STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 3: PCA in Practice

Jingshu Wang

The University of Chicago

Outline

Lecture 2 recap

For centered data matrix $\mathbf{X}_c \in \mathbb{R}^{n \times p}$:

- Sample covariance: $\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{X}_c$.
- PC loading vectors are eigenvectors of $\mathbf{S}$.
- PC scores are projections: $\mathbf{z}_{ik} = \mathbf{x}_{c,i}^\top \mathbf{v}_k$.
- Explained variance ratio: $\lambda_k / \sum_{j=1}^{p} \lambda_j$.

Today: move from derivation to analysis decisions and a deeper understanding.

PCA workflow checklist

1. **Pre-analysis diagnostics**: data quality, outliers, scale checks.
2. **Fit choices**: covariance PCA vs correlation PCA.
3. **Dimension choice**: select number of PCs with multiple criteria.
4. **Interpretation**: read loadings and scores together.
5. **Robustness**: check sensitivity and communicate uncertainty.

Principle: treat PCA as an analysis pipeline, not a single command.

Step 1: pre-analysis diagnostics

- Missingness pattern: random vs systematic missing values.
- Outliers: can rotate PC directions and dominate variance.
- Marginal distributions: heavy skew may suggest transforms.
- Variable scales/units: key for deciding scaling strategy.

Interpretation note: if one variable has much larger variance, PC1 may mostly reflect scale rather than joint structure.

Step 2: fit choices and preprocessing interpretation

Centering is usually essential; otherwise mean location can distort PCs.

Scaling decision:

- Covariance PCA: keeps original variance magnitudes.
- Correlation PCA: gives each variable unit variance before PCA.

How to interpret differences:

- if conclusions change a lot, report both and explain why;
- emphasize whether you are prioritizing absolute variability or relative structure.
- sign flips do not change the PCA subspace.

Step 3: choosing number of PCs

Use several complementary diagnostics:

- cumulative proportion of variance explained;
- elbow point on the scree plot where additional PCs add little improvement;
- optional rules (e.g., Kaiser rule) and domain goals (see later slide).

Comment:

- there is no universally correct cutoff;
- choose the smallest m that preserves patterns relevant to the task.

Step 4–5: interpretation + robustness

Interpretation:

- Loadings answer: *which variables define each PC?*
- Scores answer: *which observations are extreme/clustered?*

Robustness checks:

- compare results with/without scaling;
- re-fit after mild outlier handling;
- check whether main qualitative conclusions persist.

Additional criteria for selecting the number of PCs

Scree/cumulative PVE is useful, but can be ambiguous:

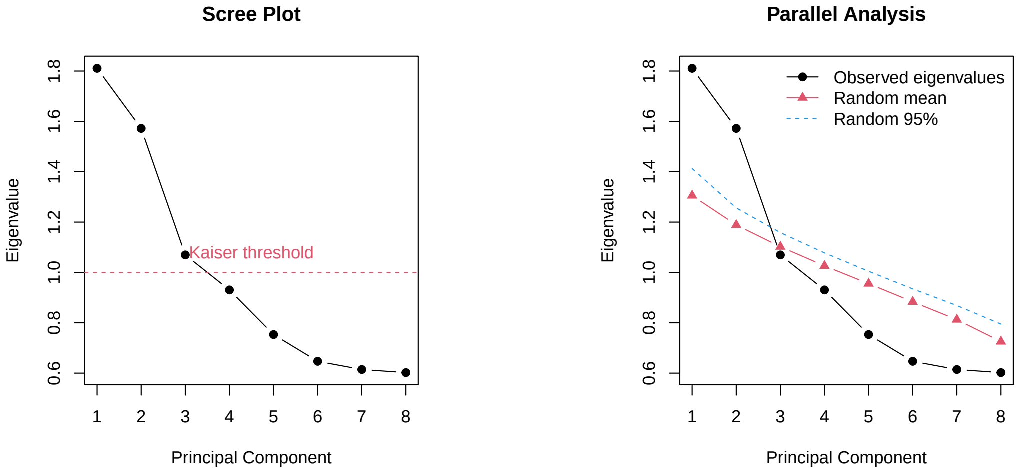
- elbow may be unclear;
- PVE thresholds (e.g., 80%, 90%) are context-specific;
- small-variance directions can still matter for downstream tasks.

Additional criteria:

- **Kaiser rule** (correlation PCA): keep PCs with $\lambda_k > 1$.
- **Parallel analysis**: compare observed eigenvalues to random-data baseline.
  - simulate many $n \times p$ datasets with independent noise (e.g., $N(0, 1)$)
  - compute their eigenvalues, and compare them with the observed ones
  - keep PCs whose eigenvalues exceed the random baseline (e.g., mean or 95% quantile).

These are complementary, not mutually exclusive.

Comparing PC-count criteria



Live R notebook

Notebook file: `Lecture3_demo.nb.html`

High-dimensional challenge: instability when p is large

Suppose we observe $n = 50$ samples with $p = 200$ variables.

- The sample covariance matrix is 200×200 but estimated from only 50 observations.
- Many sample eigenvalues can reflect noise rather than stable signal.
- Small perturbations in the data may change PC directions substantially.

