PCA on MNIST Data





Principal Axis vs. Regression Line

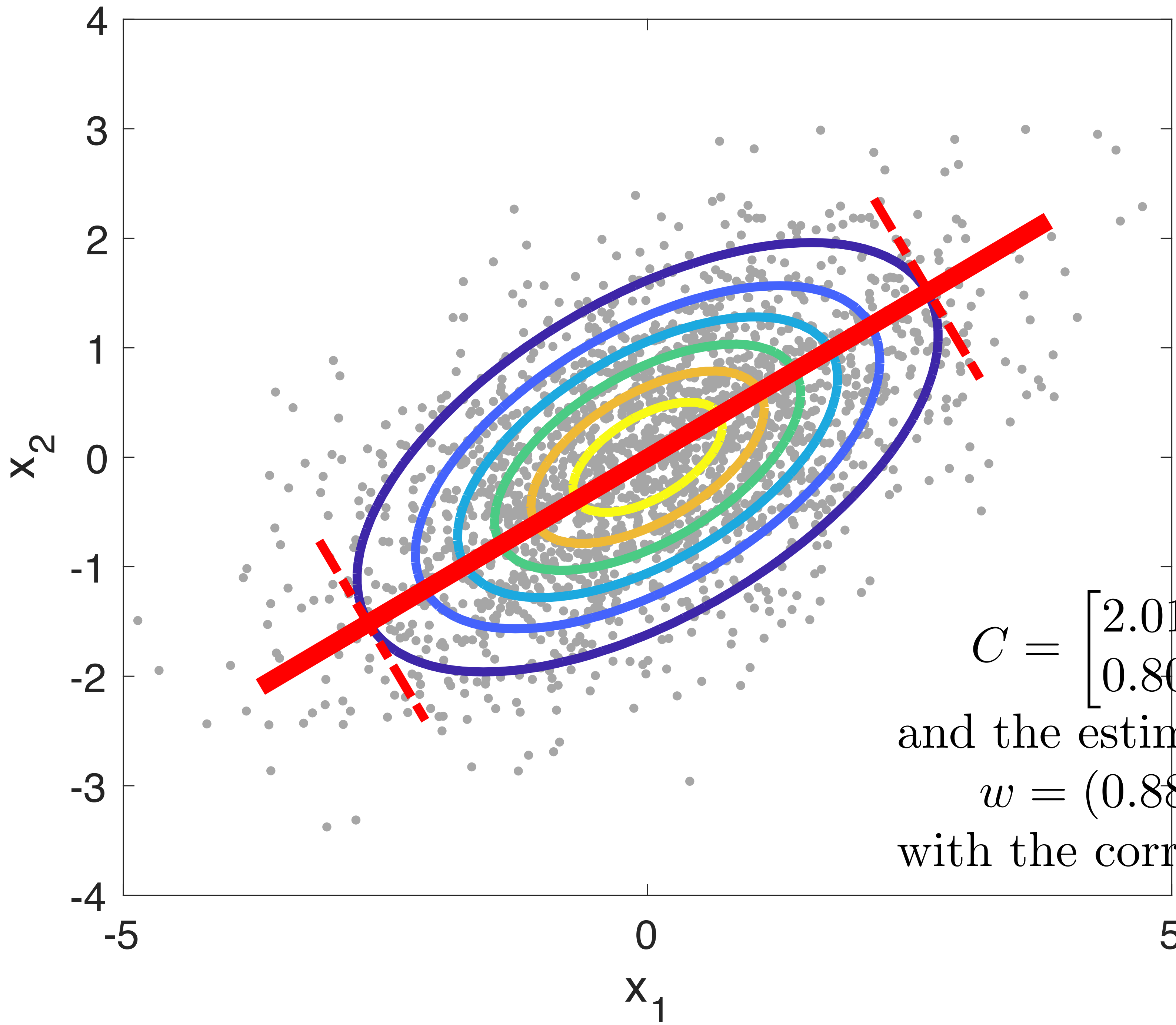


$$\boldsymbol{\mu} = 0,$$

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

Principal Axis vs. Regression Line

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



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

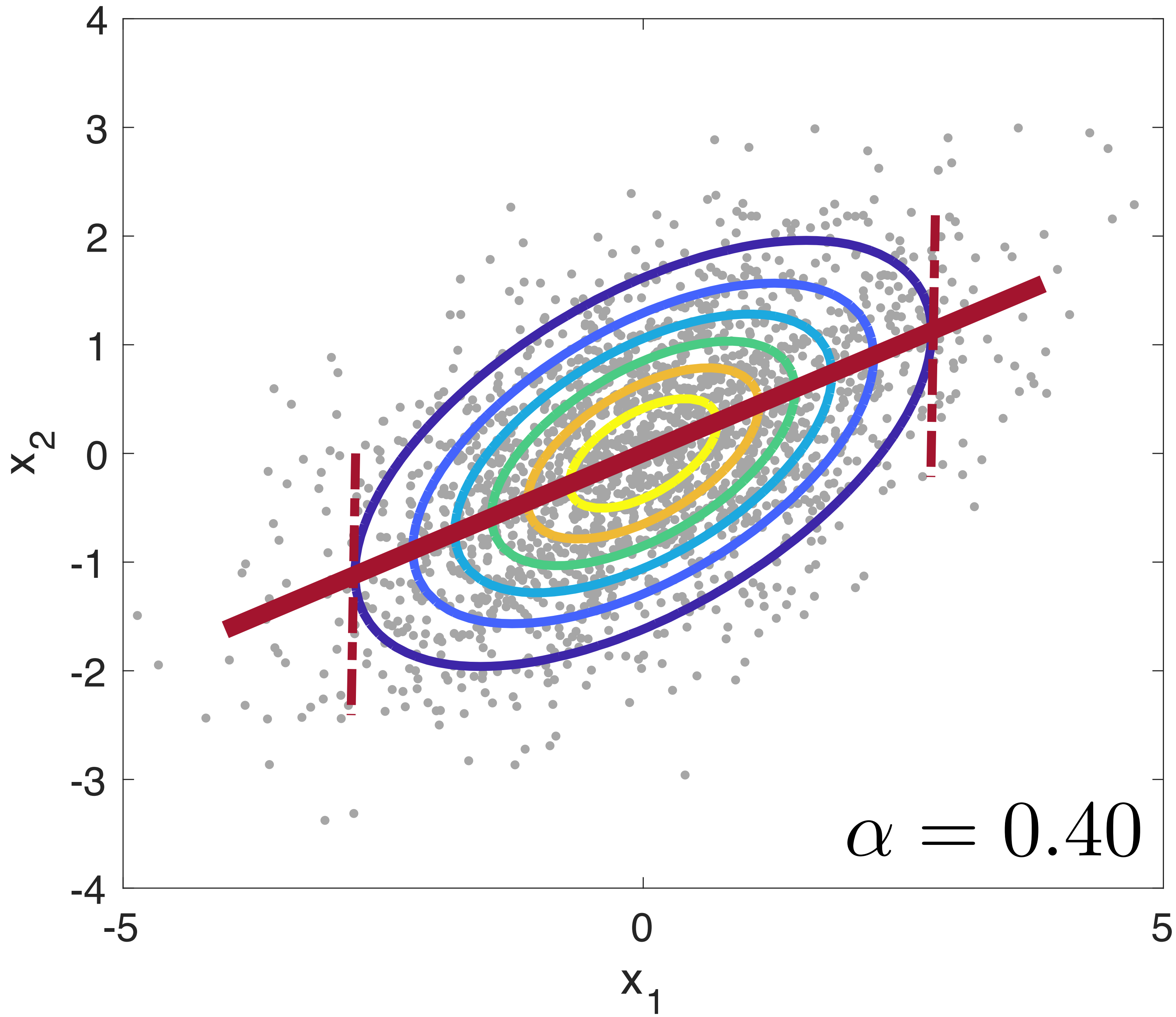
and the estimated weight vector is:

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

with the corresponding eigenvalue 2.45

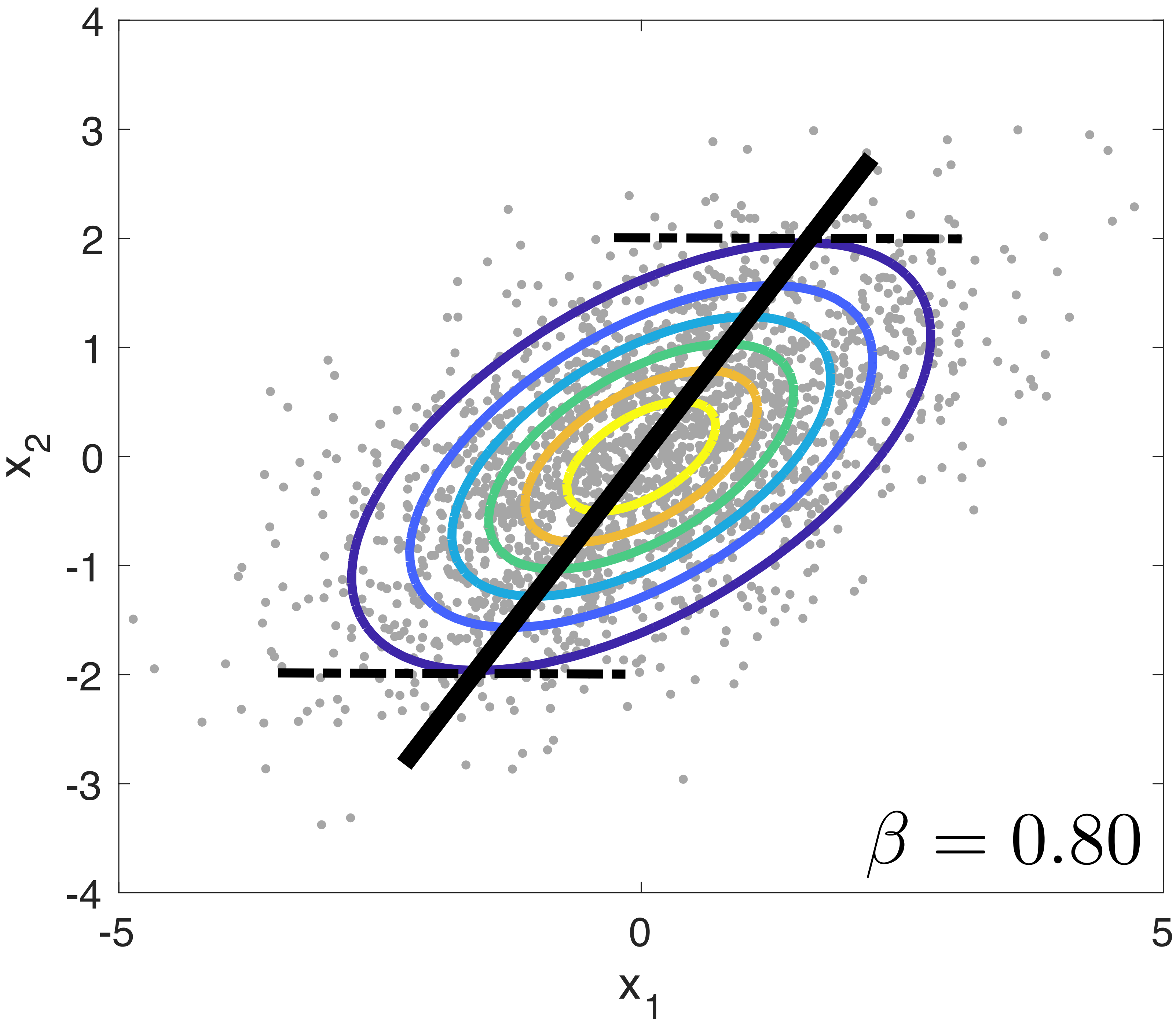
Principal Axis vs. Regression Line

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

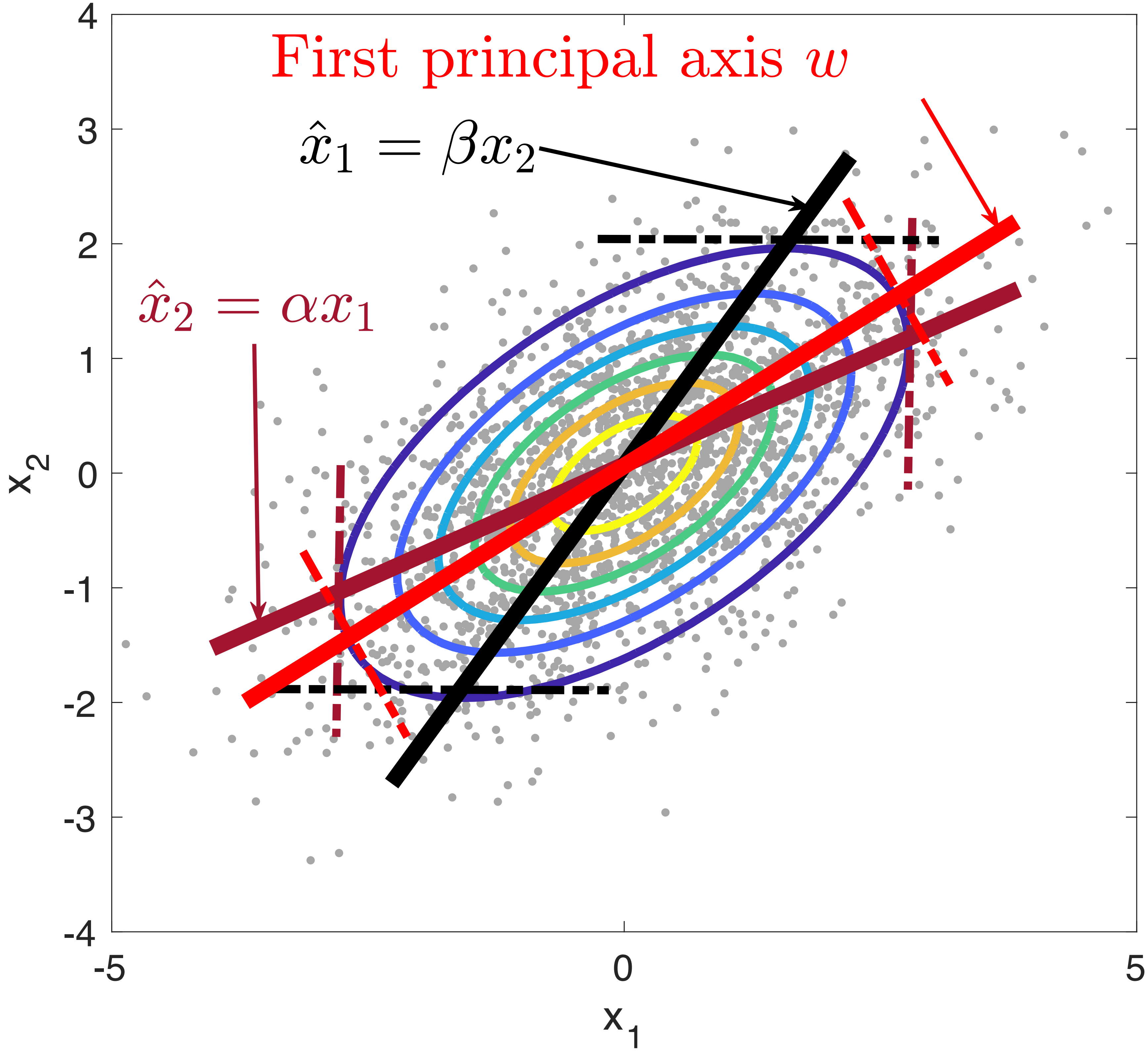


Principal Axis vs. Regression Line

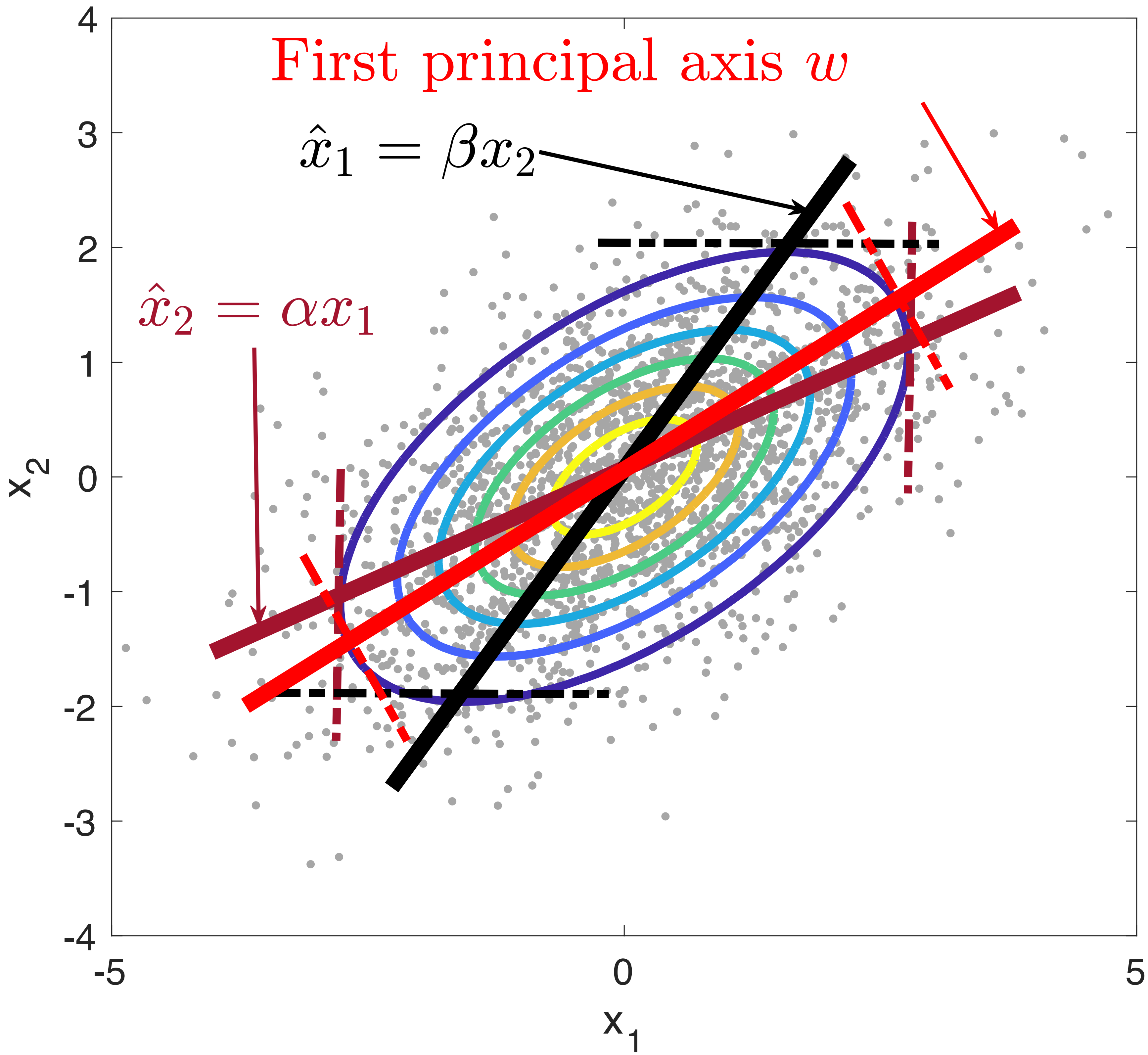
- Regression line from X_2 to X_1: $\hat{x}_1 = \beta x_2$



