On Canonical Correlation Analysis

CS189/289A: Introduction to Machine Learning

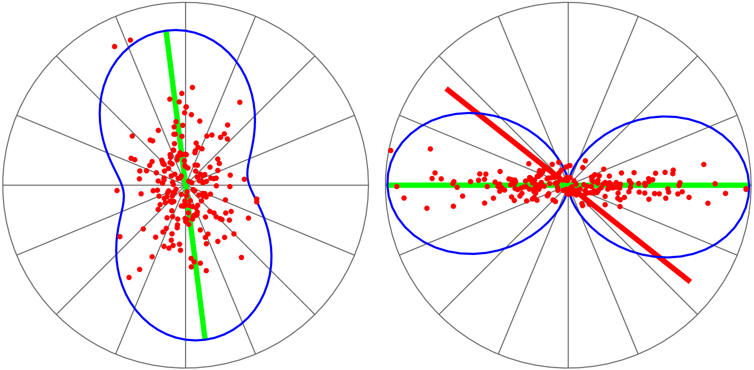
Stella Yu

UC Berkeley

26 September 2017

Why Canonical Correlation Analysis (CCA)?

- PCA varies with coordinate scaling:



(x, y) $\quad$ $(2.5x, 0.25y)$

- Need to discover cross-correlation regardless of external scaling – change of measurement units.

Why Canonical Correlation Analysis?

- Given two sets of random variables, there are correlations among the variables, CCA finds linear combinations of each set which have maximum correlation with each other.
- LS:

$$Y = Xw + N_X, \quad N_X \sim \mathcal{N}(0, \Sigma_X) \tag{1}$$

- TLS:

$$Y + N_Y = Xw + N_X, \quad N_X \sim \mathcal{N}(0, \Sigma_X), N_Y \sim \mathcal{N}(0, \Sigma_Y) \tag{2}$$

- CCA: linear dependence on a common latent space H

$$X_{n\times p} + N_x = H_{n\times k}U_{k\times p} + A \cdot N_A, \quad N_X \sim \mathcal{N}(0, \Sigma_X), N_A \sim \mathcal{N}(0, \Sigma_A) \tag{3}$$

$$Y_{n\times q} + N_y = H_{n\times k}V_{k\times q} + B \cdot N_B, \quad N_Y \sim \mathcal{N}(0, \Sigma_Y), N_B \sim \mathcal{N}(0, \Sigma_B) \tag{4}$$

Covariance and Pearson's Correlation Coefficient

- Covariance between two random variables:

$$\mathrm{cov}(X,Y) = E\left[\,(X - E[X]) \cdot (Y - E[Y])\,\right] \tag{5}$$

$$\mathrm{cov}(X,X) = E\left[\,(X - E[X])^2\,\right] = V[X] \tag{6}$$

$$\mathrm{cov}(X,Y) = 0, \text{ if } X \text{ and } Y \text{ are independent} \tag{7}$$

- Population Pearson's correlation coefficient:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X) \cdot \mathrm{cov}(Y,Y)}} = \frac{\mathrm{cov}(X,Y)}{\sqrt{V[X] \cdot V[Y]}} \tag{8}$$

$$\rho(Y,X) = \rho(X,Y) \tag{9}$$

$$-1 \leq \rho(X,Y) \leq 1 \tag{10}$$

- $\rho(X,Y)$ is not defined, when $V[X] = 0$ or $V[Y] = 0$.

Linear vs. Nonlinear Correlation vs. Independence

- Pearson's correlation detects only linear dependencies:

$$Y - E[Y] = k \cdot (X - E[X]) \tag{11}$$

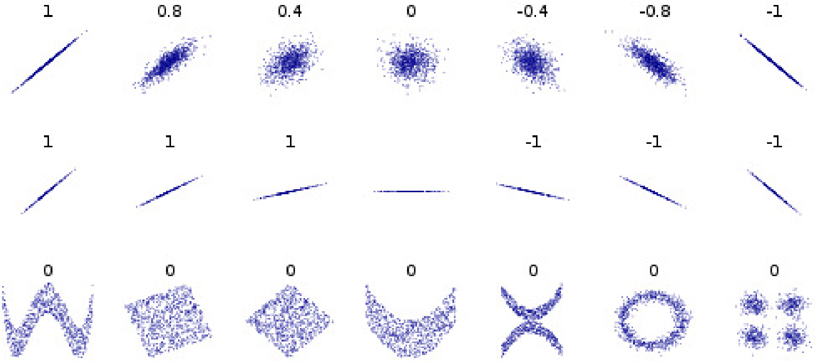
$$\Rightarrow \quad \rho(X,Y) = \frac{kV[X]}{\sqrt{V[X] \cdot k^2 V[X]}} = \pm 1, \quad \forall k. \tag{12}$$

- If X and Y are independent, then $\rho(X,Y) = 0$.

