Learning From Data
Lecture 12: Unsupervised Learning III

Yang Li **yangli@sz.tsinghua.edu.cn**

TBSI

December 16, 2022

Today's Lecture

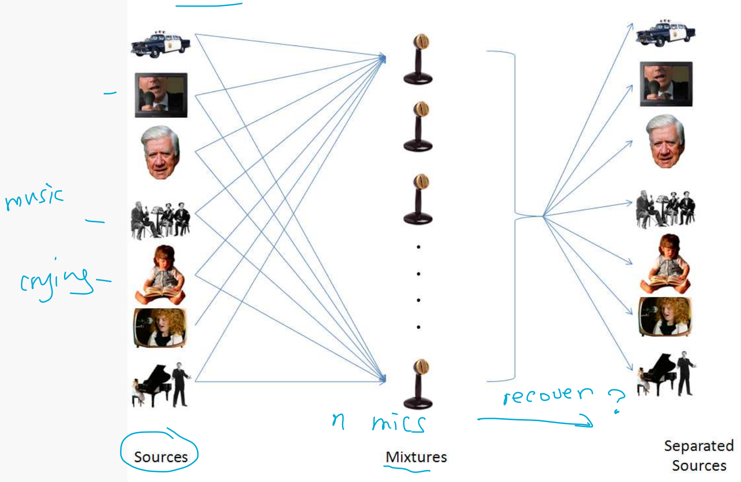
Unsupervised Learning (Part III)

- Independent Component Analysis (ICA)
- Canonical Correlation Analysis (CCA)

Independent Component Analysis

The cocktail party problem

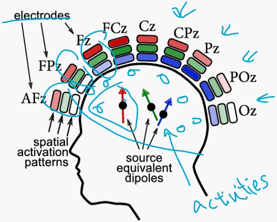
- n microphones at different locations of the room, each recording a mixture of n sound sources
- How to “unmix" the sound mixtures?

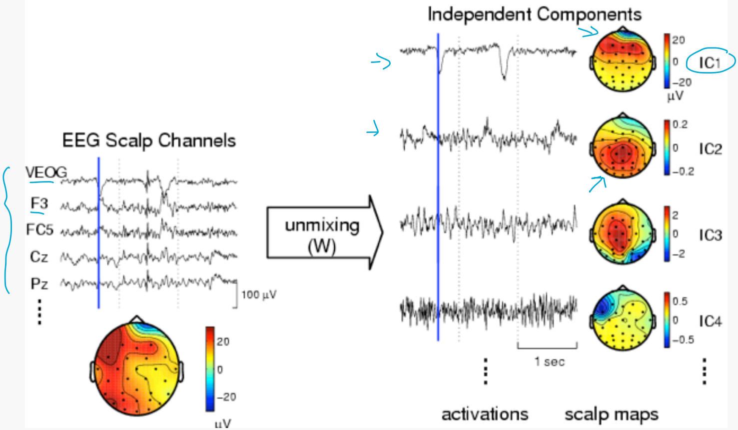


Sample audio: https://cnl.salk.edu/~tewon/Blind/blind_audio.html,
http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html

EEG Analysis

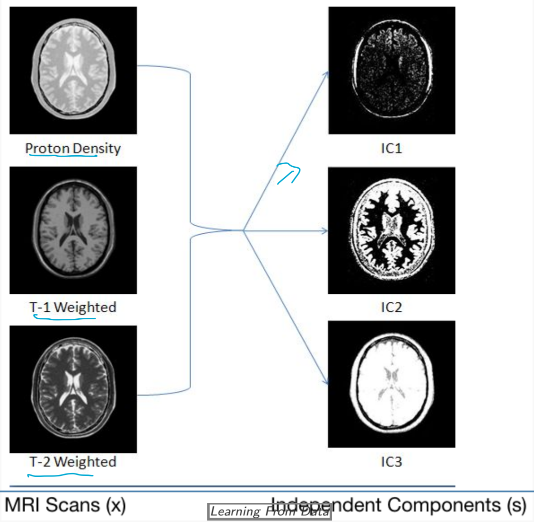
- Electrodes on patient scalp measure a mixture of different brain activations
- Finding independent activation sources helps removing artifacts in the signal





Brian imaging

- Different brain matters: gray matter, white matter, cerebrospinal fluid (CSF), fat, muscle/skin, glial matter etc.
- An MRI scan is a mixture of magnetic response signals from different brain matters



