Principal Component Analysis

Statistics · Module 8 — Multivariate Linear Models

Node 8.7

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.7 — Principal Component Analysis.

PCA is the canonical method for dimensionality reduction via variance maximization. It connects the spectral theorem from linear algebra to the statistical problem of finding the most informative low-dimensional representation of high-dimensional data. This node depends on the covariance structure (Node 8.1) and the multivariate normal (Node 8.2).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \Sigma)$ with $X_i \in \mathbb{R}^p$. The k-th **principal component direction** $v_k \in \mathbb{R}^p$ solves

$$v_k = \arg \max_{\|v\|=1,\ v \perp v_1, \ldots, v_{k-1}} v^\top \Sigma v.$$

The k-th **principal component score** is $Z_k = v_k^\top (X - \mu)$.

The solution is the eigendecomposition $\Sigma = \sum_{j=1}^{p} \lambda_j v_j v_j^\top$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, where v_k is the k-th eigenvector.

Intuition

PCA rotates the coordinate system to align with the directions of greatest variability. The first principal component captures the most variance; each subsequent component captures the maximum remaining variance orthogonal to all previous ones. The eigenvalues are the variances along these directions. If the data lies approximately in a low-dimensional subspace, the first few components explain most of the total variance and the rest can be discarded with minimal information loss. PCA is a descriptive procedure — it requires no probabilistic model — but it is the optimal linear dimensionality reduction under the mean-squared reconstruction error criterion.

Structural Role

PCA identifies the directions of greatest variance, providing a basis for low-rank representation, visualization, and preprocessing. It is the bridge between the algebraic spectral theorem and statistical data analysis. PCA connects to probabilistic PCA (a latent variable model with Gaussian noise), factor analysis, and the SVD of the data matrix. In high-dimensional statistics, PCA is both a tool and an object of study (spiked covariance models, random matrix theory).

Mathematical Development

PCA as variance maximization.

Theorem. The maximum of $v^\top \Sigma v$ subject to $\|v\| = 1$ is λ_1, attained at $v = v_1$.

Proof. By Lagrange multipliers, the stationary condition is $\Sigma v = \lambda v$, so the critical points are eigenvectors with critical values equal to eigenvalues. The maximum is the largest eigenvalue λ_1.

For the k-th component, the constraint $v \perp v_1, \ldots, v_{k-1}$ restricts to the subspace orthogonal to the first $k-1$ eigenvectors. On this subspace, Σ acts with eigenvalues $\lambda_k, \ldots, \lambda_p$, so the maximum is λ_k at v_k. $\square$

PCA as minimum reconstruction error.

Theorem. The rank-r linear subspace minimizing the expected reconstruction error is $\text{span}(v_1, \ldots, v_r)$:

$$\min_{\dim(S)=r} \mathbb{E}\left[\|X - \mu - \text{proj}_S(X-\mu)\|^2\right] = \sum_{j=r+1}^{p} \lambda_j.$$

Proof. Let $V_r = [v_1, \ldots, v_r]$. The projection is $\text{proj}_S(X-\mu) = V_r V_r^\top (X-\mu)$ and the reconstruction error is

$$\begin{aligned}\mathbb{E}[\|X - \mu - V_r V_r^\top (X-\mu)\|^2] &= \mathbb{E}[\text{tr}((I - V_r V_r^\top)(X-\mu)(X-\mu)^\top (I - V_r V_r^\top))] \\ &= \text{tr}((I - V_r V_r^\top)\Sigma) = \sum_{j=1}^{p} \lambda_j - \sum_{j=1}^{r} \lambda_j = \sum_{j=r+1}^{p} \lambda_j.\end{aligned}$$

For any other rank-r subspace, the Eckart-Young theorem (or the Fan-Poincare minimax principle) shows the error is at least $\sum_{j=r+1}^{p} \lambda_j$. $\square$

Variance explained. The proportion of variance explained by the first r components is

$$\frac{\sum_{j=1}^{r} \lambda_j}{\text{tr}(\Sigma)} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

The scree plot displays λ_j vs. j; a sharp drop ("elbow") suggests the appropriate number of components. The cumulative proportion curve $\sum_{j=1}^{r} \lambda_j / \text{tr}(\Sigma)$ provides a threshold-based selection rule (e.g., retain components until 95% of variance is explained).

Connection to SVD. Let $\tilde{X} \in \mathbb{R}^{n \times p}$ be the centered data matrix with rows $(X_i - \bar{X})^\top$. The singular value decomposition $\tilde{X} = UDV^\top$ (with $U \in \mathbb{R}^{n \times p}$, $D = \text{diag}(d_1, \ldots, d_p)$, $V \in \mathbb{R}^{p \times p}$ orthogonal) gives:

- V: columns are the principal component directions (right singular vectors = eigenvectors of $\hat{\Sigma}$).
- $d_j^2/(n-1) = \hat{\lambda}_j$: the sample eigenvalues.
- UD: the matrix of principal component scores.
- $\hat{\Sigma} = V(D^2/(n-1))V^\top$.