- If $\rho(X,Y) = 0$, then X and Y are linearly uncorrelated. They can be nonlinearly correlated and perfectly dependent, e.g. $Y = X^2, E[X] = 0$.

- If $\rho(X,Y) = 0$, when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

Sample Correlation Coefficient: $\rho \to r$



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (13)$$

$$r_{xy} = \frac{x'y}{\sqrt{x'x \cdot y'y}}, \qquad x \Leftarrow x - \bar{x}, \quad y \Leftarrow y - \bar{y} \quad (14)$$

Key: Correlation Coefficient Is Affine Invariant

$$\begin{align}
\rho(aX+c, bY+d) &= \rho(aX, bY) \tag{15}\\
&= \frac{\mathrm{cov}(aX, bY)}{\sqrt{V[aX] \cdot V[bY]}} \tag{16}\\
&= \frac{a \cdot b \cdot \mathrm{cov}(X, Y)}{\sqrt{a^2 \cdot b^2 \cdot V[X] \cdot V[Y]}} \tag{17}\\
&= \frac{\mathrm{cov}(X, Y)}{\sqrt{V[X] \cdot V[Y]}} \tag{18}\\
&= \rho(X, Y) \tag{19}
\end{align}$$

Gaussian Distribution and Correlation Coefficient

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(0, \Sigma) \tag{20}$$

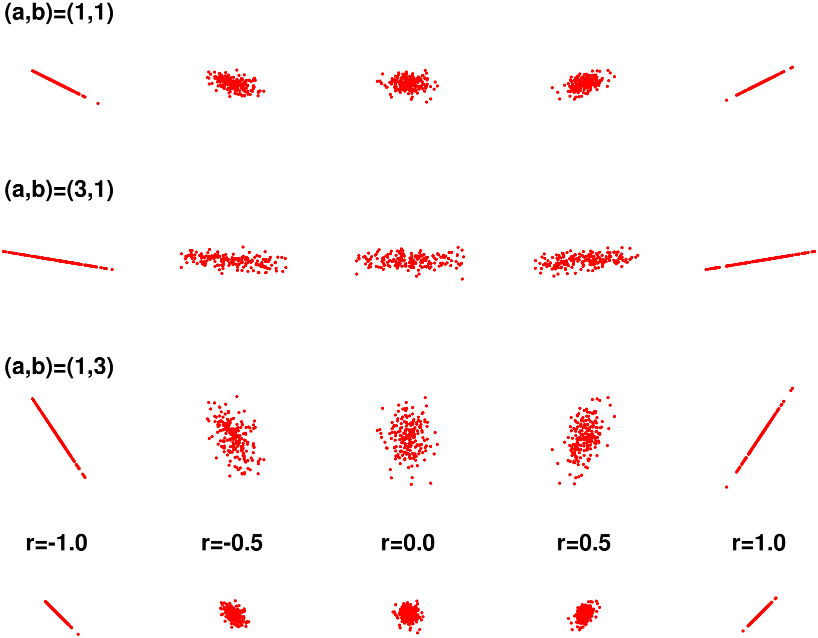
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{21}$$

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho \cdot \sigma_X \sigma_Y \\ \rho \cdot \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix} \tag{22}$$

$$\begin{bmatrix} \sigma_X^{-1} & 0 \\ 0 & \sigma_Y^{-1} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_X^{-1} & 0 \\ 0 & \sigma_Y^{-1} \end{bmatrix} \Sigma \begin{bmatrix} \sigma_X^{-1} & 0 \\ 0 & \sigma_Y^{-1} \end{bmatrix}'\right) \tag{23}$$

$$\sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \tag{24}$$

Gaussian Distribution and Correlation Coefficient



Canonical Correlation Analysis (CCA)

- Paired data matrices $(X_{n\times p}, Y_{n\times q})$, zero-mean
  n paired points in p and q dimensional spaces respectively.

- Simultaneously find projection directions $u_{p\times 1}$ in the X space and $v_{q\times 1}$ in the Y space such that the projected data onto u and v have maximal correlation:

$$\max_{u,v} \varepsilon(u,v;X,Y) = \rho(Xu, Yv) = \frac{u'X'Yv}{\sqrt{u'X'Xu \cdot v'Y'Yv}} \tag{25}$$

- In general, CCA seek a latent basis dimension k, $k \leq \min(p,q)$, where the correlation matrix between the variables is diagonal and the total correlations are maximized.

- Unlike PCA, CCA is invariant with respect to scaling or general affine transformations of the variables.