Problem Model

\# of independence source variables

Case: $n = 2$ → # of observation variables

- Observed random variables: x_1, x_2
- Independent sources: $s_1, s_2 \in \mathbb{R}$

$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

Problem Model

Case: $n = 2$

- Observed random variables: x_1, x_2
- Independent sources: $s_1, s_2 \in \mathbb{R}$

$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

A is called the **mixing matrix** $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$

$$x = As$$

Problem Model

Case: $n = 2$

- Observed random variables: x_1, x_2
- Independent sources: $s_1, s_2 \in \mathbb{R}$

$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

A is called the **mixing matrix**

$$s^{(i)} = \begin{bmatrix} s_1^{(i)} \\ s_2^{(i)} \end{bmatrix}$$

$$\begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \quad x = As$$

The blind source separation (cocktail party) problem

Given repeated observation $\{x^{(i)}; i = 1, \ldots, m\}$, recover sources $s^{(i)}$ that generated the data ($x^{(i)} = As^{(i)}$)

Independent Component Analysis (ICA)

The blind source separation (cocktail party) problem

Given repeated observation $\{x^{(i)}; i = 1, \ldots, m\}$, recover sources $s^{(i)}$ that generated the data ($x^{(i)} = As^{(i)}$)

mixing matrix

Let $W = A^{-1}$ be the **unmixing matrix**
Goal of ICA: Find W, such that given $x^{(i)}$, the sources can be recovered by $s^{(i)} = Wx^{(i)}$

$$j \rightarrow \begin{bmatrix} s_1^{(i)} \\ \vdots \\ s_n^{(i)} \end{bmatrix} = \begin{bmatrix} - w_1^T - \\ \\ - w_n^T - \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \quad W = \begin{bmatrix} - w_1^T - \\ \vdots \\ - w_n^T - \end{bmatrix}$$

$$s_j^{(i)} = w_j^T x^{(i)}$$

ICA Ambiguities

Question: Does assuming $\mathbb{E}[s_i^2]=1$ resolve the scale ambiguity in $x=As$?
A: no, because $-s_j$ gives the same result as s_j

Assume data is **non Gaussian**, ICA has two ambiguities:

scale

- Variance of the sources: *We can fix the magnitude of s_i by setting $\mathbb{E}[s_i^2] = 1$*

$x = As$

$x_j = \sum_{i=1}^{n} a_{ji} s_i$ for all j

$= \sum_{i=1}^{n} \underbrace{a_{ji}\left(\frac{1}{c_j}\right)}_{A'} \underbrace{(c_j s_i)}_{s'}$ for any $c_j \neq 0$

$W = A^{-1}$

ICA Ambiguities

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{P^{-1}} \underbrace{\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}}_{s} = \underbrace{\begin{bmatrix} s_2 \\ s_3 \\ s_1 \end{bmatrix}}_{s'}$$

Assume data is **non Gaussian**, ICA has two ambiguities:

- Variance of the sources: *We can fix the magnitude of s_i by setting $\mathbb{E}[s_i^2] = 1$*
- Order of the sources $\underline{s_1}, \ldots, \underline{s_n}$:
  Let P be a permutation matrix, then we have $\underline{x} = \underbrace{AP}_{\underline{A'}}\underbrace{P^{-1}s}_{s'}$.
  $n \times n$

ICA Ambiguities

Assume data is **non Gaussian**, ICA has two ambiguities:

- Variance of the sources: *We can fix the magnitude of s_i by setting* $\mathbb{E}[s_i^2] = 1$
- Order of the sources $s_1, \ldots, s_n$:
  Let P be a permutation matrix, then we have $x = APP^{-1}s$.

ICA Ambiguities

Assume data is **non Gaussian**, ICA has two ambiguities:

- Variance of the sources: *We can fix the magnitude of s_i by setting $\mathbb{E}[s_i^2] = 1$*
- Order of the sources $s_1, \ldots, s_n$:
  Let P be a permutation matrix, then we have $x = APP^{-1}s$.