The SVD is computationally preferable to forming $\hat{\Sigma}$ explicitly, avoiding the squaring of condition numbers.

Sample PCA. Replace Σ by $\hat{\Sigma} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top$. The sample principal directions $\hat{v}_k$ are eigenvectors of $\hat{\Sigma}$. When population eigenvalues are distinct ($\lambda_j \neq \lambda_k$ for $j \neq k$), $\hat{v}_k \xrightarrow{p} v_k$ (up to sign) by the continuous mapping theorem applied to the eigendecomposition map.

When eigenvalues coincide ($\lambda_j = \lambda_{j+1}$), the eigenvectors are not uniquely identified — any orthonormal basis of the eigenspace is valid — and the sample eigenvectors converge only to the eigenspace, not to a specific direction.

Asymptotic distribution of eigenvalues. Under normality or finite fourth moments, for distinct eigenvalues,

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0,\ 2\lambda_j^2) \quad \text{(under MVN)}.$$

The asymptotic variance of $\hat{v}_j$ depends on the eigenvalue gaps: $\mathrm{Var}(\hat{v}_j^\top v_k) \approx \lambda_j\lambda_k/(n(\lambda_j - \lambda_k)^2)$ for $k \neq j$. Small gaps produce unreliable eigenvector estimates.

Probabilistic PCA (Tipping-Bishop). Assume the generative model $X = Wz + \mu + \varepsilon$ where $z \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, and $W \in \mathbb{R}^{p\times r}$. The marginal is $X \sim \mathcal{N}(\mu, WW^\top + \sigma^2 I)$. The MLE for W satisfies $\hat{W} = V_r(\Lambda_r - \sigma^2 I)^{1/2}R$ where V_r contains the first r eigenvectors, Λ_r the corresponding eigenvalues, and R is an arbitrary orthogonal matrix. As $\sigma^2 \to 0$, this recovers classical PCA.

PCA on the correlation matrix. When variables have different units, PCA is applied to the correlation matrix R rather than Σ. This is equivalent to standardizing each variable to unit variance before computing PCA. The choice between Σ and R is a modeling decision: Σ-PCA preserves the original scale; R-PCA treats all variables equally.

High-dimensional PCA and spiked models. When $p/n \to \gamma \in (0, \infty)$, the sample eigenvalues of $\hat{\Sigma}$ are inconsistent for the population eigenvalues due to the Marchenko-Pastur law. In the spiked covariance model $\Sigma = I_p + \sum_{j=1}^{r} \ell_j v_j v_j^\top$, a phase transition occurs: the j-th sample eigenvalue separates from the bulk only when $\ell_j > \sqrt{\gamma}$ (the BBP transition). Below this threshold, the signal is undetectable by PCA.

Kernel PCA. Classical PCA finds directions in the input space $\mathbb{R}^p$. Kernel PCA maps data to a feature space $\phi(x) \in \mathcal{H}$ via a kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ and performs PCA in $\mathcal{H}$. The principal components are computed from the $n \times n$ kernel matrix $K_{ij} = K(X_i, X_j)$ without explicitly constructing ϕ. This enables nonlinear dimensionality reduction.

Relationship to factor analysis. Factor analysis posits $X = \Lambda F + \varepsilon$ with $F \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \Psi)$ where Ψ is diagonal, giving $\Sigma = \Lambda\Lambda^\top + \Psi$. Unlike PCA, factor analysis separates common variance ($\Lambda\Lambda^\top$) from unique variance (Ψ). PCA is equivalent to factor analysis only when $\Psi = \sigma^2 I$ (the probabilistic PCA model).

Courant-Fischer minimax characterization. The eigenvalues of Σ admit the variational characterization:

$$\lambda_k = \min_{\dim(S)=p-k+1} \max_{v\in S, \|v\|=1} v^\top \Sigma v = \max_{\dim(S)=k} \min_{v\in S, \|v\|=1} v^\top \Sigma v.$$

This provides the theoretical basis for PCA: the top k eigenvalues represent the maximum achievable total variance captured by any k-dimensional subspace.

Sparse PCA. In high dimensions, classical PCA eigenvectors are dense and difficult to interpret. Sparse PCA adds an ℓ_1 penalty:

$$\max_v v^\top \Sigma v - \lambda\|v\|_1 \quad \text{subject to } \|v\|_2 = 1,$$

producing sparse loading vectors. This loses optimality in reconstruction error but gains interpretability and consistency in high-dimensional settings where $p \gg n$.

Johnson-Lindenstrauss vs. PCA. Random projections preserve pairwise distances up to distortion ε in $O(\log n/\varepsilon^2)$ dimensions, regardless of the data's intrinsic structure. PCA optimally preserves variance but requires $O(np^2)$ computation. For very high-dimensional problems, random projections provide a computational alternative with different statistical guarantees.

