# FINM 331: MULTIVARIATE DATA ANALYSIS
# FALL 2021
# CANONICAL CORRELATION ANALYSIS

## 1. Population canonical correlation analysis

- in population PCA, we deal with one set of random variables
- in population CCA, we deal with two sets of random variables
- given random variables $X_1, \dots, X_p$ and $Y_1, \dots, Y_q$, we assemble them into two random vectors
$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix}$$
- without loss of generality, let $p \le q$ throughout this handout
- we introduce the following notations
$$\begin{aligned} \boldsymbol{\mu}_X &= E(\mathbf{X}) \in \mathbb{R}^p, & \boldsymbol{\mu}_Y &= E(\mathbf{Y}) \in \mathbb{R}^q, \\ \Sigma_X &= \operatorname{Cov}(\mathbf{X}) \in \mathbb{R}^{p\times p}, & \Sigma_Y &= \operatorname{Cov}(\mathbf{Y}) \in \mathbb{R}^{q\times q}, \\ \Sigma_{XY} &= \operatorname{Cov}(\mathbf{X},\mathbf{Y}) \in \mathbb{R}^{p\times q}, & \Sigma_{YX} &= \operatorname{Cov}(\mathbf{Y},\mathbf{X}) \in \mathbb{R}^{q\times p} \end{aligned}$$
- note that
$$\Sigma_{XY}^{\mathsf{T}} = \Sigma_{YX}$$
- now write
$$\mathbf{W} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix}$$
and denote
$$\begin{aligned} \boldsymbol{\mu} &= E(\mathbf{W}) = \begin{bmatrix} E(\mathbf{X}) \\ E(\mathbf{Y}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix} \in \mathbb{R}^{p+q} \\ \Sigma &= \operatorname{Cov}(\mathbf{W}) = E(\mathbf{W}-\boldsymbol{\mu})(\mathbf{W}-\boldsymbol{\mu})^{\mathsf{T}} \\ &= \begin{bmatrix} E(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{X}-\boldsymbol{\mu}_X)^{\mathsf{T}} & E(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathsf{T}} \\ E(\mathbf{Y}-\boldsymbol{\mu}_Y)(\mathbf{X}-\boldsymbol{\mu}_X)^{\mathsf{T}} & E(\mathbf{Y}-\boldsymbol{\mu}_Y)(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathsf{T}} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)} \end{aligned}$$
- consider two linear combinations
$$U = \mathbf{a}^{\mathsf{T}}\mathbf{X}, \qquad V = \mathbf{b}^{\mathsf{T}}\mathbf{Y}$$
with $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ constant vectors (note that $U$ and $V$ are random variables)

*Date*: October 26, 2021, version 1.0. Comments, bug reports: `lekheng@uchicago.edu`.

- by results in previous handouts, the variances and covariances of $U$ and $V$ are given by

$$\begin{aligned}
\operatorname{Var}(U) &= \mathbf{a}^\intercal \operatorname{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}^\intercal \Sigma_X \mathbf{a} \in \mathbb{R}, \\
\operatorname{Var}(V) &= \mathbf{b}^\intercal \operatorname{Cov}(\mathbf{Y})\mathbf{b} = \mathbf{b}^\intercal \Sigma_Y \mathbf{b} \in \mathbb{R}, \\
\operatorname{Cov}(U,V) &= \mathbf{a}^\intercal \operatorname{Cov}(\mathbf{X},\mathbf{Y})\mathbf{b} = \mathbf{a}^\intercal \Sigma_{XY} \mathbf{b} \in \mathbb{R}, \\
\operatorname{Corr}(U,V) &= \frac{\mathbf{a}^\intercal \Sigma_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^\intercal \Sigma_X \mathbf{a}}\sqrt{\mathbf{b}^\intercal \Sigma_Y \mathbf{b}}} \in \mathbb{R}
\end{aligned}$$

- note that

$$\operatorname{Cov}(U,V) = \mathbf{a}^\intercal \Sigma_{XY}\mathbf{b} = \mathbf{b}^\intercal \Sigma_{XY}\mathbf{a} = \operatorname{Cov}(V,U)$$

- the first pair of *population canonical variables* is defined by

$$\begin{aligned}
(U_1, V_1)) &= \operatorname{argmax} \operatorname{Corr}(U,V) \\
&= \operatorname{argmax}\{\operatorname{Cov}(U,V) : \operatorname{Var}(U) = 1 = \operatorname{Var}(V)\}
\end{aligned}$$

