STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 2: Principal Component Analysis: Foundations

Jingshu Wang

The University of Chicago

Outline

1. Bridge from Lecture 1 + Motivation
2. PCA Optimization Formulation
3. Variance, Geometry, and SVD
4. Practical Decisions + Financial Example
5. Wrap-up

Bridge from Lecture 1: projection language and notation

- Observed data matrix: $\mathbf{X} \in \mathbb{R}^{n \times p}$.
- Centered matrix: $\mathbf{X}_c$ (each column mean removed).
- A projection direction is denoted by $\boldsymbol{v} \in \mathbb{R}^p$ with $\|\boldsymbol{v}\|_2 = 1$.

Projected score (observation i):

$$z_i = \boldsymbol{x}_{c,i}^\top \boldsymbol{v} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{v}.$$

Key question for today:

Among infinitely many unit directions $\boldsymbol{v}$, how to choose the most informative ones?

Population PCA vs Sample PCA

Population version (theoretical target)

- Random vector $X \in \mathbb{R}^p$ with mean $\boldsymbol{\mu}$ and covariance Σ.
- Population PCs aims to maximize the true variance $\mathrm{Var}(X^\top \boldsymbol{v}) = \boldsymbol{v}^\top \Sigma \boldsymbol{v}$.

Sample version (what we actually compute)

- Observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and centered matrix X_c.
- Sample covariance:

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}_c^\top \mathbf{X}_c.$$

- Sample PCs maximizes the sample variance $\widehat{\mathrm{Var}}(X^\top \boldsymbol{v}) = \boldsymbol{v}^\top \mathbf{S} \boldsymbol{v}$.

Key idea: sample PCA is an estimate of population PCA.

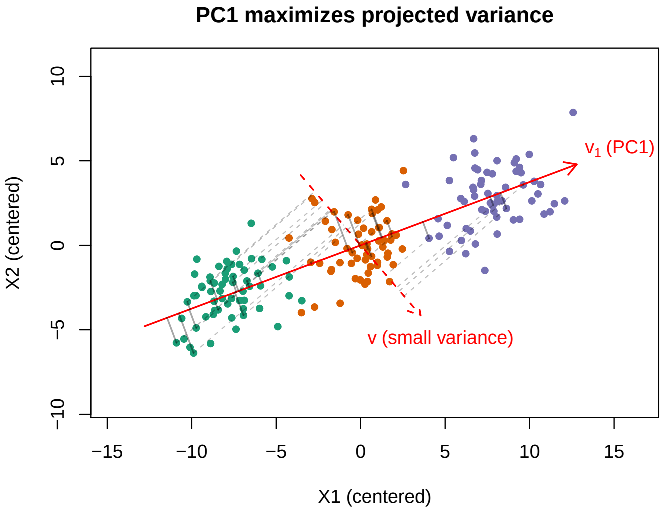
Why maximize variance?

- A direction with very small projected variance makes data look almost flat.
- A direction with large projected variance reveals major data heterogeneity.
- If we keep only a few components, we want them to retain as much information as possible.

Compression perspective:

- PCA chooses directions that minimize information loss under squared reconstruction error.
- So “maximum variance” and “best low-rank approximation” are two sides of the same coin.

Data cloud and first PC direction



Sample PCA optimization for the first component

For a unit direction $\boldsymbol{v}$, sample projected variance is

$$\widehat{\mathrm{Var}}(X^\top \boldsymbol{v}) = \boldsymbol{v}^\top \mathbf{S} \boldsymbol{v}.$$

Hence, the first sample PC solves

$$\max_{\boldsymbol{v} \in \mathbb{R}^p} \ \boldsymbol{v}^\top \mathbf{S} \boldsymbol{v} \quad \text{s.t.} \quad \|\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\top \boldsymbol{v} = 1.$$

Interpretation:

- Objective: maximize spread in projected scores.
- Constraint: avoid trivial scaling ($c\boldsymbol{v}$ would otherwise make objective arbitrarily large).

How do we find the solution?

Lagrangian method (for constrained optimization):

$$\mathcal{L}(\boldsymbol{v}, \lambda) = \boldsymbol{v}^\top \mathbf{S} \boldsymbol{v} - \lambda(\boldsymbol{v}^\top \boldsymbol{v} - 1).$$

- We combine objective + constraint using multiplier λ.

first-order condition (FOC):

- At an optimum, derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{v}$ must be zero:

$$\nabla_{\boldsymbol{v}} \mathcal{L} = 2\mathbf{S}\boldsymbol{v} - 2\lambda \boldsymbol{v} = 0.$$

- So $\mathbf{S}\boldsymbol{v} = \lambda \boldsymbol{v}$: optimal directions are eigenvectors of $\mathbf{S}$.

