FROM APPLICATION TO THEORY: SECTION I

PRINCIPAL COMPONENT ANALYSIS

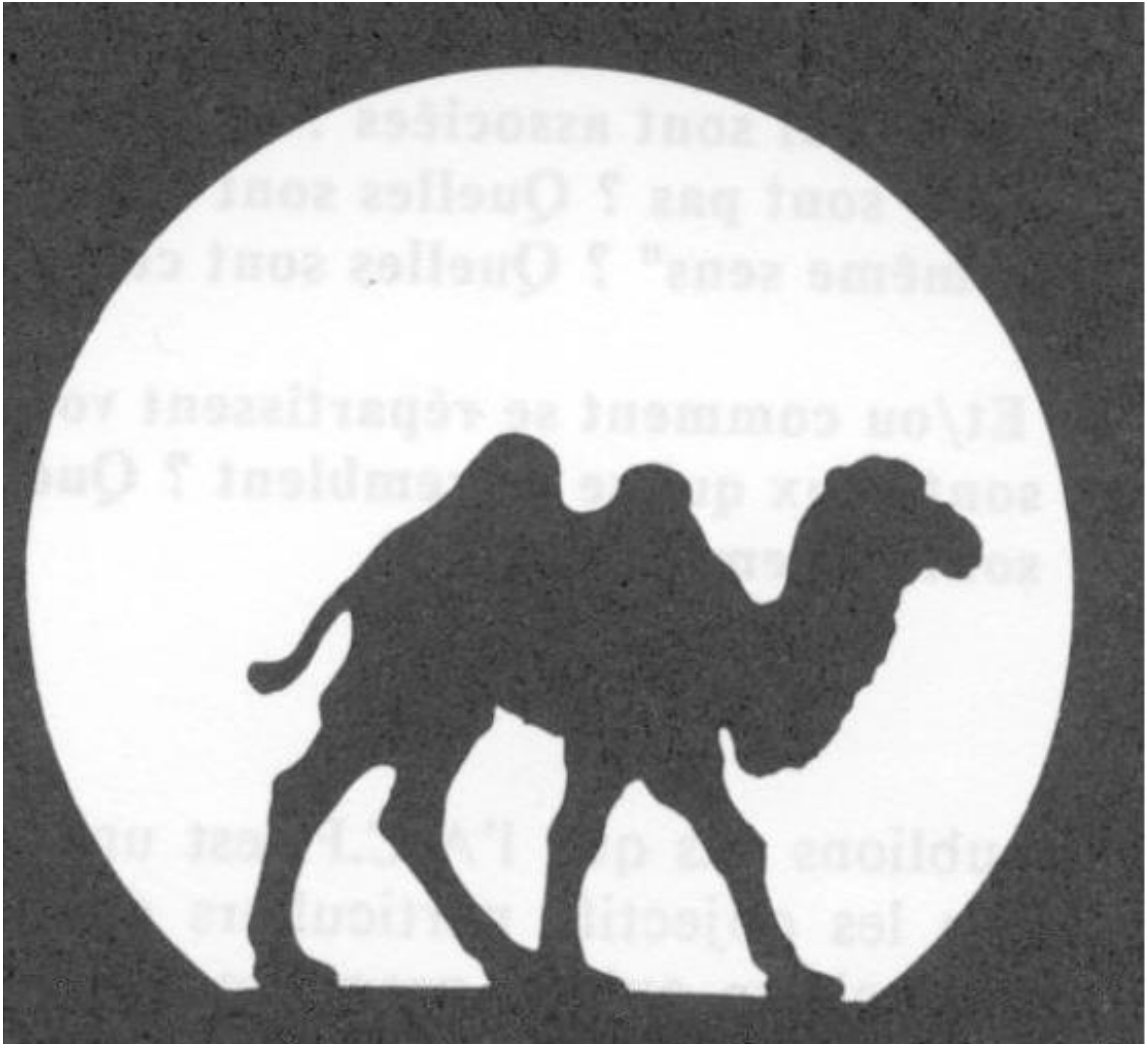
WHY USE PCA?

- **Reveal hidden structure**
  - Identify how different variables work together
  - Reduce the dimensionality
  - Decrease redundancy
  - Filter some of the noise
  - Compress data
  - Prepare the data for further

DATA IS REORGANIZED

- **The New Components:**
  - Are Independent, orthogonal, uncorrelated
  - Decrease in the amount of variance

Thus, only some will be retained for further study – Dimension Reduction



SOME MATH

A dimensionality reduction technique

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \qquad \mathrm{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

SOME MATH (CONT.)