Solution Relations to Linear Subspace Methods

PCA, PLS (partial least squares), MLR (multivariate linear regression), and CCA share the same eigensolution routine:

$$Mw = \lambda D w \tag{26}$$

method	M	D
PCA / TLS	C_{xx}	I
PLS	$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix}$	$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$
MLS	$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix}$	$\begin{bmatrix} C_{xx} & 0 \\ 0 & I \end{bmatrix}$
CCA	$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix}$	$\begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix}$

CCA in Steps: Whitening and Decorrelation

1. Whitening X and Y separately based on auto-correlation:

$$
\begin{aligned}
X'X = U_x S_x U_x' &= (U_x S_x^{\frac{1}{2}} U_x')' \cdot (U_x S_x^{\frac{1}{2}} U_x') && (27)\\
u_w = (U_x S_x^{\frac{1}{2}} U_x') u &\Rightarrow u' C_{xx} u = u_w' u_w && (28)\\
u = W_x u_w, \quad X_w = X W_x &\Leftarrow W_x = U_x S_x^{-\frac{1}{2}} U_x' && (29)\\
\rho(Xu, Yv) = \rho(X_w u_w, Y_w v_w) &= \frac{u_w' (X_w' Y_w) v}{\sqrt{u_w' u_w \cdot v_w' v_w}} && (30)
\end{aligned}
$$

2. De-correlate X_w and Y_w based on cross-correlation:

$$
\begin{aligned}
X_w' Y_w = USV' &= USV' && (31)\\
u_w = D_x u_d, \quad X_w = X_d D_x &\Leftarrow D_x = U && (32)\\
\rho(Xu, Yv) = \rho(X_d u_d, Y_d v_d) &= \frac{u_d' S v_d}{\sqrt{u_d' u_d \cdot v_d' v_d}} && (33)\\
\le S_{1,1}, \quad u_d = [1, 0, \ldots, 0], &\quad v_d = [1, 0, \ldots, 0] && (34)
\end{aligned}
$$

CCA in Two Steps: Whitening and Decorrelation

- Rayleigh quotient optimization of asymmetric matrix $X_w' Y_w$:

$$\rho(X_w u_w, Y_w v_w) = \frac{u_w'(X_w' Y_w)v}{\sqrt{u_w' u_w \cdot v_w' v_w}} \tag{35}$$

$$(u_w, v_w) = \mathrm{eig}(X_w' Y_w) \tag{36}$$

$$u_w' X_w' Y_w v_w = S_{1,1} \tag{37}$$

- Composition of transformations in the original data space:

$$u = W_x u_w = W_x D_x u_d = W_x \cdot D_x \cdot U_{(:,1)} \tag{38}$$

$$v = W_y v_w = W_y D_y v_d = \underbrace{W_y}_{\text{whitening}} \cdot \underbrace{D_y}_{\text{decorrelation}} \cdot \underbrace{V_{(:,1)}}_{\text{Rayleigh}} \tag{39}$$

- CCA bases are often not orthogonal.

Connections Between Two CCA Solutions

$$\begin{bmatrix} & C_{xy} \\ C_{yx} & \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{40}$$

$$\begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} & C_{xy} \\ C_{yx} & \end{bmatrix} \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} X \\ Y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{41}$$

$$\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} = \begin{bmatrix} C_{xx} & \\ & C_{yy} \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{42}$$

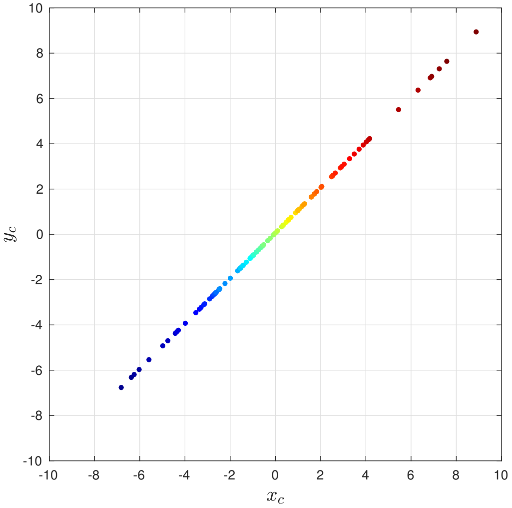
$$\begin{bmatrix} & C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}} \\ C_{yy}^{-\frac{1}{2}} C_{yx} C_{xx}^{-\frac{1}{2}} & \end{bmatrix} \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} = \lambda \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \tag{43}$$

$$C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}} \tilde{X} = \lambda \tilde{Y} \tag{44}$$

$$C_{yy}^{-\frac{1}{2}} C_{yx} C_{xx}^{-\frac{1}{2}} \tilde{Y} = \lambda \tilde{X} \tag{45}$$

