STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 8: CCA, multivariate regression, and RRR

Jingshu Wang

The University of Chicago

Outline

Lecture 7 recap: what CCA was trying to do

Given two centered random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, CCA finds linear combinations

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

that maximizes correlation.

$$\max_{a,b} \ \mathrm{Corr}(a^\top X, b^\top Y)$$

subject to variance-normalization and orthogonality constraints for later pairs.

- CCA is **symmetric** in the two blocks.
- It answers: *which linear summaries of X and Y are most strongly associated?*
- It does *not* directly optimize prediction error of Y from X.

Preprocessing matters for CCA

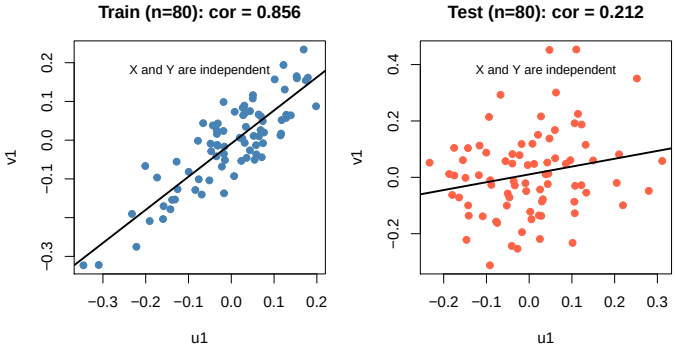
CCA depends entirely on covariance structure.

- Centering is essential.
- Standardization does not affect canonical correlations, variates, or loadings, but changes the weights and may influence numerical stability.
- Outliers may strongly distort canonical directions.
- Missing data must be handled beforehand.

Practical tip:

- center the data and check for outliers;
- assess stability of canonical correlations and directions;
- be mindful of multicollinearity and small sample issues.

Overfitting in CCA: a simple example



- X and Y are independent. Here $p = q = 20$, we split 160 observations into 80 training and 80 test observations.
- Training correlation is large; test correlation is much smaller.

CCA can find spurious relationships that do not generalize.

Why may CCA overfit?

Why does this happen?

- CCA searches over many linear combinations.
- With moderate or high dimension, noise can look structured.

Consequences

- sample canonical correlations can be misleadingly large;
- estimates become unstable when p, q are not small relative to n.

Takeaway: large canonical correlation $\neq$ real relationship.

How many canonical pairs?

Fewer standard heuristics than PCA.

- For instance, canonical correlations do not partition total variance, so there is no PVE / scree rule.

A practical workflow:

1. Look at the canonical correlations $\rho_1, \rho_2, \ldots$
2. Check stability: does the correlation persist out-of-sample?
3. Interpret loadings for the first few pairs

Optional:

- formal significance tests (require strong assumptions)

Key point: statistical significance $\neq$ practical importance.

An R demo of CCA

Notebook file: `Lecture8_demo.html`, part I

We see in the demo that CCA does not target prediction.

Question: What if our goal is to predict Y from X?

Why move beyond CCA?

Suppose we observe predictors X and responses Y.

Two different goals:

1. **Association goal**: what joint linear structure do the two blocks share?
2. **Prediction goal**: how well can we predict Y from X with a low-dimensional summary?

CCA is ideal for Goal 1.

For Goal 2, we usually want an asymmetric method:

- X plays the role of input/predictor;
- Y plays the role of output/response;
- predictive fit matters.

Multivariate linear regression as a matrix model

For centered data matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$,

$$Y = XB + E,$$

where

- $B \in \mathbb{R}^{p \times q}$ is the coefficient matrix;
- $E \in \mathbb{R}^{n \times q}$ is the noise matrix.

The least-squares estimator minimizes

$$\min_B \ \|Y - XB\|_F^2.$$

If $X^\top X$ is invertible,

$$\hat{B}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y.$$

Interpretation: each column of Y is predicted from the same predictor block X.

