CANONICAL CORRELATION ANALYSIS (CCA)

SETTING: * UNLIKE PCA WE'RE GOING TO FOCUS ON POPULATION (i.e. MEANS ; EXPECTATIONS) RATHER THAN SAMPLE AVERAGES (SAMPLE MEANS ; SAMPLE COVARIANCES)

GIVEN: A RANDOM VECTOR, $\vec{z}$, IN $(p+q)$ - DIMENSIONAL SPACE WHICH WE PARTITION INTO A p-DIMENSIONAL $\vec{x}$

q - DIMENSIONAL $\vec{y}$

LET $\vec{\mu}_x = E[\vec{x}]$ $\quad \Sigma_x = E\left[(\vec{x}-\mu_x)(\vec{x}-\mu_x)^T\right] \in \mathbb{R}^{p \times p}$

$\vec{\mu}_y = E[\vec{y}]$ $\quad \Sigma_y = E\left[(\vec{y}-\vec{\mu}_y)(\vec{y}-\mu_y)^T\right] \in \mathbb{R}^{q \times q}$

DEFINE $\Sigma_{xy} = E\left[(\vec{x}-\mu_x)(\vec{y}-\mu_y)^T\right] \in \mathbb{R}^{p \times q}$

$\Sigma_{yx} = E\left[(\vec{y}-\vec{\mu}_y)(\vec{x}-\vec{\mu}_x)^T\right] \in \mathbb{R}^{q \times p}$ $\qquad \Sigma_{xy} = \Sigma_{yx}^T$

CROSS - COVARIANCE MATRIX!

IF $Z = \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix}$ THEN

$$\mu_z = \begin{bmatrix} \vec{\mu}_x \\ \vec{\mu}_y \end{bmatrix} \qquad \Sigma_z = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \in \mathbb{R}^{(p+q) \times (p+q)}$$

GOAL: WE LOOK FOR $\vec{a} \in \mathbb{R}^p$, $b \in \mathbb{R}^q$ SO THAT:

$$\text{Corr}\left(\vec{a}^T\vec{x}, \vec{b}^T\vec{y}\right)$$

IS MAXIMIZED!

NOTE: 1) IF WE THINK OF $\vec{x}$ AS CAUSING $\vec{y}$ THEN $\vec{a}^T\vec{x}$ MAY BE CALLED THE "BEST PREDICTOR" AND $\vec{b}^T\vec{y}$ AS "MOST PREDICTABLE CRITERION". HOWEVER, THERE IS NO ASSUMPTION OF CAUSAL ASYMMETRY ; $\vec{x}$ AND $\vec{y}$ TREATED SYMMETRICALLY.

2) CONNECTIONS TO MULTIPLE REGRESSION (SEE IZENMAN CH. 7) AND REDUCED RANK REGRESSION.

CANONICAL CORRELATION ANALYSIS

THE CORRELATION BETWEEN $\eta = \vec{a}^T\vec{x}$ AND $\xi = \vec{b}^T\vec{y}$ IS

$$\rho(\vec{a},\vec{b}) = \mathrm{Corr}(\eta,\xi) = \frac{\mathrm{Cov}(\eta,\xi)}{\sqrt{\mathrm{Var}(\eta)\,\mathrm{Var}(\xi)}}$$

DEPENDS ON $\vec{a}, \vec{b}$

$$= \frac{E\left[\vec{a}^T\vec{x}\vec{y}^T b\right] - \vec{a}^T E[\vec{x}]E[\vec{y}^T]\vec{b}}{\sqrt{\mathrm{Var}(\vec{a}^T\vec{x})\,\mathrm{Var}(\vec{b}^T\vec{y})}}$$

$$= \frac{\vec{a}\left(E[\vec{x}\vec{y}^T] - E[\vec{x}]E[\vec{y}^T]\right)\vec{b}}{\sqrt{\vec{a}^T\Sigma_x\vec{a}\;\vec{b}^T\;\Sigma_y\;\vec{b}}} = \frac{\vec{a}^T\Sigma_{xy}\vec{b}}{\sqrt{\vec{a}^T\Sigma_x\vec{a}\;\vec{b}^T\Sigma_y\vec{b}}}$$