- note that the second equality holds because $\operatorname{Corr}(U,V) = \operatorname{Cov}(U,V)$ when $\operatorname{Var}(U) = 1 = \operatorname{Var}(V)$ — this applies in the following as well
- the second pair of population canonical variables is defined by

$$\begin{aligned}
(U_2, V_2) = \operatorname{argmax}\{&\operatorname{Cov}(U,V) : \operatorname{Var}(U) = \operatorname{Var}(V) = 1, \\
&\operatorname{Cov}(U,U_1) = \operatorname{Cov}(V,V_1) = \operatorname{Cov}(U,V_1) = \operatorname{Cov}(V,U_1) = 0\}
\end{aligned}$$

- the $k$th pair of population canonical variables is defined by

$$\begin{aligned}
(U_k, V_k) = \operatorname{argmax}\{&\operatorname{Cov}(U,V) : \operatorname{Var}(U) = \operatorname{Var}(V) = 1, \\
&\operatorname{Cov}(U,U_j) = \operatorname{Cov}(V,V_j) = \operatorname{Cov}(U,V_j) = \operatorname{Cov}(V,U_j) = 0,\ j = 1,\dots,k-1\}
\end{aligned}$$

  for $k = 1, \dots, p$
- if we replace argmax by max, the corresponding maximal values are called the $k$th *population canonical correlation* and denoted

$$\rho_k = \operatorname{Corr}(U_k, V_k)$$

**Theorem 1.** *Suppose $p \le q$. Let*

$$\Gamma_{XY} = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2} \in \mathbb{R}^{p\times q}$$

*and the condensed* SVD *of $\Gamma_{XY}$ be*

$$\Gamma_{XY} = [\mathbf{u}_1, \dots, \mathbf{u}_p] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\intercal \\ \vdots \\ \mathbf{v}_p^\intercal \end{bmatrix},$$

*where $\mathbf{u}_1, \dots, \mathbf{u}_p \in \mathbb{R}^p$ are the left singular vectors, $\sigma_1 \ge \cdots \ge \sigma_p$ are the singular values, and $\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^q$ are the right singular vectors. Then*

$$U_k = \mathbf{u}_k^\intercal \Sigma_X^{-1/2}\mathbf{X}, \qquad V_k = \mathbf{v}_k^\intercal \Sigma_Y^{-1/2}\mathbf{Y}$$

*are the $k$th population canonical variables and*

$$\operatorname{Corr}(U_k, V_k) = \sigma_k$$

*is the $k$th population canonical correlation for $k = 1, \dots, p$.*

- the proof will be in the homework

- the vectors
$$\mathbf{a}_k = \Sigma_X^{-1/2}\mathbf{u}_k \in \mathbb{R}^p, \qquad \mathbf{b}_k = \Sigma_Y^{-1/2}\mathbf{v}_k \in \mathbb{R}^q$$
are called the *population canonical vectors* and we have
$$U_k = \mathbf{a}_k^\mathsf{T}\mathbf{X}, \qquad V_k = \mathbf{b}_k^\mathsf{T}\mathbf{Y} \tag{1.1}$$
- every symmetric positive semidefinite matrix $A \in \mathbb{R}^{p\times p}$ has a *square root* $A^{1/2} \in \mathbb{R}^{p\times p}$ that is also symmetric positive semidefinite
- this can be defined via the EVD of $A$: if
$$A = Q\Lambda Q^\mathsf{T}$$
where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ is the diagonal matrix of nonnegative (since $A$ is symmetric positive semidefinite) eigenvalues, then
$$A^{1/2} = Q\Lambda^{1/2}Q^\mathsf{T}$$
where $\Lambda^{1/2} = \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$
- you can check that the population canonical variables in Theorem 1 satisfy
$$\operatorname{Cov}(U_i, U_j) = \operatorname{Cov}(V_i, V_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$
for $i, j = 1, \dots, p$, and
$$\operatorname{Cov}(U_i, V_j) = \operatorname{Cov}(V_i, U_j) = 0$$
for $i \neq j$, $i, j = 1, \dots, p$
- recall from the properties of SVD that $\sigma_1^2 \geq \cdots \geq \sigma_p^2$ are the eigenvalues of the matrix
$$\Gamma_{XY}\Gamma_{XY}^\mathsf{T} = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1/2} \in \mathbb{R}^{p\times p}$$
with corresponding eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_p$
- $\sigma_1^2 \geq \cdots \geq \sigma_p^2$ are also the $p$ largest eigenvalues of the matrix
$$\Gamma_{XY}^\mathsf{T}\Gamma_{XY} = \Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2} \in \mathbb{R}^{q\times q}$$
with corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_p$
- CCA has a nice property not found in PCA and FA