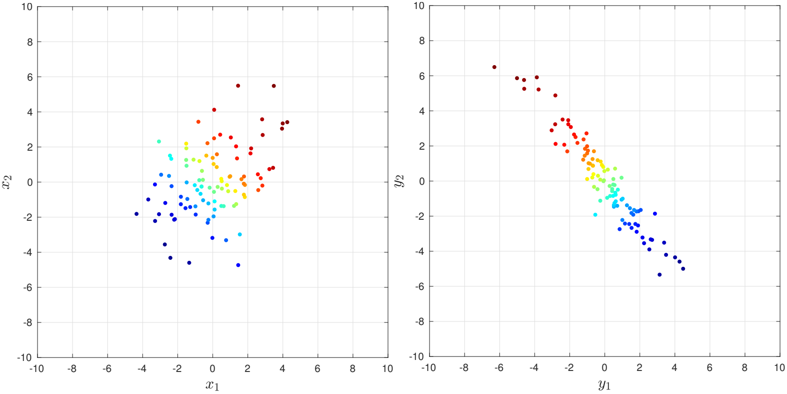
$$(\tilde{X}, \tilde{Y}) = \mathrm{eig}(C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}}) \tag{46}$$

Point Set #1: Hidden Correlation Between Spaces



$$y_c = f(x_c) = k \cdot x_c \tag{47}$$

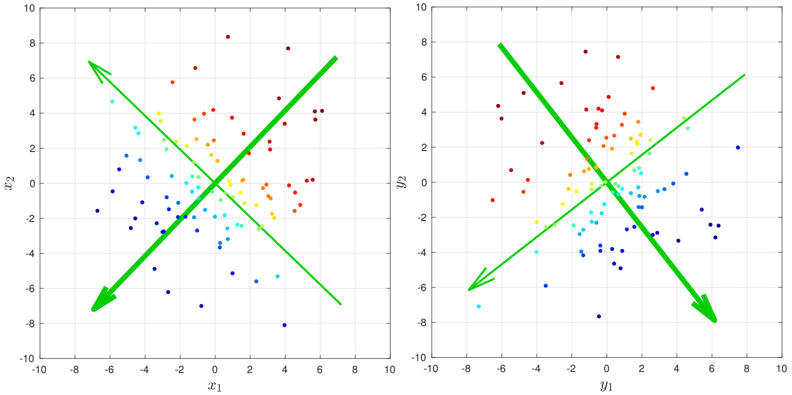
Irrelevant Orthogonal Components in Each Space



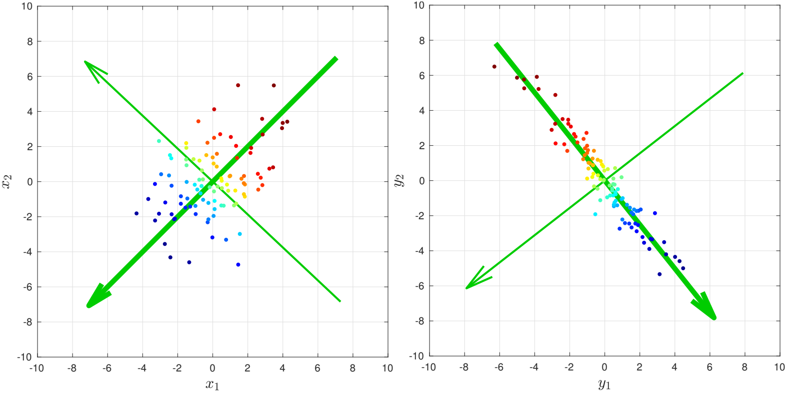
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} x_c \\ x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos(\beta) & \sin(\beta) \\ -\sin(\beta) & \cos(\beta) \end{bmatrix} \begin{bmatrix} y_c \\ y_n \end{bmatrix}$$

$$x_n \sim \mathcal{N}(0, \sigma_x^2) \qquad y_n \sim \mathcal{N}(0, \sigma_y^2)$$

