Improved Estimation of Canonical Vectors in Canonical Correlation Analysis

Nicholas Asendorf, Ph.D.
asendorf@umich.edu

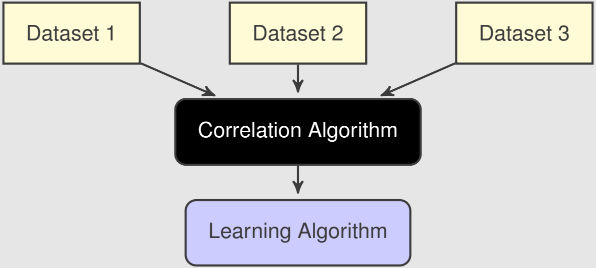
Prof. Raj Nadakuditi
rajnrao@umich.edu

Department of Electrical Engineering and Computer Science
University of Michigan

Asilomar Conference on Signals, Systems, and Computers

November 11, 2015

Motivation



Goal

Develop theoretically justified, robust correlation algorithms for multi-dataset fusion

Multiple Datasets

- Audio-Video
- Audio-Audio

Machine Learning

- emotion identification
- shopping predictions
- music genre classification

Medical Signal Processing

- MRI, fMRi, EEG, MEG, etc.

Multiple Datasets

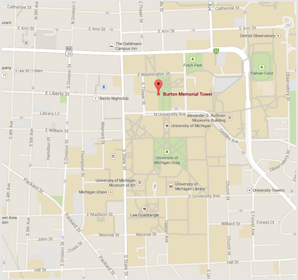
- Audio-Video
- Audio-Audio

Machine Learning

- emotion identification
- shopping predictions
- music genre classification

Medical Signal Processing

- MRI, fMRi, EEG, MEG, etc.



A Myriad of Applications

Multiple Datasets

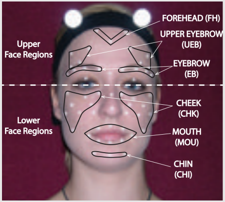
- Audio-Video
- Audio-Audio

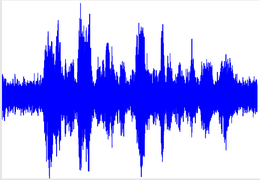
Machine Learning

- emotion identification
- shopping predictions
- music genre classification

Medical Signal Processing

- MRI, fMRi, EEG, MEG, etc.





Multiple Datasets

- Audio-Video
- Audio-Audio

Machine Learning

- emotion identification
- shopping predictions
- music genre classification

Medical Signal Processing

- MRI, fMRi, EEG, MEG, etc.

Multiple Datasets

- Audio-Video
- Audio-Audio

Machine Learning

- emotion identification
- shopping predictions
- music genre classification

Medical Signal Processing

- MRI, fMRi, EEG, MEG, etc.

- disco influences
- danceable grooves
- repetitive melodic phrasing
- extensive vamping
- minor key tonality

Canonical Correlation Analysis

What is it?

- Dimensionality reduction algorithm for exactly 2 datasets, X, Y
- Correlation coefficients, linear transformations

What is it not?

- Data fusion algorithm

Covariance matrices

- $R_{xx} = \mathbb{E}\left[xx^H\right]$
- $R_{yy} = \mathbb{E}\left[yy^H\right]$
- $R_{xy} = \mathbb{E}\left[xy^H\right]$

Optimization problem

$$\begin{aligned} \underset{w_x, w_y}{\operatorname{argmax}} \quad & \rho = w_x^H R_{xy} w_y \\ \text{subject to} \quad & w_x^H R_{xx} w_x = 1 \\ & w_y^H R_{yy} w_y = 1 \end{aligned}$$

Variable Transformation

- $f = R_{xx}^{1/2} w_x$
- $g = R_{yy}^{1/2} w_y$

What is it?

- Dimensionality reduction algorithm for exactly 2 datasets
- Correlation coefficients, linear transformations

What is it not?