NOTE:

1) NEED $\left.\begin{array}{l}\vec{a}^T\Sigma_x\vec{a} \neq 0 \\ \vec{b}^T\Sigma_y\vec{b} \neq 0\end{array}\right\}$ ASSUME Σ_x, Σ_y HAVE FULL RANK

2) IF SCALARS $c, d \neq 0$ THEN

$$\rho(\vec{a},\vec{b}) = \rho(c\vec{a}, d\vec{b})$$

SO WE'LL FOCUS ON CONSTRAINTS $\vec{a}^T\Sigma_x\vec{a} = \vec{b}^T\Sigma_y\vec{b} = 1$

$$* \quad \boxed{(\vec{a},\vec{b})^* = \underset{\vec{a}^T\Sigma_x\vec{a} = \vec{b}^T\Sigma_y\vec{b} = 1}{\mathrm{argmax}} \; \rho(a,b)}$$

3) CAN MINIMIZE CORRELATION $\vec{a} \longmapsto -a$ OR $\vec{b} \longmapsto -\vec{b}$

$\rho(-\vec{a},\vec{b}) = -\rho(\vec{a},\vec{b})$ $\qquad$ $\rho(\vec{a},-\vec{b}) = -\rho(\vec{a},\vec{b})$

BUT NOT BOTH $\vec{a} \to -\vec{a}$, $\vec{b} \to -\vec{b}$

4) WE CAN SOLVE $*$ USING i) LAGRANGE MULTIPLIERS LIKE PCA TO GET GENERALIZED EIGENVALUE PROBLEM (MESSY & ~~VERY~~ TECHNICAL) OR USING ii) SVD (JUST MESSY)

ASIDE: FRACTIONAL POWERS OF A MATRIX

LET A BE A DIAGONALIZABLE $n \times n$ MATRIX SO THAT THERE EXISTS INVERTIBLE $P \in \mathbb{R}^{n \times n}$ AND DIAGONAL $D \in \mathbb{R}^{n \times n}$ SUCH THAT

$$A = P D P^{-1} \qquad , \quad D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

FOR $n = 0, 1, 2, \ldots$ WE DEFINE

$$A^n = P D^n P^{-1} \qquad D^\alpha = \begin{bmatrix} \lambda_1^n & & 0 \\ & \ddots & \\ 0 & & \lambda_n^n \end{bmatrix}$$

IF $\lambda_1, \ldots \lambda_n$ ARE > 0 WE CAN EXTEND THIS TO NEGATIVE AND FRACTIONAL POWERS, e.g.

$$A^{-1} = P \begin{bmatrix} 1/\lambda_1 & & 0 \\ & \ddots & \\ 0 & & 1/\lambda_n \end{bmatrix} P^{-1}$$

$$A^{1/2} = P \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix} P^{-1}$$

SINCE Σ_x, Σ_y ARE FULL RANK COVARIANCE MATRICES (POSITIVE DEFINITE)

$$\Sigma_x = \Sigma_x^{1/2} \Sigma_x^{1/2} \qquad \Sigma_x = P_x \Lambda_x P_x^T \qquad \Sigma_x^{1/2} = P_x \Lambda^{1/2} P_x^T$$

$$\Sigma_y = \Sigma_y^{1/2} \Sigma_y^{1/2} \qquad \Sigma_y = P_y \Lambda_y P_y^T \qquad \Sigma_y^{1/2} = P_y \Lambda_y^{1/2} P_y^T$$