Worked Examples

Example 1. Let $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$. The eigenvalues satisfy $\lambda^2 - 7\lambda + 1 = 0$, giving $\lambda_1 = (7 + 3\sqrt{5})/2 \approx 6.854$, $\lambda_2 = (7 - 3\sqrt{5})/2 \approx 0.146$. The first PC explains $6.854/7 \approx 97.9\%$ of the variance. The eigenvector v_1 satisfies $(5 - \lambda_1)v_{11} + 3v_{12} = 0$, giving $v_1 \propto (3, \lambda_1 - 5)^\top \approx (3, 1.854)^\top$, normalized to unit length. The data is nearly one-dimensional: the principal direction captures almost all variability.

Example 2 (SVD computation). Given centered data matrix $\tilde{X} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \\ -1 & -2 \end{pmatrix}$, compute $\tilde{X}^\top \tilde{X} = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$. Eigenvalues: 9 and 3, with eigenvectors $(1,1)^\top/\sqrt{2}$ and $(1,-1)^\top/\sqrt{2}$. Sample variances: $\hat{\lambda}_1 = 9/2 = 4.5$, $\hat{\lambda}_2 = 3/2 = 1.5$. First PC direction: $\hat{v}_1 = (1,1)^\top/\sqrt{2}$, the 45-degree line. Scores: $Z_1 = \tilde{X}\hat{v}_1 = (3/\sqrt{2}, 0, -3/\sqrt{2})^\top$. Proportion of variance: $4.5/6 = 75\%$.

Cross-Links

AI. PCA is the basis for whitening, feature extraction, and linear autoencoders. A linear autoencoder with r hidden units recovers the PCA projection. In NLP, PCA applied to word embedding matrices identifies semantic dimensions. Kernel PCA extends to nonlinear dimensionality reduction via the kernel trick.

Economics. PCA on macroeconomic time series extracts latent factors (diffusion indices) used in forecasting. In finance, PCA on bond yield curves identifies level, slope, and curvature factors that explain >95% of yield variation.

Linear Algebra. PCA is the spectral theorem applied to the covariance matrix. The Eckart-Young theorem establishes the optimality of the truncated SVD for low-rank approximation in Frobenius norm. The Davis-Kahan theorem quantifies how perturbations in Σ affect the eigenvectors.

Quantum Mechanics. PCA on quantum state tomography data decomposes the noise covariance of reconstructed states. The eigenvalues of the density matrix ρ are the quantum analogue of PCA eigenvalues: they determine the effective dimensionality (rank) of the quantum state.

Structural Compression

Invariant. The principal components are the eigenvectors of the covariance matrix; they diagonalize Σ and provide the variance-maximizing (equivalently, reconstruction-error-minimizing) orthonormal basis.

Mechanism. The spectral theorem for symmetric positive semidefinite matrices decomposes Σ into rank-one projections $\lambda_j v_j v_j^\top$; each successive eigenvector maximizes residual variance subject to orthogonality. The SVD of the data matrix provides a computationally stable path to the same decomposition.

Cross-System Echo. PCA is the method of moments estimator in probabilistic PCA (Tipping-Bishop). In information theory, PCA maximizes the mutual information between the data and its linear projection under Gaussianity. The connection between variance maximization and reconstruction error minimization is an instance of the duality between primal and dual optimization.

Exercises

1. Prove that the k-th principal component direction maximizes $v^\top \Sigma v$ subject to $\|v\| = 1$ and $v \perp v_1, \ldots, v_{k-1}$ using Lagrange multipliers.
2. Show that PCA on Σ and PCA on $\hat{\Sigma}$ are equivalent to performing the SVD of the centered data matrix, by relating the eigendecomposition of $\hat{\Sigma}$ to the right singular vectors of $\tilde{X}$.

3. For a 3×3 covariance matrix with eigenvalues $\lambda_1 = 10$, $\lambda_2 = 2$, $\lambda_3 = 0.5$, compute the proportion of variance explained by the first two components and the reconstruction error of the rank-2 approximation.
4. Prove that the total variance $\mathrm{tr}(\Sigma)$ is invariant under orthogonal transformations $Q^\top \Sigma Q$, explaining why PCA redistributes but does not create or destroy variance.
5. In the probabilistic PCA model $X = Wz + \mu + \varepsilon$, derive the marginal distribution of X and show that the MLE for μ is $\bar{X}$.
6. Explain why PCA on the correlation matrix vs. the covariance matrix can give qualitatively different results, and give a concrete example where one is appropriate and the other is not.
7. Argue, structurally, why PCA is simultaneously the optimal linear dimensionality reduction under two different criteria (maximum variance and minimum reconstruction error), and explain why this duality breaks down for nonlinear methods.

Statistics · Module 8 — Multivariate Linear Models · Node 8.7
Concepts: eigenvalue-decomposition, variance-and-covariance, dimensionality, spectral-analysis
Prerequisites: 8.1, 8.2

hbar.university — own.your.cognition