Why is Gaussian data problematic?

independent sources $s = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$

Given A, $x = As$

$s \sim \mathcal{N}(0, I)$

$\boxed{x \sim \mathcal{N}(0, AA^T)}$

$\mathbb{E}[As] = A\mathbb{E}[s] = A \cdot 0 = 0.$

$\mathbb{E}[xx^T] = \mathbb{E}[(As)(As)^T] = \mathbb{E}[Ass^TA^T]$

$= A\underbrace{\mathbb{E}[ss^T]}_{I}A^T = AA^T$

$x' = (A \cdot R)s$ rotation (orthogonal)

$R^{-1} = R^T$

$\boxed{x' \sim \mathcal{N}(0, \underbrace{(AR)(AR)^T}) = \mathcal{N}(0, AA^T)}$

$A\underbrace{RR^T}_{I}A^T = AA^T.$

ICA Ambiguities

Assume data is **non Gaussian**, ICA has two ambiguities:

- Variance of the sources: *We can fix the magnitude of s_i by setting* $\mathbb{E}[s_i^2] = 1$
- Order of the sources $s_1, \ldots, s_n$:
  Let P be a permutation matrix, then we have $x = APP^{-1}s$.

Why is Gaussian data problematic?

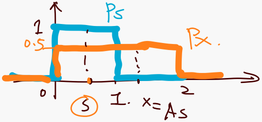
- The distribution of any rotation of Gaussian x has the same distribution as x.
- As long as at least one s_j is non-Gaussian, given enough data, we can recover the n independent sources.

Densities and Linear Transformations

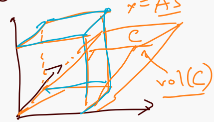
let $s \sim \text{Uniform}(0,1)$

let $A = 2$

$x = \underbrace{As}_{2s.} \sim \text{Uniform}(0,2)$



$s \sim U(0,1)\times U(0,1)$



$p_s(s) = 1.$

$$\left\{ \begin{array}{l} p_x(x) = p_x(As) = 0.5 \\ p_s(s) \cdot \underbrace{|A^{-1}|}_{\frac{1}{2}} = 1 \cdot \frac{1}{2} = 0.5 \end{array} \right\}$$

$p_x = \frac{1}{vol(C)} 1\{x \in C\}$

Theorem 1

If random vector s has density p_s, and $x = As$ for a square, invertible matrix A, then the density of x is

$$p_x(x) = p_s(\overset{s}{Wx}) \cdot \underbrace{|W|}_{\text{determinant}}$$

where $W = A^{-1}$.

ICA Algorithm

The joint distribution of *independent* sources $s = \{s_1, \ldots, s_n\}$:

$$p(s) = \prod_{j=1}^{n} p_s(s_j)$$

ICA Algorithm

The joint distribution of *independent* sources $s = \{s_1, \dots, s_n\}$:

$$p(s) = \prod_{j=1}^{n} p_s(s_j)$$

The density of observation $x = As$ is:

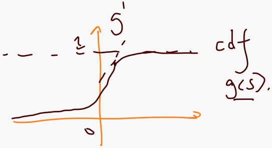
$s_j = w_j^T x.$

$$p_x(x) = p_s(s)|W| = \prod_{j=1}^{n} p_s(s_j)|W| = \prod_{j=1}^{n} p_s(w_j^T x)|W|$$

ICA Algorithm

The joint distribution of *independent* sources $s = \{s_1, \ldots, s_n\}$:

$$p(s) = \prod_{j=1}^{n} p_s(s_j)$$



The density of observation $x = As$ is:

$$p_x(x) = p_s(s)|W| = \prod_{j=1}^{n} p_s(s_j)|W| = \prod_{j=1}^{n} p_s(w_j^T x)|W|$$

Choose the sigmoid function $g(s) = \frac{1}{1+e^{-s}}$ as the *non-Gaussian* cdf for p_s, then

$$p_s(s) = g'(s)$$

ICA Algorithm

The joint distribution of *independent* sources $s = \{s_1, \dots, s_n\}$:

$$p(s) = \prod_{j=1}^{n} p_s(s_j)$$

The density of observation $x = As$ is:

$$p_x(x) = p_s(s)|W| = \prod_{j=1}^{n} p_s(s_j)|W| = \prod_{j=1}^{n} p_s(w_j^T x)|W|$$

Choose the sigmoid function $g(s) = \frac{1}{1+e^{-s}}$ as the *non-Gaussian* cdf for p_s, then

$$p_s(s) = g'(s)$$

This appears to be a heuristic choice, yet it can be justified rigorously in other interpretations.

