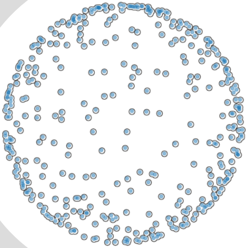
WiSe 2024/25

**Machine Learning for Data Science**

Lecture by G. Montavon



# Lecture 6a | Correlation

# Outline

**Recap: Data science vs. experimentation**

**Correlation between single variables**

- Pearson's correlation
- Connection to predictability

**Correlation between multiple variables**

- Motivations
- Canonical correlation analysis (CCA)

**Application 1**

- Learning equations with CCA

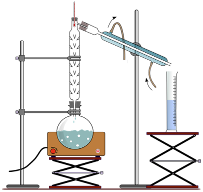
Part 1

# Data Science vs. Experimentation (Recap)

# Data Science vs. Experimentation

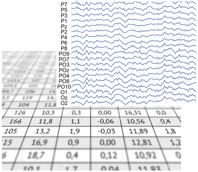
Two main approaches to empirical science:

### Experimentation



- The system of interest is available, and can be manipulated.
- Hypothesis are first stated and then guide the experimental protocol, i.e. which manipulation is applied to the system and which observables are measured.

### Data science

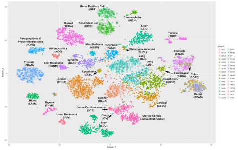



- Data is simply there, before any hypothesis has been stated.
- Correlation between features in the data can be measured, and can be used to suggest mechanistic/causal relation between variables.

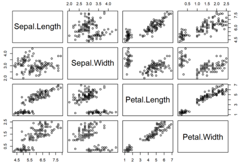
# Data Science Examples

**Examples:**

- *Cluster Analysis:* Find cluster structures in the data can suggest meaningful categories or a taxonomy.



- *Correlation Analysis:* Find that two variables are correlated can suggest a causal or mechanistic link between these two variables, so that changing one variable would incur a change on the other variable.



# Data Science vs. Experimentation

**Advantages of Data Science:**

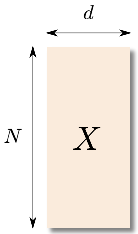
- No need to manipulate the system at hand (less time-consuming, avoids ethical issues).
- Can cover a much broader hypothesis space than a single experiment.

**Limitations of Data Science:**

- Bound to the data that is currently available (usually not possible to collect further observations from the system of interest).
- Identified correlations do not necessarily imply a causal link between the two variables. Actual experiments need to be performed to establish this causal relationship.

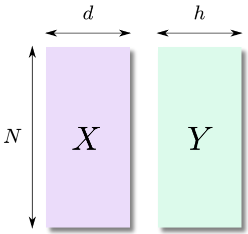
# Data Science

**Dispersion:** (Lectures 3–5)




- Identifying combination of measurements along which data varies the most (PCA).
- Find clusters that capture most variation in the data (k-Means, etc.).

**Correlation:** (Lectures 6–7)




- Find the level of correlation between two data modalities (Pearson's correlation, CCA).
- Find how predictable one modality is from the other (regression, discriminant analysis).

Part 2

# Pearson's Correlation

# Pearson's Correlation

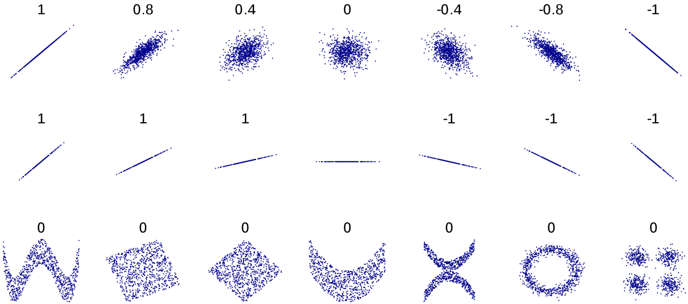
- Let $(x_i, y_i)_{i=1}^N$ be our dataset where each instance $i$ has two associated features $x_i$ and $y_i$. Let $\mu_x = \mathrm{E}[x]$ and $\mu_y = \mathrm{E}[y]$ be the dataset means for these two features.
- The *Pearson's correlation* associated to this dataset is defined as:

$$\rho = \frac{\overbrace{\mathrm{E}[(x-\mu_x)(y-\mu_y)]}^{\sigma_{xy}}}{\underbrace{\sqrt{\mathrm{E}[(x-\mu_x)^2]}}_{\sigma_x}\underbrace{\sqrt{\mathrm{E}[(y-\mu_y)^2]}}_{\sigma_y}}$$

