

# Principal Component Analysis: From Classical Ideas to Geometric and Statistical Analysis

Diogo Ribeiro

TSIW / Department of Informatics

November 16, 2025

# Outline

- 1 Motivation and Intuition
- 2 Classical PCA: Algebra and Optimization
- 3 Population vs Sample PCA
- 4 Grassmann Manifold View
- 5 Asymptotic Performance of PCA
- 6 Excess Risk Quantiles and Non-Asymptotics
- 7 Example: Spiked Covariance Model
- 8 Summary and References

## Why PCA? (1/2)

- Modern datasets:  $X_1, \dots, X_n \in \mathbb{R}^d$  with
  - $d$  large (tens, hundreds, thousands).
  - Strong correlations between coordinates.
- Problems:
  - Visualization is impossible in high dimensions.
  - Learning algorithms may overfit or become unstable.
  - Computation and storage become expensive.
- Idea: replace the original variables by a smaller set of *linear combinations* that capture most of the structure.

## Why PCA? (2/2)

- PCA is an unsupervised method.
- No labels  $Y$ ; we use only the geometry of  $X$  (through covariance).
- Typical goals:
  - Data compression / dimensionality reduction.
  - Visualization in 2D/3D.
  - Denoising.
  - As a preprocessing step before regression / classification.
- For MSc/PhD level:
  - PCA is also a key example of spectral methods.
  - It is a clean model to study statistical risk and geometry.

# Toy Geometric Picture (2D/3D)

- Imagine points in  $\mathbb{R}^2$  forming an elongated cloud.
- First principal component (PC1):
  - A unit vector  $u_1$  along the longest direction of variability.
  - Projections  $\langle u_1, X \rangle$  have maximal variance.
- Second principal component (PC2):
  - A unit vector  $u_2$  orthogonal to  $u_1$ .
  - Captures the next largest variance, orthogonal to PC1.
- In  $\mathbb{R}^3$ , PCA often finds a plane where the cloud lies approximately.

**Key intuition:** PCA finds a low-dimensional *linear* subspace that best fits the data in a mean squared error sense.

# Two Classical Viewpoints of PCA

## ① Max-variance view:

- First PC: direction of maximal variance.
- Next PCs: directions of maximal variance orthogonal to previous ones.

## ② Min-error view:

- Find a  $k$ -dimensional subspace such that orthogonal projection onto that subspace minimizes the mean squared reconstruction error.

These views are equivalent and both will be useful:

- For implementation and interpretation: max-variance/SVD view.
- For theory and geometry: reconstruction-error view.

# Centering and Covariance

- Observations:  $X_1, \dots, X_n \in \mathbb{R}^d$ .
- Empirical mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Centered observations:

$$\tilde{X}_i = X_i - \bar{X}.$$

- Empirical covariance:

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top \in \mathbb{R}^{d \times d}.$$

- PCA is typically applied to  $\tilde{X}_i$  and  $\Sigma_n$ .

# Eigen-Decomposition of $\Sigma_n$

- $\Sigma_n$  is symmetric positive semidefinite.
- It has eigen-decomposition

$$\Sigma_n = V \Lambda V^\top,$$

where

$$V = [v_1, \dots, v_d], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d),$$

and

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0.$$

- $v_j$  are called *principal directions* (or loadings).
- $\lambda_j$  are variances along these directions.

# First Principal Component as a Rayleigh Quotient

- Let  $u \in \mathbb{R}^d$  with  $\|u\|_2 = 1$ .
- Variance of projection:

$$\text{Var}(\langle u, X \rangle) \approx u^\top \Sigma_n u.$$

- First principal component solves:

$$u_1 \in \arg \max_{\|u\|_2=1} u^\top \Sigma_n u.$$

- By Rayleigh quotient theory:

$u_1 = v_1$ , the eigenvector with largest eigenvalue  $\lambda_1$ .

- Next components:

$$u_k \in \arg \max_{\|u\|_2=1, u \perp u_1, \dots, u_{k-1}} u^\top \Sigma_n u,$$

giving  $u_k = v_k$ .