ICA Algorithm

pdf of s: $g'(s) = g(s)(1 - g(s))$

$\frac{d}{ds} g(s)$

Given i.i.d. training samples $\{x^{(1)}, \ldots, x^{(m)}\}$, the log likelihood is

$$l(W) = \sum_{i=1}^{m} log(p_x(x^{(i)})) = \sum_{i=1}^{m} log(\prod_{j=1}^{n} p_s(w_j^T x)|W|)$$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$

ICA Algorithm

Given i.i.d. training samples $\{x^{(1)}, \ldots, x^{(m)}\}$, the log likelihood is

$$l(W) = \sum_{i=1}^{m} log(p_x(x^{(i)})) = \sum_{i=1}^{m} log(\prod_{j=1}^{n} p_s(w_j^T x)\underline{|W|})$$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log \underbrace{g'(w_j^T x^{(i)})}_{g(w_j^T x^i)(1-g(w_j^T x^i))} + \underline{\log |W|} \right)$$

Stochastic gradient ascent learning rule for sample $x^{(i)}$:

$$W := W + \alpha \left(\underbrace{\begin{bmatrix} 1 - 2g(w_1{}^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n{}^T x^{(i)}) \end{bmatrix} x^{(i)}{}^T + (W^T)^{-1}}_{\nabla W.} \right)$$

Check this at home!

Theoretical Motivation of ICA

- Originally proposed by Jutten & Herault (1991) [1] *90 years later than PCA*

$W = A^{-1}$

kurtosis

- Equivalent to learning projection directions $w_1, \ldots, w_n$ that
  - maximize the **sum of non-gaussianity** of the projected signals
  - minimize the **mutual information** of the projected signals → independent sources

  under the constraint that $w_1^T x, \ldots, w_n^T x$ are uncorrelated. [2]

$s = Wx$

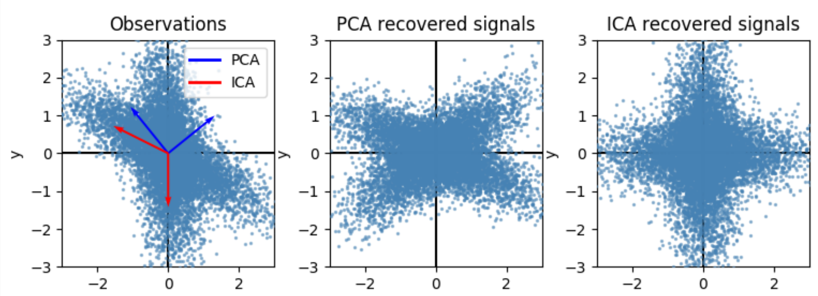
Mixture problem → Nonlinear Mixture.

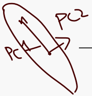
nonlinear-mixture

[1] Christian Jutten, Jeanny Herault, Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture, Signal Processing, Vol 24:1, 1991

[2] Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." Neural networks 13.4-5 (2000): 411-430.

ICA vs PCA





PCA	ICA
approximately Gaussian data	non-Gaussian data
removes correlation (low order dependence)	removes correlations and higher order dependence $\mathbb{E}[x^4]$
ordered importance	all components are equally important
orthogonal	not orthogonal

Canonical Correlation Analysis

Canonical Correlation Analysis

Canonical correlation analysis (CCA) finds the associations among two sets of variables.

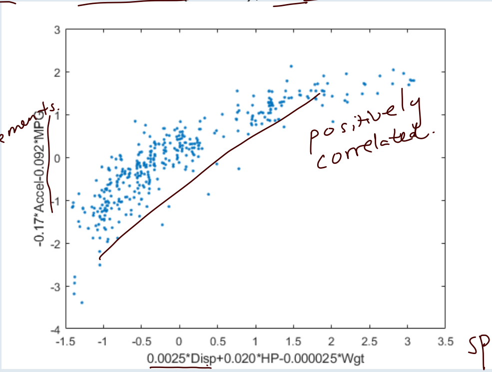
Canonical Correlation Analysis

Canonical correlation analysis (CCA) finds the associations among two sets of variables.

Example: two sets of measurements of 406 cars:

- Specification: Engine displacement (Disp), horsepower (HP), weight (Wgt)
- Measurement: Acceleration (Accel), MPG

measurements



specification

find important features that explain covariation between sets of variables

CCA Definitions

- Random vectors $X = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_2} \end{bmatrix}$

$cov(X) = cov(X, X)$ Σ_{XX}

$cov(Y) = \Sigma_{YY}$.