Λ_x, Λ_y DIAGONAL

$$\frac{\vec{a}^T \Sigma_{xy} \vec{b}}{\sqrt{\vec{a}^T \Sigma_x \vec{a} \; \vec{b}^T \Sigma_y \vec{b}}} \xrightarrow[\substack{\vec{a} = \Sigma_x^{-1/2} \vec{\alpha} \\ \vec{b} = \Sigma_y^{-1/2} \vec{\beta}}]{\substack{\text{CHANGE OF VARIABLE} \\ \vec{\alpha} = \Sigma_x^{1/2} \vec{a} \\ \vec{\beta} = \Sigma_y^{1/2} \vec{b}}} \frac{\vec{\alpha}^T \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} \vec{\beta}}{\sqrt{\vec{\alpha}^T \Sigma_x^{-1/2} \Sigma_x \Sigma_x^{-1/2} \vec{\alpha} \; \vec{\beta}^T \Sigma_y^{-1/2} \Sigma_y \Sigma_y^{-1/2} \vec{\beta}}}$$

$$= \frac{\vec{\alpha}^T \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} \vec{\beta}}{\sqrt{\vec{\alpha}^T \vec{\alpha} \; \vec{\beta}^T \vec{\beta}}}$$

SUPPOSE RANK $(\Sigma_{xy}) = k \leq \min(p,q)$ SO THAT $\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}$ HAS RANK k AS WELL.

$$\Rightarrow \quad \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} = \ddot{U} \ddot{D} \ddot{V}^T$$

$$= \left[\vec{u}_1 | \ldots | \vec{u}_k\right] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \left[\vec{v}_1 | \ldots \vec{v}_k\right]^T$$

$\vec{u}_1, \ldots \vec{u}_k$ ORTHONORMAL VECTORS IN $\mathbb{R}^p$

$\vec{v}_1, \ldots \vec{v}_k$ " " , $\mathbb{R}^q$

TO MAXIMIZE

$$\frac{\vec{\alpha}^T \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}}{\sqrt{\vec{\alpha}^T\vec{\alpha} \; \vec{\beta}^T\vec{\beta}}} = \frac{\vec{\alpha}^T \left[\vec{u}_1, \ldots \vec{u}_k\right] \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_k \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vdots \\ \vec{v}_k^T \end{bmatrix} \vec{\beta}}{\sqrt{\vec{\alpha}^T\vec{\alpha} \; \vec{\beta}^T\vec{\beta}}}$$

WE CHOOSE $\vec{\alpha}_1 = \vec{u}_1 \quad \vec{\beta}_1 = \vec{v}_1$ TO ATTAIN (GLOBAL) MAXIMUM σ_1

SIMILARLY WE COULD CHOOSE $\vec{\alpha}_j = \vec{u}_j, \vec{\beta}_j = \vec{v}_j$ TO ATTAIN (LOCAL) MAXIMUM $\sigma_j \quad j=1,\ldots k$

<u>DEF</u>: LET $\vec{a}_j = \Sigma_x^{-1/2} \vec{\alpha}_j \;, \; \vec{b}_j = \Sigma_y^{-1/2} \vec{\beta}_j$ AND $\sigma_1, \ldots \sigma_k$ BE AS ABOVE. FOR $j=1,\ldots k$

(a) THE VECTORS $\vec{a}_j, \vec{b}_j$ ARE CALLED THE j^{th} CANONICAL CORRELATION VECTORS!

(b) THE RANDOM VARIABLES $\eta_j = \vec{a}_j^T \vec{x}$ AND $\xi_j = \vec{b}_j^T \vec{y}_j$ ARE CALLED THE j^{th} CANONICAL CORRELATION VARIBLES

(c) σ_j IS CALLED THE j^{th} CANONICAL CORRELATION

<u>NOTE</u>: 1) THE NONZERO SINGULAR VALUES OF $\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}$ ARE THE SQUARE ROOTS OF THE NONZERO EIGENVALUES OF

$$\left(\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}\right)\left(\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}\right)^T = \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1/2} = N$$

BY SIMILARITY THE EIGENVALUES OF N ARE THE SAME AS

$$M = \Sigma_x^{-1/2} N \Sigma_x^{1/2} = \boxed{\Sigma_x^{-1} \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}}$$