**Theorem 2.** *Let* $\mathbf{X} = [X_1, \dots, X_p]^\mathsf{T}$ *and* $\mathbf{Y} = [Y_1, \dots, Y_q]^\mathsf{T}$ *be random vectors. If*
$$\mathbf{X}' = M\mathbf{X} + \mathbf{c}, \qquad \mathbf{Y}' = N\mathbf{Y} + \mathbf{d}$$
*for some nonsingular matrices* $M \in \mathbb{R}^{p\times p}$, $N \in \mathbb{R}^{q\times q}$ *and some vectors* $\mathbf{c} \in \mathbb{R}^p$, $\mathbf{d} \in \mathbb{R}^q$, *then*

(i) *the canonical correlations between* $\mathbf{X}'$ *and* $\mathbf{Y}'$ *are the same as the canonical correlations between* $\mathbf{X}$ *and* $\mathbf{Y}$*;*

(ii) *the canonical correlation vectors of* $\mathbf{X}'$ *and* $\mathbf{Y}'$ *are*
$$\mathbf{a}_k' = M^{-\mathsf{T}}\mathbf{a}_k \in \mathbb{R}^p, \qquad \mathbf{b}_k' = N^{-\mathsf{T}}\mathbf{b}_k \in \mathbb{R}^q$$
*where* $\mathbf{a}_k \in \mathbb{R}^p$ *and* $\mathbf{b}_k \in \mathbb{R}^q$ *are the canonical vectors between* $\mathbf{X}$ *and* $\mathbf{Y}$*.*

- note that if $M$ is nonsingular, then $M^\mathsf{T}$ must also be nonsingular and
$$(M^\mathsf{T})^{-1} = (M^{-1})^\mathsf{T}$$
we denote this matrix (inverse of transpose or transpose of inverse) by $M^{-\mathsf{T}}$

- just as PCA and FA apply with correlation matrix in place of covariance matrix, the same is true for CCA, we may use
$$P_X^{-1/2} P_{XY} P_Y^{-1/2} \in \mathbb{R}^{p\times q}$$
in place of
$$\Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} \in \mathbb{R}^{p\times q}$$
in Theorem 1
- thanks to Theorem 2, there is a straightforward relation between using correlation and covariance matrices
- if $\mathbf{a}_k' \in \mathbb{R}^p$ and $\mathbf{b}_k' \in \mathbb{R}^q$ are the canonical vectors of $\mathbf{X}$ and $\mathbf{Y}$ coming the correlation matrix version of CCA and $\mathbf{a}_k \in \mathbb{R}^p$ and $\mathbf{b}_k \in \mathbb{R}^q$ are the canonical vectors of $\mathbf{X}$ and $\mathbf{Y}$ coming the covariance matrix version, then
$$\mathbf{a}_k' = V_X^{1/2}\mathbf{a}_k, \qquad \mathbf{b}_k' = V_Y^{1/2}\mathbf{b}_k \tag{1.2}$$
where $V_X^{1/2} \in \mathbb{R}^{p\times p}$ and $V_Y^{1/2} \in \mathbb{R}^{q\times q}$ are the population standard deviation matrices, i.e.,
$$V_X = \operatorname{diag}(\operatorname{Var}(X_1), \dots, \operatorname{Var}(X_p)) \in \mathbb{R}^{p\times p}, \qquad V_Y = \operatorname{diag}(\operatorname{Var}(Y_1), \dots, \operatorname{Var}(Y_q)) \in \mathbb{R}^{q\times q}$$
- this follows from applying Theorem 2 with $M = V_X^{-1/2}$ and $N = V_Y^{-1/2}$
- note that there is no equivalent for PCA — there is no simple relationship in general between the principal components obtained from the covariance matrix and those obtained from the correlation matrix
- Theorem 2 and the discussion that follows applies also to sample CCA

