Learning From Data
Lecture 8: ICA, CCA, & HGR Maximal Correlation

Yang Li yangli@sz.tsinghua.edu.cn

TBSI

11/14/2019

Today's Lecture

Unsupervised Learning (Part II)

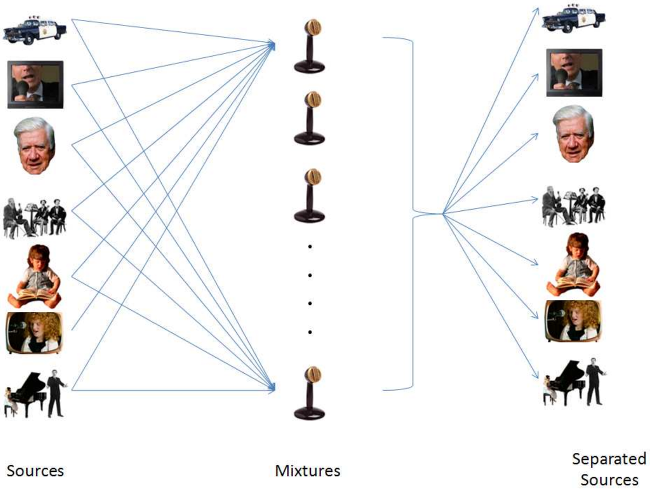
- Independent Component Analysis (ICA)
- Canonical Correlation Analysis (CCA)
- HGR Maximal Correlation

Written Assignment 3 is due next Saturday.
Programming Assignment 4 and project will be released next week.

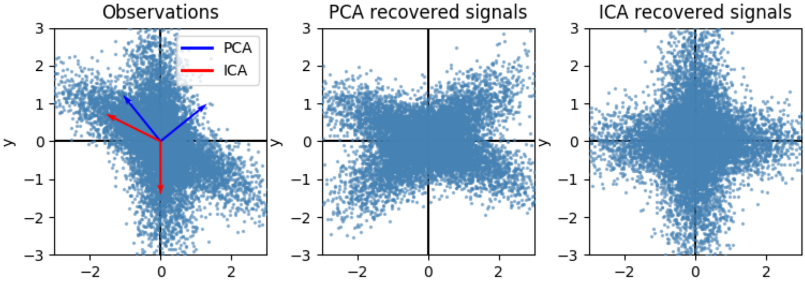
Independent Component Analysis

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?

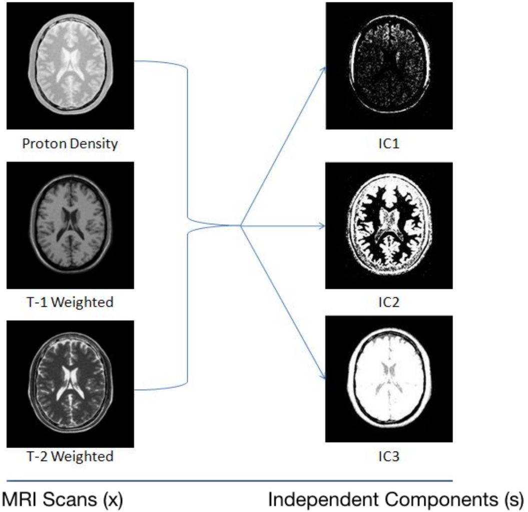


ICA vs PCA


Observations
PCA
ICA
PCA recovered signals
ICA recovered signals
y
3
2
1
0
−1
−2
−3
−2
0
2

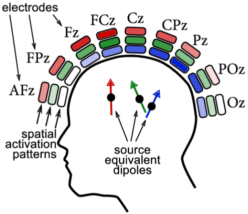
Brian imaging

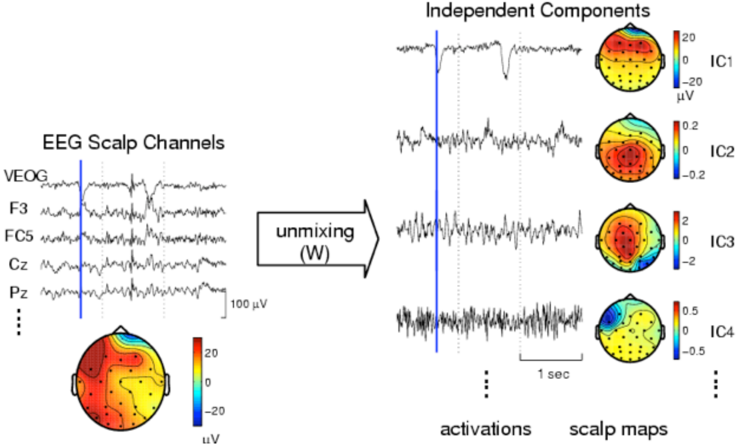
- Different brain matters: gray matter, white matter, cerebrospinal fluid (CSF), fat, muscle/skin, glial matter etc.
- An MRI scan is a mixture of different brain matters