## Optimization View: Reconstruction Error

- Let  $U \in \mathbb{R}^{d \times k}$  with  $U^\top U = I_k$ .
- Projection of  $\tilde{X}_i$  onto the subspace spanned by columns of  $U$ :

$$P_U(\tilde{X}_i) = UU^\top \tilde{X}_i.$$

- Reconstruction error for point  $i$ :

$$\left\| \tilde{X}_i - UU^\top \tilde{X}_i \right\|_2^2.$$

- Empirical risk:

$$\tilde{R}_n(U) := \frac{1}{2n} \sum_{i=1}^n \left\| \tilde{X}_i - UU^\top \tilde{X}_i \right\|_2^2.$$

- PCA subspace of dimension  $k$ :

$$U_n = [v_1, \dots, v_k] \in \arg \min_{U^\top U = I_k} \tilde{R}_n(U).$$

# Quick Derivation: Minimizing Reconstruction Error

**Sketch:**

$$\begin{aligned}\tilde{R}_n(U) &= \frac{1}{2n} \sum_{i=1}^n \|\tilde{X}_i - UU^\top \tilde{X}_i\|_2^2 \\ &= \frac{1}{2n} \sum_{i=1}^n \left( \|\tilde{X}_i\|_2^2 - \|U^\top \tilde{X}_i\|_2^2 \right) \\ &= \frac{1}{2n} \sum_{i=1}^n \|\tilde{X}_i\|_2^2 - \frac{1}{2n} \sum_{i=1}^n \tilde{X}_i^\top UU^\top \tilde{X}_i \\ &= \text{constant} - \frac{1}{2} \text{Tr}(U^\top \Sigma_n U).\end{aligned}$$

- So minimizing  $\tilde{R}_n(U)$  is the same as maximizing  $\text{Tr}(U^\top \Sigma_n U)$ .
- The maximizer is  $U_n = [v_1, \dots, v_k]$  by a “block Rayleigh quotient” argument.
- This expression (risk = constant – block Rayleigh quotient) is key for the geometric analysis later.

# SVD View and Scores

- Data matrix  $X \in \mathbb{R}^{n \times d}$  with rows  $\tilde{X}_i^\top$ .
- SVD:

$$X = WSV^\top,$$

with  $W \in \mathbb{R}^{n \times d}$ ,  $S = \text{diag}(s_1, \dots, s_d)$ ,  $V = [v_1, \dots, v_d]$ .

- Then

$$\Sigma_n = \frac{1}{n} X^\top X = V \frac{S^2}{n} V^\top.$$

- Principal component scores:

$$Z = XV_k = WS_k, \quad V_k = [v_1, \dots, v_k].$$

# Explained Variance and Choosing $k$

- Total variance:

$$\text{tr}(\Sigma_n) = \sum_{j=1}^d \lambda_j.$$

- Variance explained by first  $k$  PCs:

$$\sum_{j=1}^k \lambda_j.$$

- Proportion of variance explained (PVE):

$$\text{PVE}(k) = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^d \lambda_j}.$$

- Practical heuristics:

- Scree plot (look for “elbow”).
- Choose smallest  $k$  with  $\text{PVE}(k) \geq$  (e.g.) 0.9.
- Application-specific trade-off: model complexity vs information retained.

# Board / Discussion Break (Optional)

## Suggestion for 90-minute slot:

- On the board, work through a tiny numerical example:
  - $d = 2, n = 3$  or  $4$ .
  - Compute  $\Sigma_n$ , eigenvalues, eigenvectors.
  - Show projections and reconstruction.
- Ask students:
  - What happens if we do not center the data?
  - When might PCA be a bad idea (nonlinear structure, heavy tails, etc.)?

# Population PCA

- Random vector  $X \in \mathbb{R}^d$  with mean  $m = \mathbb{E}[X]$  and covariance

$$\Sigma = \mathbb{E}[(X - m)(X - m)^\top].$$

- Eigen-decomposition:

$$\Sigma = U\Lambda U^\top, \quad U = [u_1, \dots, u_d], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d).$$

- The population  $k$ -dimensional PCA subspace:

$$U^* = [u_1, \dots, u_k].$$

- Sample PCA uses  $U_n = [v_1, \dots, v_k]$  based on  $\Sigma_n$  as an estimator of  $U^*$ .

# Population Reconstruction Risk

- For any  $U \in \mathbb{R}^{d \times k}$  with  $U^\top U = I_k$ , define the population risk:

$$R(U) := \frac{1}{2} \mathbb{E} \left[ \left\| X - \mathbb{E}[X] - UU^\top(X - \mathbb{E}[X]) \right\|_2^2 \right].$$

- One can show:

$$R(U) = \frac{1}{2} \text{tr}(\Sigma) - \frac{1}{2} \text{tr}(U^\top \Sigma U).$$

- Therefore,  $U^* = [u_1, \dots, u_k]$  minimizes  $R(U)$ .
- Now we start to think of PCA as choosing a *parameter* (a subspace) to minimize a risk functional.

# Redundancy and the Grassmannian

- The risk  $R(U)$  depends only on the subspace  $\mathcal{S} = \text{span}(U)$ : if  $V = UQ$  with  $Q$  orthogonal, then

$$VV^\top = UU^\top, \quad R(V) = R(U).$$

- We want a space where each point is a  $k$ -dimensional subspace:
  - Stiefel manifold: orthonormal frames.
  - Grassmann manifold: subspaces (frames modulo rotations).