## 2. SAMPLE CANONICAL CORRELATION ANALYSIS

- in sample PCA, we deal with one data matrix
- in sample CCA, we deal with two data matrices
$$X = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \mathbf{x}_2^\intercal \\ \vdots \\ \mathbf{x}_n^\intercal \end{bmatrix} \in \mathbb{R}^{n\times p}, \qquad Y = \begin{bmatrix} \mathbf{y}_1^\intercal \\ \mathbf{y}_2^\intercal \\ \vdots \\ \mathbf{y}_n^\intercal \end{bmatrix} \in \mathbb{R}^{n\times q}$$
- the relation between population and sample CCA, again, is the following: the $(i, j)$th entry of the data matrix
$$x_{ij} = X_j(\omega_i) = i\text{th observation of the random variable } X_j,$$
$$y_{ik} = Y_k(\omega_i) = i\text{th observation of the random variable } Y_k,$$
where $i = 1, \dots, n$, $j = 1, \dots, p$, $k = 1, \dots, q$
- if you do not get this, go review previous handouts
- so each row vector $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^\intercal \in \mathbb{R}^p$ is an observation of the random vector $\mathbf{X} = [X_1, \dots, X_p]^\intercal$, each row vector $\mathbf{y}_i = [y_{i1}, \dots, y_{iq}]^\intercal \in \mathbb{R}^q$ is an observation of the random vector $\mathbf{Y} = [Y_1, \dots, Y_p]^\intercal$
- note that in CCA, the observations must come in pairs, so that the number of rows of the two data matrices $X$ and $Y$ must be the same
- we will assemble the data matrices into a single one
$$W = [X, Y] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix} \in \mathbb{R}^{n\times(p+q)}$$

- the sample mean of $W$ is

$$\overline{\mathbf{w}} = \begin{bmatrix} \overline{\mathbf{x}} \\ \overline{\mathbf{y}} \end{bmatrix} \in \mathbb{R}^{p+q}, \qquad \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \in \mathbb{R}^p, \qquad \overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \in \mathbb{R}^q$$

- the sample covariance of $W$ is

$$S = \begin{bmatrix} S_X & S_{XY} \\ S_{YX} & S_Y \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)},$$

$$S_X = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \qquad S_Y = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^{\mathsf{T}} \in \mathbb{R}^{q\times q},$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})^{\mathsf{T}} = S_{YX}^{\mathsf{T}} \in \mathbb{R}^{p\times q}$$

- for any $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{b} \in \mathbb{R}^q$, the *sample correlation* is a function of $\mathbf{a}$ and $\mathbf{b}$ defined by

$$r_{XY}(\mathbf{a}, \mathbf{b}) := \frac{\mathbf{a}^{\mathsf{T}} S_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^{\mathsf{T}} S_X \mathbf{a}}\sqrt{\mathbf{b}^{\mathsf{T}} S_Y \mathbf{b}}} \in \mathbb{R}$$

- note that $r_{XY}(\mathbf{a}, \mathbf{b})$ is the correlation between the list of values $\mathbf{a}^{\mathsf{T}}\mathbf{x}_1, \dots, \mathbf{a}^{\mathsf{T}}\mathbf{x}_n$ and the list of values $\mathbf{b}^{\mathsf{T}}\mathbf{y}_1, \dots, \mathbf{b}^{\mathsf{T}}\mathbf{y}_n$
- the first pair of *sample canonical vectors* is defined by

$$\begin{aligned} (\widehat{\mathbf{a}}_1, \widehat{\mathbf{b}}_1) &= \operatorname{argmax} \frac{\mathbf{a}^{\mathsf{T}} S_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^{\mathsf{T}} S_X \mathbf{a}}\sqrt{\mathbf{b}^{\mathsf{T}} S_Y \mathbf{b}}} \\ &= \operatorname{argmax}\{\mathbf{a}^{\mathsf{T}} S_{XY} \mathbf{b} : \mathbf{a}^{\mathsf{T}} S_X \mathbf{a} = 1 = \mathbf{b}^{\mathsf{T}} S_Y \mathbf{b}\} \end{aligned}$$

- as in the population CCA case, sample correlation and sample covariance are identical under the conditions $\mathbf{a}^{\mathsf{T}} S_X \mathbf{a} = 1 = \mathbf{b}^{\mathsf{T}} S_Y \mathbf{b}$
- the second pair of sample canonical vectors is defined by

$$\begin{aligned} (\widehat{\mathbf{a}}_2, \widehat{\mathbf{b}}_2) = \operatorname{argmax}\{&\mathbf{a}^{\mathsf{T}} S_{XY} \mathbf{b} : \mathbf{a}^{\mathsf{T}} S_X \mathbf{a} = \mathbf{b}^{\mathsf{T}} S_Y \mathbf{b} = 1, \\ &\mathbf{a}^{\mathsf{T}} S_X \widehat{\mathbf{a}}_1 = \mathbf{a}^{\mathsf{T}} S_{XY} \widehat{\mathbf{b}}_1 = \mathbf{b}^{\mathsf{T}} S_{YX} \widehat{\mathbf{a}}_1 = \mathbf{b}^{\mathsf{T}} S_Y \widehat{\mathbf{b}}_1 = 0\} \end{aligned}$$

