The Canonical Correlatoin Analysis Family

Xunyi Jiang, Langtian Ma

2023 Summer Seminar on SSL - Week 7

malt2020@mail.sustech.edu.cn

August 3, 2023

Generalized CCA Framework

For random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, we seek two transformations $f : \mathbb{R}^p \to \mathbb{R}^d$ and $g : \mathbb{R}^q \to \mathbb{R}^d$:

$$\max_{f,g} \; \mathbb{E}[f(X)^T g(Y)]$$
$$\text{Subject to } \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = \mathbf{0},$$
$$Cov[f(X)] = Cov[g(Y)] = \mathbf{I}$$

with $d \leq \min\{p, q\}$.

> Traditional CCA: f and g are linear.

Traditional Nonlinear CCA

> EstimatinOgptimalTransformatiofnosrMultiple Regressionand Correlation
> LEO BREIMANand JEROMEH Friedman

Univariate Setting:

$$Y \in \mathbb{R}, f(X) = (f_1(X_1), \ldots, f_p(X_p))$$

Alternating Conditional Expectations: nonlinear least square with objective function:

$$\mathcal{L}(f, g) = \frac{\mathbb{E}[g(Y) - \mathbf{1}^T f(X)]^2}{\mathbb{E}g^2(Y)}$$

Traditional Nonlinear CCA

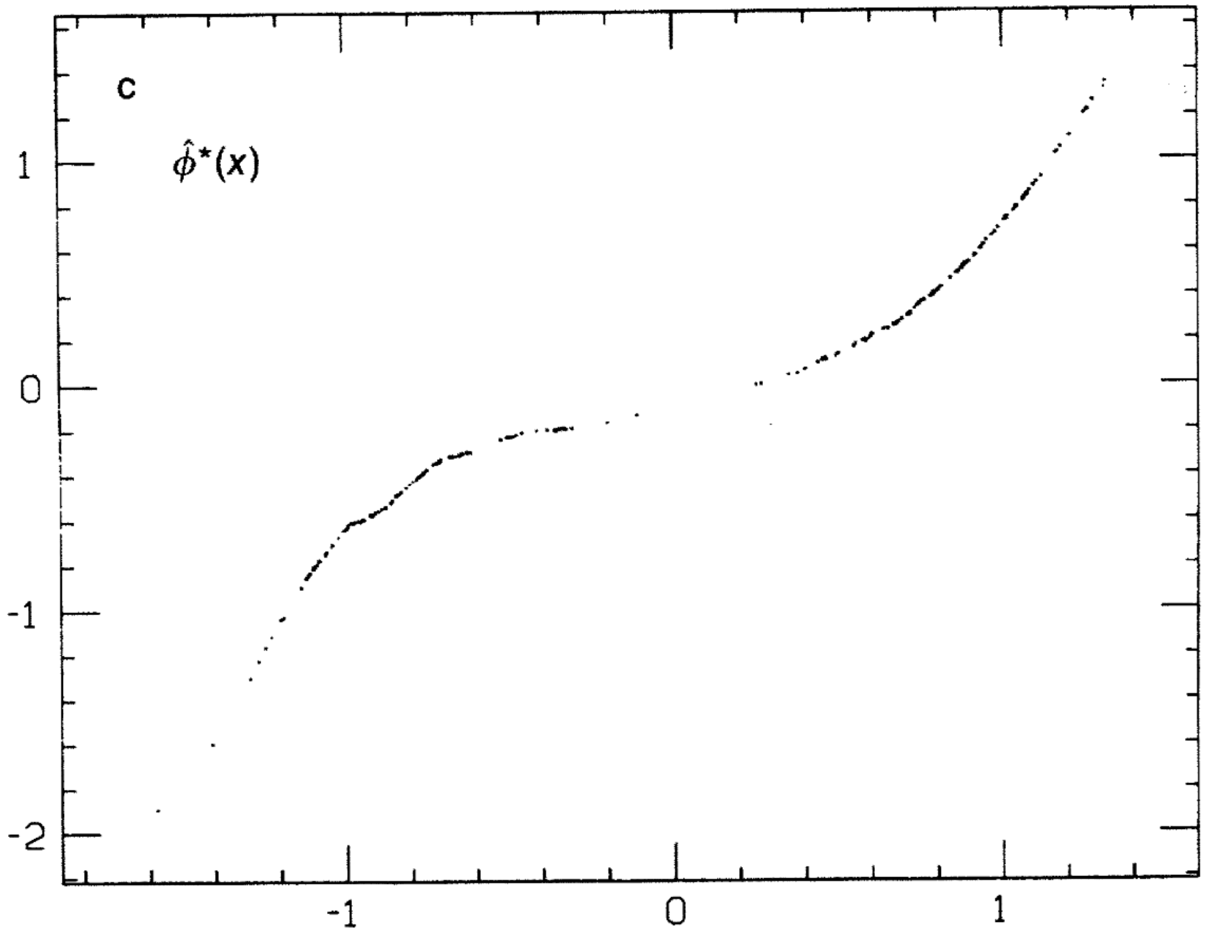
> For any random variable X and Y, the best predictor for Y given X is $\mathbb{E}[Y|X]$