- Data fusion algorithm

Optimization problem

$$\underset{f,g}{\operatorname{argmax}} \quad \rho = f^H \underbrace{R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2}}_{C_{\text{cca}}} g$$

$$\text{subject to} \quad \|f\|_2 = 1\,,\ \|g\|_2 = 1$$

Canonical Vectors

- $w_x = R_{xx}^{-1/2} f$
- $w_y = R_{yy}^{-1/2} g$

Insight

$\#$ correlated signals $= k = \text{rank}(C_{\text{cca}})$

Empirical CCA

Training Datasets

- $X = [x_1, \ldots, x_n]$
- $Y = [y_1, \ldots, y_n]$

Sample Covariance Matrices

- $\widehat{R}_{xx} = \frac{1}{n} XX^H$
- $\widehat{R}_{yy} = \frac{1}{n} YY^H$
- $\widehat{R}_{xy} = \frac{1}{n} XY^H$

Estimate

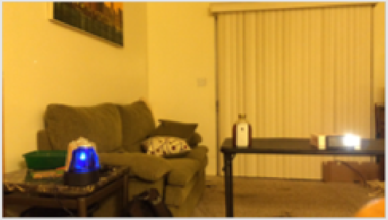
$$\boxed{\begin{aligned} \widehat{C}_{\text{cca}} &= \widehat{R}_{xx}^{-1/2} \widehat{R}_{xy} \widehat{R}_{yy}^{-1/2} \\ &= \widehat{F} \widehat{K} \widehat{G}^H \end{aligned}}$$

Data SVDs

- $X = \widehat{U}_x \widehat{\Sigma}_x \widehat{V}_x^H$
- $Y = \widehat{U}_y \widehat{\Sigma}_y \widehat{V}_y^H$
- $\sigma\left(\widehat{C}_{\text{cca}}\right) = \sigma\left(\widehat{V}_x^H \widehat{V}_y\right)$

Motivational Example - Flashing Light Video

Left Camera



Right Camera

System parameters

- Vectorize 135×240 image $\Rightarrow p = q = 32400$ pixels
- 30 fps @ 30 seconds $\Rightarrow n = 900$ frames

Goal

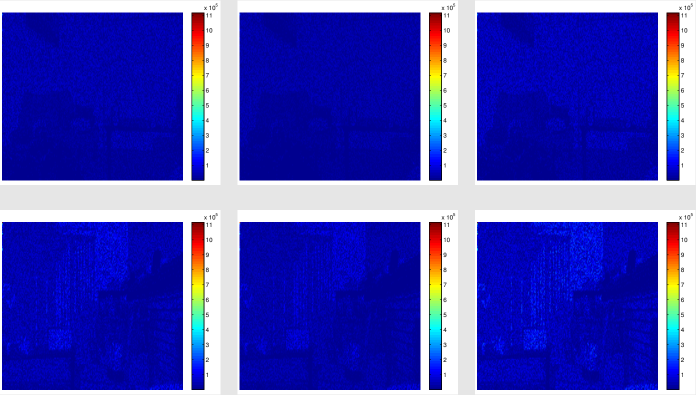
Identify correlated pixels between camera views

- After 900 frames = 30 seconds of video

First **Second** **Third**

Left

Right



Two-Dataset Model

Linear Subspace Model

$$x_i = U_x s_{x,i} + z_{x,i}$$
$$y_i = U_y s_{y,i} + z_{y,i}$$

Parameters

- $U_x^H U_x = I_{k_x}$, $U_y^H U_y = I_{k_y}$
- $z_{x,i} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_p)$, $\;z_{y,i} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_q)$
- $\mathbb{E}\left[\begin{bmatrix} s_{x,i} \\ s_{y,i} \end{bmatrix}\begin{bmatrix} s_{x,i}^H & s_{y,i}^H \end{bmatrix}\right] = \begin{bmatrix} \Theta_x & K_{xy} \\ K_{xy}^H & \Theta_y \end{bmatrix}$
- $K_{xy} = \Theta_x^{1/2} P_{xy} \Theta_y^{1/2}$
- $\Theta_x = \textbf{diag}\left(\left(\theta_1^{(x)}\right)^2, \ldots, \left(\theta_{k_x}^{(x)}\right)^2\right)$, $\;\Theta_y = \textbf{diag}\left(\left(\theta_1^{(y)}\right)^2, \ldots, \left(\theta_{k_y}^{(y)}\right)^2\right)$
- P_{xy} contains correlations ρ_{kj} between signals of x_i and y_i
- $\widetilde{K}_{xy} = \left(\Theta_x + I_{k_x}\right)^{-1/2} K_{xy} \left(\Theta_y + I_{k_y}\right)^{-1/2}$, with singular values $\kappa_1, \ldots, \kappa_{\min(k_x, k_y)}$

Informative CCA (ICCA)

Not all singular vectors are informative! (Nadakuditi, 2011)

- $\sigma_i\left(\widehat{C}_{\text{cca}}\right) = \sigma_i\left(\widehat{V}_x^H \widehat{V}_y\right)$
- Trim data SVD's to only use informative components