- It is always contained in the interval $[-1, 1]$. A value of 1 indicates that $x$ and $y$ are perfectly correlated, $-1$ indicates perfect negative correlation, and 0 indicates that $x$ and $y$ are decorrelated.
- Unlike the variance of the data (lectures 3–5), the Pearson's correlation is invariant to the scaling of the data.

# Pearson's correlation

**Examples:**




- Pearson's correlation can only detect certain types of correlations (linear correlations) and may consequently underestimate the true level of correlation.

# Correlation as Mutual Predictability

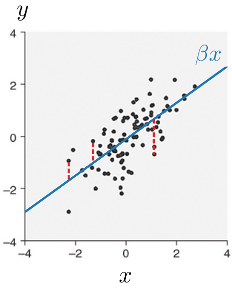
Assume the data is centered ($\mu_x = \mu_y = 0$). We build a homogeneous linear model that predicts $y$ from $x$ in a way that minimizes square errors:

$$\begin{aligned}
&\arg\min_\beta \mathrm{E}[(y - \beta x)^2] \\
&\quad = \arg\min_\beta \mathrm{E}[\beta^2 x^2 - 2\beta y x] \\
&\quad = \arg\min_\beta \beta^2 \sigma_x^2 - 2\beta\sigma_{xy}
\end{aligned}$$

This is a convex optimization problem with solution found at:

$$\frac{\partial}{\partial\beta}\left(\beta^2\sigma_x^2 - 2\beta\sigma_{xy}\right) = 0$$

which gives us the closed form $\beta = \sigma_{xy}/\sigma_x^2$.



# Correlation as Mutual Predictability

Injecting the closed form $\beta = \sigma_{xy}/\sigma_x^2$ in the error function gives us the error at the optimum (or some measure of predictivity of $y$ from $x$).

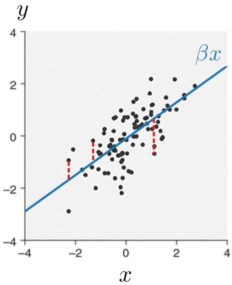
$$\min_\beta \mathrm{E}[(y - \beta x)^2]$$

$$= \mathrm{E}[(y - \frac{\sigma_{xy}}{\sigma_x^2} x)^2]$$

$$= \mathrm{E}[y^2] - 2\frac{\sigma_{xy}}{\sigma_x^2}\mathrm{E}[xy] + \Big(\frac{\sigma_{xy}}{\sigma_x^2}\Big)^2 \mathrm{E}[x^2]$$

$$= \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

$$= \sigma_y^2 \cdot \Big(1 - \Big(\underbrace{\frac{\sigma_{xy}}{\sigma_x \sigma_y}}_{\rho}\Big)^2\Big)$$



# Correlation as Mutual Predictability

We found that the residual prediction error of the optimal linear model is given by:

$$\begin{aligned}\min_\beta \mathrm{E}[(y - \beta x)^2] \\ = \sigma_y^2 \cdot (1 - \rho^2)\end{aligned}$$

Conversely, we can predict $x$ from $y$ with error given by

$$\begin{aligned}\min_\beta \mathrm{E}[(\beta y - x)^2] \\ = \sigma_x^2 \cdot (1 - \rho^2)\end{aligned}$$

**Conclusion:** There is a direct relation between Pearson's correlation and mutual predictibility (in the least squares sense, using a linear model).

Part 3

# Canonical Correlation Analysis

**Suggested reading:**
**Borga (2001): Canonical Correlation a Tutorial**

# Motivations: Multivariate Data

**Question:**

- Are images and their associated text correlated?