To maximize $\boldsymbol{v}^\top \mathbf{S} \boldsymbol{v} = \lambda \boldsymbol{v}^\top \boldsymbol{v} = \lambda$, we take the largest eigenvalue λ_1.

For PC1: principal components, loadings, and scores

For the first component:

- **Principal component**: the eigenvector $\boldsymbol{v}_1$ that satisfies $\mathbf{S}\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$.
  - $\boldsymbol{v}_1$ may not be unique. (Why?)
- **Loading**: entries of $\boldsymbol{v}_1$; they show how original variables contribute to PC k.
- **Scores**: projected coordinates of observations on PC direction,

$$z_{i1} = \boldsymbol{x}_{c,i}^\top \boldsymbol{v}_1.$$

Interpretation: loadings describe variables; scores describe observations.

Why do we need a 2nd component? Why orthogonal?

Why 2nd component?

- PC1 is one-dimensional summary only.
- Residual variation remains after removing PC1 information.
- PC2 captures the strongest remaining variation.

Why orthogonality to PC1?

- To avoid rediscovering the same direction/information.
- Orthogonality ensures non-redundant factors in Euclidean geometry.
- With centered data, orthogonal loading vectors imply uncorrelated PC scores.

Optimization for PC2:

$$\max_{\boldsymbol{v}} \ \boldsymbol{v}^\top \mathbf{S} \boldsymbol{v} \quad \text{s.t.} \quad \|\boldsymbol{v}\|_2 = 1, \ \boldsymbol{v}^\top \boldsymbol{v}_1 = 0.$$

How to find PC2? (short derivation)

let $\mathbf{S} = V \Lambda V^\top$ with

$$V = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_p], \quad V^\top V = V V^\top = I_p,$$

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p), \quad \lambda_1 \geq \cdots \geq \lambda_p \geq 0.$$

So V is the orthogonal matrix whose columns are exactly the eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_p$.

- Expand any unit $\boldsymbol{v}$ orthogonal to $\boldsymbol{v}_1$ in eigenbasis:

$$\boldsymbol{v} = \sum_{j=2}^{p} c_j \boldsymbol{v}_j, \quad \sum_{j=2}^{p} c_j^2 = 1.$$

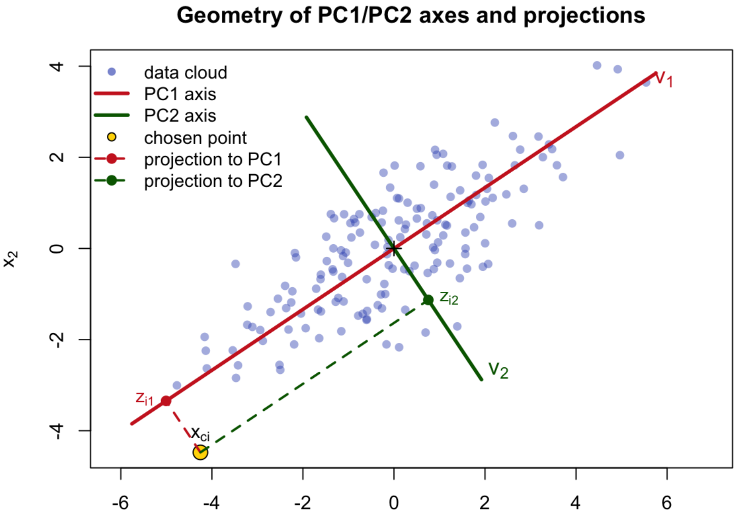
- Then

$$\boldsymbol{v}^\top \mathbf{S} \boldsymbol{v} = \sum_{j=2}^{p} \lambda_j c_j^2 \leq \lambda_2.$$

- Equality holds at $\boldsymbol{v} = \boldsymbol{v}_2$.

Therefore PC2 is an eigenvector for λ_2 (and similarly for later PCs).

Geometry of PC1 and PC2 axes



Total variance and explained variance

For centered data $\boldsymbol{x}_{c,1}, \ldots, \boldsymbol{x}_{c,p}$:

$$\text{Total sample variance} = \sum_{j=1}^{p} s_{jj} = \mathrm{tr}(\mathbf{S}).$$

- Because diagonal entries s_{jj} are sample variances of each variable.