Principal Axis vs. Regression Line

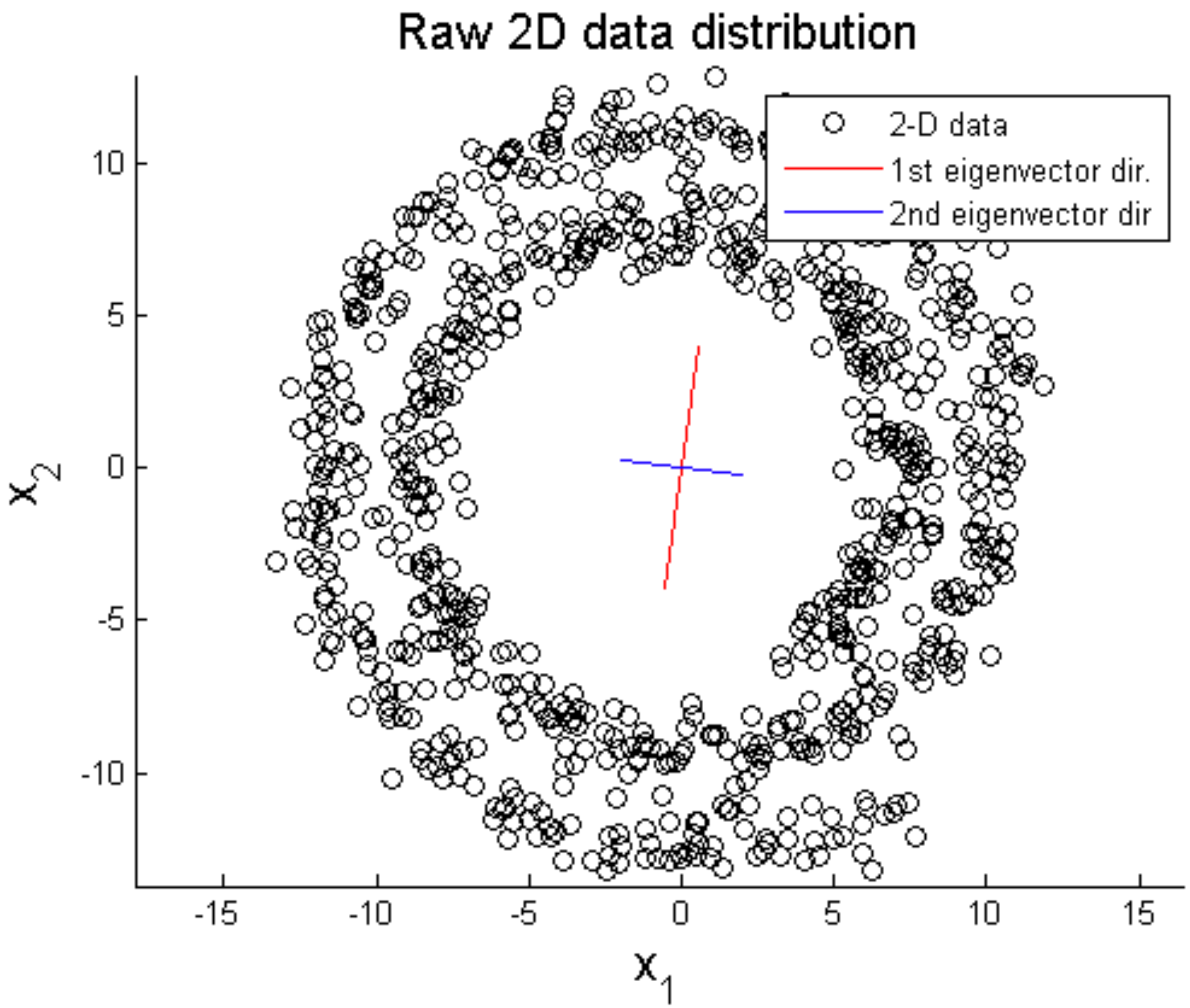


Principal Axis vs. Regression Line



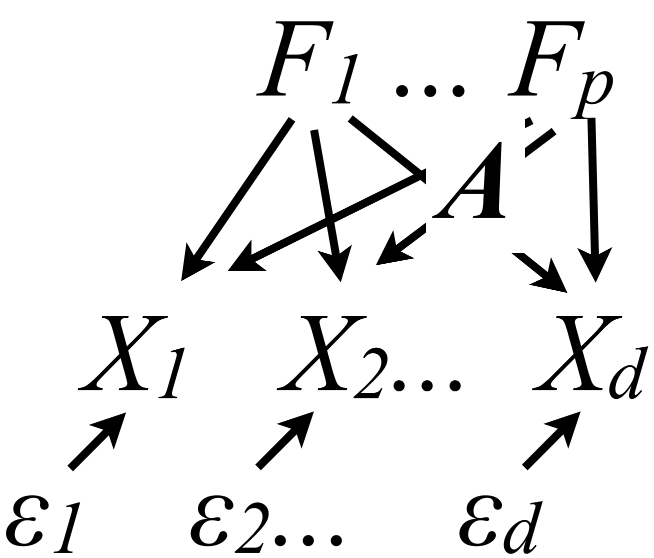
Nonlinear PCA

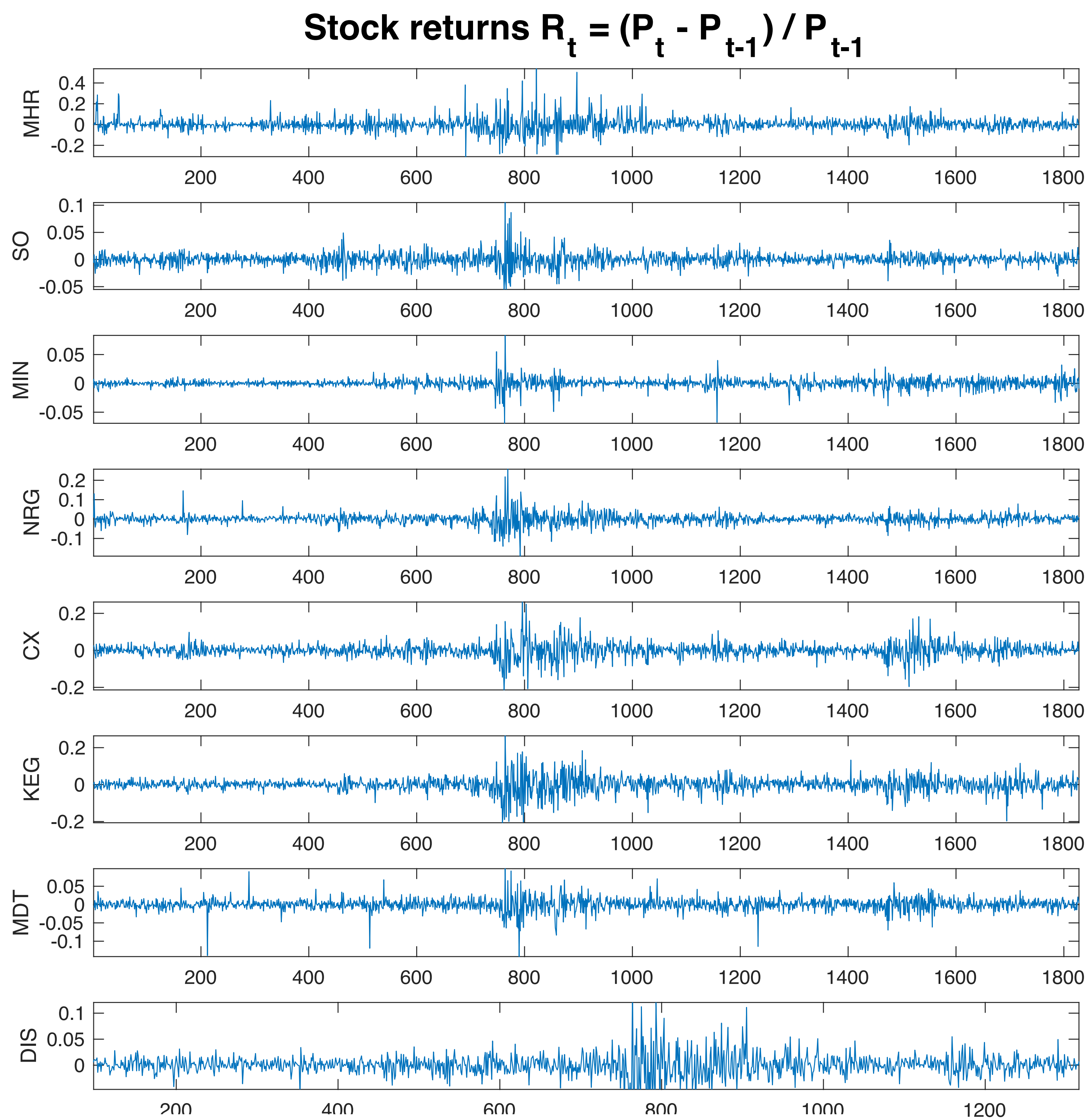
- Projections onto nonlinear manifold instead...
- Easily kernelized

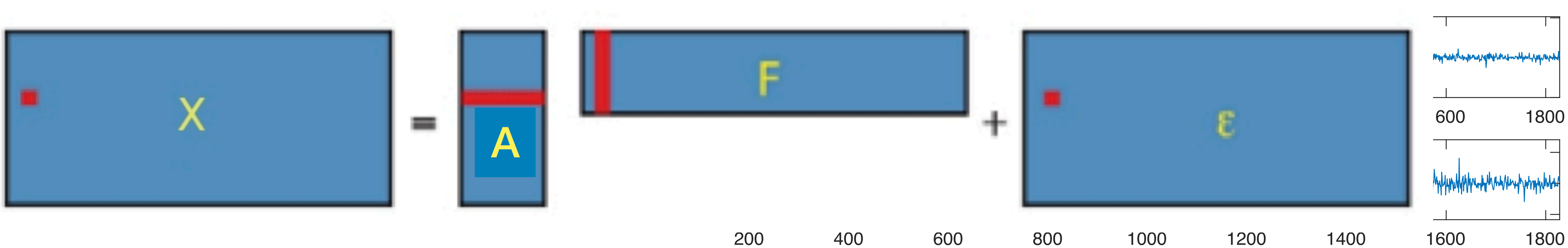


Underlying Factors?

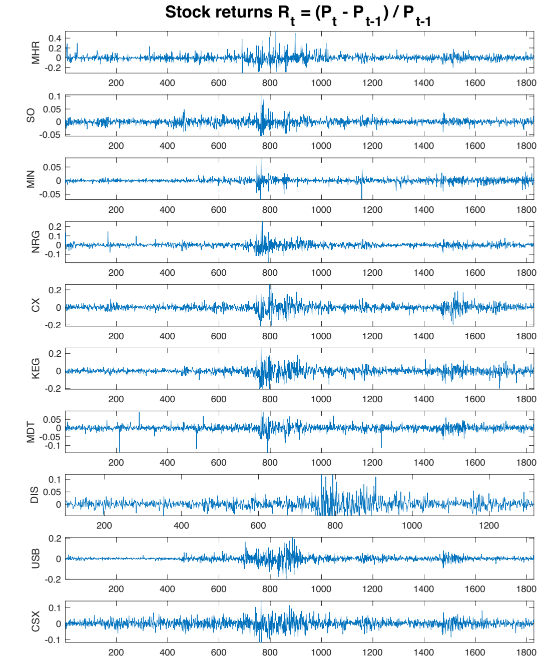
- Major information in the NYSE stock market?