CCA and regression share the same building blocks

Both methods are based on the same covariance matrices:

$$S_{XX}, \quad S_{YY}, \quad S_{XY}.$$

Multivariate regression (after centering X and Y):

$$\hat{B}_{\text{OLS}} = S_{XX}^{-1} S_{XY}.$$

CCA is based on the SVD of the whitened cross-covariance:

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

- Both use the same cross-covariance S_{XY}.
- Both try to capture relationships between X and Y.

But they use it in different ways.

Same ingredients, but different goals: association vs. prediction.

Connection: when Y is one-dimensional

Suppose Y is scalar.

CCA objective

$$\max_{w} \ \mathrm{Corr}(Xw, Y)$$

Since $\mathrm{Var}(Y)$ is fixed, this is equivalent to

$$\max_{w} \ \frac{\mathrm{Cov}(Xw, Y)^2}{\mathrm{Var}(Xw)} = \max_{w} \ \frac{(w^\top S_{XY})^2}{w^\top S_{XX} w}.$$

Solution is

$$w \propto S_{XX}^{-1} S_{XY}.$$

Compare with OLS: regression of Y on X gives

$$\hat{\beta}_{\mathrm{OLS}} = S_{XX}^{-1} S_{XY}.$$

Conclusion:CCA and regression produce the same direction when Y is scalar.

CCA and regression answer different questions

CCA

- symmetric in X and Y;
- normalizes both sides;
- maximizes correlation between linear combinations;
- extracts paired low-dimensional summaries;
- focuses on shared structure.

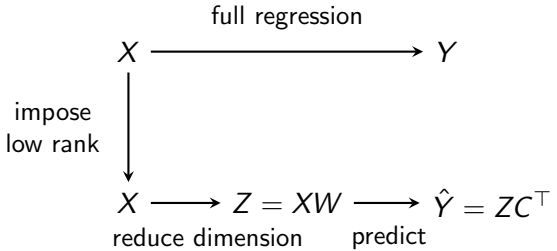
Multivariate regression

- asymmetric: $X \to Y$;
- minimizes squared prediction error;
- coefficients depend on the scale of Y;
- fits a full linear map from X to Y;
- focuses on prediction.

CCA finds low-dimensional structure in association; multivariate regression uses all directions for prediction.

From regression to reduced-rank regression

- **CCA:** finds low-dimensional directions that capture shared variation between X and Y.
- **Regression:** uses all directions in X to predict Y.
- **Idea:** can we predict Y using only a few informative directions of X?



Reduced-Rank Regression (RRR): combine dimension reduction (like CCA) with prediction (like regression).

Reduced-rank regression: two equivalent views

Low-dimensional representation view

- Compress predictors into r scores:

$$Z = XA, \quad A \in \mathbb{R}^{p \times r}$$

- Then predict:

$$Y \approx ZC^\top$$

Equivalent formulation

$$Y \approx XAC^\top = XB, \quad \text{where } B = AC^\top$$

- This implies $\operatorname{rank}(B) \leq r$.

Reduced-rank regression (RRR)

$$\min_B \ \|Y - XB\|_F^2 \quad \text{s.t.} \quad \operatorname{rank}(B) \leq r.$$

Low-dimensional predictors $\Longleftrightarrow$ low-rank coefficient matrix.

How to compute reduced-rank regression

$$\min_{B:\ \mathrm{rank}(B)\le r} \|Y - XB\|_F^2$$

Key reduction:

$$\|Y - XB\|_F^2 = \|\hat{Y}_{\mathrm{OLS}} - XB\|_F^2 + \text{constant}$$

So we approximate $\hat{Y}_{\mathrm{OLS}}$ by a rank-r matrix.

Solution:

- SVD: $\hat{Y}_{\mathrm{OLS}} = UDV^\top$
- Keep top r: $\hat{Y}_{\mathrm{RRR}} = U_r D_r V_r^\top = \hat{Y}_{\mathrm{OLS}} V_r V_r^\top$

Resulting coefficient matrix:

$$\hat{B}_{\mathrm{RRR}} = \hat{B}_{\mathrm{OLS}} V_r V_r^\top$$

RRR = best low-rank approximation of $\hat{Y}_{\mathrm{OLS}}$.