Basic Algorithm (For illustration):

- Set $g(Y) = Y/\|Y\|$
- Iterate until $\mathcal{L}(f, g)$ fails to decrease;
  - $f(X) = \mathbb{E}[g(Y)|X]$
  - $g(Y) = \mathbb{E}[f(X)|Y]$
- End Iteration Loop

Remark: Smoothing is applied repeatedly throughout the algorithm.

Traditional Nonlinear CCA



Problem Formulation

> Nonlinear Canonical Correlation Analysis:A Compressed Representation Approach
> 2020; Amichai Painsky, Meir Feder, Naftali Tishby

Additional mutual information constraints:

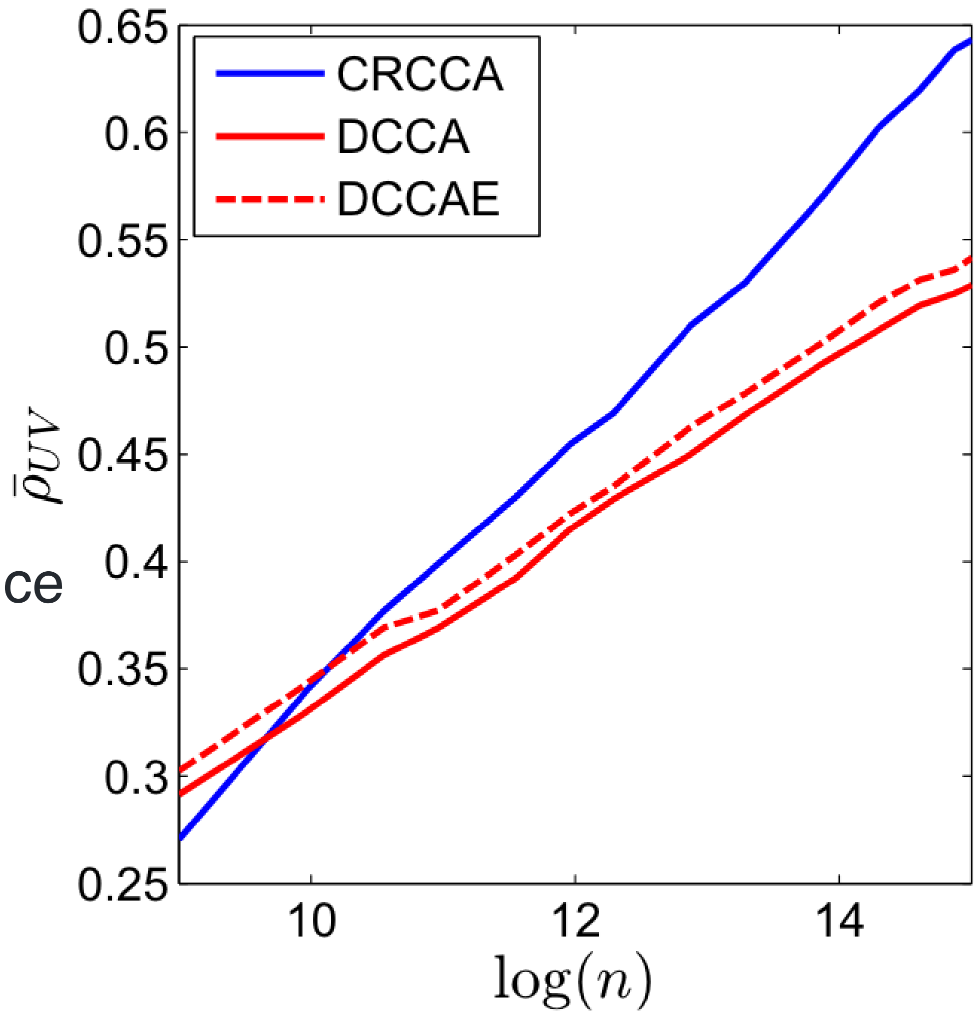
$$I(X, f(X)) \leq R_X, \quad I(Y, g(Y)) \leq R_Y$$