CCA in the Whitened Spaces: Orthogonal



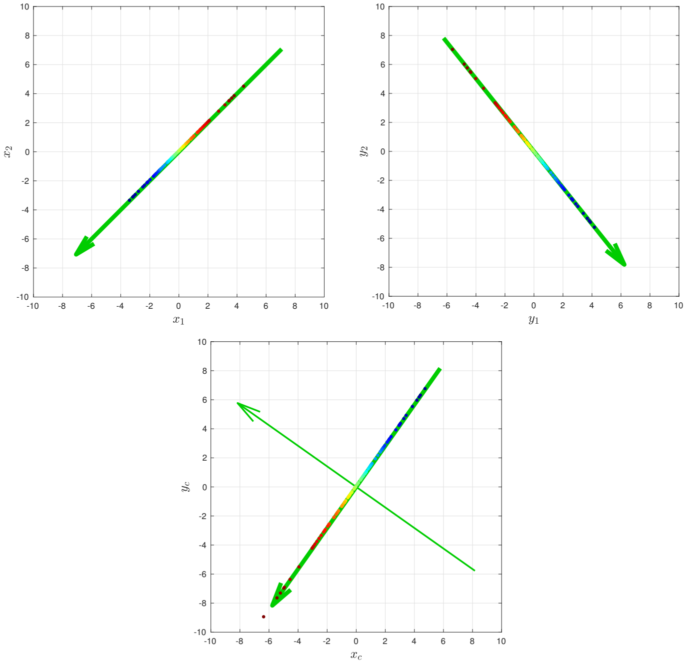
CCA in the Original Spaces: Orthogonal



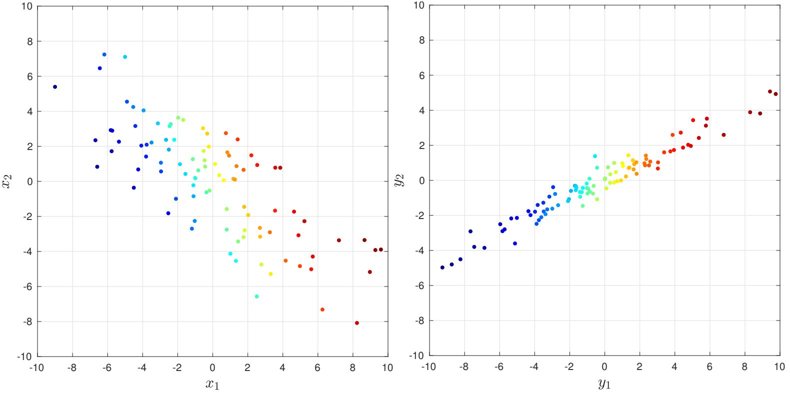
$$\rho_1 = 1.000 \tag{48}$$

$$\rho_2 = 0.211 \tag{49}$$

CCA Projection: Irrelevant, Orthogonal



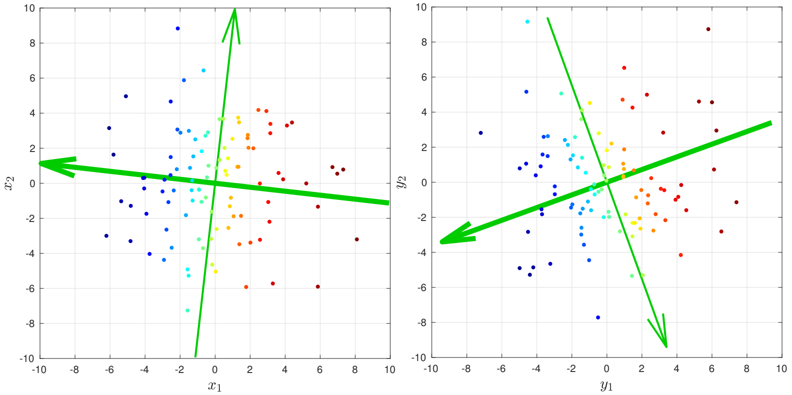
Point Set #2: Oblique Irrelevant Components



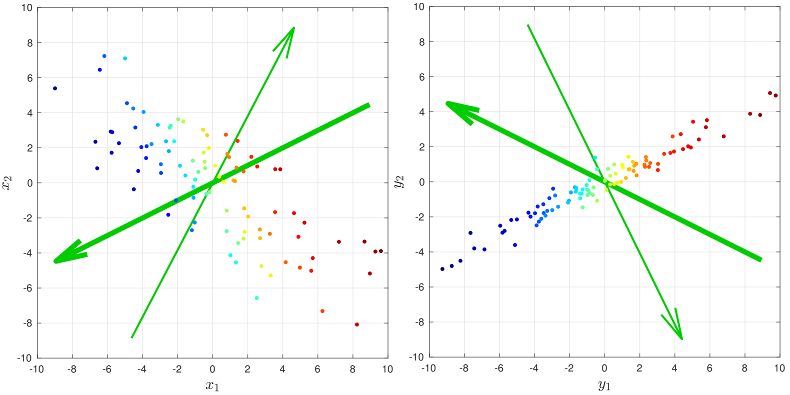
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ x_n \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \begin{bmatrix} y_c \\ y_n \end{bmatrix}$$

$$x_n \sim \mathcal{N}(0, \sigma_x^2) \qquad y_n \sim \mathcal{N}(0, \sigma_y^2)$$