What is RRR doing?

- OLS predicts Y using all directions.
- RRR keeps only the top r directions in the predicted responses.

Interpretation:

- Find a low-dimensional subspace of Y that is predictable from X.
- Project $\hat{Y}_{\text{OLS}}$ onto that subspace.

RRR = PCA applied to the fitted values $\hat{Y}_{\text{OLS}}$ from X to Y, not to raw Y nor X.

Connection between CCA and RRR

Both methods are based on the same core matrix:

$$S_{YX} S_{XX}^{-1} S_{XY}.$$

- RRR (after centering) uses its eigenvectors directly:

$$\hat{Y}_{\text{OLS}}^{\top} \hat{Y}_{\text{OLS}} = S_{YX}(S_{XX})^{-1} X^{\top} X (S_{XX})^{-1} S_{XY} = S_{YX} S_{XX}^{-1} S_{XY}.$$

- CCA solves a generalized eigenproblem:

$$S_{YX} S_{XX}^{-1} S_{XY} v = \lambda S_{YY} v.$$

- Difference: CCA normalizes by S_{YY}^{-1}.
- If Y is whitened ($S_{YY} = I$), the two methods coincide.

RRR focuses on prediction; CCA focuses on scale-free association.

When should I prefer RRR to CCA?

Prefer reduced-rank regression when:

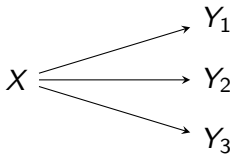
- Y is clearly the response block and prediction matters;
- responses are correlated and may share a few latent predictive factors;
- you want dimension reduction tailored to supervised prediction.

Prefer CCA when:

- the two blocks play symmetric scientific roles;
- the main goal is interpretation of shared structure;
- scale-normalized association is more important than prediction error.

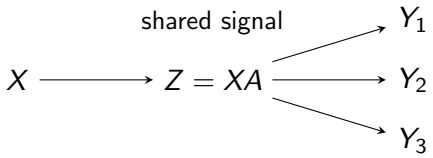
RRR v.s. multivariate regression

Multivariate regression



- separate models for each response
- ignores shared structure
- many parameters $\Rightarrow$ may overfit noise

Reduced-rank regression



- shared low-dimensional predictors
- links responses through Z
- reduces variance and improves prediction

RRR improves prediction by exploiting shared structure in Y.

Choosing the rank via cross-validation

Goal: estimate out-of-sample prediction error for each rank r.

K-fold cross-validation:

1. Split the data into K folds.
2. For each fold k:
   - fit RRR with rank r on the training data;
   - predict responses on the held-out fold.
3. Compute prediction error:
$$\| Y_{\text{test}} - \hat{Y}_{\text{test}} \|_F^2$$
4. Average error across folds.

Final step:

- repeat for different r;
- choose r with smallest CV error.

High-dimensional setting: opportunity and challenges

In modern data, p and/or q can be large relative to n.

Opportunity:

- High-dimensional Y often has shared low-dimensional structure.
- RRR exploits this by learning a few predictive factors.

Challenge:

- Estimation becomes unstable when p or q is large.
- Sample covariance matrices may be ill-conditioned or singular.

Potential fixes (generalize ideas will be discussed later in the course):

- add regularization (ridge, sparsity) on B to stabilize estimation;
- impose additional structural assumptions on A or C where $B = AC^T$.

RRR is especially useful in high dimensions, but often needs regularization.

Live R notebook: comparing CCA with multivariate regression and RRR

Notebook file: `Lecture8_demo.html`, part II

Lecture 8 summary

1. CCA studies shared linear association between two variable blocks.
2. Multivariate regression predicts Y from X and is inherently asymmetric.
3. Reduced-rank regression imposes a low-dimensional predictive structure on the coefficient matrix.
4. The right method depends on whether the scientific question is about association, prediction, or both.