- the $k$th pair of sample canonical variables is defined by

$$\begin{aligned} (\widehat{\mathbf{a}}_k, \widehat{\mathbf{b}}_k) = \operatorname{argmax}\{&\mathbf{a}^{\mathsf{T}} S_{XY} \mathbf{b} : \mathbf{a}^{\mathsf{T}} S_X \mathbf{a} = \mathbf{b}^{\mathsf{T}} S_Y \mathbf{b} = 1, \\ &\mathbf{a}^{\mathsf{T}} S_X \widehat{\mathbf{a}}_j = \mathbf{a}^{\mathsf{T}} S_{XY} \widehat{\mathbf{b}}_j = \mathbf{b}^{\mathsf{T}} S_{YX} \widehat{\mathbf{a}}_j = \mathbf{b}^{\mathsf{T}} S_Y \widehat{\mathbf{b}}_j = 0,\ j = 1, \dots, k-1\} \end{aligned}$$

  for $k = 1, \dots, p$
- if we replace argmax by max, the corresponding maximal values are called the $k$th *sample canonical correlation* and denoted

$$r_k = \widehat{\mathbf{a}}_k^{\mathsf{T}} S_{XY} \widehat{\mathbf{b}}_k$$

- we have the following sample analogue of Theorem 1

**Theorem 3.** *Suppose $p \le q$. Let*

$$G_{XY} = S_X^{-1/2} S_{XY} S_Y^{-1/2} \in \mathbb{R}^{p\times q}$$

*and the condensed* SVD *of $G_{XY}$ be*

$$G_{XY} = [\mathbf{u}_1, \dots, \mathbf{u}_p] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{v}_p^{\mathsf{T}} \end{bmatrix},$$

*where* $\mathbf{u}_1, \ldots, \mathbf{u}_p \in \mathbb{R}^p$ *are the left singular vectors,* $\sigma_1 \geq \cdots \geq \sigma_p$ *are the singular values, and* $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^q$ *are the right singular vectors. Then*

$$\widehat{\mathbf{a}}_k = S_X^{-1/2}\mathbf{u}_k \in \mathbb{R}^p, \qquad \widehat{\mathbf{b}}_k = S_Y^{-1/2}\mathbf{v}_k \in \mathbb{R}^q$$