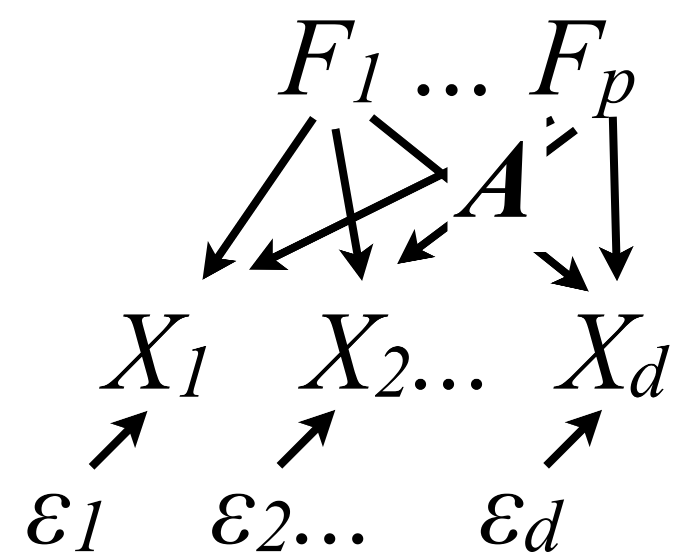


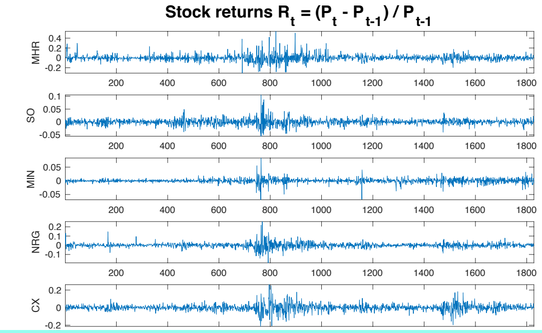


Factor Analysis



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

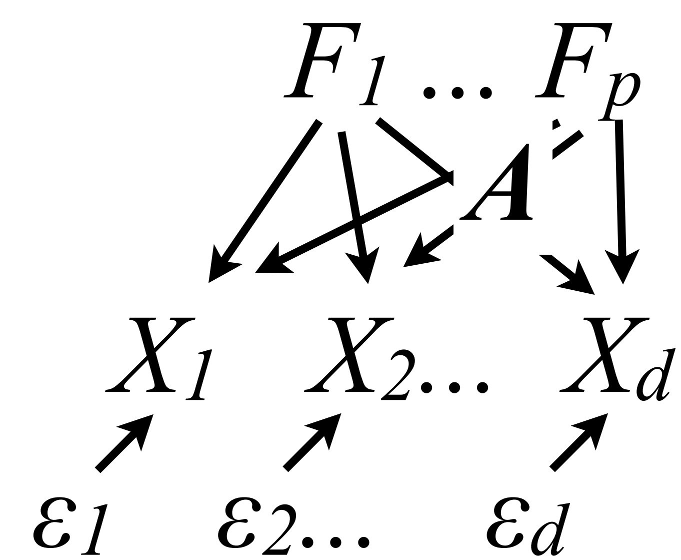




Factor Analysis

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

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

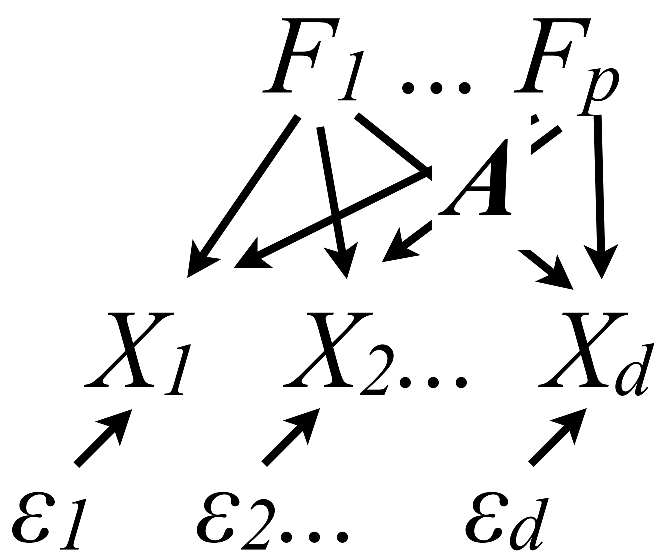


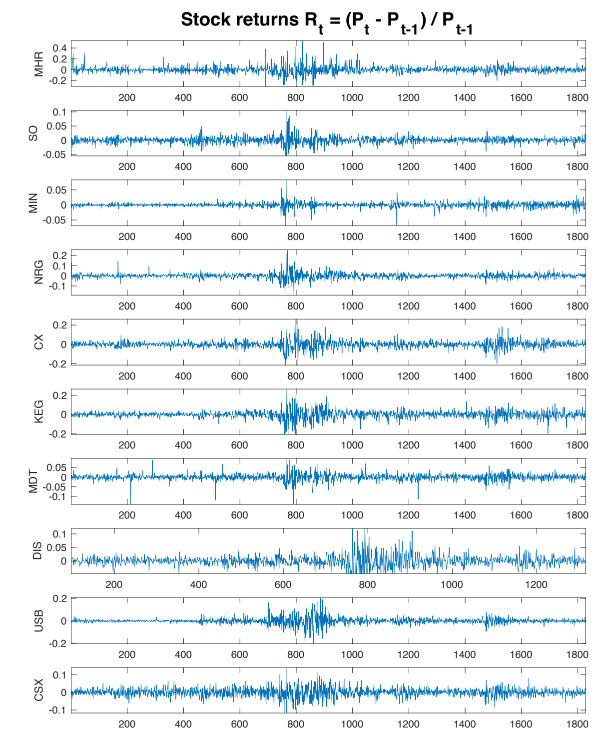
$\mathbf{A}\mathbf{A}^\intercal + \boldsymbol{\Psi} = \mathbf{A}\mathbf{U}\mathbf{U}^\intercal\mathbf{A}^\intercal + \boldsymbol{\Psi}$,
where $\mathbf{U}$ is an orthogonal matrix.

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

- Bekker, P.A. and ten Berge, J. M. F., Generic global indentification in factor analysis. Linear Algebra and its Applications, 264:255–263, 1997.

Factor Analysis on the Returns





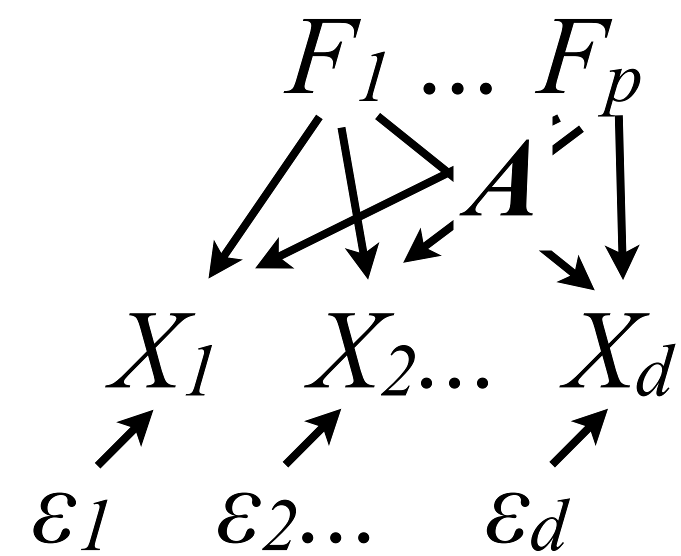
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$; $\mathrm{Cov}[\boldsymbol{F}]=\mathrm{I}$.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

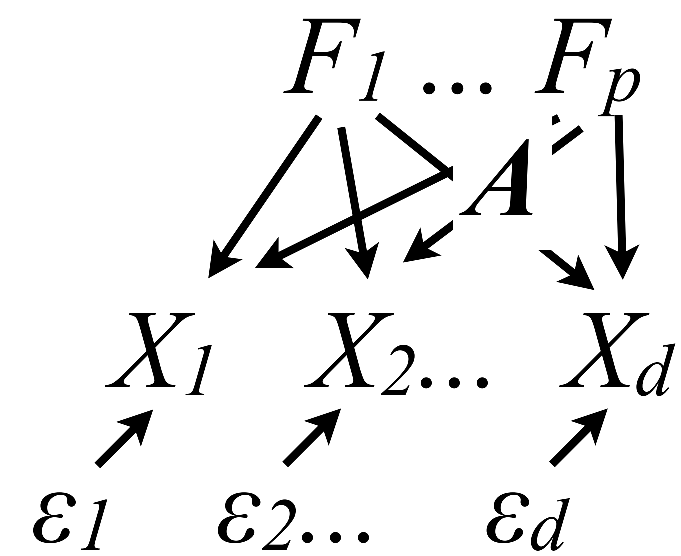
- Assume a generating model
- $\mathbf{X} = \mathbf{A}\boldsymbol{F} + \boldsymbol{\varepsilon}$
  - $\mathbf{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$.
  - $\boldsymbol{F} = [F_1, \ldots, F_p]$, $p<n$.
  - $\boldsymbol{F} \perp\!\!\!\perp \boldsymbol{\varepsilon}$
  - $\mathrm{E}[\boldsymbol{F}]=\mathbf{0}$;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



Relationship between FA and PCA?
-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 $\mathbf{A}$
- Estimation: MLE

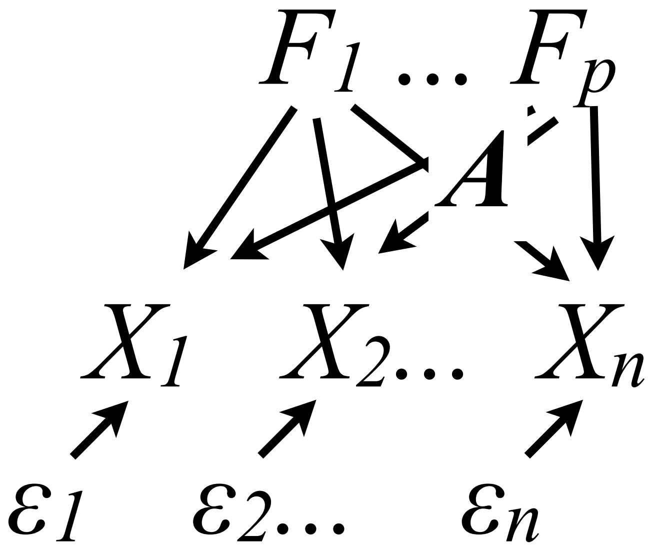


Relationship between FA and PCA:

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

Factor Analysis: A Bit History

- Charles Spearman was the first psychologist to discuss common factor analysis, in a 1904 paper that provided few details about his methods and was concerned with single-factor models.
  - discovered that school children's scores on a wide variety of seemingly unrelated subjects were positively correlated, which led him to postulate that a single general mental ability underlies and shapes human cognitive performance.