"motorcycle front wheel"

"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"wild boar head portrait forest creature boar"

"st oswalds way and shops"

Image source: `https://ai.googleblog.com/2021/05/align-scaling-up-visual-and-vision.html`

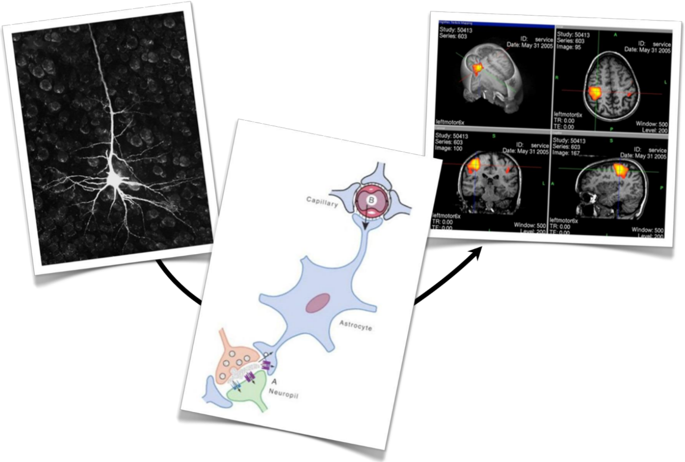
**Problem:**

- Images and text are multivariate (composed of multiple visual features or words).
- We need to generalize correlation to modalities of more than one dimension.

# Motivations: Neuro-Vascular Coupling

intercortical neural activity

FMRI/BOLD signal




*haemodynamic response*

# Motivations: Climate Dynamics

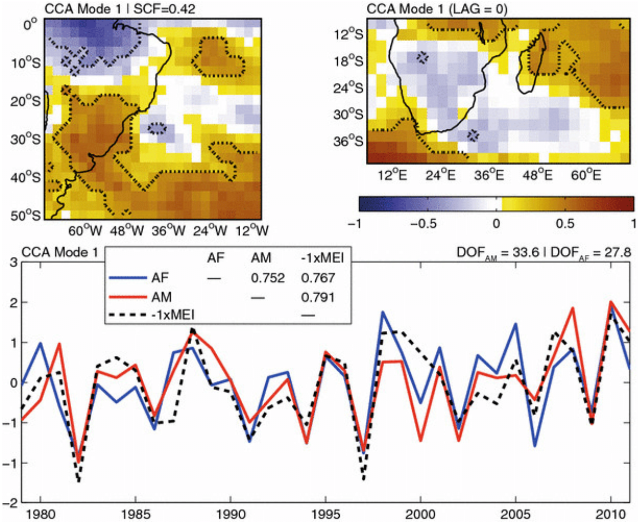
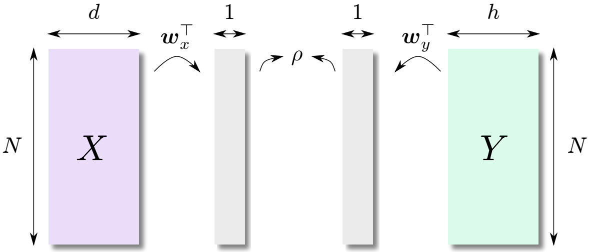



Image source: Climate co-variability between South America and Southern Africa at interannual, intraseasonal and synoptic scales. DOI:10.1007/s00382-016-3318-x

# Canonical Correlation Analysis




**Formalization:**

- Let $\boldsymbol{x} \in \mathbb{R}^d$ be our first modality and $\boldsymbol{y} \in \mathbb{R}^h$ be our second modality.
- Find projections $\boldsymbol{w}_x \in \mathbb{R}^d$ and $\boldsymbol{w}_y \in \mathbb{R}^h$ of the two modalities in which the Pearson's correlation is maximized:

$$\arg\max_{\boldsymbol{w}} \ \mathrm{Corr}(\boldsymbol{w}_x^\top \boldsymbol{x}, \boldsymbol{w}_y^\top \boldsymbol{y})$$

$(\boldsymbol{w} = (\boldsymbol{w}_x, \boldsymbol{w}_y))$. *Note: Because Pearson's correlation is invariant to the scale of the data, the vectors $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ don't need to be constrained to be e.g. of unit norm.*