- Covariance matrix $\Sigma_{XY} = cov(X, Y)$

CCA Definitions

- Random vectors $X = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_2} \end{bmatrix}$
- Covariance matrix $\Sigma_{XY} = cov(X, Y)$
- CCA finds vectors a and b such that the random variables $a^T X$ and $b^T Y$ maximize the correlation

$$\rho = corr(a^T X, b^T Y)$$

CCA Definitions

- Random vectors $X = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_2} \end{bmatrix}$
- Covariance matrix $\Sigma_{XY} = cov(X, Y)$
- CCA finds vectors a and b such that the random variables $a^T X$ and $b^T Y$ maximize the correlation

$$\rho = corr(\underbrace{a^T X}_{u}, \underbrace{b^T Y}_{V})$$

- $U = a^T X$ and $V = b^T Y$ are called **the first pair of canonical variables**

CCA Definitions

- Random vectors $X = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_2} \end{bmatrix}$
- Covariance matrix $\Sigma_{XY} = cov(X, Y)$
- CCA finds vectors a and b such that the random variables $a^T X$ and $b^T Y$ maximize the correlation

$$\rho = corr(a^T X, b^T Y)$$

$$a_1, b_1 = \underset{a,b}{\operatorname{argmax}}\ corr(a^T X, b^T Y)$$

s.t. . . . .

- $U = a^T X$ and $V = b^T Y$ are called **the first pair of canonical variables**
- Subsequent pairs of canonical variables maximizes ρ while being *uncorrelated* with all previous pairs

$a_2 \perp a_1$

Review: Singular Value Decomposition

A generalization of eigenvalue decomposition to rectangle ($m \times n$) matrices M.

$(m\times n)\ (m\times m)\ (m\times n)\ (n\times n)$ rank

$$M = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

$u_1 \cdots u_m$ $\quad$ $-v_1-$ $\cdots$ $-v_n-$

- $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices
- $\Sigma \in \mathbb{R}^{m\times n}$ is a **rectangular diagonal matrix**.
  Examples:

  $r = \min(m, n)$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \\ 0 & 0 & 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \end{bmatrix}$$

Diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$, $k = \min(n, m)$ are called **singular values of** M.

Review: Singular Value Decomposition

parallel to eigenvector function

$Av = \lambda v$

A non-negative real number σ is a singular value for $M \in \mathbb{R}^{m \times n}$ if and only if there exist unit-length $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that

$$Mv = \sigma u$$
$$M^T u = \sigma v$$

u is called the **left singular vector** of σ, v is called the **right singular vector** of σ

Review: Singular Value Decomposition

A non-negative real number σ is a singular value for $M \in \mathbb{R}^{m\times n}$ if and only if there exist unit-length $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that

$$Mv = \sigma u$$
$$M^T u = \sigma v$$

u is called the **left singular vector** of σ, v is called the **right singular vector** of σ

$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}$

Connection to eigenvalue decomposition

symmetric

$\begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix}$

Given SVD of matrix $M \overset{I}{=} U\Sigma V^T$,

- $M^T M = (V\Sigma^T U^T)(U\Sigma V^T) = V(\Sigma^T\Sigma)V^T \leftarrow v_i$ *is an eigenvector of* $M^T M$ *with eigenvalue* σ_i^2
- $MM^T = (U\Sigma V^T)(V^T\Sigma^T U) = U(\Sigma\Sigma^T)U^T \leftarrow u_i$ *is an eigenvector of* MM^T *with eigenvalue* σ_i^2

$\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix}$

CCA Derivations

The original problem:

u v

$$(a_1, b_1) = \underset{a \in \mathbb{R}^{n_1}, b \in R^{n_2}}{\operatorname{argmax}} \; corr(a^T X, b^T Y) \qquad (1)$$

CCA Derivations

The original problem:

$$(a_1, b_1) = \underset{a \in \mathbb{R}^{n_1}, b \in R^{n_2}}{\operatorname{argmax}} \; corr(\underset{u}{a^T X}, \underset{v}{b^T Y}) \tag{1}$$

Assume $\mathbb{E}[x_1] = \cdots = \mathbb{E}[x_{n_1}] = \mathbb{E}[y_1] = \cdots = \mathbb{E}[y_{n_2}] = 0$,

$a^T \mathbb{E}[X Y^T] b.$

$$\begin{aligned} corr(a^T X, b^T X) &= \frac{\mathbb{E}[(a^T X)(b^T Y)]}{\sqrt{\mathbb{E}[(a^T X)^2]\mathbb{E}[(a^T Y)^2]}} \\ &= \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \end{aligned}$$

