STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 7: Canonical Correlation Analysis — Foundations

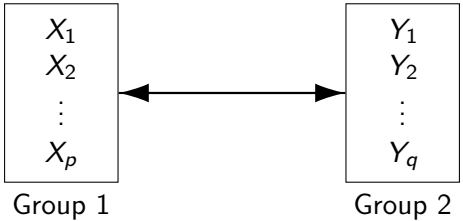
Jingshu Wang

The University of Chicago

Outline

From structure within variables to relationships between variables

- So far, we have focused on structure within a single set of variables:
  - PCA: directions of large variation
  - Factor analysis: latent structure
  - Clustering / mixtures: grouping observations
- But in many problems, variables naturally split into two groups.



- Goal: find linear combinations of X and Y that are maximally correlated.

Motivating Example 1: Macro variables vs asset returns

Suppose each time period i contains two vectors:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) = \text{macro variables}, \qquad \mathbf{y}_i = (y_{i1}, \ldots, y_{iq}) = \text{asset returns}.$$

Examples of variables:

- $\mathbf{x}_i$: inflation, interest rates, unemployment, GDP growth;
- $\mathbf{y}_i$: stock market return, bond returns, sector returns.

CCA idea:

- Find

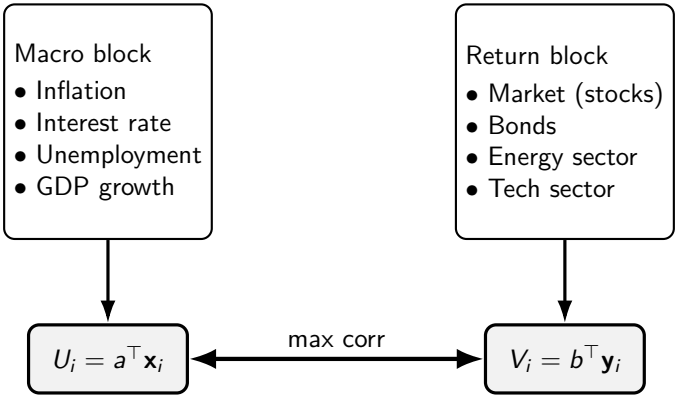
$$U_i = a^\top \mathbf{x}_i, \qquad V_i = b^\top \mathbf{y}_i$$

  so that U_i and V_i are as correlated as possible.

Interpretation:

- U_i: a macro index;
- V_i: a market-return index;
- CCA finds the strongest link between the two blocks.

Motivating Example 1: Macro variables vs asset returns



Motivating Example 2: Test scores in two domains

Suppose each student i has two sets of scores:

$$\mathbf{x}_i = \text{math-related scores}, \qquad \mathbf{y}_i = \text{verbal-related scores}.$$

Examples of variables:

- $\mathbf{x}_i$: algebra, geometry, calculus;
- $\mathbf{y}_i$: reading, writing, vocabulary.

CCA idea:

- Construct

$$U_i = a^\top \mathbf{x}_i, \qquad V_i = b^\top \mathbf{y}_i$$

  so that the two summaries are maximally correlated.

Interpretation:

- First CCA pair may reflect general academic strength;
- later pairs may reflect differences in math vs verbal profile.
- CCA replaces many pairwise correlations with a few paired scores.