*are the* $k$*th sample canonical vectors and*

$$r_k = \sigma_k$$

*is the* $k$*th sample canonical correlation for* $k = 1, \ldots, p$*.*

- as in the PCA case, sometimes people like to view sample canonical components as random variables
- the random variables
$$\widehat{U}_k := \widehat{\mathbf{a}}_k^{\mathsf{T}}\mathbf{X}, \qquad \widehat{V}_k := \widehat{\mathbf{b}}_k^{\mathsf{T}}\mathbf{Y} \tag{2.1}$$
are called the $k$th *sample canonical variables*, in analogy with (1.1)
- in this case, we may evaluate these random variables[1] on the $\omega_i \in \Omega$ to get
$$\widehat{U}_k(\omega_i) = \widehat{\mathbf{a}}_k^{\mathsf{T}}\mathbf{X}(\omega_i) = \widehat{\mathbf{a}}_k^{\mathsf{T}}\mathbf{x}_i, \qquad \widehat{V}_k(\omega_i) = \widehat{\mathbf{b}}_k^{\mathsf{T}}\mathbf{Y}(\omega_i) = \widehat{\mathbf{b}}_k^{\mathsf{T}}\mathbf{y}_i,$$
which are values in $\mathbb{R}$ called *scores*
- the *score vectors* in the context of CCA are
$$X\mathbf{a} = \begin{bmatrix} \mathbf{a}^{\mathsf{T}}\mathbf{x}_1 \\ \vdots \\ \mathbf{a}^{\mathsf{T}}\mathbf{x}_n \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad Y\mathbf{b} = \begin{bmatrix} \mathbf{b}^{\mathsf{T}}\mathbf{y}_1 \\ \vdots \\ \mathbf{b}^{\mathsf{T}}\mathbf{y}_n \end{bmatrix} \in \mathbb{R}^n$$
where the entries are scores
- sample CCA could of course also be used with sample correlation matrices in place of sample covariance matrices, i.e.,
$$R_X^{-1/2}R_{XY}R_Y^{-1/2} \in \mathbb{R}^{p\times q}$$
in place of
$$S_X^{-1/2}S_{XY}S_Y^{-1/2} \in \mathbb{R}^{p\times q}$$
in Theorem 3
- Theorem 2 also holds for sample CCA and so does (1.2)
- as in the case of population CCA, $\sigma_1^2 \geq \cdots \geq \sigma_p^2$ are the eigenvalues of the matrix
$$G_{XY}G_{XY}^{\mathsf{T}} = S_X^{-1/2}S_{XY}S_Y^{-1}S_{YX}S_X^{-1/2} \in \mathbb{R}^{p\times p}$$
with corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$
- $\sigma_1^2 \geq \cdots \geq \sigma_p^2$ are also the $p$ largest eigenvalues of the matrix
$$G_{XY}^{\mathsf{T}}G_{XY} = S_Y^{-1/2}S_{YX}S_X^{-1}S_{XY}S_Y^{-1/2} \in \mathbb{R}^{q\times q}$$
with corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$
- sometimes, especially in instances when you need calculate things by hand, it can be more covenient to work with the matrices
$$H_{XY} = S_X^{-1/2}G_{XY}G_{XY}^{\mathsf{T}}S_X^{1/2} = S_X^{-1}S_{XY}S_Y^{-1}S_{YX} \in \mathbb{R}^{p\times p}$$
and
$$H_{YX} = S_Y^{-1/2}G_{XY}^{\mathsf{T}}G_{XY}S_Y^{1/2} = S_Y^{-1}S_{YX}S_X^{-1}S_{XY} \in \mathbb{R}^{q\times q}$$
- the eigenvalues of $H_{XY}$ are exactly those of $G_{XY}G_{XY}^{\mathsf{T}}$ and the eigenvalues of $H_{YX}$ are exactly those of $G_{XY}^{\mathsf{T}}G_{XY}$ since these matrices are related by similarity transformations
- to see this note that $A\mathbf{x} = \lambda\mathbf{x}$ iff $(BAB^{-1})(B\mathbf{x}) = \lambda B\mathbf{x}$ so
  (i) $\lambda$ is an eigenvalue of $A$ iff $\lambda$ is an eigenvalue of $BAB^{-1}$

[1] Recall that a random variable is a function.

(ii) $\mathbf{x}$ is a $\lambda$-eigenvector of $A$ iff $B\mathbf{x}$ is a $\lambda$-eigenvector of $BAB^{-1}$

- hence the sample canonical correlations can be calculated without having to compute square roots of matrices, i.e., from the eigenvalues of
$$S_X^{-1}S_{XY}S_Y^{-1}S_{YX} \in \mathbb{R}^{p\times p}, \qquad S_Y^{-1}S_{YX}S_X^{-1}S_{XY} \in \mathbb{R}^{q\times q}$$
- however you should only do this if you're calculating things by hand (like in the first example in the next section)
- as in the case of PCA, when you do sample CCA on a computer, you should always use SVD of $G_{XY}$ and avoid EVD altogether
- note that if you find the sample canonical vectors from the eigenvectors of $H_{XY}$ and $H_{YX}$, you would still need to involve square roots of matrices (why?)
- what we say above apply to population CCA as well, i.e., the population canoical correlations are the eigenvalues of the following square-root free expressions
$$\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} \in \mathbb{R}^{p\times p}, \qquad \Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \in \mathbb{R}^{q\times q}$$
- they also apply to using population or sample correlation matrices in place of covariance matrices