CCA in the Whitened Spaces: Oblique



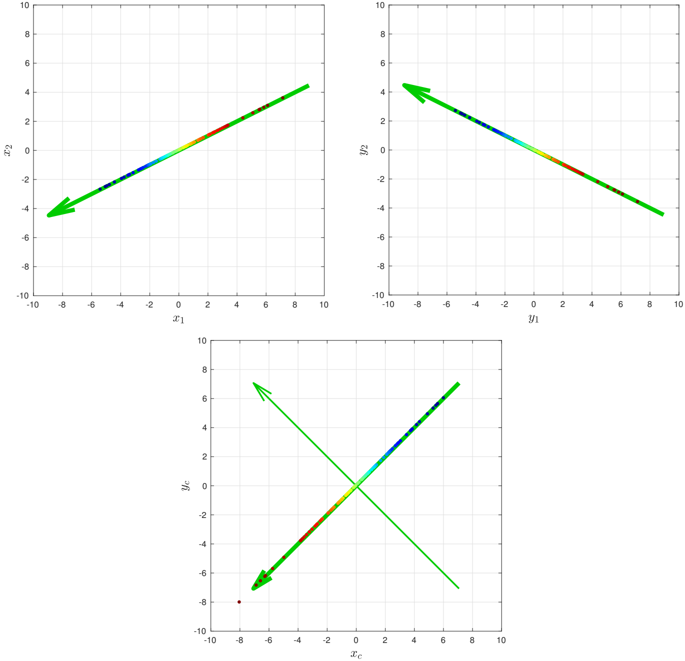
CCA in the Original Spaces: Oblique



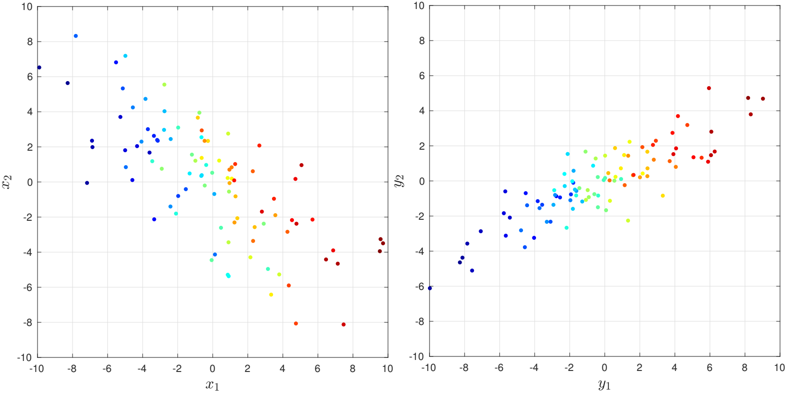
$$\rho_1 = 1.000 \tag{50}$$

$$\rho_2 = 0.211 \tag{51}$$

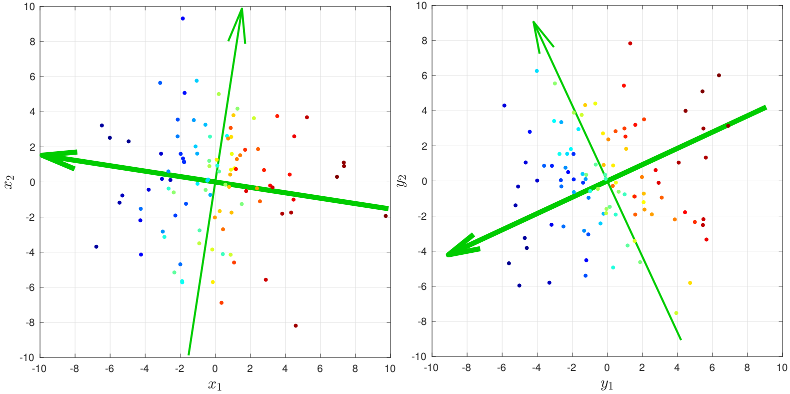
CCA Projection: Irrelevant, Oblique



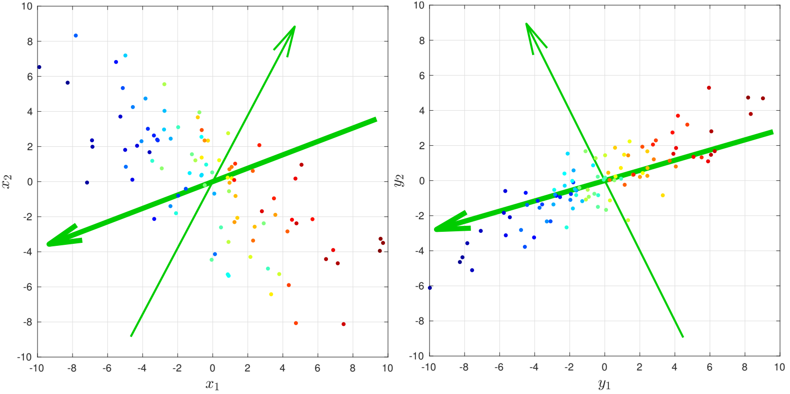
Point Set #3: Additive Noise in Each Space