- f and g are not required to be deterministic.
- $f(X)$ and $g(Y)$ are also restricted to be compressed representations of X and Y.
- R_X and ,R_Y define the amount of information preserved from the original vectors.

Mutual information constraint controls the generalization gap, and it can be viewed as a soft dimensionality reduction: restrict the level of information allowed to represent the data.

Information-theoretic Compressed Representation

Comparation of generalization performance



Kernel CCA

Kernel functions $\kappa(\cdot,\cdot)$ can be expressed as an inner product in a representation space:

$$k(x,y) = \langle \phi(x), \phi(y) \rangle,$$

Kernel CCA is equivalent to conducting linear CCA on the representation space.

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the data matrices, $\kappa(\cdot,\cdot)$ be aspecified kernel function

- Sample version of Covariance matrix $\hat{Cov}(X) = \mathbf{X}^T\mathbf{X}, \hat{Cov}(Y) = \mathbf{Y}^T\mathbf{Y}$
- Let $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ be the kernel Gram matirces defined as $(K_x)_{ij} = \kappa(x_i, x_j)$ and $(K_y)_{ij} = \kappa(y_i, y_j)$

Find vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ such that

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^m}{\operatorname{argmax}} \boldsymbol{\alpha}'\mathbf{K}_{\mathcal{X}}\mathbf{K}_{\mathcal{Y}}\boldsymbol{\beta}$$

$$\text{subject to } \boldsymbol{\alpha}'\mathbf{K}_{\mathcal{X}}\mathbf{K}_{\mathcal{X}}\boldsymbol{\alpha} = \boldsymbol{\beta}'\mathbf{K}_{\mathcal{Y}}\mathbf{K}_{\mathcal{Y}}\boldsymbol{\beta} = 1$$

Deep CCA

> Deep Canonical Correlation Analysis
>
> 2013 Galen Andrew, Raman Arora, Jeff Bilmes, Karen Livescu

Idea: Let f and g be neural networks.

- Initialize the parameters of each layer with a denoising autoencoder
  - Input data: $\mathbf{X} \in \mathbb{R}^{n \times m}$,
  - Adding i.i.d zero-mean Guassian noise to obtain distorted matrix $\tilde{\mathbf{X}}$
  - Learn denoising auto encoder by minimizing reconstruction loss
- Updating parameters by maximizing correlation:

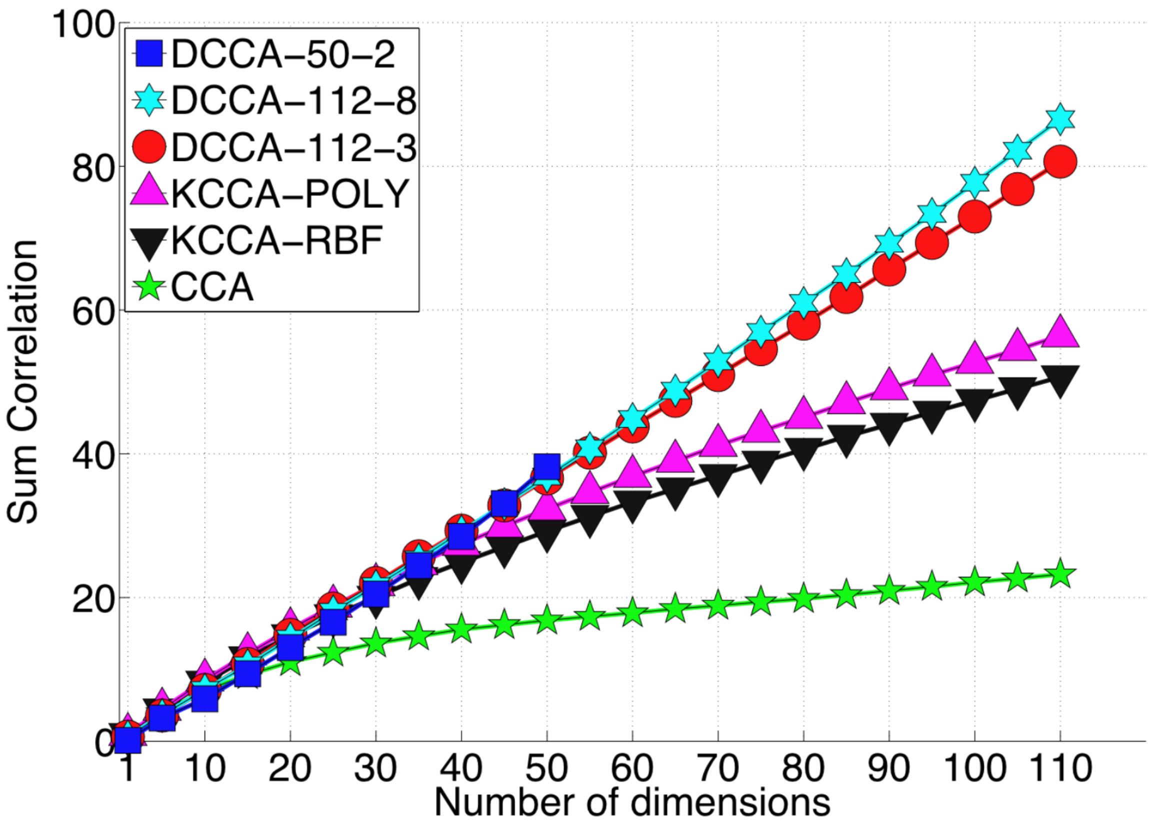
$$\max_{\theta_1,\theta_2} \operatorname{corr}(f(X;\theta_1), g(Y;\theta_2))$$

MNIST handwritten image

Each image is splited along the central axis to form two views.