Consider the linear combinations

- PCA projects p-dimensional data into a q-dimensional sub-space ($q<=p$)

$$
\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

$$\mathrm{var}(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i$$

$$\mathrm{cov}(Y_i, Y_j) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{jl}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

$$\boxed{\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}}$$

SOME MATH (CONT.)

1^{st} Principal Component:

- The linear combination of X, i.e., Y_1 or PC_1, that has maximum variance, subject to the constrain that the sum of all e_{ij}^2 over $j = 1, \ldots p$ is 1.

$$\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$$

More formally, select e_{11}, e_{12}, ... , e_{1p} that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} \sigma_{kl} = \mathbf{e}_1' \Sigma \mathbf{e}_1$$

Subject to:

$$\mathbf{e}_1' \mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1$$

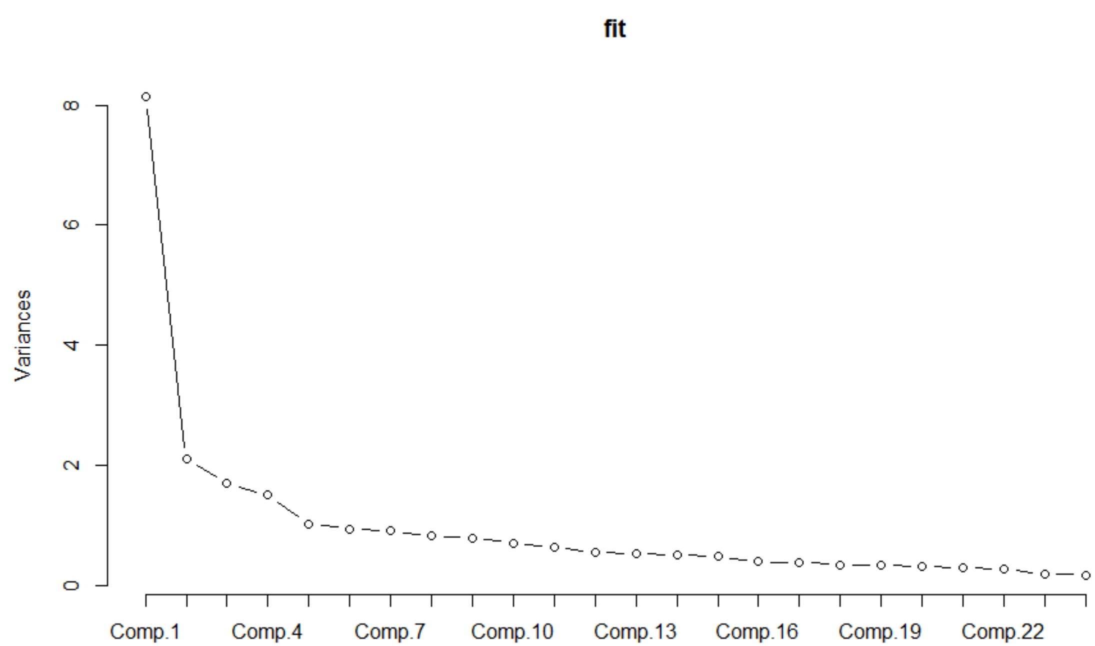
Correction: This is to ensure an unique answer

PCA

- First q Principal Component:
  - projected our p-dimensional data into a q-dimensional sub-space
  - We use the ratio of variance "explained" by the projected data to help us decide how many (q) PCs to retain → this can also be done/assisted with a Scree plot (next slide)