1. Trim data SVD's: $X = \widehat{U}_x \widehat{\Sigma}_x \widehat{V}_x^H$ and $Y = \widehat{U}_y \widehat{\Sigma}_y \widehat{V}_y^H$
   - $\mathring{U}_x = \widehat{U}_x\left(:, 1 : \widehat{k}_x\right)$, $\mathring{U}_y = \widehat{U}_y\left(:, 1 : \widehat{k}_y\right)$
   - $\mathring{V}_x = \widehat{V}_x\left(:, 1 : \widehat{k}_x\right)$, $\mathring{V}_y = \widehat{V}_y\left(:, 1 : \widehat{k}_y\right)$
2. Form $\widehat{C}_{\text{icca}} = \mathring{U}_x \mathring{V}_x^H \mathring{V}_y \mathring{U}_y$
3. Take SVD: $\widehat{C}_{\text{icca}} = \widetilde{F}\widetilde{K}\widetilde{G}^H$
4. $\widehat{\rho}_{\text{icca}}^{(i)} = \widetilde{k}_i$
5. $\widetilde{w}_x^{(i)} = \widehat{R}_{xx}^{-1/2} \widetilde{f}_i$
6. $\widetilde{w}_y^{(i)} = \widehat{R}_{yy}^{-1/2} \widetilde{g}_i$

New Algorithm: ICCA+

Motivation

- We expect ICCA to outperform CCA
- However, we expect ICCA to be suboptimal because we substitute $\widehat{U}_x$ and $\widehat{U}_y$ without considering accuracy

Proposed Estimate

$$\widehat{w}_{x,i}^{\text{icca+}} = \mathring{U}_x \,\textbf{diag}\left(\lambda_{x,i}^{\text{opt}}\right) \widehat{U}_{\widetilde{K}}\,(:,i)$$

$$\widehat{w}_{y,i}^{\text{icca+}} = \mathring{U}_y \,\textbf{diag}\left(\lambda_{y,i}^{\text{opt}}\right) \widehat{V}_{\widetilde{K}}\,(:,i)\,,$$

Optimization Problem

$$\lambda_{x,i}^{\text{opt}} = \underset{\lambda_x}{\operatorname{argmin}} \left\| w_x^{(i)} - \widehat{U}_x \,\textbf{diag}(\lambda_x) \widehat{U}_{\widetilde{K}}\,(:,i) \right\|_F$$

$$\lambda_{y,i}^{\text{opt}} = \underset{\lambda_y}{\operatorname{argmin}} \left\| w_y^{(i)} - \widehat{U}_y \,\textbf{diag}(\lambda_x) \widehat{V}_{\widetilde{K}}\,(:,i) \right\|_F .$$

Closed Form Soluation

Proposition

The solutions to the previous optimization problem is given by

$$\lambda_{x,i}^{opt} = \textbf{diag}\left(\mathring{U}_x^H U_x \left(\Theta_x + I_{k_x}\right)^{-1/2}\right)$$

$$\lambda_{y,i}^{opt} = \textbf{diag}\left(\mathring{U}_y^H U_y \left(\Theta_y + I_{k_y}\right)^{-1/2}\right).$$

Results from random matrix theory: Eigenvector accuracy

$$\boxed{|\langle \widehat{u}_i, u_i \rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \dfrac{-2\varphi_{\mu_Z}(\rho)}{\theta_i^2 D'_{\mu_Z}(\rho)} & \theta_i^2 > 1/D_{\mu_Z}\left(b^+\right) \\ 0 & \text{otherwise} \end{cases}}$$

$$D_\mu(z) =: \left[\int \frac{z}{z^2 - t^2} d_\mu(t)\right] \times \left[c_x \int \frac{z}{z^2 - t^2} d_{\mu(t)} + \frac{1 - c_x}{z}\right]$$

$$\varphi_\mu(z) =: \int \frac{z}{z^2 - t^2} d\mu(t)$$

Main Theorem

Theorem

The solution to the optimal ICCA+weights exhibits the following behavior in the asymptotic regime where $p, q, n \to \infty$ with $p/n \to c_x$ and $q/n \to c_y$.
For $i = 1, \ldots, k_x$,

$$\lambda_{x,opt}^{(i)} \xrightarrow{a.s.} \begin{cases} D_{\mu_{Z_x}}\left(\sigma_x^{(i)}\right)\sqrt{\dfrac{-2\varphi_{\mu_{Z_x}}\left(\sigma_x^{(i)}\right)}{D'_{\mu_{Z_x}}\left(\sigma_x^{(i)}\right)\left(1+D_{\mu_{Z_x}}\left(\sigma_x^{(i)}\right)\right)}} & \text{if } \left(\theta_i^{(x)}\right)^2 > 1/D_{\mu_{Z_x}}(b_x^+) \\ 0 & \text{otherwise} \end{cases}$$

where $\sigma_x^{(i)} = D_{\mu_{Z_x}}^{-1}\left(1/\left(\theta_i^{(x)}\right)^2\right)$ and the D-transform. A similar expression for $\lambda_{y,opt}$ exists by replacing subscripts of x with y.

