PRINCIPAL COMPONENT ANALYSIS

Authored by

mohammad looti

October 11, 2025

PRINCIPAL COMPONENT ANALYSIS

Primary Disciplinary Field(s): Statistics, Data Science, Machine Learning, Psychometrics

1. Core Definition and Objective

Principal Component Analysis (PCA) is a powerful statistical procedure utilized primarily for dimensionality reduction. Its fundamental goal is to simplify the complexity inherent in high-dimensional datasets by transforming a large set of potentially correlated variables into a smaller set of variables, known as **principal components**. This transformation is linear, meaning the principal components are linear combinations of the original variables. PCA achieves its simplification objective by identifying the directions, or axes, in the data space along which the variance is maximal, effectively preserving the most significant information while discarding noise and redundancy. The technique is foundational to data exploration, preprocessing for machine learning models, and visualization in multivariate statistics.

The core objective of PCA is to find an optimal low-dimensional representation of the data that captures the greatest possible variability. If a dataset has p variables, PCA generates p principal components. These components are mathematically constructed such that the first component accounts for the largest possible variance in the data; the second component, which is mathematically orthogonal (uncorrelated) to the first, accounts for the next largest variance; and so on. This hierarchical structure allows researchers to select only the top k components (where $k < p$), thereby achieving the crucial aim of **data reduction**, as explicitly noted in the source material. By focusing on variance maximization, PCA ensures that the majority of the signal residing in the original, often noisy, high-dimensional space is efficiently summarized.

A defining characteristic of the resulting principal components is their **mutual independence**. Unlike the original variables, which are often highly correlated, the new components are constructed to be orthogonal, meaning they carry unique, non-overlapping information about the structure of the data. This decorrelation is immensely valuable in subsequent analyses, particularly in regression or classification tasks, where highly correlated predictors (multicollinearity) can destabilize model estimates. By replacing the original correlated variables with a smaller set of uncorrelated components, PCA provides a more stable and computationally efficient foundation for downstream statistical modeling.

2. Mathematical Foundation: Eigenvectors and Eigenvalues

The theoretical backbone of Principal Component Analysis rests firmly on concepts from linear algebra, specifically the eigen-decomposition of the data's covariance matrix. Before transformation, the relationships between the original variables are summarized in the covariance

matrix (or correlation matrix, if the data is standardized). The elements of this matrix indicate how each pair of variables co-varies. The computational task of PCA is to find a set of orthogonal basis vectors that can efficiently represent the scatter of the data points described by this covariance structure.

These sought-after basis vectors are the **eigenvectors** of the covariance matrix. Each eigenvector defines a direction in the original p-dimensional space, which corresponds precisely to a principal component. Since PCA aims to maximize variance, the eigenvectors are ordered based on the amount of variance they explain. The eigenvector associated with the largest possible eigenvalue represents the first principal component (PC1), defining the axis along which the data exhibits the most spread. Subsequent eigenvectors define PC2, PC3, and so forth, each orthogonal to the previous ones and capturing the maximum remaining variance.

The corresponding **eigenvalues** quantify the magnitude of the variance along their respective eigenvector directions. In essence, an eigenvalue represents the relative importance of its corresponding principal component. A large eigenvalue signifies that the component explains a substantial portion of the total variability in the dataset, justifying its retention. Conversely, components with very small eigenvalues explain negligible variance and are often interpreted as noise, making them ideal candidates for removal during the dimensionality reduction process. The entire variance of the original dataset is equal to the sum of all eigenvalues, providing a clear metric for assessing the percentage of total variance retained after selecting a subset of components.

3. The PCA Algorithm (Step-by-Step)

The standard implementation of the PCA algorithm follows a rigorous sequence of steps designed to ensure the accurate identification and extraction of principal components. The initial step is almost always the **standardization** of the input data. If variables are measured in different units (e.g., kilograms vs. meters), standardization is critical. By centering the data (subtracting the mean) and scaling it (dividing by the standard deviation), the algorithm ensures that all variables contribute equally to the calculation of variance, preventing variables with larger numerical ranges from disproportionately dominating the first principal component.

Following standardization, the next crucial step is the calculation of the **covariance matrix**. This square matrix captures the joint variability between all pairs of variables. If the data has been standardized (Z-scores), the resulting matrix is equivalent to the correlation matrix. This matrix encapsulates the geometric shape of the data cloud and is the input required for the core linear algebra computation. Once the covariance matrix is derived, the process moves to the **eigen-decomposition** phase, where the eigenvectors and corresponding eigenvalues are computed numerically. This intensive computation identifies the optimal linear combinations that maximize

variance.

The final two steps involve component selection and data projection. After computing the eigenvalues, they are sorted in descending order, along with their associated eigenvectors. The researcher must then decide on the optimal number of components (k) to retain (a process discussed in Section 6). Once k is chosen, the dataset is **projected** onto the subspace spanned by the selected top k eigenvectors. This projection yields the final principal component scores, where each data point is now represented by k values instead of p original values. These new scores retain the maximum possible variance from the original data while operating in a much lower-dimensional space.

4. Relationship to Factor Analysis and Other Techniques

The source material accurately points out that PCA is "akin to **factor analysis** in its aims." Both are multivariate statistical techniques designed to explore the structure of complex data by reducing the number of variables. However, despite this shared high-level objective, their underlying statistical models and philosophical goals are distinct. The distinction lies primarily in their treatment of variance and their modeling assumptions. PCA is a descriptive technique focusing purely on data summarization, while Factor Analysis (FA) is an inferential technique focused on modeling latent theoretical constructs.