# Canonical Correlation Analysis

**Observation:**

- The correlation can be rewritten as:

$$
\begin{aligned}
&\mathrm{Corr}(\boldsymbol{w}_x^\top \boldsymbol{x}, \boldsymbol{w}_y^\top \boldsymbol{y}) \\
&\quad = \frac{\mathrm{E}[(\boldsymbol{w}_x^\top (\boldsymbol{x} - \boldsymbol{\mu}_x) \cdot (\boldsymbol{w}_y^\top (\boldsymbol{y} - \boldsymbol{\mu}_y))]}{\sqrt{\mathrm{E}[(\boldsymbol{w}_x^\top (\boldsymbol{x} - \boldsymbol{\mu}_x))^2]}\sqrt{\mathrm{E}[(\boldsymbol{w}_y^\top (\boldsymbol{y} - \boldsymbol{\mu}_y))^2]}} \\
&\quad = \frac{\boldsymbol{w}_x^\top C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x}\sqrt{\boldsymbol{w}_y^\top C_{yy} \boldsymbol{w}_y}}
\end{aligned}
$$

where

$$
\begin{aligned}
C_{xy} &= \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top] \\
C_{xx} &= \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top] \\
C_{yy} &= \mathrm{E}[(\boldsymbol{y} - \boldsymbol{\mu}_y)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top]
\end{aligned}
$$

are cross-covariance and auto-covariance matrices that can be precomputed.

# Canonical Correlation Analysis

**Observation (2):**

- Remember that the optimum of the problem:

$$\arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}_x^\top C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x}\sqrt{\boldsymbol{w}_y^\top C_{yy} \boldsymbol{w}_y}}$$

  is specified up to a scaling of $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$.

**Idea:**

- Force a particular scale of $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, and get in turn a simpler optimization problem.
- In particular, the optimization problem can be reformulated as a standard quadratic optimization problem with quadratic constraints (like for PCA):

$$\arg\max_{\boldsymbol{w}} \; \boldsymbol{w}_x^\top C_{xy} \boldsymbol{w}_y \quad \text{s.t.} \quad \begin{aligned} \boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x &= 1 \\ \boldsymbol{w}_y^\top C_{yy} \boldsymbol{w}_y &= 1 \end{aligned}$$

# Canonical Correlation Analysis

**Idea (2):**

- For the constrained optimization problem

$$\arg\max_{\boldsymbol{w}} \; \boldsymbol{w}_x^\top C_{xy} \boldsymbol{w}_y \quad \text{s.t.} \quad \begin{aligned} \boldsymbol{w}_x^\top C_{xx} \boldsymbol{w}_x &= 1 \\ \boldsymbol{w}_y^\top C_{yy} \boldsymbol{w}_y &= 1 \end{aligned}$$

  one can obtain using the method of Lagrange multipliers a closed form solution ($\rightarrow$ Lecture 6b for the derivation).

**Final formulation:**

- One gets that the solution of CCA is the first eigenvector (i.e. with highest eigenvalue) of the generalized eigenvalue problem:

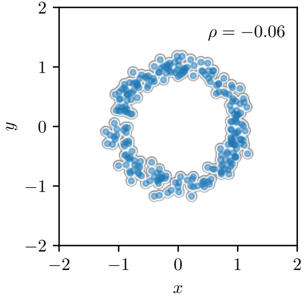
$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix}$$

  and the latter can be solved with common libraries (e.g. scipy).
- The eigenvalue $\lambda$ represent the correlation $\rho$ of data in the two projected spaces.

Part 4

# Application: Nonlinear Correlations

# Application: Finding Nonlinear Correlations



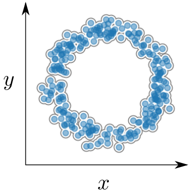

**Observation:**

- Negligible Pearson's correlation coefficient.
- However, it is clear visually that $x$ and $y$ are still related in some way.

**Idea:**

- Nonlinearly expand $x$ and $y$, and apply CCA to find correlation between $x$ and $y$ in expanded space.