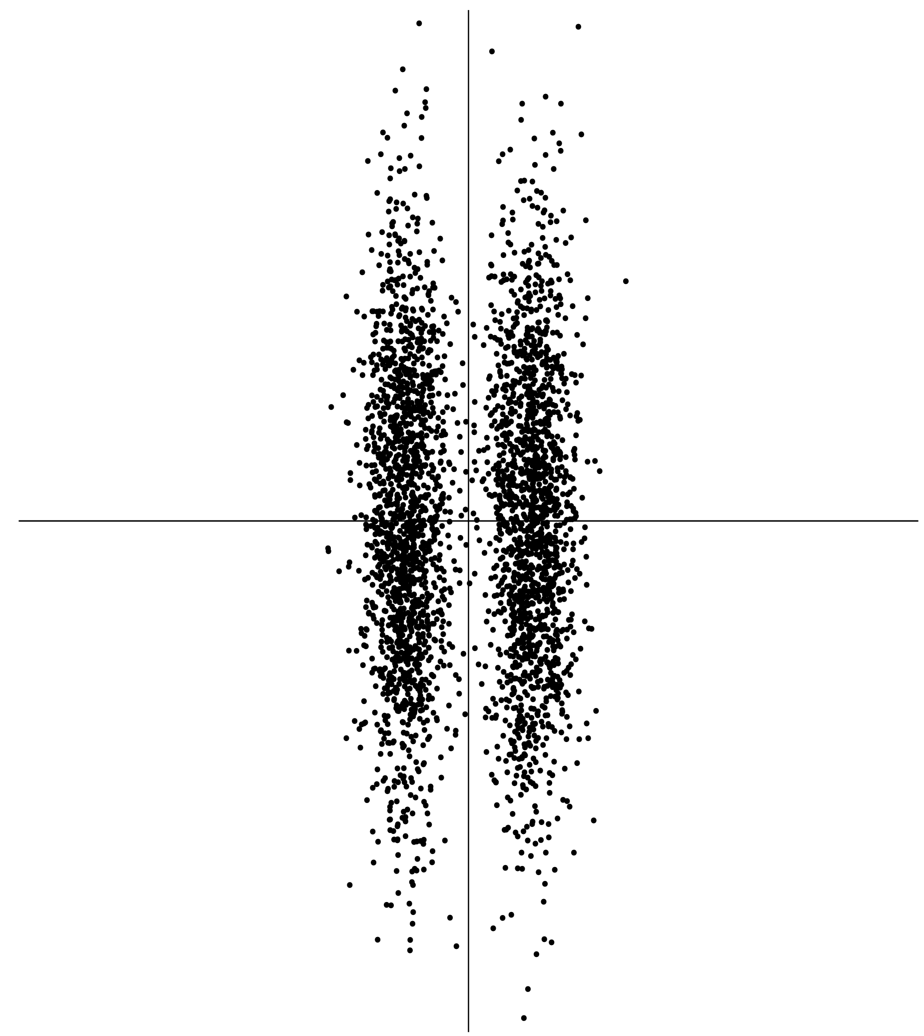


- The initial development of common factor analysis with multiple factors was given by Louis Thurstone in two papers in the early 1930s. Thurstone introduced several important factor analysis concepts, including uniqueness and rotation…

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

- Smaller entropy, more structural, more interesting

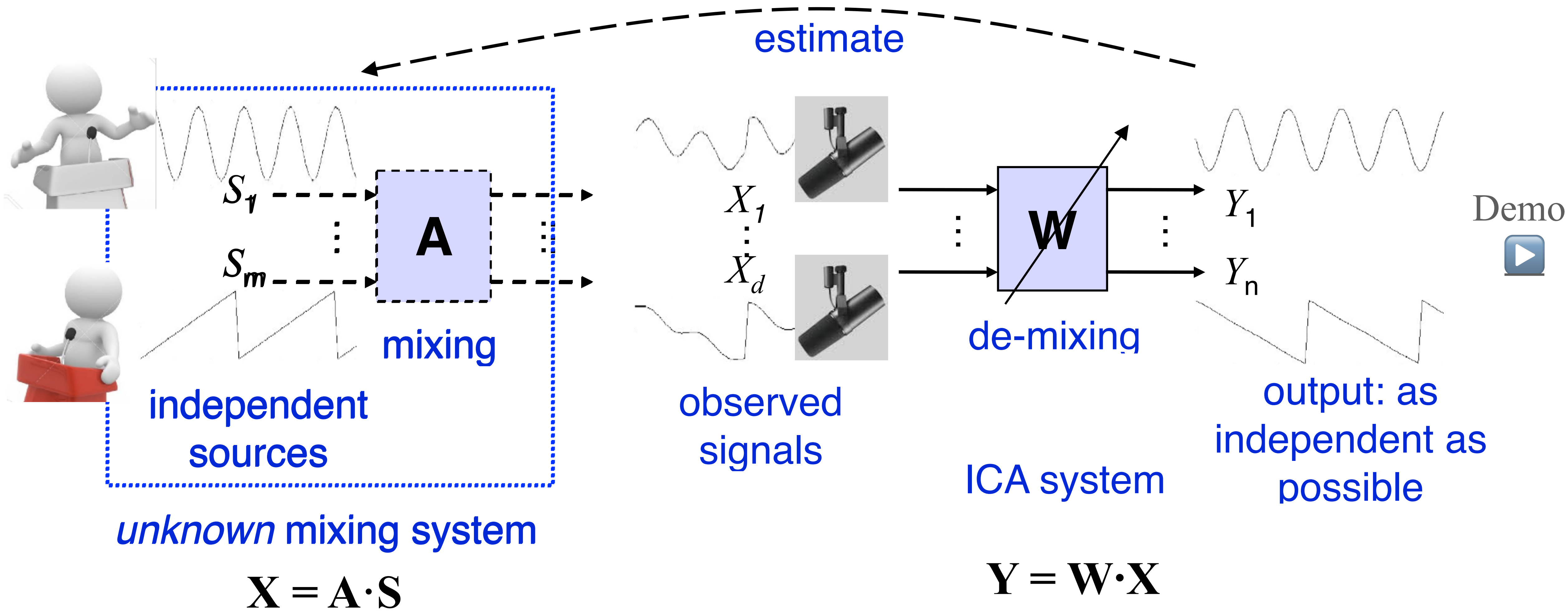
- "Purer" according to the central limit theorem



Which direction is more interesting?

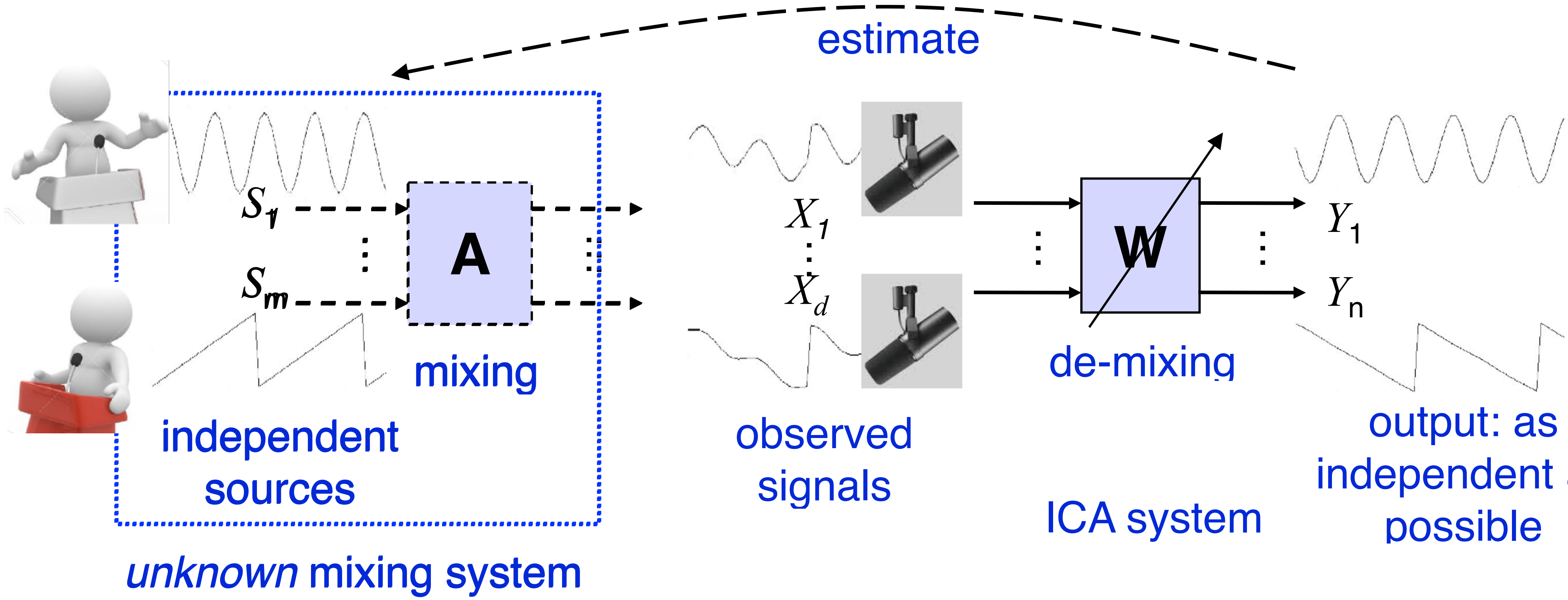
Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis



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

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

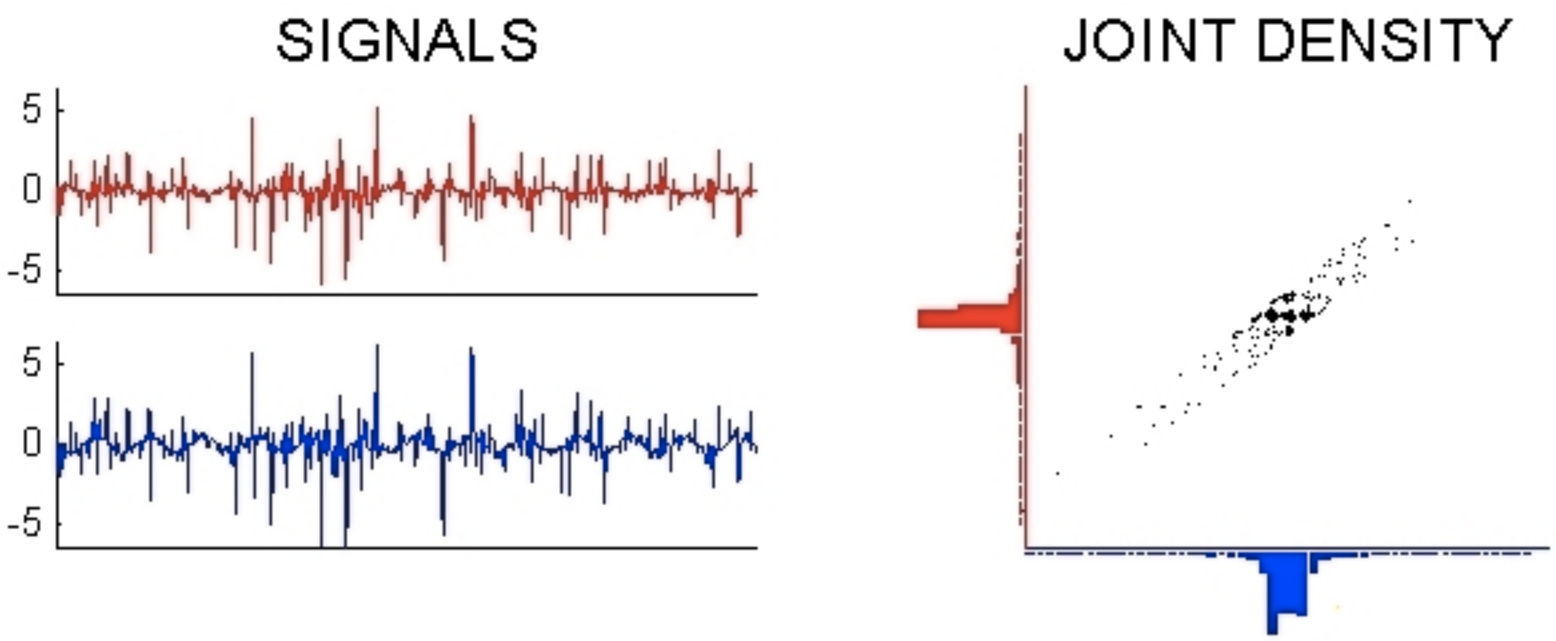
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \end{array}$$

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

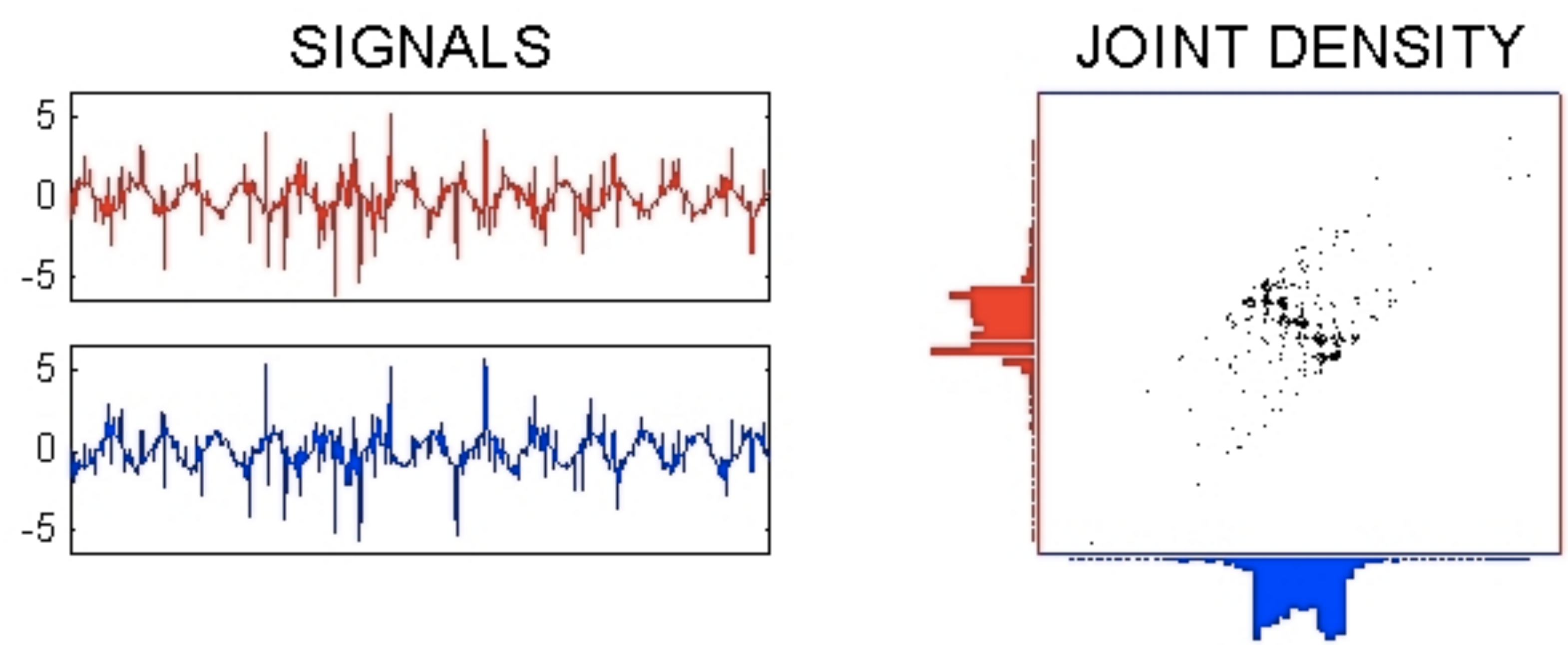
Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