	CCA	KCCA (RBF)	DCCA (50-2)
Dev	28.1	33.5	**39.4**
Test	28.0	33.0	**39.7**

Wisconsin X-ray Microbeam Database



Deep canonically correlated autoencoders

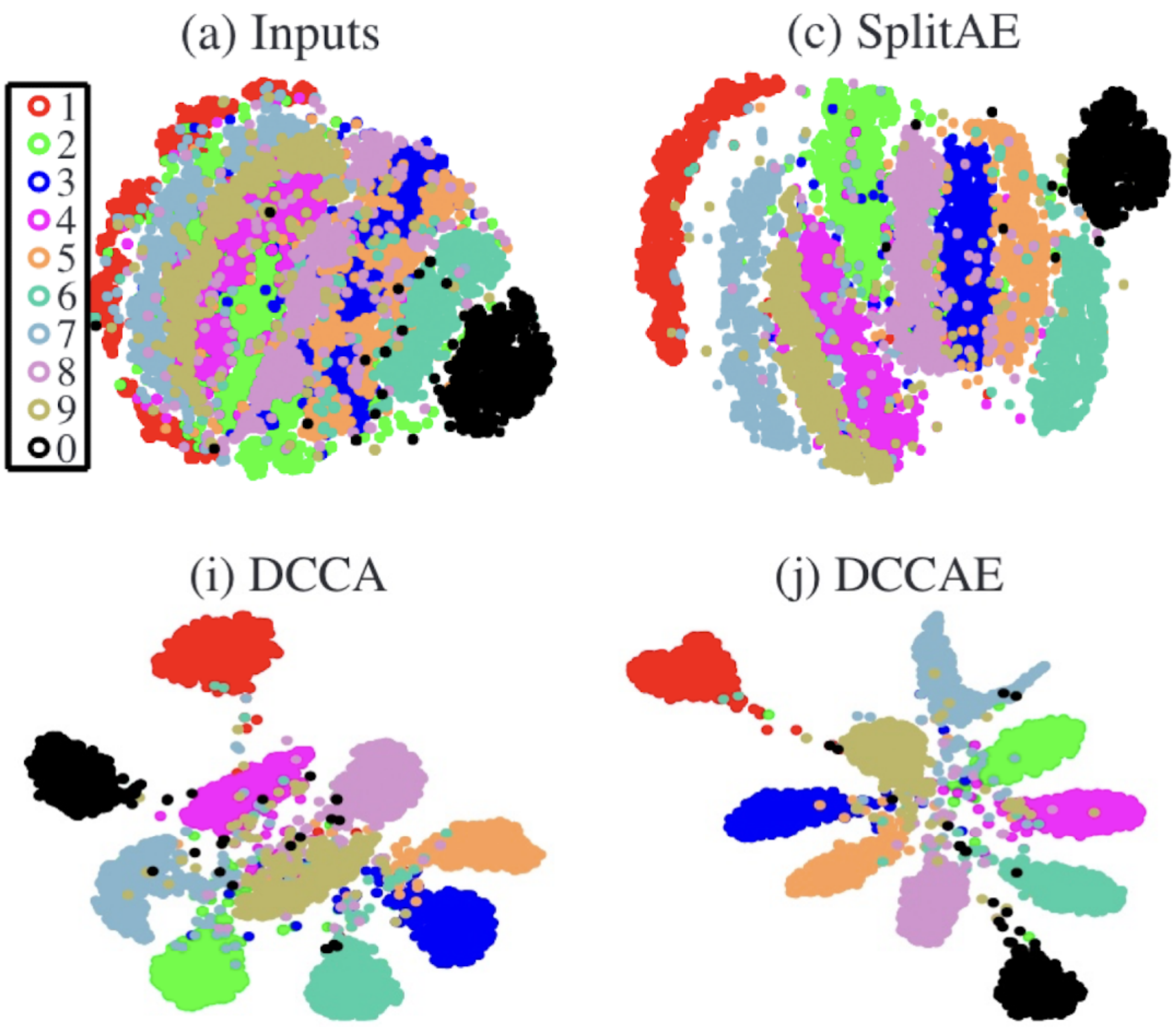
> On Deep Multi-View Representation Learning
>
> Weiran Wang, Raman Arora, Karen Livescu, Jeff Bilmes

Two autoencoders, optimizethe combination of canonical correlation and the reconstruction errors. For illustration, we write:

$$\min -\mathrm{Corr}(f(X), g(Y)) + \frac{\lambda}{N}\sum_{i=1}^{N}(\|x_i - p(f(x_i))\|^2 + \|y_i - q(g(y_i))\|^2)$$

where p and q are decoders for X and Y, respectively.

- CCA: maximizes the mutual information between the transformed views.
- Reconstruction error: maximizes the mutual information between inputs and learned features.


(a) Inputs
(c) SplitAE
(i) DCCA
(j) DCCAE
1
2
3
4
5
6
7
8
9
0

Summary

- Linear CCA: Linear transformation.
- Nonlinear CCA: Conditional Expectation & Smoothing.
- Information Compressed CCA: constrain the level of information allowed to represent the data.
- Kernel CCA: use kernel function to seek for nonlinear representation
- Deep CCA: Use correlation as objective functions
- DCCAE: Combination of Deep CCA and autoencoders.