The idea of CCA

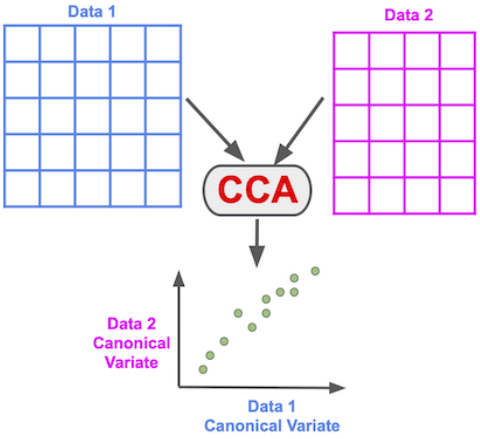


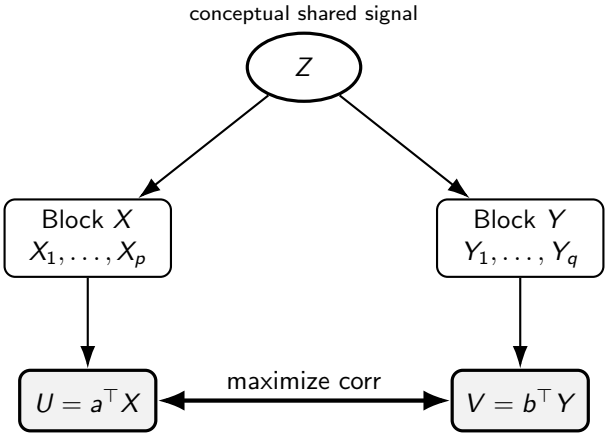
Figure: Source: `https://zia207.quarto.pub/canonical-correlation%20-analysis.html`

Why maximize correlation?

- we expect the two blocks to share some underlying signal;
- each block contains a noisy, high-dimensional view of that signal;
- CCA finds linear summaries that best capture shared variation.

A useful intuition is that both blocks may be influenced by an underlying factor Z. However, CCA does *not* generally recover the true latent variable Z. It finds maximally correlated linear combinations of X and Y.

A useful intuition of CCA



Notation and covariance partition

Let the centered random vectors be

$$X \in \mathbb{R}^p, \qquad Y \in \mathbb{R}^q,$$

with joint covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}.$$

- Σ_{XX}: within-X covariance.
- Σ_{YY}: within-Y covariance.
- $\Sigma_{XY} = \Sigma_{YX}^\top$: cross-covariance between the two blocks.

CCA searches for linear combinations (canonical variates)

$$U = \mathbf{a}^\top X, \qquad V = \mathbf{b}^\top Y,$$

whose correlation is as large as possible.

First canonical variates: optimization problem

Choose weight vectors $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ to maximize

$$\mathrm{Corr}(U, V) = \mathrm{Corr}(\mathbf{a}^\top X, \mathbf{b}^\top Y) = \frac{\mathbf{a}^\top \Sigma_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{XX} \mathbf{a}} \sqrt{\mathbf{b}^\top \Sigma_{YY} \mathbf{b}}}.$$

Because correlation is scale-invariant, we can impose normalization constraints:

$$\max_{\mathbf{a},\mathbf{b}} \ \rho \stackrel{\Delta}{=} \mathbf{a}^\top \Sigma_{XY} \mathbf{b} \quad \text{s.t.} \quad \mathbf{a}^\top \Sigma_{XX} \mathbf{a} = 1, \ \mathbf{b}^\top \Sigma_{YY} \mathbf{b} = 1.$$

Reading the constraints:

- $\mathrm{Var}(U) = 1$ and $\mathrm{Var}(V) = 1$;
- after standardizing the two scores, maximize their covariance = correlation.

Why the constraints matter

Without constraints, the objective

$$\mathbf{a}^\top \Sigma_{XY} \mathbf{b}$$

can be made arbitrarily large by replacing $(\mathbf{a}, \mathbf{b})$ with $(c\mathbf{a}, d\mathbf{b})$.

The CCA constraints therefore play the same role as the unit-norm constraint in PCA:

- they remove trivial rescaling;
- they put $U = \mathbf{a}^\top X$ and $V = \mathbf{b}^\top Y$ on a common variance scale;
- they make the optimization target interpretable as a correlation.

Key difference from PCA:

- PCA uses one block and maximizes variance;
- CCA uses two blocks and maximizes cross-block correlation.