CCA in the Whitened Spaces: Noise



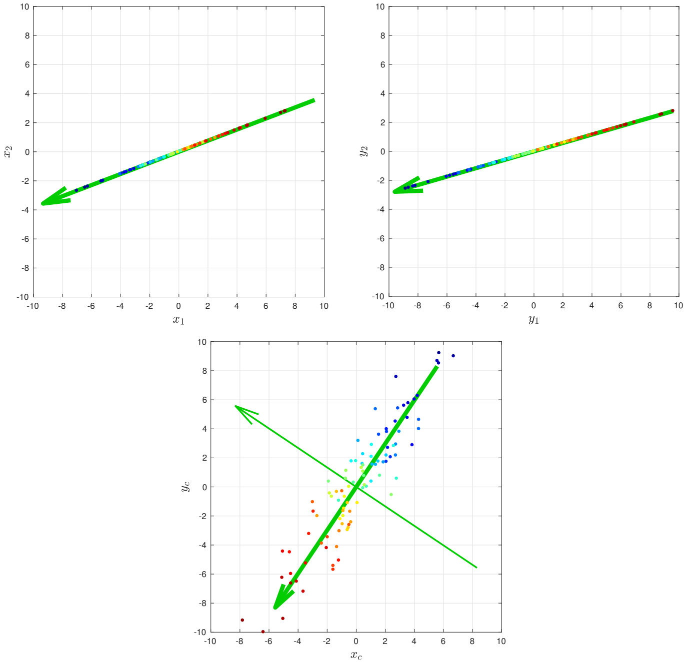
CCA in the Original Spaces: Noise



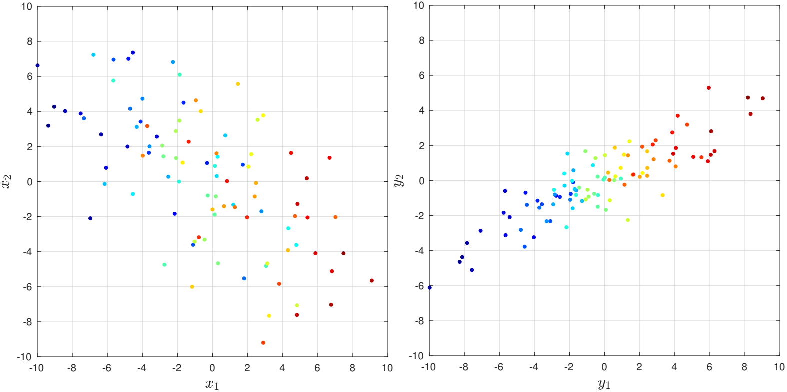
$$\rho_1 = 0.941 \tag{52}$$

$$\rho_2 = 0.050 \tag{53}$$

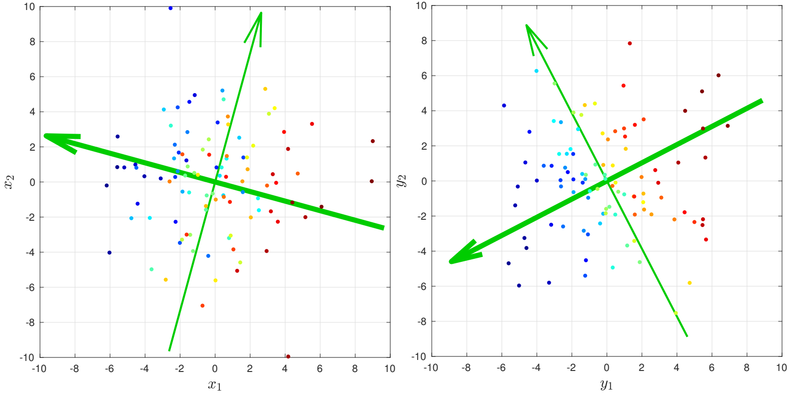
CCA Projection: Additive Noise



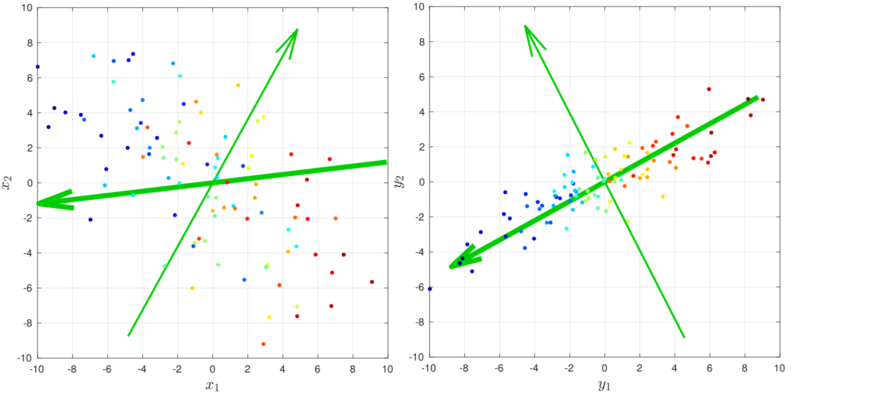
Point Set #4: Different Additive Noises



CCA in the Whitened Spaces: Different Noises



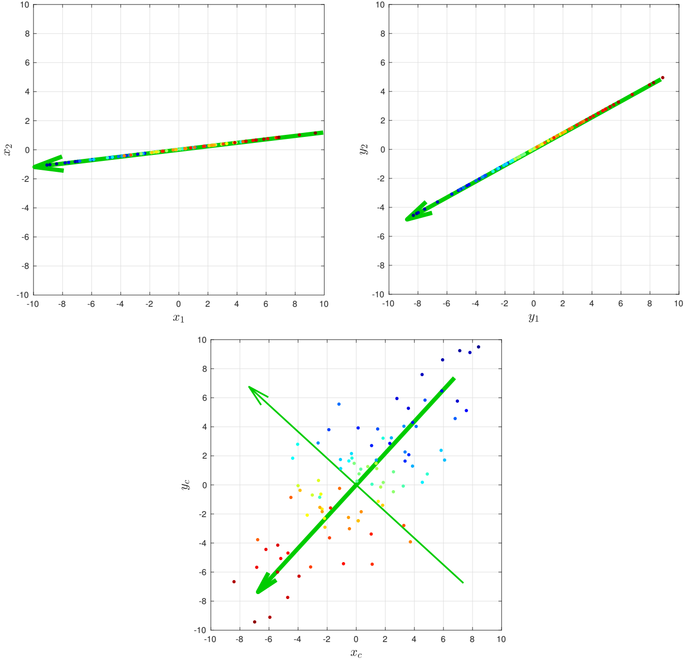
CCA in the Original Spaces: Different Noises



$$\rho_1 = 0.827 \tag{54}$$

$$\rho_2 = 0.077 \tag{55}$$

CCA Projection: Different Additive Noises



CCA and Coefficient Regression: Training

1. Given zero-mean training data (X, Y), compute CCA (U, V):

$$(U_{p\times k}, V_{q\times k}) = \mathrm{CCA}(X, Y) \tag{56}$$

2. Given (X, Y, U, V), compute their individual CCA coefficients:

$$X_c = X_{n\times p} U_{p\times k} \tag{57}$$

$$Y_c = Y_{n\times q} V_{q\times k} \tag{58}$$

   Note that (X_c, Y_c) is also of zero mean.

3. Given (X_c, Y_c), fit a $k \times k$ linear regressor A

$$Y_c = X_c A_{k\times k} \tag{59}$$

$$A = (X_c' X_c)^{-1}(X_c' Y_c) \tag{60}$$

$$= (U'X'XU)^{-1}(U'X'YV) \tag{61}$$