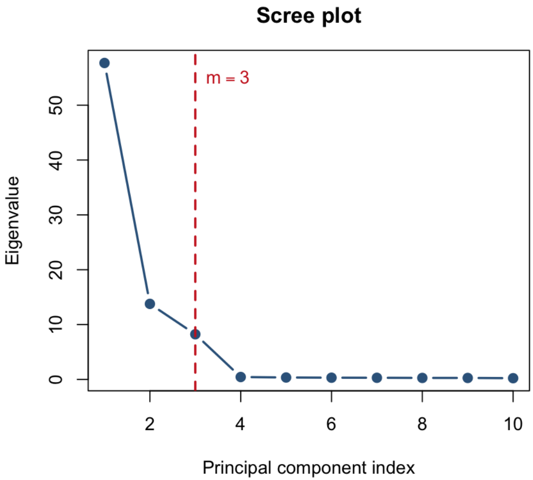
Using spectral decomposition $\mathbf{S} = V \Lambda V^\top$:

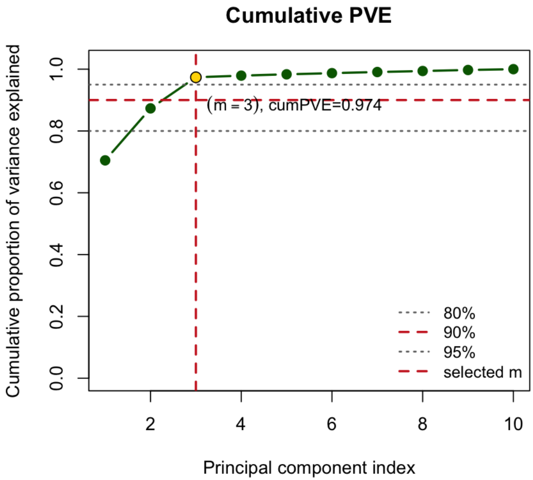
$$\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\Lambda) = \sum_{j=1}^{p} \lambda_j.$$

So variance explained by the first m PCs is

$$\mathrm{cumPVE}(m) = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

How many PC to select? scree plot + cumulative PVE





Review of SVD

For centered data matrix $\mathbf{X}_c \in \mathbb{R}^{n \times p}$ with rank $r \leq \min(n, p)$:

$$\mathbf{X}_c = UDV^\top,$$

where (thin SVD form)

$$U \in \mathbb{R}^{n \times r}, \quad D \in \mathbb{R}^{r \times r}, \quad V \in \mathbb{R}^{p \times r}.$$

- $U^\top U = I_r$, $V^\top V = I_r$ (orthonormal columns).
- $D = \mathrm{diag}(d_1, \ldots, d_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ (singular values).
- Equivalent full SVD uses $U \in \mathbb{R}^{n \times n}$, $D \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$.

Connection between SVD and PCA

Using $\mathbf{X}_c = UDV^\top$,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{X}_c = \frac{1}{n-1} VD^2V^\top.$$

- So columns of V are eigenvectors of $\mathbf{S}$ (PC directions / loadings).
- Eigenvalues of $\mathbf{S}$ are

$$\lambda_j = \frac{d_j^2}{n-1}, \quad j = 1, \dots, r.$$

- Scores are

$$\mathbf{Z} = \mathbf{X}_c V = UD,$$

  score variance along PC j is λ_j.
- **Solving PCA using SVD:** SVD is numerically stable / faster computation when $n << p$.
- PCA naturally connects to low-rank approximations of the centered data matrix

Covariance PCA vs correlation PCA

- **Covariance PCA**: run PCA on $\mathbf{X}_c$.
- **Correlation PCA**: standardize each column to have variance 1 first, then PCA.

When to use which?

- Similar units/scale (e.g., yield changes in bps): covariance PCA often preferred.
- Mixed units/scales (returns, volume, macro levels): correlation PCA usually safer.

Reminder: choice changes the meaning of “variance explained”.

Financial data example: treasury yield changes

Dataset: daily changes in zero-coupon treasury yields at maturities

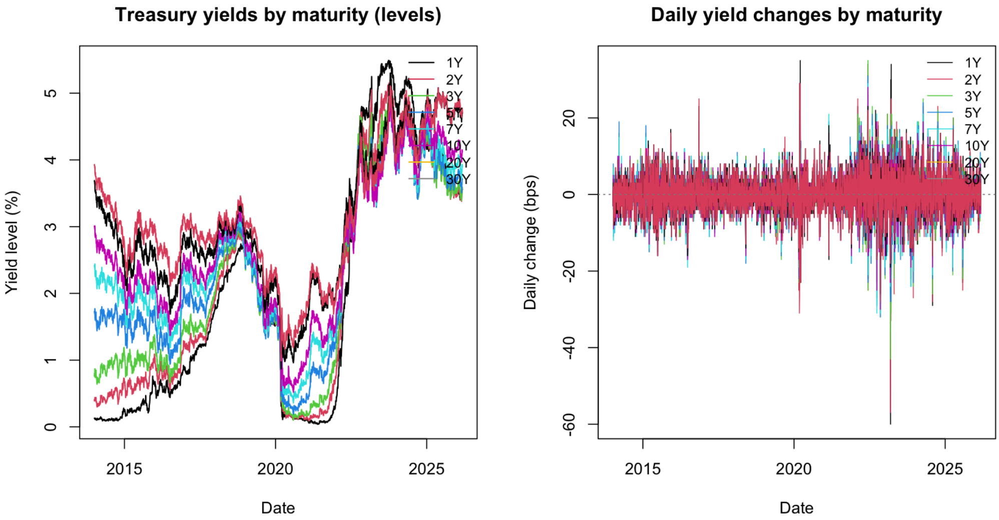
$$\{1\mathrm{Y}, 2\mathrm{Y}, 3\mathrm{Y}, 5\mathrm{Y}, 7\mathrm{Y}, 10\mathrm{Y}, 20\mathrm{Y}, 30\mathrm{Y}\}.$$