□

CCA Derivations

The original problem:

$$(a_1, b_1) = \underset{a \in \mathbb{R}^{n_1}, b \in R^{n_2}}{\operatorname{argmax}} \; corr(\underbrace{a}_{u}{}^T X, \underbrace{b}_{v}{}^T Y) \tag{1}$$

Assume $\mathbb{E}[x_1] = \cdots = \mathbb{E}[x_{n_1}] = \mathbb{E}[y_1] = \cdots = \mathbb{E}[y_{n_2}] = 0$,

$$\begin{aligned} corr(a^T X, b^T X) &= \frac{\mathbb{E}[(a^T X)(b^T Y)]}{\sqrt{\mathbb{E}[(a^T X)^2]\mathbb{E}[(a^T Y)^2]}} \\ &= \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \end{aligned}$$

(1) is equivalent to:

$$(a_1, b_1) = \underset{\substack{a \in \mathbb{R}^{n_1}, b \in R^{n_2} \\ a^T \Sigma_{XX} a = b^T \Sigma_{YY} b = 1}}{\operatorname{argmax}} \; a^T \Sigma_{XY} b \tag{2}$$

CCA Derivations

$$\max_{\substack{a\in\mathbb{R}^{n_1}\\ b\in\mathbb{R}^{n_2}}} a^T\Sigma_{XY}b. \qquad (1)$$

$$\text{st.}\quad a^T\Sigma_{XX}a=1,\quad b^T\Sigma_{YY}b=1.$$

$$a^T\Sigma_{XY}b = a^T\overbrace{\Sigma_{XX}^{\frac12}\Sigma_{XX}^{-\frac12}}^{I}\Sigma_{XY}\Sigma_{YY}^{-\frac12}\Sigma_{YY}^{\frac12}b,$$

$$= \left(\left(\Sigma_{XX}^{-\frac12}\right)^{\cancel{T}}a\right)^T\Sigma_{XX}^{-\frac12}\Sigma_{XY}\Sigma_{YY}^{-\frac12}\Sigma_{YY}^{\frac12}b.$$

$$= \underbrace{\left(\Sigma_{XX}^{\frac12}a\right)^T}_{c^T}\underbrace{\left(\Sigma_{XX}^{-\frac12}\Sigma_{XY}\Sigma_{YY}^{-\frac12}\right)}_{\Omega}\underbrace{\left(\Sigma_{YY}^{\frac12}b\right)}_{d.}$$

$$a^T\Sigma_{XX}a = a^T\Sigma_{XX}^{\frac12}\Sigma_{XX}^{\frac12}a = c^Tc = \|c\|^2=1$$

$$b^T\Sigma_{YY}b = d^Td=\|d\|^2=1$$

Therefore, (1) is equivalent to $\displaystyle\max_{\substack{c\in\mathbb{R}^{n_1}\\ d\in\mathbb{R}^{n_2}}} c^T\Omega d$

$$\text{st.}\quad \|c\|^2=1,\quad \|d\|^2=1$$

CCA Derivations

Define $\Omega \in R^{n_1 \times n_2}$, $c \in \mathbb{R}^{n_1}$ and $d \in \mathbb{R}^{n_2}$,

$$\Omega = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$$
$$c = \Sigma_{XX}^{\frac{1}{2}} a$$
$$d = \Sigma_{YY}^{\frac{1}{2}} b$$

(2) can be written as

given

$$(c_1, d_1) = \underset{\substack{c \in \mathbb{R}^{n_1}, d \in \mathbb{R}^{n_2} \\ ||c||^2 = ||d||^2 = 1}}{\operatorname{argmax}} c^T \Omega d \tag{3}$$

CCA Derivations

Define $\Omega \in R^{n_1 \times n_2}$, $c \in \mathbb{R}^{n_1}$ and $d \in \mathbb{R}^{n_2}$,

$$\Omega = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$$
$$c = \Sigma_{XX}^{\frac{1}{2}} a \Rightarrow a = \Sigma_{xx}^{-\frac{1}{2}} c$$
$$d = \Sigma_{YY}^{\frac{1}{2}} b$$