EEG Analysis

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





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**

$$x = As$$

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 $(x^{(i)} = As^{(i)})$

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

$$W = \begin{bmatrix} -w_1^T- \\ \vdots \\ -w_n^T- \end{bmatrix}$$

ICA Ambiguities

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

- Permutation of original sources $s_1, \ldots, s_n$
- Scaling of w_i

Why is Gaussian data problematic?

As long as the data is non-Gaussian, given enough data, we can recover the n independent sources.

Densities and Linear Transformations

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

$$p_x(x) = p_s(Wx)|W|,$$

where $W = A^{-1}$

ICA Algorithm

Joint distributions of *independent* sources $s = \{s_1, \ldots, s_n\}$:

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

The density on $x = As = W^{-1}s$:

$$p(x) = \prod_{i=1}^{n} p_s(w_i^T x)|W|$$

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

Given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$, the log likelihood is

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

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

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

Check this at home!

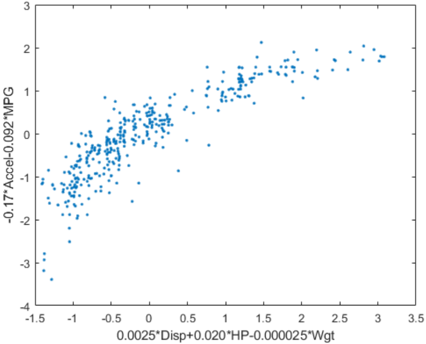
Canonical Correlation Analysis

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



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 $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_2} \end{bmatrix}$
- 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

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

- $U = a^T X$ 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

Review: Singular Value Decomposition

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

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

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

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

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

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 value** of σ, v is called the **right singular value** of σ

Connection to eigenvalue decomposition

Given SVD of matrix $M = U\Sigma V^T$,

- $M^T M = (V\Sigma^T U^T)(U\Sigma V^T) = V(\Sigma^T\Sigma)V^T \leftarrow$ *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*

CCA Derivations

The original problem:

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

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

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

(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

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

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

(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

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

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.

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:

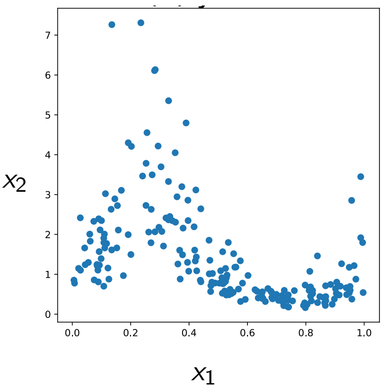
- Compute $\Omega = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$
- Compute SVD decomposition of Ω

$$\Omega = \begin{bmatrix} | & \dots & | \\ c_1 & \dots & c_{n_1} \\ | & \dots & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & 0 & \end{bmatrix} \begin{bmatrix} -d_1^T- \\ \vdots \\ -d_{n_2}^T- \end{bmatrix}$$

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

Discussion of CCA

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



Non-linear dependency between x_1 and x_2

Maximal HGR Correlation Analysis

A Non-linear Measure of Dependence

Hirschfeld-Gebelein-Renyi (HGR) maximal correlation

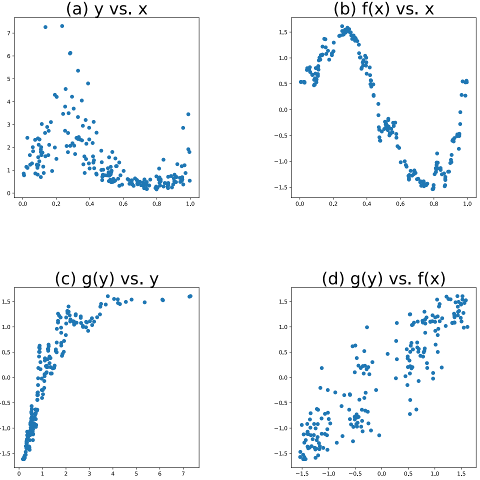
Given random variables X, Y, the HGR maximal correlation is

$$\rho(X;Y) = \max_{f(X),g(Y)} \mathbb{E}[f(X)g(Y)]$$
$$s.t. \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$$
$$\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$$

where $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ are real-valued functions