3) NOTE THAT $\vec{\alpha}_1, \ldots, \vec{\alpha}_K$ ARE ORTHONORMAL, e.g. $\vec{\alpha}_j^T \vec{\alpha}_k = \delta_{jk}$

BUT $\vec{a}_1, \ldots, \vec{a}_K$ ARE IN GENERAL NEITHER PERPENDICULAR NOR UNIT LENGTH, e.g. $\vec{a}_j^T \vec{a}_k \neq \delta_{jk}$. HOWEVER,

$$\vec{a}_j^T \Sigma_x \vec{a}_k = \vec{\alpha}_j^T \vec{\alpha}_k = \delta_{jk}$$

SIMILAR STATEMENT HOLDS FOR $\vec{\beta}_1, \ldots, \vec{\beta}_K, \vec{b}_1, \ldots, \vec{b}_K, \Sigma_y$

4) SINCE $\vec{\alpha}_j, \vec{\beta}_j$ WERE CHOSEN AS SINGULAR VECTORS OF $\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2}$

$$\Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} \vec{\beta}_j = \sigma_j \vec{\alpha}_j$$

$$\vec{\alpha}_j^T \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} = \sigma_j \vec{\beta}_j^T$$

CCA AND DATA

GIVEN INDEPENDENT SAMPLES OF $\begin{pmatrix} \vec{x}_i \\ \vec{y}_i \end{pmatrix}$ $i = 1, \ldots, N$ WE CAN COMPUTE THE SAMPLE COVARIANCE MATRICES

$$\hat{\Sigma}_x = \left(\frac{1}{N} \sum_{i=1}^{N} \vec{x}_i \vec{x}_i^T \right) - \bar{x}\bar{x}^T \qquad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$$

$$\hat{\Sigma}_y = \left(\frac{1}{N} \sum_{i=1}^{N} \vec{y}_i \vec{y}_i^T \right) - \bar{y}\bar{y}^T \qquad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} \vec{y}_i$$

AND THE SAMPLE CROSS-COVARIANCE MATRIX

$$\hat{\Sigma}_{xy} = \left(\frac{1}{N} \sum_{i=1}^{N} \vec{x}_i \vec{y}_i^T \right) - \bar{x}\bar{y}^T$$

$$= \frac{1}{N} \sum_{i=1}^{N} (\vec{x}_i - \bar{x})(\vec{y}_i - \bar{y})^T$$

AND REPEAT THE ANALYSIS

NOTE: NEED $\hat{\Sigma}_x, \hat{\Sigma}_y$ TO BE POSITIVE DEFINITE

$\Rightarrow$ NEED $N \geq \max\{p, q\}$

Given $\hat{\Sigma}_x, \hat{\Sigma}_y, \hat{\Sigma}_{xy}$ we can compute

a) Sample canonical correlations $\hat{\sigma}_1, \dots, \hat{\sigma}_k$ $\quad k = \text{rank}\left(\hat{\Sigma}_{xy}\right) \le \min\{p,q\}$

b) Sample canonical vectors $\vec{a}_1, \dots \vec{a}_k, \vec{b}_1, \dots \vec{b}_k$

Then for each sample $(\vec{x}_i, \vec{y}_i)$ we can compute the canonical variables (scores)

$$\eta_{ij} = \vec{a}_j^T x_i = \vec{x}_i^T \vec{a}_j$$
$$\xi_{ij} = \vec{b}_j \vec{y}_i = \vec{y}_i \vec{b}_j \qquad j = 1, \dots k$$

Note:

1) Let $\vec{\eta} = \begin{bmatrix} \vec{a}_1^T \vec{x} \\ \vdots \\ \vec{a}_k \vec{x} \end{bmatrix} = \begin{bmatrix} \vec{a}_1^T \\ \vdots \\ \vec{a}_k^T \end{bmatrix} \vec{x}$ and $\vec{\xi} = \begin{bmatrix} \vec{b}_1^T \vec{y} \\ \vdots \\ \vec{b}_k \vec{y} \end{bmatrix} = \begin{bmatrix} \vec{b}_1^T \\ \vdots \\ \vec{b}_k^T \end{bmatrix} \vec{y}$

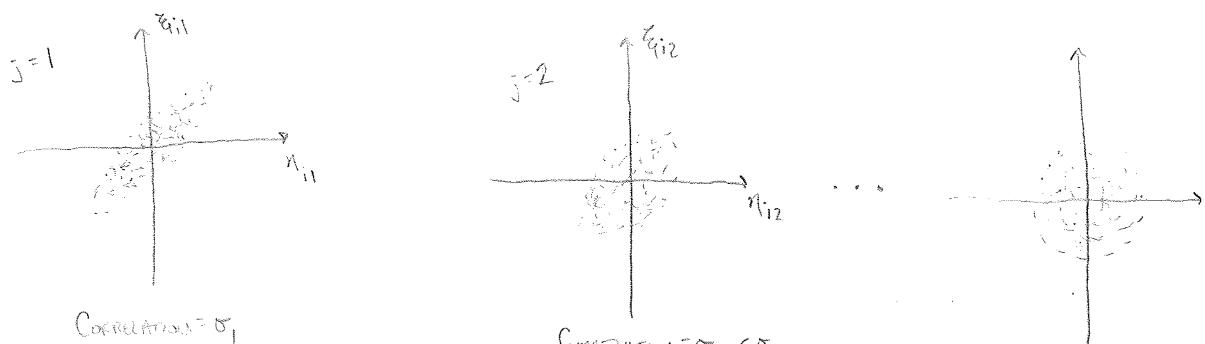
One can show (homework)

$$\Sigma_{\eta\xi} = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{bmatrix} \qquad \Sigma_\eta = \Sigma_\xi = I_k$$

Same statement holds for sample covariances and cross-covariance of

$$\vec{\eta}_i = \begin{bmatrix} \vec{a}_1^T \\ \vdots \\ \vec{a}_k \end{bmatrix} \vec{x}_i \qquad \hat{\xi}_i = \begin{bmatrix} \vec{b}_1^T \\ \vdots \\ \vec{b}_k \end{bmatrix} \vec{y}_i$$

2) If we plot $\left\{(\eta_{ij}, \xi_{ij})\right\}_{i=1}^{N}$ for $j = 1, \dots, k$ the plots will show decreasing positive correlation



Correlation $= \sigma_1$ $\qquad$ Correlation $= \sigma_2 < \sigma_1$ $\qquad$ Correlation $\sigma_k \le \sigma_{k-1} \dots$

3) IN PCA, DIMENSION REDUCTION FOCUSED ON MINIMIZING THE LOSS OF 'INFORMATION' (VARIANCE/CORRELATION) WITHIN THE DATA, $\underline{X}$.
CCA REDUCES DIMENSIONS WITH THE GOAL OF MINIMIZING THE LOSS OF 'INFORMATION' (CORRELATION) BETWEEN DATA $\underline{X}$ AND $\underline{Y}$.

- IF $\sigma_j \approx 0$ FOR $j \geq s$ AND $s \ll k$, WE CAN DISCARD η_{ij}, ξ_{ij} FOR $j \geq s$ WITHOUT LOSING MUCH OF THE INFORMATION IN $\vec{x}_i$ ABOUT $\vec{y}_i$
$\vec{y}_i$ ABOUT $\vec{x}_i$!

- $$\vec{x}_i^{(k)} = \sum_{s=1}^{k} \eta_{ij} \vec{a}_s \in \text{SPAN}\{\vec{a}_1, \ldots \vec{a}_k\} \subseteq \mathbb{R}^p$$
$$\vec{y}_i^{(k)} = \sum_{s=1}^{k} \xi_{ij} \vec{b}_s \in \text{SPAN}\{\vec{b}_1, \ldots, \vec{b}_k\} \subseteq \mathbb{R}^q$$
DIFFERENT SPACES!