HOW DO WE CHOOSE *Q*? - **VISUALIZATION**

- Screeplot – help to find the cutting point of choosing the number of PCs



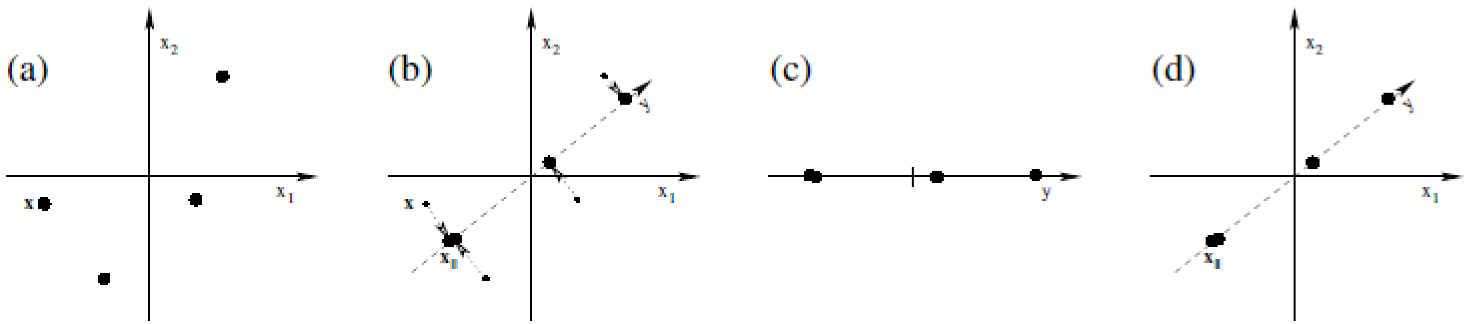
HOW DO WE CHOOSE *Q*?

- How many principal components to retain will depend on the specific application.

e.g. I choose the first 20 PCs as my candidate predictors in one of my studies because they together explain 87% of total variance in the original space.

(*Lu et al., 2013, Precipitation predictability associated with tropical moisture exports and circulation patterns for a major flood in France in 1995*)

PROJECTION & RECONSTRUCTION ERROR



data points in 2D | projection onto 1D | data points in 1D | reconstruction in 2D

Figure 1: Projection of 2D data points onto a 1D subspace and their reconstruction.

$$\mathbf{x} = (x_1, x_2) \qquad \mathbf{x}_\parallel := \mathbf{v}\mathbf{v}^T\mathbf{x} \qquad y := \mathbf{v}^T\mathbf{x} \qquad \mathbf{x}_\parallel \overset{(1,2)}{=} \mathbf{v}y$$

a unit vector $\mathbf{v}$

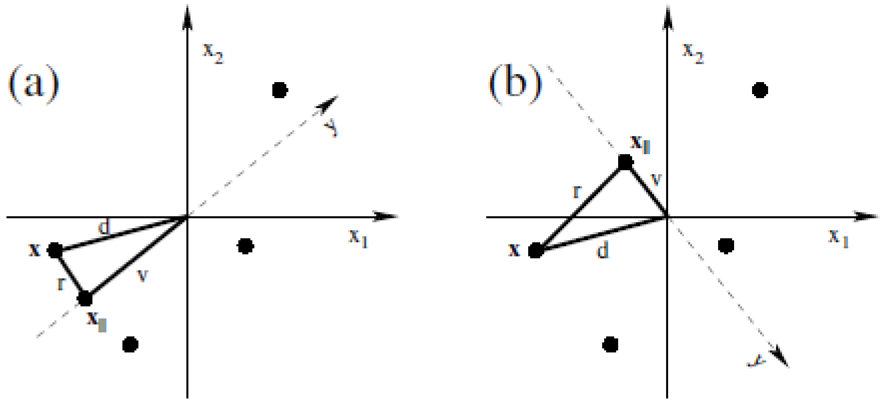
the unit vector $\mathbf{v}$

Laurenz Wiskott
Institut für Neuroinformatik, Ruhr-Universität Bochum

STILL REMEMBER?

"By minimizing the reconstruction error, we achieve maximizing the variance of the projected data too! Win-Win! And I have a better explanation than mathematical expression."

RECONSTRUCTION ERROR & VARIANCE



$$r^2 + v^2 = d^2$$

Reconstruction Error	Variance of the PC	Variance of the Data
Minimized	Maximized	Constant

DIRECTION OF MAXIMAL VARIANCE

- **By Covariance Matrix**

I: $X = (x_1, x_2)^T \; assume \quad zero \quad mean$

$$\begin{cases} Var(x_1) = C_{11} = < x_1 x_1 > \\ Var(x_2) = C_{22} = < x_2 x_2 > \end{cases}$$

If $C_{11} > C_{22}$, maximal direction is close to $(1,0)^T$ or $(-1,0)^T$

If $C_{22} > C_{11}$, maximal direction is close to $(0,1)^T$ or $(0,-1)^T$

What if $C_{11} \cong C_{22}$? $\longrightarrow$ Off-diagonal parts of the covariance matrix