## 3. Example: population CCA

- here is an example where population CCA is useful in yielding insights
- suppose we have $p = q = 2$ and we are interested in population CCA on correlation matrices of a pair of bivariate random vectors $\mathbf{X}$ and $\mathbf{Y}$
- note that the $2\times 2$ correlation matrices must necessarily have the forms
$$P_X = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \qquad P_Y = \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix}$$
- we will just assume, for non-triviality, that $|\alpha| < 1$ and $|\gamma| < 1$
- in which case $\det(P_X) = 1 - \alpha^2 \neq 0$ and so $P_X$ is nonsinuglar, likewise for $P_Y$
- suppose the cross-correlation matrix has the form
$$P_{XY} = \begin{bmatrix} \beta & \beta \\ \beta & \beta \end{bmatrix} = P_{YX}$$
- in this example, we will calculate everything by hand, starting with
$$P_X^{-1}P_{XY} = \frac{\beta}{1-\alpha^2}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{\beta}{1+\alpha}\mathbf{1}\mathbf{1}^{\mathsf{T}} \in \mathbb{R}^{2\times 2}$$
where $\mathbf{1} = [1,1]^{\mathsf{T}} \in \mathbb{R}^2$
- likewise
$$P_Y^{-1}P_{YX} = \frac{\beta}{1+\gamma}\mathbf{1}\mathbf{1}^{\mathsf{T}} \in \mathbb{R}^{2\times 2}$$
- since $(\mathbf{1}\mathbf{1}^{\mathsf{T}})^2 = 2\mathbf{1}\mathbf{1}^{\mathsf{T}}$, we obtain
$$H_{XY} = P_X^{-1}P_{XY}P_Y^{-1}P_{YX} = \frac{2\beta^2}{(1+\alpha)(1+\gamma)}\mathbf{1}\mathbf{1}^{\mathsf{T}}$$
- since the eigenvalues of the matrix $\mathbf{1}\mathbf{1}^{\mathsf{T}}$ are 2 and 0 (check it), the eigenvalues of $H_{XY}$ are
$$\frac{4\beta^2}{(1+\alpha)(1+\gamma)} \quad \text{and} \quad 0$$
- thus the first canonical correlation coefficient is
$$\rho_1 = \frac{2\beta}{\sqrt{(1+\alpha)(1+\gamma)}}$$

- this is the famous Frets's head data set (`fretsheads` in R) comprising measurements of head length and head breath for the first and 2nd sons of 25 families
- so we regard the measurements as observations of the random vectors

$$\mathbf{X} = \begin{bmatrix} \texttt{head length of 1st son} \\ \texttt{head breath of 1st son} \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} \texttt{head length of 2nd son} \\ \texttt{head breath of 2nd son} \end{bmatrix}$$

- so $p = q = 2$, $n = 25$ and with data matrices $X, Y \in \mathbb{R}^{25\times 2}$ and

$$W = [X, Y] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \\ 18 \\ 19 \\ 20 \\ 21 \\ 22 \\ 23 \\ 24 \\ 25 \end{array} \begin{array}{c} \begin{array}{cccc} l_1 & b_1 & l_2 & b_2 \end{array} \\ \begin{bmatrix} 191 & 155 & 179 & 145 \\ 195 & 149 & 201 & 152 \\ 181 & 148 & 185 & 149 \\ 183 & 153 & 188 & 149 \\ 176 & 144 & 171 & 142 \\ 208 & 157 & 192 & 152 \\ 189 & 150 & 190 & 149 \\ 197 & 159 & 189 & 152 \\ 188 & 152 & 197 & 159 \\ 192 & 150 & 187 & 151 \\ 179 & 158 & 186 & 148 \\ 183 & 147 & 174 & 147 \\ 174 & 150 & 185 & 152 \\ 190 & 159 & 195 & 157 \\ 188 & 151 & 187 & 158 \\ 163 & 137 & 161 & 130 \\ 195 & 155 & 183 & 158 \\ 186 & 153 & 173 & 148 \\ 181 & 145 & 182 & 146 \\ 175 & 140 & 165 & 137 \\ 192 & 154 & 185 & 152 \\ 174 & 143 & 178 & 147 \\ 176 & 139 & 176 & 143 \\ 197 & 167 & 200 & 158 \\ 190 & 163 & 187 & 150 \end{bmatrix} \end{array} \in \mathbb{R}^{25\times 4}$$

  where $l_1, b_1, l_2, b_2$ denote respectively the head length of 1st son son, head breath of 1st son son, head length of 2nd son, head breath of 2nd son; the rows are indexed by the family number (arbitrary)
- the sample correlation matrices are

$$R_X = \begin{bmatrix} 1 & 0.7346 \\ 0.7346 & 1 \end{bmatrix}, \quad R_Y = \begin{bmatrix} 1 & 0.8392 \\ 0.8392 & 1 \end{bmatrix}, \quad R_{XY} = \begin{bmatrix} 0.7108 & 0.7040 \\ 0.6932 & 0.7086 \end{bmatrix} = R_{YX}^{\mathsf{T}}$$