1) $\vec{a}_1, \ldots \vec{a}_k$ NOT AN ORTHONORMAL BASIS FOR SPAN$\{\vec{a}_1, \ldots \vec{a}_k\}$
$\vec{b}_1, \ldots \vec{b}_k$ " " SPAN$\{\vec{b}_1, \ldots \vec{b}_k\}$

2) CCA DOES NOT FIND A k-DIMENSIONAL SUBSPACE OF $\mathbb{R}^{(p+q)}$!

LIMITATIONS OF PCA

1) LIKE PCA, THE CANONICAL CORRELATION VECTORS AND VARIABLES (SCORES) REPRESENT LINEAR COMBINATIONS OF COMPONENTS OF $\vec{x}, \vec{y}$

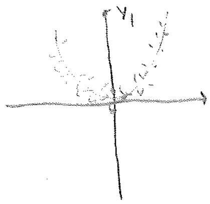
$\Rightarrow$ INTERPRETABILITY IS A ISSUE

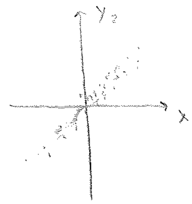
2) CORRELATION, CANONICAL OR OTHERWISE, IS A MEASUREMENT OF LINEAR DEPENDENCE

Ex: $x \sim N(0,1) \in \mathbb{R}^1$

$$\vec{y}\,|\,x \sim \begin{bmatrix} x^2 \\ x \end{bmatrix} + \vec{w}, \quad \vec{w} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 I\right)$$

DATA DRAWN FROM ↑





FIND CANONICAL CORRELATIONS AND VECTORS

- $Ex = 0, \quad \Sigma_x = 1$
- $Ey = E\left[\begin{bmatrix} x^2 \\ x \end{bmatrix} + \vec{w}\right] = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \Sigma_y = E\, y\vec{y}^T - \mu_y \mu_y^T$

$$= E\left[\begin{bmatrix} x^2 \\ x \end{bmatrix}\begin{bmatrix} x^2 & x \end{bmatrix}\right] + \sigma^2 I - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= E\begin{bmatrix} x^4 & x^3 \\ x^3 & x^2 \end{bmatrix} + \sigma^2 I - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2+\sigma^2 & 0 \\ 0 & 1+\sigma^2 \end{bmatrix}$$

- $\Sigma_{xy} = E\left[x\vec{y}\right] - \cancel{(Ex)}(E\vec{y})^{=0} = E\left[[x^3, x^2] + x\vec{w}\right] = [0, 1]$

$$\Rightarrow \Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2} = [1][0, 1]\begin{bmatrix} \frac{1}{\sqrt{2+\sigma^2}} & 0 \\ 0 & \frac{1}{\sqrt{1+\sigma^2}} \end{bmatrix} = \left[0, \frac{1}{\sqrt{1+\sigma^2}}\right]$$

$$= [1]\left[\frac{1}{\sqrt{1+\sigma^2}}, 0\right]\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\Rightarrow \sigma_1 = \frac{1}{\sqrt{1+\sigma^2}}$$

$\alpha_1 = 1 \Rightarrow a_1 = \Sigma_x^{-1/2}\alpha_1 = 1$

$\vec{\beta}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow b_1 = \Sigma_y^{-1/2}\vec{\beta}_1 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{1+\sigma^2}} \end{bmatrix}$

NO WEIGHT TO $y_1 = x^2 + \sigma w_1$!

* CORRELATION BETWEEN x^2 AND x IS 0!

8

* CANONICAL CORRELATIONS MAY ALL BE NEAR BUT THERE MAY BE A STRONG DEPENDENCE BETWEEN (COMPONENTS OF) $\vec{x}$ AND (COMPONENTS OF) $\vec{y}$!