- FRED maturity series from 2014-2026

Modeling goal:

- Understand day-to-day curve changes for risk management.
- Use PCA to explain yield-curve movement.

Financial data example: treasury yield changes

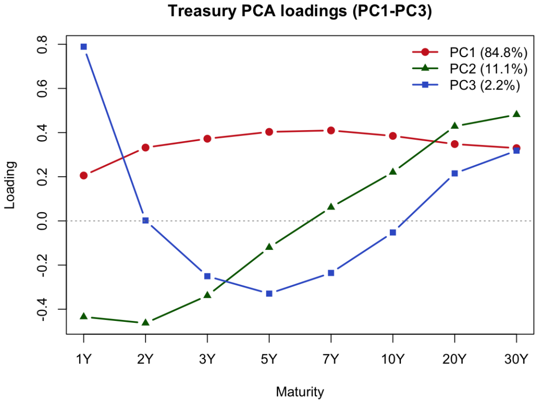


Financial data example: PCA loadings

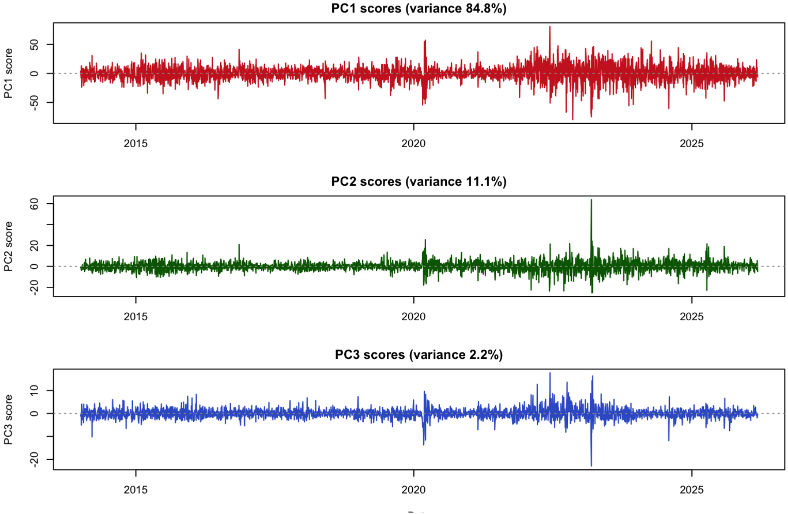
Interpretation of three PCs:

- PC1 (Level): nearly same-sign loadings across maturities.
- PC2 (Slope): short-end and long-end opposite signs.
- PC3 (Curvature): middle maturities opposite to ends.

These are called yield curve risk factors.



Financial data example: PCA scores



PC1 scores (variance 84.8%)

PC2 scores (variance 11.1%)

PC3 scores (variance 2.2%)

R code template (sample PCA on centered data X_c)

```
# X: n x p raw matrix/data.frame (rows = days, cols = maturities)
X_c <- scale(X, center = TRUE, scale = FALSE)

# Sample PCA via SVD (prcomp internally uses SVD)
pca <- prcomp(X_c, center = FALSE, scale. = FALSE)

eig <- pca$sdev^2
pve <- eig / sum(eig)
V   <- pca$rotation      # loading vectors v1, v2, ...
Z   <- pca$x             # scores = X_c %*% V
```

Python code template (sample PCA with consistent notation)

```python
import numpy as np
from sklearn.decomposition import PCA

# X: n x p raw array
X_c = X - X.mean(axis=0, keepdims=True)

pca = PCA().fit(X_c)
eig = pca.explained_variance_
pve = pca.explained_variance_ratio_
V   = pca.components_.T   # columns are loading vectors v1, v2, ...
Z   = X_c @ V             # scores
```

Lecture 2 summary

- Population PCA (on Σ) v.s. sample PCA (on $\mathbf{S}$)
- PCA directions maximize projected variance under unit-norm and orthogonality constraints
- Lagrangian + FOC lead naturally to eigenvectors/eigenvalues.
- Solving PCA using SVD
- In finance, yield-curve PCA connects directly to level/slope/curvature factors

Suggested reading

- Johnson & Wichern (6th Edition), Chapter 8.1-8.4.