Consequences:

- Scree plots may be less reliable.
- Leading PCs may partially capture noise.
- Regularized methods are often needed (details will be introduced in later lectures).

PCA as low-rank matrix approximation

Let the centered data matrix be $\mathbf{X}_c \in \mathbb{R}^{n\times p}$ with SVD

$$X_c = UDV^\top, \qquad d_1 \geq d_2 \geq \cdots .$$

The rank-m PCA approximation is

$$\hat{\mathbf{X}}^{(m)} = U_m D_m V_m^\top,$$

which keeps only the top m singular values/components.

Key optimization view:

$$\hat{\mathbf{X}}^{(m)} = \arg \min_{\mathrm{rank}(M)\leq m} \|\mathbf{X}_c - M\|_F^2.$$

So PCA gives the best low-dimensional linear compression under squared error.

Connection to high-dimensionality: when p is large, this provides a principled denoising/compression perspective before introducing regularized variants.

Probabilistic PCA (PPCA)

For each observation $\mathbf{x}_i \in \mathbb{R}^p$, a latent variable model:

$$\mathbf{x}_i = W\mathbf{z}_i + \mu + \boldsymbol{\epsilon}_i, \quad \mathbf{z}_i \sim \mathcal{N}(0, I_q), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 I_p),$$

with $q < p$.

- $\mathbf{x}_i \in \mathbb{R}^{p \times 1}$ observed vector.
- $\mathbf{z}_i \in \mathbb{R}^{q \times 1}$ latent factor score.
- $W \in \mathbb{R}^{p \times q}$ loading matrix.
- $\boldsymbol{\mu} \in \mathbb{R}^{p \times 1}$ mean vector.
- $\boldsymbol{\epsilon}_i \in \mathbb{R}^{p \times 1}$ isotropic Gaussian noise.

Hence $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \ \Sigma)$ with

$$\Sigma = WW^\top + \sigma^2 I_p.$$

We can show that the maximum-likelihood solution recovers the PCA subspace.

Maximum likelihood estimation (MLE)

MLE chooses parameters that make observed data most probable under the model.
For PPCA, parameters are

$$\theta = (W, \boldsymbol{\mu}, \sigma^2).$$

Given centered data, we maximize

$$\ell(\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i \mid W, \boldsymbol{\mu}, \sigma^2),$$

where $p(\mathbf{x}_i)$ is Gaussian with covariance $\Sigma = WW^\top + \sigma^2 I_p$.

Interpretation:

- MLE chooses parameters so that the model covariance $WW^\top + \sigma^2 I_p$ resembles the sample covariance.
- It decomposes variability into a low-rank structure ($WW^\top$) and isotropic noise ($\sigma^2 I_p$).

Why PPCA MLE gives the PCA subspace

Let sample covariance be $\mathbf{S}$ with eigenpairs $(\lambda_j, \mathbf{v}_j)$. PPCA MLE has closed form:

$$\hat{W} = V_q \left(\Lambda_q - \hat{\sigma}^2 I_q\right)^{1/2} R,$$

where

- $V_q = [\mathbf{v}_1, \ldots, \mathbf{v}_q]$ contains top q eigenvectors of $\mathbf{S}$,
- $\Lambda_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$,
- R is any $q \times q$ orthogonal rotation,
- $\hat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j$.

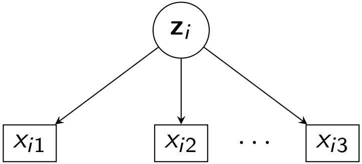
Therefore, the column space of $\hat{W}$ is exactly the PCA principal subspace spanned by top sample PCs.

PCA vs PPCA vs Factor Analysis

Method	Latent factors	Noise structure	Main goal
PCA	implicit	none explicit	variance summary
PPCA	Gaussian latent	isotropic $\sigma^2 I_p$	probabilistic PCA
Factor analysis	Gaussian latent	diagonal Ψ	common vs unique variance

Lecture 4: will move from PCA geometry to full factor modeling assumptions.

PPCA

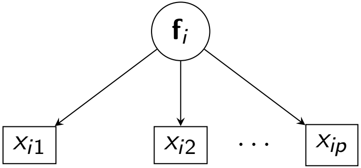


$$\mathbf{x}_i = W\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 I_p)$$

same noise variance for all variables

Factor Analysis



$$\mathbf{x}_i = L\mathbf{f}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Psi)$$

$$\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$$

different noise variance for different variables

Lecture 3 summary

- PCA is a full workflow: diagnostics, preprocessing, fitting, interpretation, and robustness checks.
- Number of PCs should be chosen by combining criteria (scree/PVE, Kaiser threshold, parallel analysis, and context).
- PCA can provide a low-rank approximation of the centered data matrix.
- PPCA adds a probabilistic latent-variable model and recovers the same principal subspace as sample PCA.
- This provides a natural bridge to Lecture 4 on factor analysis.

Suggested reading

- MVnormal.pdf
- James, Witten, Hastie & Tibshirani (2nd edition), Chapter 12.1-12.3
- M. E. Tipping and C. M. Bishop (1999), *Probabilistic Principal Component Analysis*, JRSS-B.