In PCA, the fundamental assumption is that the total variance observed in a variable is relevant and should be accounted for. The goal is simply to find the axes that best represent the overall spread of the data. Conversely, Factor Analysis operates on the premise that the observed variance can be partitioned into two types: **common variance** (variance shared with other variables, driven by underlying latent factors) and **unique variance** (error and variance specific to that single variable). FA seeks to explain only the common variance, identifying unobservable factors that causally influence the observed variables, making it more suited for confirmatory hypothesis testing in fields like psychology or sociology.

Furthermore, while PCA components are mathematically orthogonal (uncorrelated) and derived through linear transformation of the covariance matrix, Factor Analysis often utilizes rotation methods (like Varimax or Oblimin) to produce factors that are more theoretically meaningful and easier to interpret, though this may result in non-orthogonal (correlated) factors. While PCA is often used interchangeably with Factor Analysis in exploratory contexts, especially in older literature, modern statistical practice differentiates them: PCA is for data compression and simplification; FA is for theory testing and measurement development.

Other related techniques include Singular Value Decomposition (SVD), which is mathematically equivalent to PCA when applied to centered data and is often the preferred computational method in large-scale machine learning. Another related method is Independent Component Analysis

(ICA). Unlike PCA, which enforces orthogonality (uncorrelation), ICA aims to find components that are statistically **independent**, focusing on non-Gaussian sources in the data, which is especially useful in signal separation problems like the "cocktail party problem."

5. Applications Across Disciplines

The ability of PCA to reduce complexity while retaining maximum information has made it an indispensable tool across numerous scientific and technical disciplines. In **Machine Learning** and Data Science, PCA is predominantly used as a pre-processing step. High-dimensional data, such as images or genomic features, can lead to the "curse of dimensionality," increasing computational time and the risk of overfitting. Applying PCA reduces the feature space, accelerates training times for algorithms like support vector machines or neural networks, and often improves model generalization by filtering out noisy, low-variance dimensions.

In fields dealing with complex sensory data, such as **image processing and computer vision**, PCA has specialized applications. One classic example is the "Eigenfaces" technique used in facial recognition systems. By applying PCA to a large collection of face images, the algorithm identifies the most significant principal components (the Eigenfaces) that represent the fundamental variation in human faces. Any new face can then be represented by a concise set of scores along these primary axes, dramatically speeding up comparison and classification tasks. Similarly, PCA is vital in spectroscopy and chemical analysis for distinguishing complex mixture compositions.

Within the **Social Sciences and Financial Modeling**, PCA is used to construct parsimonious indices. For example, in econometrics, a large set of financial indicators (interest rates, inflation measures, unemployment figures) can be collapsed into a small number of uncorrelated principal components, which may be interpreted as underlying economic factors (e.g., "growth factor" or "liquidity factor"). In psychology and psychometrics, PCA can reduce the scores from a battery of personality tests into a limited set of key psychological constructs, enabling easier interpretation and modeling of human behavior based on core traits.

6. Model Evaluation and Component Selection

One of the most critical and often subjective steps in applying PCA is deciding how many principal components (k) to retain. The goal is to strike a balance: retaining enough components to explain a high percentage of the total variance (fidelity) while achieving substantial data reduction (parsimony). Several quantitative and heuristic methods are employed to guide this selection process, each offering a different perspective on the trade-off.

A widely used and simple criterion is the **Kaiser Criterion**, which dictates that only components with an eigenvalue greater than 1 should be retained. The logic behind this rule stems from the

idea that if the data is standardized, any component with an eigenvalue less than 1 explains less variance than a single standardized original variable. While simple, this rule is often criticized for being arbitrary and can sometimes lead to either over- or under-extraction depending on the dataset complexity and structure.

More nuanced methods include analyzing the **Cumulative Variance Explained** and the use of the **Scree Plot**. The cumulative variance approach involves calculating the percentage of total variance accounted for by the first k components and retaining components until a predetermined threshold is met (e.g., 80% or 90%). The Scree Plot, named for its resemblance to a geological scree slope, graphs the eigenvalues in descending order. Analysts look for the "elbow"--the point where the slope of the plot levels off dramatically. Components before the elbow are typically retained as they contribute meaningfully to variance, while those after it are usually considered noise. Ultimately, domain expertise often dictates the final choice, ensuring the components retained are interpretable and relevant to the research question.

7. Limitations and Criticisms

Despite its utility, PCA is subject to several important limitations that must be considered when interpreting results. A primary criticism is that PCA is fundamentally a **linear technique**. It assumes that the relationships between variables and the underlying structure of the data can be accurately described by straight lines. If the data points lie on a complex, non-linear manifold (such as a spiral or a curved surface), PCA will fail to capture the true low-dimensional structure, necessitating the use of advanced techniques like Kernel PCA or manifold learning algorithms (e.g., t-SNE, LLE).

Furthermore, while the mathematical construction of principal components ensures maximum variance extraction, it often sacrifices **interpretability**. Each principal component is a weighted sum of all original variables, which can make it difficult to assign a meaningful name or conceptual label to the resulting factor. For instance, a component might highly weight 10 different, seemingly disparate financial metrics. While mathematically sound, interpreting this component as a single economic factor requires significant inferential leaps and domain knowledge, a challenge that Factor Analysis often addresses more effectively through rotation methods designed to simplify the component loadings.

Finally, PCA is highly sensitive to the initial scaling of the data and the presence of **outliers**. Because the algorithm relies on the calculation of the covariance matrix (which involves squared distances), extreme outliers can dramatically skew the orientation of the principal axes, potentially leading the first component to primarily capture the variance due to the outlier rather than the underlying structure of the bulk of the data. Therefore, robust preprocessing, including outlier detection and careful standardization, is essential for obtaining reliable PCA results.

Further Reading

Dimensionality reduction (Wikipedia)
Singular Value Decomposition (Wikipedia)
Psychometrics (Wikipedia)
Machine Learning (Wikipedia)