Lagrangian and first-order conditions

We maximize

$$\rho = \mathbf{a}^\top \Sigma_{XY} \mathbf{b} \quad \text{subject to} \quad \mathbf{a}^\top \Sigma_{XX} \mathbf{a} = 1,\ \mathbf{b}^\top \Sigma_{YY} \mathbf{b} = 1.$$

Lagrangian:

$$\mathcal{L}(\mathbf{a}, \mathbf{b}, \lambda, \mu) = \mathbf{a}^\top \Sigma_{XY} \mathbf{b} - \frac{\lambda}{2}(\mathbf{a}^\top \Sigma_{XX} \mathbf{a} - 1) - \frac{\mu}{2}(\mathbf{b}^\top \Sigma_{YY} \mathbf{b} - 1).$$

First-order conditions:

$$\Sigma_{XY} \mathbf{b} = \lambda \Sigma_{XX} \mathbf{a}, \qquad \Sigma_{YX} \mathbf{a} = \mu \Sigma_{YY} \mathbf{b}.$$

Multiply the first equation by $\mathbf{a}^\top$:

$$\mathbf{a}^\top \Sigma_{XY} \mathbf{b} = \rho = \lambda \mathbf{a}^\top \Sigma_{XX} \mathbf{a} = \lambda.$$

Multiply the second equation by $\mathbf{b}^\top$:

$$\mathbf{b}^\top \Sigma_{YX} \mathbf{a} = \rho = \mu \mathbf{b}^\top \Sigma_{YY} \mathbf{b} = \mu.$$

From multipliers to canonical correlation

Assume Σ_{XX} and Σ_{YY} are invertible, multiply the first equation of FOC by Σ_{XX}^{-1}:

$$\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b} = \rho\mathbf{a}.$$

Multiply the second equation of FOC by Σ_{YY}^{-1}:

$$\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} = \rho\mathbf{b}.$$

Substitute back:

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} = \rho^2\mathbf{a},$$

$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b} = \rho^2\mathbf{b}.$$

Conclusion: Squared canonical correlations are eigenvalues.

How to obtain the first canonical correlation:

- solve the eigenvalue problem

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} = \rho^2\mathbf{a};$$

- the largest eigenvalue $\lambda_{\max}$ gives

$$\rho_1 = \sqrt{\lambda_{\max}}.$$

Why this works:

- no need to use both matrices: The matrices

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \quad \text{and} \quad \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

  share the same nonzero eigenvalues.

- although not symmetric, this matrix has the same eigenvalues as

$$KK^\top, \quad K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2},$$

  which is symmetric positive semidefinite.

Whitened view: understanding the first CCA

Define whitened variables

$$\tilde{X} = \Sigma_{XX}^{-1/2} X, \qquad \tilde{Y} = \Sigma_{YY}^{-1/2} Y.$$

Then both have identity covariance, and their cross-covariance becomes

$$K = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}.$$

Key idea:

- CCA reduces to finding linear combinations of $\tilde{X}$ and $\tilde{Y}$ with maximal correlation;
- this is equivalent to finding directions (satisfying $\mathbf{u}^\top \mathbf{u} = \mathbf{v}^\top \mathbf{v} = 1$) that maximize

$$\mathbf{u}^\top K \mathbf{v}.$$

Result:

- the maximum correlation is the largest singular value of K: $\rho_1 = \sigma_{\max}(K)$;
- Directions: $\mathbf{a} = \Sigma_{XX}^{-1/2} \mathbf{u}_1$ and $\mathbf{b} = \Sigma_{YY}^{-1/2} \mathbf{v}_1$ where $\mathbf{u}_1, \mathbf{v}_1$ are the first left/right singular vectors of K.
- canonical variates: $U_1 = \mathbf{u}_1^\top \tilde{X}$, $V_1 = \mathbf{v}_1^\top \tilde{Y}$.

Successive canonical pairs: CCA as an SVD problem

After obtaining the first pair (U_1, V_1), CCA looks for new directions that:

- require orthogonality within each block:

$$\mathrm{Cov}(U_k, U_\ell) = 0, \qquad \mathrm{Cov}(V_k, V_\ell) = 0, \qquad k \neq \ell;$$

- and maximize remaining cross-block correlation.

Key fact (from SVD):

- $K = PDQ^\top$, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots$;
- the k-th canonical correlation is

$$\rho_k = \sigma_k(K);$$

- corresponding directions are given by columns of $P = (\mathbf{u}_1, \mathbf{u}_2, \dots)$ and $Q = (\mathbf{v}_1, \mathbf{v}_2, \dots)$.

Conclusion:

- $\rho_1 \geq \rho_2 \geq \cdots \geq 0$;
- at most $\min(p, q)$ nonzero canonical correlations;
- each pair reveals a new mode of association: $U_k = \mathbf{u}_k^\top \tilde{X}$, $V_k = \mathbf{v}_k^\top \tilde{Y}$.