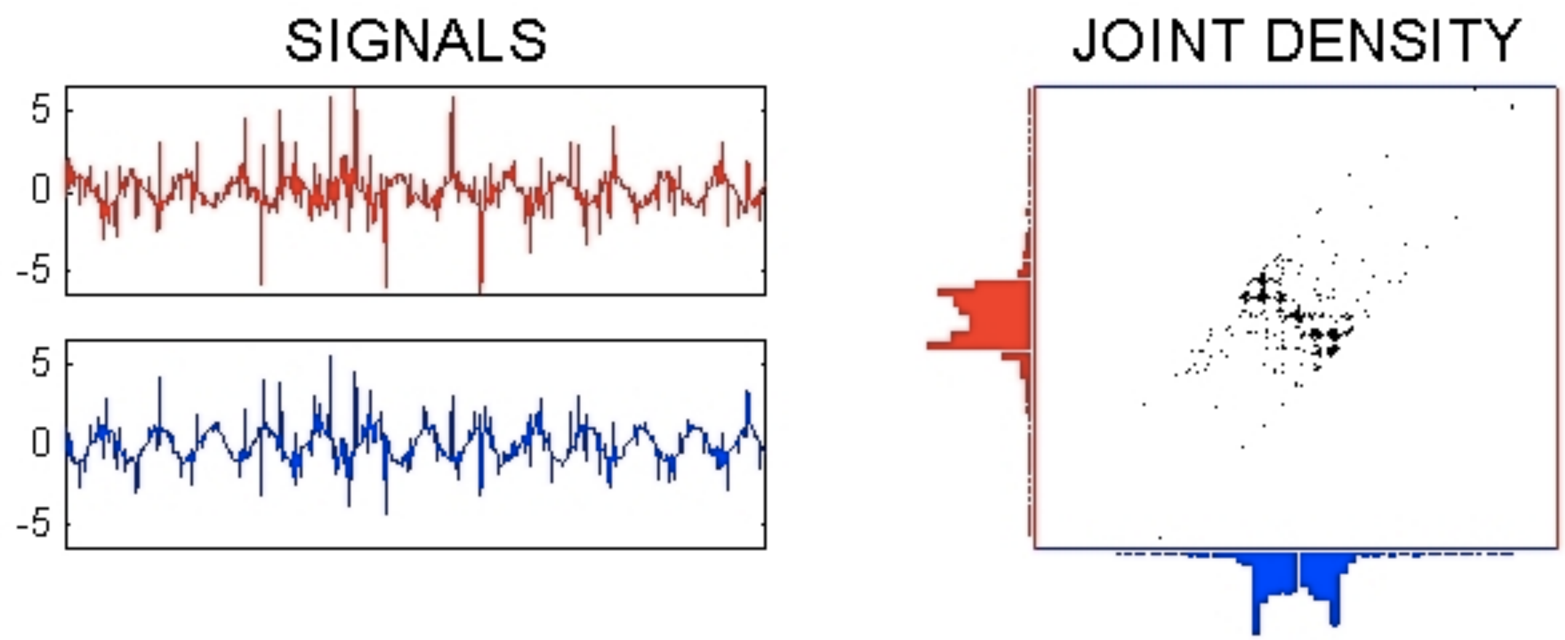
A Demo of the ICA Procedure



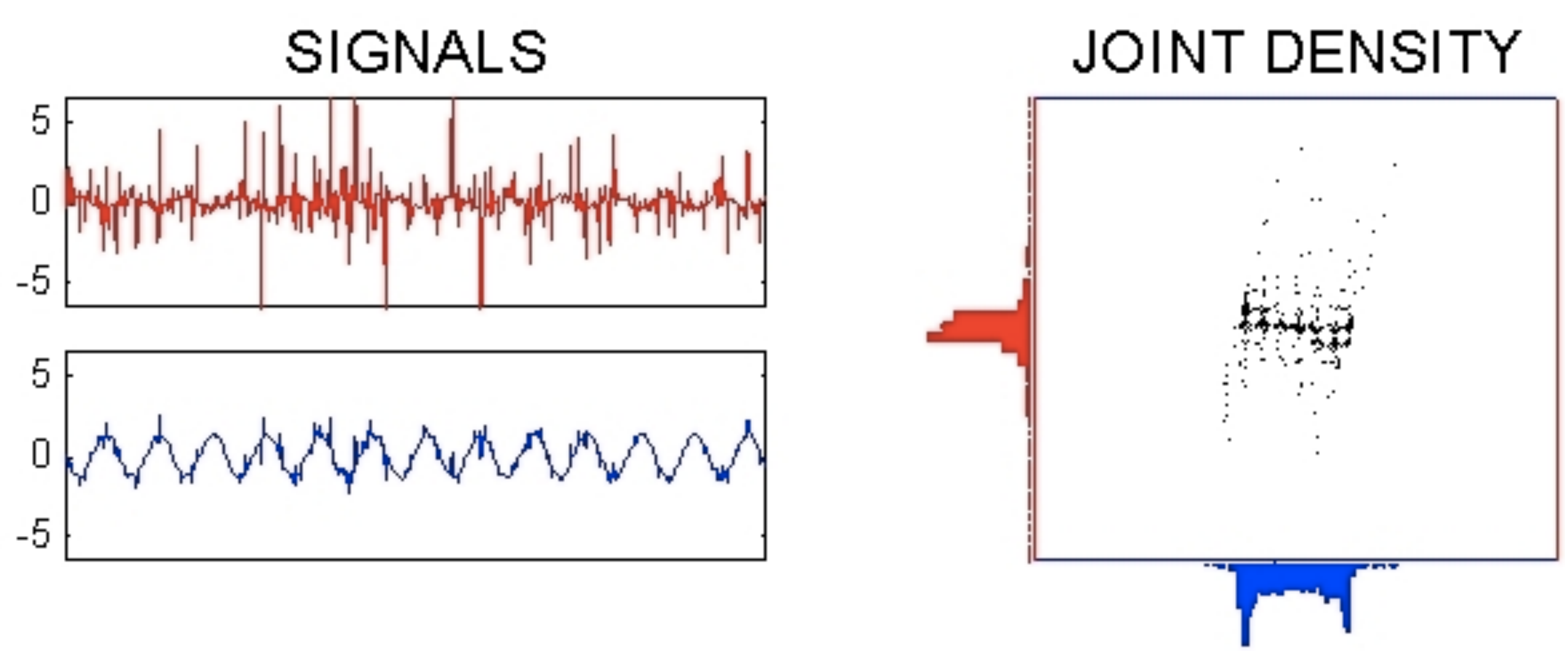
Input signals and density



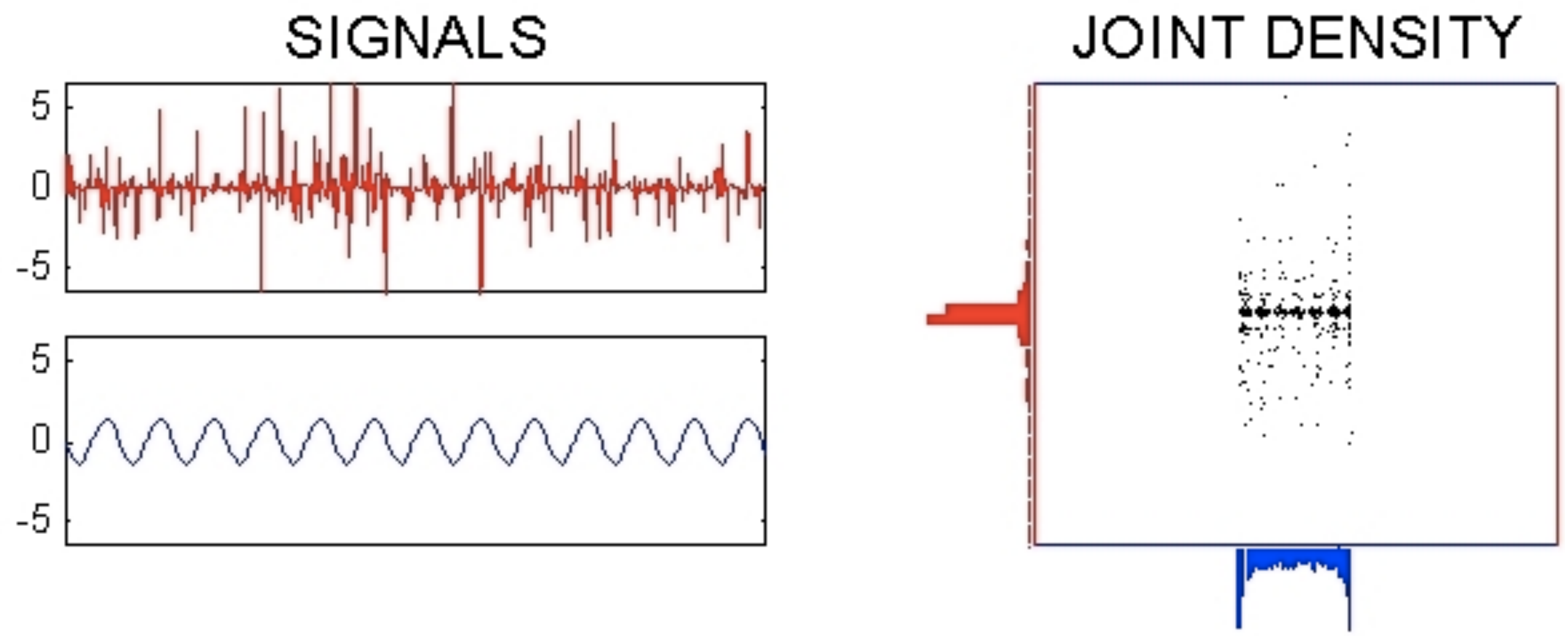
Whitened signals and density



Separated signals after 1 step of FastICA



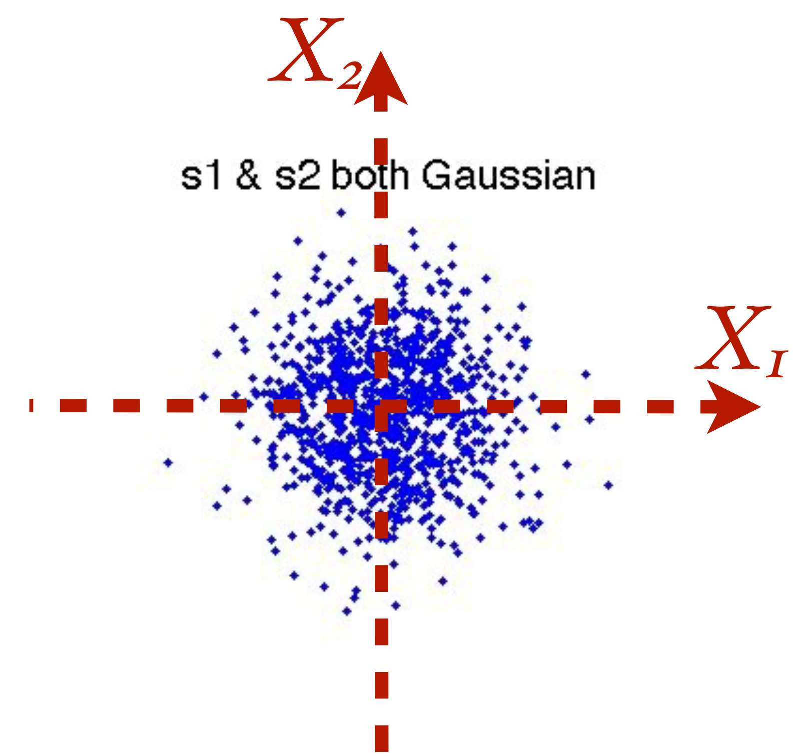
Separated signals after 3 steps of FastICA