Example of HGR maximal correlation

Synthesized data: $y^{(i)} = \exp\left(\sin\left(2\pi x^{(i)} + \frac{\epsilon^{(i)}}{2}\right)\right)$, $e^{(i)} \approx \mathcal{N}(0,1)$ for $i = 1, \ldots, 200$



$$\rho(X;Y) = 0.902$$

Example of HGR maximal correlation

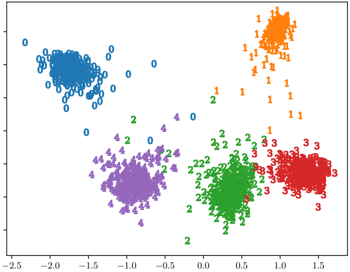
Use multi-dimensional HGR maximal correlation to learn unsupervised features from MNIST.

"$\mathbf{y^{(i)}} = \mathbf{x^{(i)}} + \mathbf{4}$"

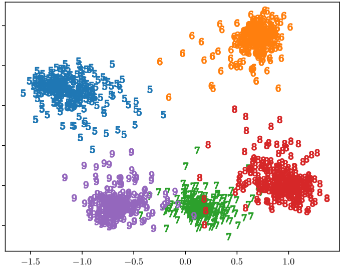
Example of HGR maximal correlation

Use multi-dimensional HGR maximal correlation to learn unsupervised features from MNIST.

$f_1(x)$ vs $f_2(x)$



$g_1(y)$ vs $g_2(y)$



How to solve it?

Assume X and Y are both discrete with alphabet $\mathcal{X}$, $\mathcal{Y}$.

$$\mathbb{E}[f(x)g(y)] = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{X,Y}(x,y)f(x)g(y)$$

Define $\phi(x) \triangleq \sqrt{P_X(x)}f(x)$, $\psi(y) \triangleq \sqrt{P_Y(y)}g(y)$, then

$$\mathbb{E}[f(x)g(y)] = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}\phi(x)\psi(y) = \psi^T B\phi$$

- Matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, where $B(y,x) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}$
- Vectors $\phi \in \mathbb{R}^{|\mathcal{X}|}, \psi \in \mathbb{R}^{|\mathcal{Y}|}$

How to represent the constraints using ϕ and ψ?

How to solve it?

Given $\phi(x) = \sqrt{P_X(x)} f(x)$, $\psi(y) = \sqrt{P_Y(y)} g(y)$

Unit-variance constraints

- $\mathbb{E}[f(x)^2] = 1 \implies$
  $\sum_x P_X(x) \left(\frac{\phi(x)}{\sqrt{P_X(x)}} \right)^2 = \sum_x \phi(x)^2 = ||\phi||^2 = 1$
- Similarly, $\mathbb{E}[g(y)^2] = 1 \implies ||\psi||^2 = 1$

Zero-mean constraints

- $\mathbb{E}[f(x)] = 0 \implies$
  $\sum_x P_X(x) \frac{\phi(x)}{\sqrt{P_X(x)}} = \sum_x \phi(x) \sqrt{P_X(x)} = \langle \phi, \sqrt{P_X} \rangle = 0$, i.e. $(\phi \perp \sqrt{P_X})$
- Similarly, $\mathbb{E}[g(y)] = 0 \implies \langle \psi, \sqrt{P_Y} \rangle = 0$, i.e. $(\psi \perp \sqrt{P_Y})$

HGR Maximal Correlation as an SVD problem

Alternative definition for HGR Maximal Correlation

$$\rho(X, Y) = \max_{\phi \in \mathbb{R}^{|\mathcal{X}|}, \psi \in \mathbb{R}^{|\mathcal{Y}|}} \psi^T B \phi$$
$$s.t. ||\phi||^2 = ||\psi||^2 = 1$$
$$\phi \perp \sqrt{P_X}, \psi \perp \sqrt{P_Y}$$

Proposition 2

$(u_1, v_1) = \text{argmax}_{||u||=||v||=1} u^T B v$ *are the largest left and right singular vector of* B.