What is orthogonal to what?

A common source of confusion:

- Within the X side, different canonical variates are uncorrelated:

$$\mathrm{Cov}(U_k, U_\ell) = \mathbf{u}_k^\top \mathrm{Cov}(\tilde{X})\mathbf{u}_\ell = \mathbf{u}_k^\top \mathbf{u}_\ell = 0 \quad (k \neq \ell).$$

- Within the Y side, the same is true:

$$\mathrm{Cov}(V_k, V_\ell) = \mathbf{v}_k^\top \mathrm{Cov}(\tilde{Y})\mathbf{v}_\ell = \mathbf{v}_k^\top \mathbf{v}_\ell = 0 \quad (k \neq \ell).$$

- Across sides, only matched pairs are generally nonzero:

$$\mathrm{Corr}(U_k, V_k) = \mathbf{u}_k^\top K \mathbf{v}_k = \rho_k,$$

$$\mathrm{Cov}(U_k, V_\ell) = \mathbf{u}_k^\top K \mathbf{v}_\ell = 0 \text{ for } k \neq \ell.$$

Interpretation: CCA creates paired latent axes linking the two blocks one mode at a time.

Weights, loadings, and cross-loadings

There are several different objects to interpret:

- **Canonical weights**: $\mathbf{a}_k = \Sigma_{XX}^{-1/2}\mathbf{u}_k$ and $\mathbf{b}_k = \Sigma_{YY}^{-1/2}\mathbf{v}_k$.
- **Canonical variates (scores)**: $U_k = \mathbf{a}_k^\top X = \mathbf{u}_k^\top \tilde{X}$ and $V_k = \mathbf{b}_k^\top Y = \mathbf{v}_k^\top \tilde{Y}$.
- **Structure correlations / loadings**: correlations between original variables and their own canonical variate.

$$\mathrm{Corr}(X_j, U_k), \qquad \mathrm{Corr}(Y_\ell, V_k).$$

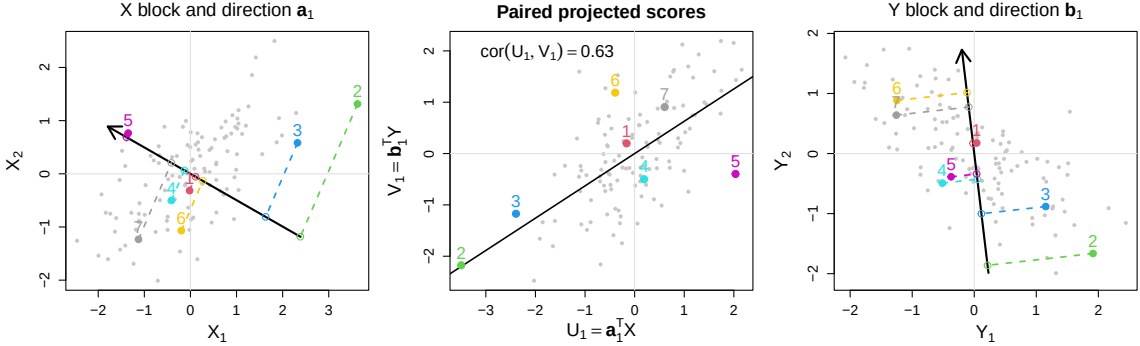
- **Cross-loadings**: correlations with the opposite-side canonical variate.

$$\mathrm{Corr}(X_j, V_k), \qquad \mathrm{Corr}(Y_\ell, U_k).$$

Weights define the direction; loadings tell us how variables align with it.

Geometry of CCA

CCA chooses directions so paired scores are maximally associated



PCA vs FA vs CCA

PCA

- one block;
- maximize variance;
- orthogonal directions.

FA

- one block;
- covariance model;
- latent factors + idiosyncratic noise;
- interpretation via rotation.

CCA

- two blocks;
- maximize cross-block correlation;
- paired canonical variates;
- symmetric in X and Y.

Bridge idea: PCA summarizes one cloud; CCA summarizes *how two clouds move together*.