Estimation using training data

- Using data X, Y, we can estimate $\widehat{D}$, $\widehat{D}'$, $\widehat{\varphi}_\mu$, and $\widehat{\varphi}_{\widetilde{\mu}}$

Canonical Vector Estimation

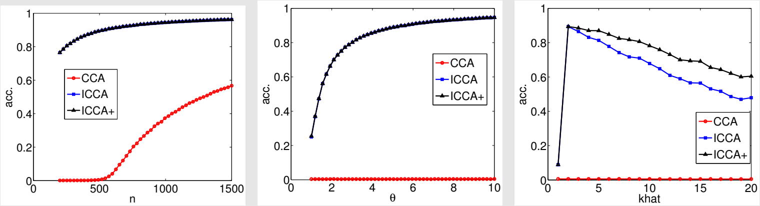
Data SVDs

- $X = \widehat{U}_x \widehat{\Sigma}_x \widehat{V}_x^H$ with trimmed versions $\mathring{U}_x$, $\mathring{\Sigma}_x$, $\mathring{V}_x$
- $U_{\widetilde{K}}$ left singular vectors of $\widetilde{K}_{xy}$
- $\widehat{U}_{\widetilde{K}}$ left singular vectors of $\widehat{C}_{\text{cca}}$
- $\mathring{U}_{\widetilde{K}}$ left singular vectors of $\widehat{C}_{\text{icca}}$

Estimate	
Population	$W_x = U_x \left(\Theta_x + I_{k_x}\right)^{-1/2} U_{\widetilde{K}}$
Empirical CCA	$\widehat{W}_x^{\text{cca}} = \widehat{U}_x \left(\widehat{\Sigma}_x\right)^{-1} \widehat{U}_{\widetilde{K}}$
ICCA	$\widehat{W}_x^{\text{icca}} = \mathring{U}_x \mathring{\Sigma}_x^{-1} \mathring{U}_{\widetilde{K}}$
ICCA+	$\widehat{W}_x^{\text{icca+}} = \mathring{U}_x \Lambda_x^{\text{opt}} \mathring{U}_{\widetilde{K}}$

Numerical Simulations

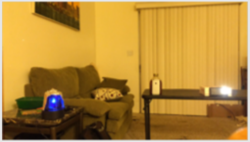
- Rank-2 setting $k_x = k_y = 2$, $p = 200$, $q = 250$,
- $\Theta_x = \Theta_y = \textbf{diag}(16, 4)$, $P_{xy} = \textbf{diag}(0.9, 0.5)$
- Plot the accuracy of the first canonical vector for empirical CCA, ICCA, and ICCA+



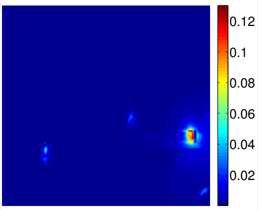
(a) n sweep　(b) θ sweep　(c) $\widehat{k}$ sweep

Canonical Vector Accuracy Experiment

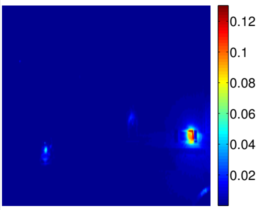
Original Scene



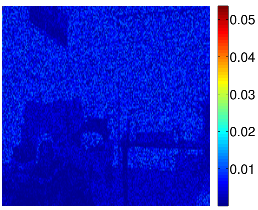
ICCA



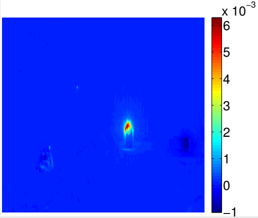
ICCA+



Empirical CCA



ICCA - ICCA+



Conclusion

- Improved canonical vector accuracy in low-sample, low-SNR regime in Canonical Correlation Analysis (CCA)
- Proposed a new algorithm: ICCA+
- ICCA+ uses insights from random matrix theory to optimally scale sample eigenvectors setting
- ICCA+ is more robust to over estimation of number of signals