# Application: Finding Nonlinear Correlations




$$\begin{tikzcd}
x \arrow[rr] \arrow[dr] & & \phi(x) = [x, x^2, x^3] \arrow[dr] & \\
& \rho(x,y) \approx 0 & & \mathrm{CCA}(\phi(x), \phi(y)) > 0\ ? \\
y \arrow[rr] \arrow[ur] & & \phi(y) = [y, y^2, y^3] \arrow[ur] &
\end{tikzcd}$$

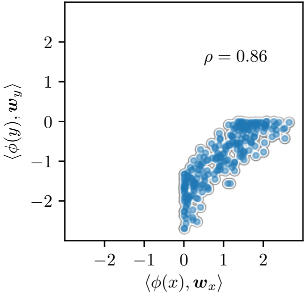
- CCA can explore possible nonlinear correlations.
- Correlation found by CCA should be in theory at least as good original correlation (if including $x$ and $y$ in the feature map).

# Application: Finding Nonlinear Correlations

**CCA Result:**

$$\lambda_1 = 0.86 \qquad \boldsymbol{u}_1 = [\underbrace{-0.13, 1.84, 0.15}_{\boldsymbol{w}_x}, \underbrace{0.09, -1.88, -0.07}_{\boldsymbol{w}_y}]$$

- Projecting expanded modalities on this eigenvector, gives a new scatter plot, where a linear correlation can be observed and measured:




- Correlation strength was readily given by the first eigenvalue of the CCA generalized eigenvalue problem.

# Application: Finding Nonlinear Correlations

**CCA Result:**

$$\lambda_1 = 0.86 \qquad \boldsymbol{u}_1 = \big[\underbrace{-0.13, 1.84, 0.15}_{\boldsymbol{w}_x}, \underbrace{0.09, -1.88, -0.07}_{\boldsymbol{w}_y}\big]$$

- An inspection of this result suggests that approximately:

  $$\text{Corr}(0 + x^2 + 0, 0 - y^2 + 0) \approx 0.86$$

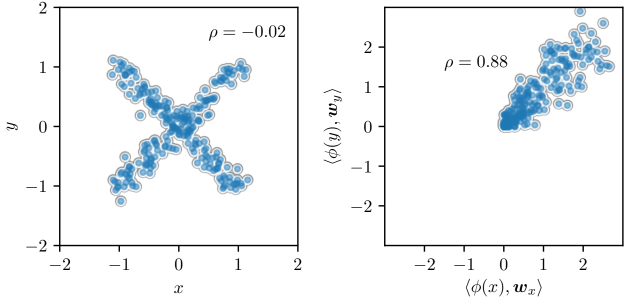
  In other words, it suggests that the data is governed by the equation:

  $$x^2 = \alpha(-y^2) + \beta$$

  The true equation that describes the data would be obtained by setting $\alpha = 1$ and $\beta = 1$, i.e. $x^2 + y^2 = 1$.

# Application: Finding Nonlinear Correlations

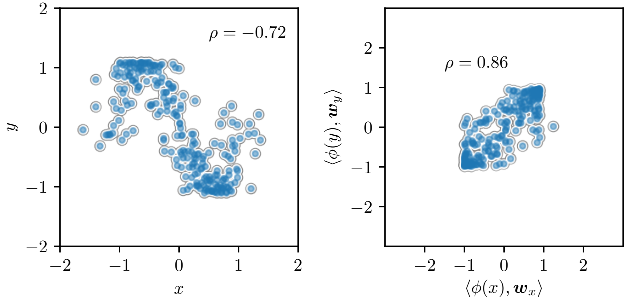
*Example 2:*




$$\lambda_1 = 0.88 \qquad \boldsymbol{u}_1 = [\underbrace{0.16, 2.04, -0.12}_{\boldsymbol{w}_x}, \underbrace{-0.09, 2.00, 0.16}_{\boldsymbol{w}_y}]$$

# Application: Finding Nonlinear Correlations

*Example 3:*




$$\lambda_1 = 0.86 \qquad u_1 = \big[\underbrace{1.84, -0.11, -1.06}_{\boldsymbol{w}_x}, \underbrace{-1.06, 0.00, 0.15}_{\boldsymbol{w}_y}\big]$$

## Summary

# Summary

- Finding correlations is an important component of data science as it allows to identify potential causal or mechanistic relations in some system of interest.
- Pearson's correlation is the most established measure of correlation and it has many connections to other statistical properties (e.g. explained variance, error of a linear model, etc.).
- Canonical correlation analysis generalizes Pearson's correlation to the case where the two modalities one would like to correlate are multidimensional.
- Canonical correlation analysis can also be used to discover nonlinear relations in bivariate data.