Prediction from CCA Coefficients: Testing

- Given CCA basis (U, V) and coefficient regressor A from the training data, given zero-mean test data X, predict Y:

$$\hat{Y}_c = XUA \tag{62}$$

$$\hat{Y} = \hat{Y}_c(V'V)^{-1}V' \tag{63}$$

- Equivalent linear predictor A_{eq} from X to $\hat{Y}$:

$$\hat{Y} = XA_{\mathrm{eq}} \tag{64}$$

$$A_{\mathrm{eq}} = UA(V'V)^{-1}V' \tag{65}$$

$$= \underbrace{U}_{\text{projection}} \underbrace{(U'X'XU)^{-1}}_{\text{whitening}} \underbrace{(U'X'YV)}_{\text{decorrelation}} \underbrace{(V'V)^{-1}V'}_{\text{projection back}} \tag{66}$$

Summary

- CCA investigates the relationships between two sets of variables, whereas PCA investigates the relationships within a single set of variables.
- CCA simultaneously find projection directions in the two spaces such that the projected data have maximal correlation, whereas PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.
- CCA is limited to the minimal dimension of the two spaces.
- PCA and CCA are comptued using SVD of correlation matrices.
- Unlike PCA, CCA is invariant with respect to scaling or general affine transformations of the variables.