- so

$$G_{XY} = R_X^{-1/2} R_{XY} R_Y^{-1/2} = \begin{bmatrix} 0.4217 & 0.3761 \\ 0.3586 & 0.4206 \end{bmatrix}$$

  and its singular values are

$$r_1 = 0.7886, \qquad r_2 = 0.0539$$

- if you want to do this by hand, you can compute the eigenvalues of

$$H_{XY} = R_X^{-1} R_{XY} R_Y^{-1} R_{YX} \in \mathbb{R}^{2\times 2},$$

which are 0.6218 and 0.0029, taking square roots then gives you $r_1$ and $r_2$

- the left and right singluar vectors of $G_{XY}$ are

$$\mathbf{u}_1 = \begin{bmatrix} 0.7149 \\ 0.6992 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 0.6992 \\ -0.7149 \end{bmatrix}, \quad \mathbf{v}_1 = \begin{bmatrix} 0.7003 \\ 0.7138 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0.7138 \\ 0.7003 \end{bmatrix}$$

- so the first pair of sample canonical vectors are

$$\widehat{\mathbf{a}}_1 = R_X^{-1/2}\mathbf{u}_1 = \begin{bmatrix} 0.552 \\ 0.522 \end{bmatrix}, \qquad \widehat{\mathbf{b}}_1 = R_Y^{-1/2}\mathbf{v}_1 = \begin{bmatrix} 0.505 \\ 0.538 \end{bmatrix}$$

- the second pair of sample canonical vectors are

$$\widehat{\mathbf{a}}_2 = R_X^{-1/2}\mathbf{u}_2 = \begin{bmatrix} 1.367 \\ -1.378 \end{bmatrix}, \qquad \widehat{\mathbf{b}}_2 = R_Y^{-1/2}\mathbf{v}_2 = \begin{bmatrix} 1.767 \\ -1.757 \end{bmatrix}$$

- equivalently we may express these as pairs of sample canonical variables and interpret them as follows:

  $\widehat{U}_1 = \widehat{\mathbf{a}}_1^{\mathsf{T}}\mathbf{X} = 0.552 \cdot \texttt{head length of 1st son} + 0.522 \cdot \texttt{head breath of 1st son}$

  is roughly the sum[2] of head length and breadth of the first brother since $0.552 \approx 0.522$

  $\widehat{V}_1 = \widehat{\mathbf{b}}_1^{\mathsf{T}}\mathbf{Y} = 0.505 \cdot \texttt{head length of 2nd son} + 0.538 \cdot \texttt{head breath of 2nd son}$

  is roughly the sum of head length and breadth of the second brother since $0.505 \approx 0.538$

  $\widehat{U}_2 = \widehat{\mathbf{a}}_2^{\mathsf{T}}\mathbf{X} = 1.367 \cdot \texttt{head length of 1st son} - 1.378 \cdot \texttt{head breath of 1st son}$

  is roughly the difference[3] of head length and breadth of the first brother since $1.367 \approx 1.378$

  $\widehat{V}_2 = \widehat{\mathbf{b}}_2^{\mathsf{T}}\mathbf{Y} = 1.767 \cdot \texttt{head length of 2nd son} - 1.757 \cdot \texttt{head breath of 2nd son}$

  is roughly the difference of head length and breadth of the second brother since $1.767 \approx 1.757$

- we will call the sum of length and breadth *girth* and the difference of length and breadth *shape*
- since $r_1 = 0.7886 \approx 0.8$, the head girths of first and second brothers are highly correlated
- since $r_2 = 0.0539 \approx 0.05$, the head shapes of first and second brothers has little correlation

## 5. Real-world example from Luis Moreira da Costa

- this is an example from a former student who encountered it in his work (many thanks to him for sharing it)

```
We wanted to calculate a proxy for the commodities part of Chinese GDP - a
Chinese commodities GDP, if you will. We tried to do this using CCA. On one
side, we gathered a variety of Chinese economic activity numbers related to
commodities (industrial production, some surveys, construction figures,
etc...). This would be our X_1, X_2,..., X_n. On the other side, we had
the prices of five commodities. This was our Y_1, Y_2,..., Y_n. The
variables in both X and Y are time-series, naturally. They were both monthly
in frequency, and were log-differenced to ensure stationarity. We then ran a
CCA on these two sets. The first canonical covariate related to X (U_1) was
our proxy.
```

- very nice and not as surprising as it first appears — my favorite example along the same lines is `garbage.pdf` in the handouts folder

[2] Strictly speaking a constant multiple of the sum.
[3] Strictly speaking a constant multiple of the difference.