(2) can be written as

$$(c_1, d_1) = \underset{\substack{c \in \mathbb{R}^{n_1}, d \in \mathbb{R}^{n_2} \\ ||c||^2 = ||d||^2 = 1}}{\operatorname{argmax}} \quad c^T \Omega d \tag{3}$$

(c_1, d_1) can be solved by SVD, then the first pair of canonical variables are

$$a_1 = \Sigma_{XX}^{-\frac{1}{2}} c_1, \quad b_1 = \Sigma_{YY}^{-\frac{1}{2}} d_1$$

CCA Derivations

$$(c_1, d_1) = \underset{\substack{c \in \mathbb{R}^{n_1}, d \in \mathbb{R}^{n_2} \\ ||c||^2 = ||d||^2 = 1}}{\operatorname{argmax}} c^T \Omega d$$

Proposition 1

c_1 and d_1 are the left and right unit singular vectors of Ω with the largest singular value.

CCA Derivations

$$(c_1, d_1) = \underset{\substack{c \in \mathbb{R}^{n_1}, d \in \mathbb{R}^{n_2} \\ ||c||^2 = ||d||^2 = 1}}{\operatorname{argmax}} \; c^T \Omega d$$

Proposition 1

c_1 and d_1 are the left and right unit singular vectors of Ω with the largest singular value.

Theorem 2

c_i and d_i are the left and right unit singular vectors of Ω with the ith largest singular value.

$\Omega = U \Sigma V^T.$

CCA Algorithm

Input: Covariance matrices for centered data X and Y:

- Σ_{XY} , invertible Σ_{XX} and Σ_{YY}
- Dimension $k \leq \min(n_1, n_2)$

Output: CCA projection matrices A_k and B_k:

- Compute $\Omega = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$
- Compute SVD decomposition of Ω

$$\Omega = \overset{u}{\underbrace{\begin{bmatrix} | & \ldots & | \\ c_1 & \ldots & c_{n_1} \\ | & \ldots & | \end{bmatrix}}_{k.}} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & 0 & \end{bmatrix} \overset{V^T}{\begin{bmatrix} -d_1^T- \\ \vdots \\ -d_{n_2}^T- \end{bmatrix}} \Big\} k$$

- $A_k = \Sigma_{XX}^{-\frac{1}{2}} [c_1, \ldots, c_k]$ and $B_k = \Sigma_{YY}^{-\frac{1}{2}} [d_1, \ldots, d_k]$

Discussion of CCA

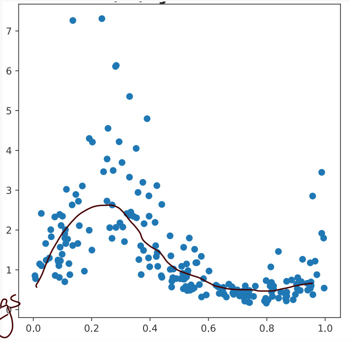
- CCA only measures linear dependencies
- Non-linear generalizations:
  - Kernel CCA (KCCA)
  - Deep CCA (DCCA)
  - Maximal HGR Correlation

non-parametric mappings

$x \to f(x)$

$y \to g(y)$

$\phi(x_2)$



$x_1 \to \phi(x_1)$

Non-linear dependency between x_1 and x_2

$$\begin{cases} \max\limits_{f,g} \mathbb{E}[f(x)^T g(y)] \\ \text{s.t. } \mathbb{E}[f]=\mathbb{E}[g]=0;\ \mathbb{E}[f^2]=\mathbb{E}[g^2]=1. \end{cases}$$

X, Y are discrete
→ Alternating Direction
Conditional expectation
(ACE).

PCA, ICA and CCA

Linear Subspace Learning

Given high dimensional random vector **x**, transform it to a low-dimensional vector **y** through a projection matrix U:

$$y = U^T x$$

PCA, ICA and CCA

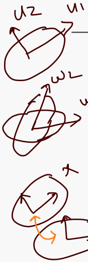
Linear Subspace Learning

$\left\{\begin{matrix} x \\ z. \end{matrix}\right.$

Given high dimensional random vector **x**, transform it to a low-dimensional vector **y** through a projection matrix U:

$$y = U^T x$$

- PCA, ICA and CCA are all unsupervised linear subspace learning methods.



Name	What is U ?	goal	subspace
PCA	principal component (U)	remove (low order) correlation	single
ICA	unmixing matrix (W)	remove (high order) correlation	single
CCA	canonical projection matrices (A, B)	maximize correlation between feature pairs	paired