Sample CCA

Given centered data matrices

$$\mathbf{X}_c \in \mathbb{R}^{n\times p}, \qquad \mathbf{Y}_c \in \mathbb{R}^{n\times q},$$

form sample covariance blocks

$$\mathbf{S}_{XX} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{X}_c, \quad \mathbf{S}_{YY} = \frac{1}{n-1}\mathbf{Y}_c^\top \mathbf{Y}_c, \quad \mathbf{S}_{XY} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{Y}_c.$$

Then the sample version solves

$$\max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^\top \mathbf{S}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{S}_{XX}\mathbf{a}}\sqrt{\mathbf{b}^\top \mathbf{S}_{YY}\mathbf{b}}}.$$

This leads to the same eigenvalue/SVD forms as in the population case, with Σ replaced by $\mathbf{S}$.

Mini example: two blocks of student skills

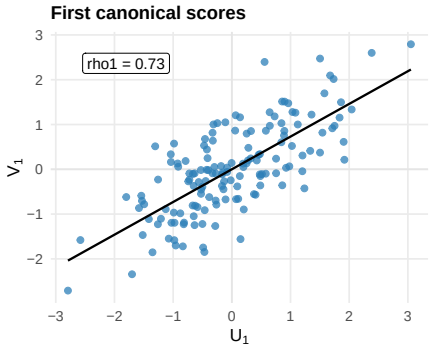
Suppose each student has two groups of variables:

- X block (quantitative): algebra, calculus, mechanics;
- Y block (communication): writing, reading, presentation.

We simulate a small dataset, and the first canonical correlation is $\rho_1 \approx 0.73$

Combinations:

- $U_1 = 0.28\,\text{algebra} + 0.35\,\text{calculus} + 0.45\,\text{mechanics}$, summarizes quantitative strength;
- $V_1 = 0.32\,\text{writing} + 0.36\,\text{reading} + 0.35\,\text{presentation}$, summarizes communication strength;
- students strong in one profile also tend to be strong in the other.

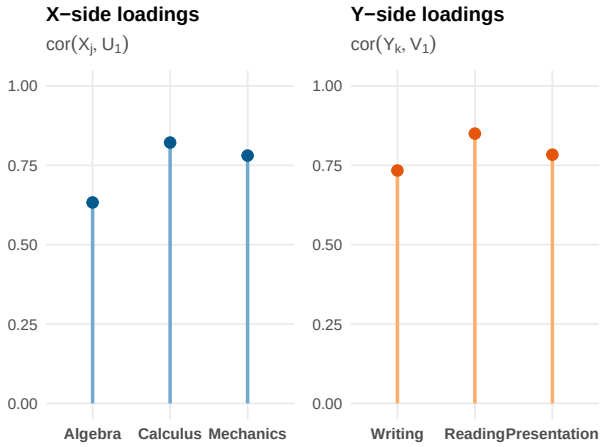


How to interpret the first canonical pair

- the X-side loadings show which quantitative variables align with U_1;
- the Y-side loadings show which communication variables align with V_1.

Main message

- CCA is usually interpreted through the canonical scores and loadings;
- loadings often tell a cleaner story than raw weights.
  - The absolute values of weights are not meaningful.



Pairwise correlations can miss the bigger pattern

	Writing	Reading	Presentation
Algebra	0.52	0.41	0.18
Calculus	0.48	0.46	0.48
Mechanics	0.27	0.51	0.55

No single pairwise correlation is especially large, but the pattern is consistent across many variable pairs.

- largest pairwise correlation above: 0.55;
- first canonical correlation: $\rho_1 \approx 0.73$;
- CCA combines many moderate relationships into one stronger cross-block summary.

Lecture 5 summary

- CCA studies relationships between two variable blocks by finding paired linear combinations with maximal correlation.
- The optimization problem leads to generalized eigenvalue equations, or equivalently, an SVD of a whitened cross-covariance matrix.
- Canonical correlations come in decreasing order and define successive modes of association.
- Interpretation requires more than weights alone: use loadings and score plots.

Suggested reading

- Johnson & Wichern (6th Edition), Chapter 10.1-10.4.