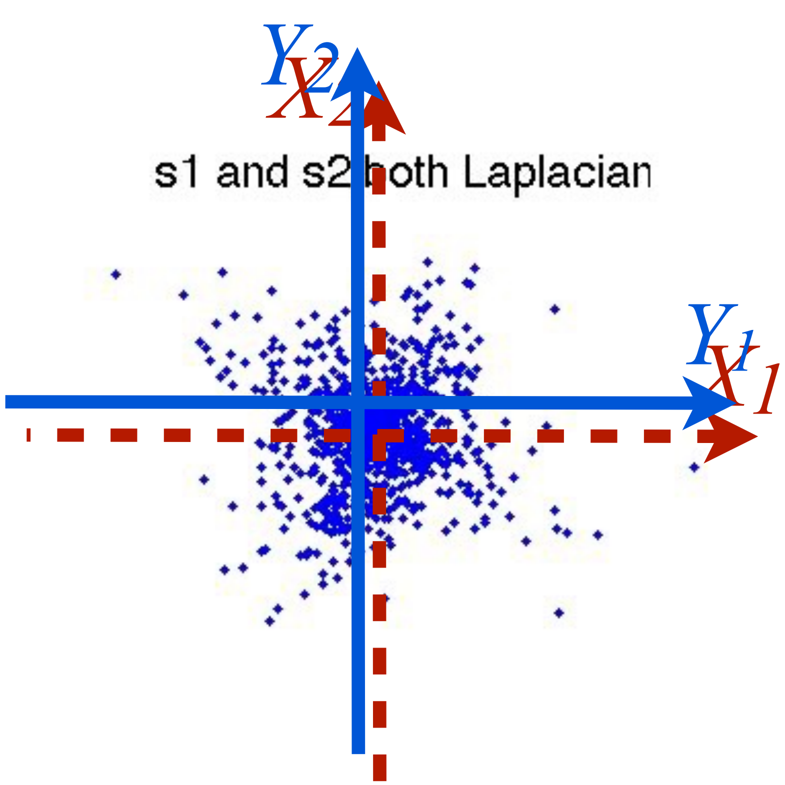


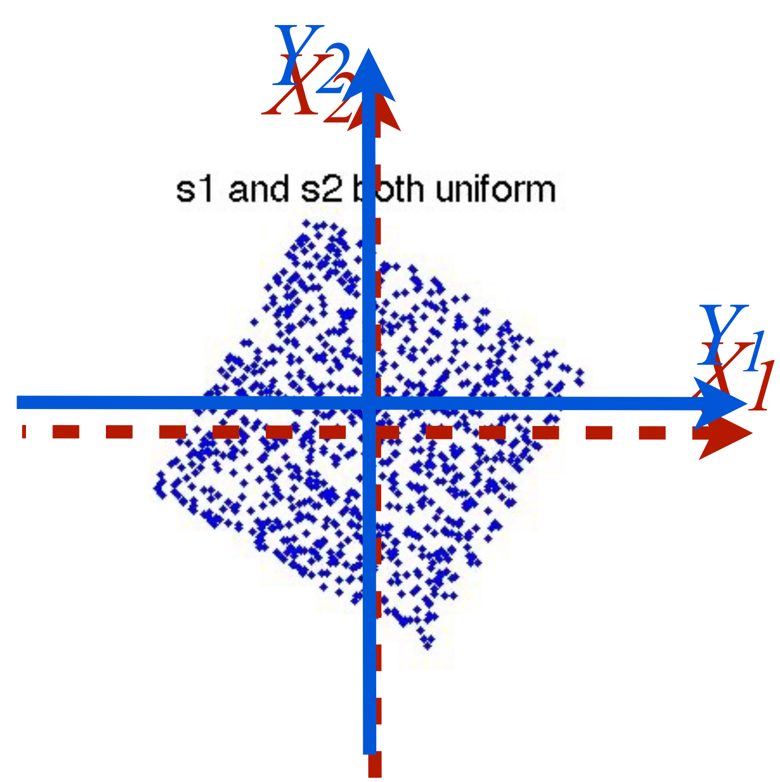
Separated signals after 5 steps of FastICA

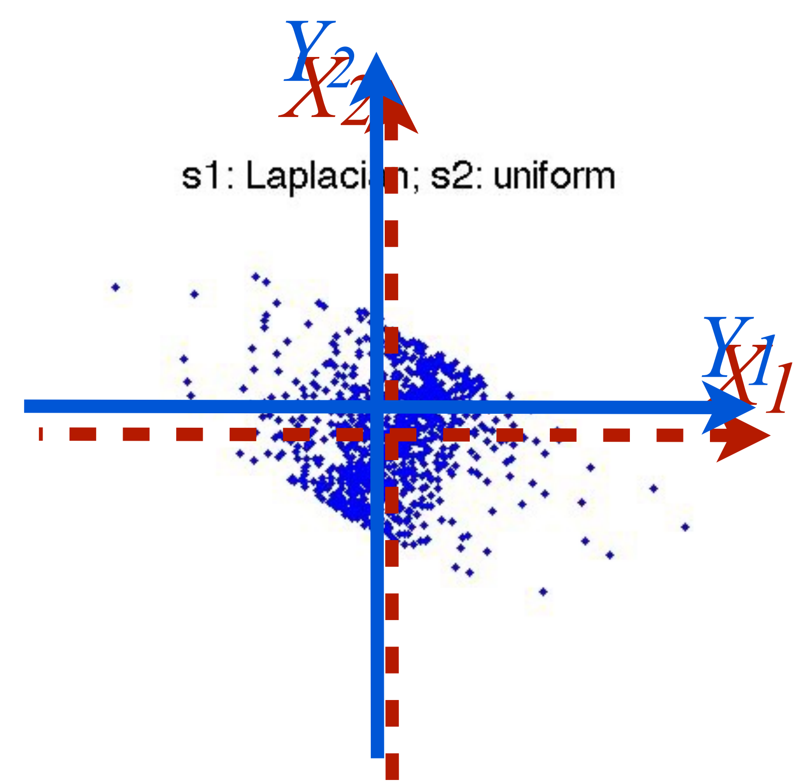
Intuition: Why ICA works?

- (After preprocessing with **Z=QX**) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{U} \cdot \mathbf{Z}$ to making Y_i independent
  - How 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, $i = 1, \ldots, n$:

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

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

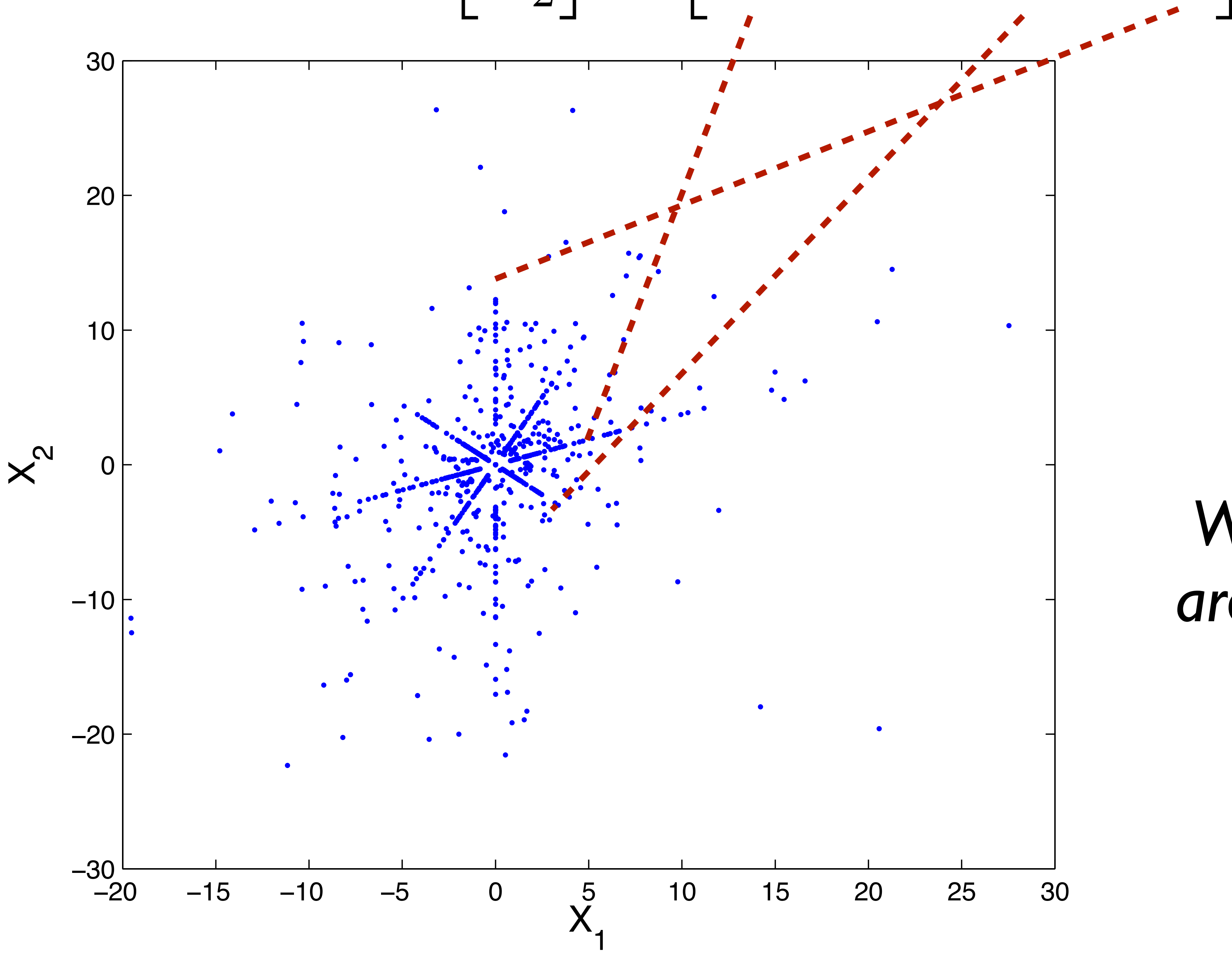
Cool! Can you then see the identifiability of the ICA problem?

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

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

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

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

$(\mathbf{x}_t:$ the t-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 $\prod_i P_{Yi}$:

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

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

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

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

- Nonnegative and zero iff Y_i are independent

- H(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 $|\mathbf{W}| = 1$

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

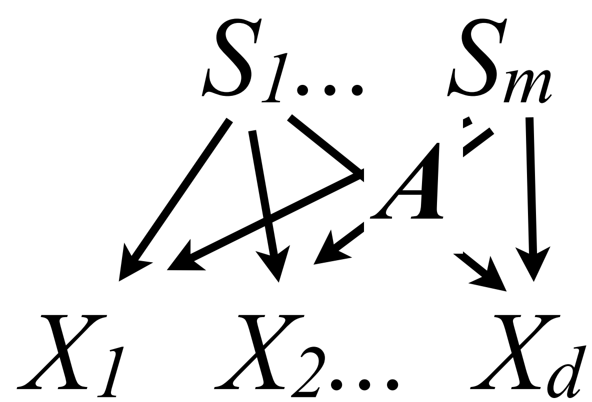
- Minimizing I $\Leftrightarrow$ minimizing the entropies

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

- Maximizing non-Gaussianity !

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

Connecting ICA to Causal Analysis



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

Summary: Class 5

- Typical unsupervised multivariate analysis methods: goals, models, assumptions, solutions, and relations to causal modeling
  - Principal component analysis
  - Factor analysis
  - Independent component analysis
- Graphical models
  - Local and global Markov property
  - Markov factorization of
  - d-separation
  - Causal graphical models