# Stiefel and Grassmann Manifolds

## Stiefel Manifold

$$\text{St}(d, k) := \{U \in \mathbb{R}^{d \times k} : U^\top U = I_k\}.$$

## Grassmann Manifold

$$\text{Gr}(d, k) := \text{St}(d, k) / \sim,$$

where  $U \sim V$  if  $V = UQ$  for some orthogonal  $Q \in \mathbb{R}^{k \times k}$ .

- Each  $[U] \in \text{Gr}(d, k)$  is a  $k$ -dimensional subspace of  $\mathbb{R}^d$ .
- PCA:

$$[U^*] \in \arg \min_{[U] \in \text{Gr}(d, k)} R([U]), \quad [U_n] \in \arg \min_{[U] \in \text{Gr}(d, k)} R_n([U]).$$

# Geometry of $\text{Gr}(d, k)$ (Sketch)

- Tangent space at  $[U]$ :

$$T_{[U]}\text{Gr}(d, k) \simeq \{\Delta \in \mathbb{R}^{d \times k} : U^\top \Delta = 0\}.$$

- Principal angles  $\theta_1, \dots, \theta_k$  between  $[U]$  and  $[V]$ :

- Compute SVD of  $U^\top V$ .
  - Cosines of the angles are the singular values.

- Riemannian distance:

$$\text{dist}^2([U], [V]) = \sum_{j=1}^k \theta_j^2.$$

- For statistical analysis:

- Use exponential map  $\exp_{[U]}$  to move along geodesics.
  - Use logarithm map  $\text{Log}_{[U]}$  to map nearby subspaces to the tangent space.

# Two Notions of Error

- **Geometric error:**

$$\text{dist}([U_n], [U^*]) \quad (\text{distance on } \text{Gr}(d, k)).$$

- **Excess risk:**

$$E_n = R([U_n]) - R([U^*]).$$

- Both are random (depend on the sample).
- We are interested in:
  - Convergence:  $[U_n] \rightarrow [U^*]$  as  $n \rightarrow \infty$ .
  - Rate and limiting distribution (CLT-type results).
  - How  $E_n$  behaves and what determines it.

# Theorem (Informal): Asymptotic Normality on $\text{Gr}(d, k)$

Under assumptions such as:

- Eigengap:  $\lambda_k > \lambda_{k+1}$ .
- Finite moments (up to order 4) of projections  $\langle u_j, X \rangle$ .

one can show:

- **Consistency:**

$$\text{dist}([U_n], [U^*]) \xrightarrow{P} 0.$$

- **Asymptotic normality:**

$$\sqrt{n} \text{Log}_{[U^*]}([U_n]) \xrightarrow{d} G,$$

where  $G$  is a Gaussian matrix in the tangent space  $T_{[U^*]}\text{Gr}(d, k)$  with explicitly described covariance.

- **Excess risk:**

$$nE_n \xrightarrow{d} \frac{1}{2} \|H\|_F^2,$$

where  $H$  is another Gaussian matrix with covariance linked to  $G$  and the curvature of the risk.

# Interpretation of the Asymptotic Result

- This is an analogue of the CLT for estimators:
  - In Euclidean settings,  $\sqrt{n}(\hat{\theta}_n - \theta^*)$  converges to a Gaussian.
  - Here  $\hat{\theta}_n$  is a point on  $\text{Gr}(d, k)$ .
  - We compare it to  $\theta^* = [U^*]$  via the log map into the tangent space.
- The distribution of  $G$  and  $H$  depends on:
  - Eigengaps  $\lambda_j - \lambda_{k+i}$ .
  - Fourth-order moments  $\mathbb{E}[\langle u_{k+i}, X \rangle^2 \langle u_j, X \rangle^2]$ .
- So PCA risk is not determined solely by  $\Sigma$ ; higher moments matter.

# Excess Risk Quantiles

- For  $\delta \in (0, 1)$  define the  $(1 - \delta)$ -quantile of  $E_n$ :

$$Q_{E_n}(1 - \delta) := \inf\{t \in \mathbb{R} : \mathbb{P}(E_n \leq t) \geq 1 - \delta\}.$$

- As  $n \rightarrow \infty$ ,  $nQ_{E_n}(1 - \delta)$  converges to a quantity depending on:
  - Mixed moments  $\mathbb{E}[\langle u_{k+i}, X \rangle^2 \langle u_j, X \rangle^2]$ .
  - Eigengaps  $\lambda_j - \lambda_{k+i}$ .
- This gives distribution-aware asymptotic confidence bands for PCA excess risk.