DIRECTION OF MAXIMAL VARIANCE

- **By Covariance Matrix**

II: $X = (x_1, x_2)^T$ *assume zero mean*

$$Cov(x_1, x_2) = C_{12} = C_{21} = < x_1 x_2 >$$

If C_{12} (large) positive, data cloud stretched along $(1,1)^T$
If C_{12} (large) negative, data cloud stretched along $(-1,1)^T$

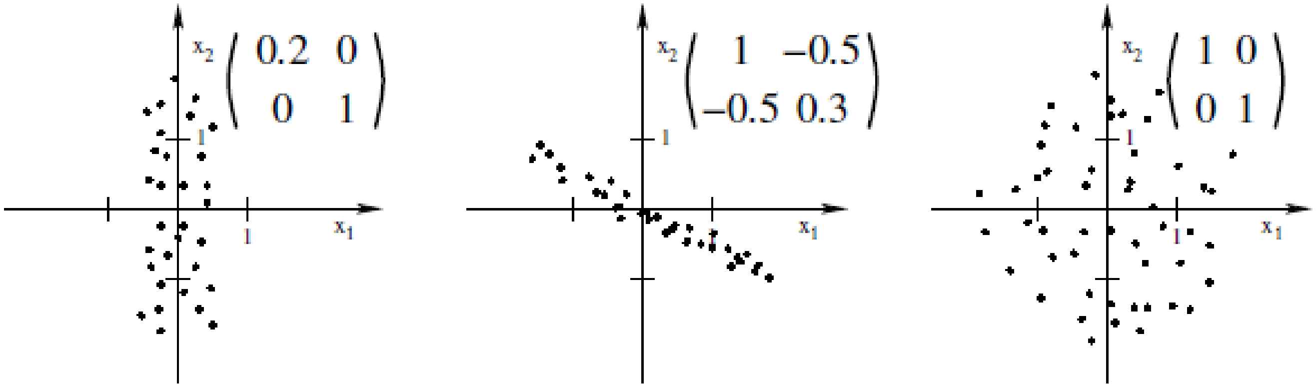
What if C_{12} is small and $C_{11} \cong C_{22}$ → No correlation and no prominent direction of maximal variance

DIRECTION OF MAXIMAL VARIANCE

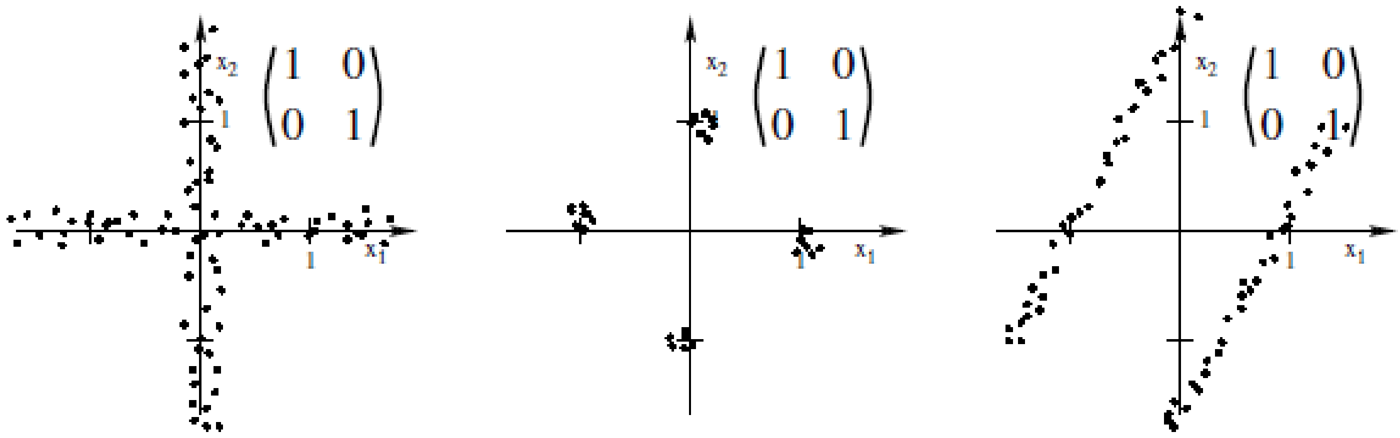
- **By Covariance Matrix**

III: $X = (x_1, x_2)^T \; assume \quad zero \quad mean$

$$C_{ij} = < x_i x_j >, \quad i = 1, 2; j = 1, 2$$



COVARIANCE ≠ DATA STRUCTURE



- ❑ The covariance matrix only gives you information about this general extent of the data, no higher-order structure of the data.