Proposition 3

The largest left and right singular vectors are $\sqrt{P_Y}$ *and* $\sqrt{P_X}$

Proposition 4

ψ^* *and* ϕ^* *are the 2nd largest left and right singular vectors of* B*, respectively.*

Alternating Condition Expectation (ACE)

A generalization of power iteration for finding singular vectors:

ACE algorithm for 1d data [Breiman & Friedman 1985]

Data: Discrete data samples $x^{(1)}, \ldots, x^{(m)}$
Result: compute $f^*(x), g^*(y)$
Randomly choose $g(y), y \in \mathcal{Y}$ such that $\mathbb{E}[g(Y)] = 0$;
while *σ not converged* **do**
    $f(x) \leftarrow \mathbb{E}_m[g(Y)|X = x]$
    Normalize $f(x)$ $\forall x \in \mathcal{X}$;
    $g(y) \leftarrow \mathbb{E}_m[f(X)|Y = y]$;
    Normalize $g(y)$ $\forall y \in \mathcal{Y}$;
    $\sigma \leftarrow \mathbb{E}_m[f(X)g(Y)]$;
end

Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc., 80(391),1985b

Alternating Condition Expectation (ACE)

A generalization of power iteration for finding singular vectors:

ACE algorithm for 1d data [Breiman & Friedman 1985]

Data: Discrete data samples $x^{(1)}, \ldots, x^{(m)}$
Result: compute $f^*(x), g^*(y)$
Randomly choose $g(y), y \in \mathcal{Y}$ such that $\mathbb{E}[g(Y)] = 0$;
while σ *not converged* **do**
    $f(x) \leftarrow \mathbb{E}_m[g(Y)|X = x]$ // $\mathbb{E}_m[\cdot]$: sample expectation ;
    Normalize $f(x)$ $\forall x \in \mathcal{X}$;
    $g(y) \leftarrow \mathbb{E}_m[f(X)|Y = y]$;
    Normalize $g(y)$ $\forall y \in \mathcal{Y}$;
    $\sigma \leftarrow \mathbb{E}_m[f(X)g(Y)]$;
end

Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc., 80(391),1985b

Extension to high dimension case

k-dimensional HGR Maximal Correlation

$$\rho(X;Y) = \max_{\substack{f:\mathcal{X}\to\mathbb{R}^k \\ g:\mathcal{Y}\to\mathbb{R}^k}} \mathbb{E}[f(X)^T g(Y)] \leftarrow \text{ optimize k values in parallel}$$

$$s.t.\ \mathbb{E}[f_i(X)] = \mathbb{E}[g_i(Y)] = 0,\ \forall i = 1,\dots,k$$

$$\mathbb{E}[f_i(X)^T f_j(X)] = \mathbb{E}[g_i(Y)^T g_j(Y)] = \mathbf{1}\{i=j\},\ \forall i,j = 1,\dots,k$$

ACE algorithm for k-d data

Data: Discrete data samples $x^{(1)},\dots,x^{(m)}$
Result: compute $f^*(x), g^*(y)$
Randomly choose $g(y), y \in \mathcal{Y}$ such that $\mathbb{E}[g(Y)] = 0$;
while *σ not converged* **do**
  $f(x) \leftarrow \mathbb{E}_m[g(Y)|X=x]$;
  Normalize $f(x)$ $\forall x \in \mathcal{X}$;
  $g(y) \leftarrow \mathbb{E}_m[f(X)|Y=y]$;
  Normalize $g(y)$ $\forall y \in \mathcal{Y}$;
  $\sigma \leftarrow \mathbb{E}_m[f(X)^T g(Y)]$;
end

Normalize k-d feature: for all $x \in \mathcal{X}$,

- $f(x) \leftarrow f(x) - \mathbb{E}_m[f(X)]$
- $f(x) \leftarrow f(x)\mathbb{E}_m[f(X)f(X)^T]^{-\frac{1}{2}}$

$g(y)$ is normalized similarly for all $y \in \mathcal{Y}$.

Discussion on HGR Maximal Correlation

- Useful for modal estimation from data
- ACE in Python: `https://github.com/mace-cream/xyace` (limited to discrete X and Y)
- Extension to continuous case: a deep neural network implementation of HGR maximal correlation [Wang et. al. 2018]

An Efficient Approach to Informative Feature Extraction from Multimodal Data, Wang, Lichen, et al. AAAI (2018).