# Block Rayleigh Quotient

- Recall:

$$R([U]) = \frac{1}{2}\mathbb{E}[\|X\|_2^2] - \frac{1}{2}\text{Tr}(U^\top \Sigma U).$$

- Define

$$F([U]) = -\frac{1}{2}\text{Tr}(U^\top AU), \quad A = \Sigma.$$

- This is the *block Rayleigh quotient*.
- $[U^*]$  is a minimizer of  $F$  (equivalently of  $R$ ).
- To control non-asymptotic behavior, we study the curvature of  $F$  on  $\text{Gr}(d, k)$  near  $[U^*]$ .

# Self-Concordance Along Geodesics (Idea)

- Consider a geodesic  $\gamma(t)$  between  $[U^*]$  and another point  $[U]$ .
- Study the 1D function  $g(t) = F(\gamma(t))$ .
- Proposition (informal):
  - If the maximum principal angle between  $[U]$  and  $[U^*]$  is  $< \pi/4$ , then  $g$  satisfies a generalized self-concordance condition of the form

$$|g'''(t)| \leq C(\theta) g''(t)$$

along the geodesic.

- Consequence:
  - Second-order Taylor expansions of  $F$  around  $[U^*]$  are well controlled within a neighborhood (no wild changes in the Hessian).
  - This is analogous to Bach's self-concordant analysis of logistic regression, but now on a manifold.

# Non-Asymptotic Excess Risk (High-Level Statement)

Under eigengap and moment assumptions, and for  $n$  above a certain threshold, one can prove that, with high probability  $1 - \delta$ ,

$$E_n = R([U_n]) - R([U^*]) \leq \frac{C}{n} \sum_{i=1}^{d-k} \sum_{j=1}^k \frac{\mathbb{E}[\langle u_{k+i}, X \rangle^2 \langle u_j, X \rangle^2]}{\lambda_j - \lambda_{k+i}},$$

where

- $C$  depends on  $\delta$  and constants in the self-concordance and concentration arguments.
- The right-hand side matches (up to constants) the asymptotic limit.

## Interpretation:

- For large enough  $n$ , PCA's excess risk decays as  $1/n$ .
- The leading constant encodes both spectral structure and higher moments.

# Spiked Covariance Model

- A simple but important model:

$$X = Z + \varepsilon,$$

where

- $Z$  lives in a  $k$ -dimensional subspace (low-rank signal),
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$  (isotropic noise),
- $Z$  and  $\varepsilon$  are independent.

- Then covariance

$$\Sigma = S + \sigma^2 I_d,$$

where  $S$  has rank  $k$  and encodes the signal.

- The top- $k$  eigenvectors of  $\Sigma$  recover the signal subspace.

# Asymptotic Behavior in the Spiked Model

- In this model, the Gaussian matrices appearing in the asymptotic characterization of PCA simplify.
- The variances of entries of the limiting Gaussian  $G$  and  $H$  can be written in closed form in terms of:
  - noise variance  $\sigma^2$ ,
  - signal eigenvalues (non-zero eigenvalues of  $S$ ).
- This yields explicit formulas for the asymptotic distribution of:
  - geometric error  $\text{dist}([U_n], [U^*])$ ,
  - scaled excess risk  $nE_n$ .
- Good sandbox model to connect theory and simulation.

# Possible Exercise / Project Idea

- Simulate the spiked covariance model:
  - Fix  $d, k, \sigma^2$ , and signal eigenvalues.
  - Generate  $X_1, \dots, X_n$  and compute sample PCA.
- Empirically estimate:
  - distribution of  $\text{dist}([U_n], [U^*])$ ,
  - distribution of  $nE_n$ .
- Compare with asymptotic predictions from the theory.
- Explore the impact of:
  - decreasing eigengap,
  - increasing noise  $\sigma^2$ ,
  - heavier tails for  $Z$  (to see role of fourth moments).

# Summary of the Lecture

- Classical PCA:
  - Eigen-decomposition / SVD framework.
  - Equivalent max-variance and min-error formulations.
- Geometric viewpoint:
  - Parameter is a subspace  $\Rightarrow$  point on  $\text{Gr}(d, k)$ .
  - Use distances and tangent spaces to measure errors.
- Statistical performance:
  - PCA behaves as an  $M$ -estimator on a manifold.
  - Asymptotic normality of subspace error.
  - Excess risk  $E_n$  decays like  $1/n$  with a distribution-aware constant.
- Tools:
  - Principal angles, Grassmannian geometry.
  - Block Rayleigh quotient and generalized self-concordance.

# References (Core Reading)

## Classical PCA and spectral methods

- I.T. Jolliffe. *Principal Component Analysis*. Springer.
- T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*, Ch. 14.

## Geometric and statistical analysis of PCA

- A. El Hanchi, M. A. Erdogan, C. J. Maddison. *A Geometric Analysis of PCA*.

# Thank You

Questions?