PCA, que significa Análise de Componentes Principais (Principal Component Analysis, em inglês), é uma técnica em estatística multivariada utilizada para simplificar a complexidade em conjuntos de dados de alta dimensão, preservando as informações mais importantes presentes nos dados. Ela é frequentemente aplicada para redução de dimensionalidade, visualização de dados, compressão de dados e, em alguns casos, para eliminar a multicolinearidade em conjuntos de dados.

A ideia principal por trás do PCA é transformar as variáveis originais em um novo conjunto de variáveis (as chamadas componentes principais) que são combinações lineares das variáveis originais. Essas componentes principais são ordenadas de forma que a primeira componente capture a maior variação nos dados, a segunda componente capture a segunda maior variação, e assim por diante.

Os passos principais para realizar a PCA são:

Padronização dos dados: Antes de aplicar o PCA, é comum padronizar as variáveis para garantir que todas tenham a mesma escala. Isso é importante porque o PCA é sensível às escalas das variáveis.

Cálculo da matriz de covariância ou matriz de correlação: A matriz de covariância ou correlação é utilizada para entender as relações entre as variáveis originais.

Cálculo dos autovetores e autovalores: A partir da matriz de covariância (ou correlação), calcula-se os autovetores e autovalores. Os autovetores representam as direções principais dos dados, e os autovalores indicam a quantidade de variância ao longo dessas direções.

Seleção das componentes principais: As componentes principais são escolhidas com base nos autovalores, começando pela que tem o maior autovalor e assim por diante.

Projeção dos dados: Os dados originais são projetados nas novas coordenadas definidas pelas componentes principais.

A principal vantagem do PCA é a capacidade de reduzir a dimensionalidade do conjunto de dados, preservando a maior parte da informação. Isso é útil em situações em que há muitas variáveis, e a redução de dimensionalidade pode facilitar a análise e a interpretação dos dados. Além disso, o PCA é uma ferramenta valiosa para identificar padrões e estruturas nos dados, facilitando a visualização e compreensão dos relacionamentos entre as variáveis.

Autoanálise: Refere-se ao fato de que a PCA é uma técnica que, de certa forma, permite que os dados se "analisem automaticamente". Isso é alcançado resolvendo os autovetores e autovalores de uma matriz quadrada simétrica.

Autovetores e Autovalores: Na PCA, uma matriz de covariância (ou correlação) dos dados é calculada. Os autovetores dessa matriz indicam as direções principais (ou componentes principais) dos dados, enquanto os autovalores representam a variância ao longo dessas direções. Os autovetores são as "setas" que apontam na direção dos máximos valores de variância.

Primeiro Componente Principal: O autovetor associado ao maior autovalor aponta na direção do primeiro componente principal. Em termos simples, ele representa a direção ao longo da qual os dados têm a maior variação.

Segundo Componente Principal: O autovetor associado ao segundo autovalor determina a direção do segundo componente principal, que é ortogonal ao primeiro. Cada componente principal subsequente é ortogonal aos anteriores e captura a maior variação restante.

Traço da Matriz Quadrada: A soma dos autovalores é equivalente ao traço da matriz quadrada. O traço é a soma dos elementos da diagonal principal de uma matriz. No contexto da PCA, isso significa somar a quantidade de variância total nos dados.

Número Máximo de Autovetores: O número máximo de autovetores que podem ser extraídos é igual ao número de linhas ou colunas na matriz quadrada. Em termos práticos, isso significa que o número de componentes principais não pode exceder o número de variáveis originais nos dados.

Em resumo, a PCA é uma técnica que utiliza a álgebra linear para transformar os dados em um novo conjunto de variáveis (os componentes principais) que são combinações lineares das variáveis originais, preservando a maior variação nos dados. Isso é alcançado através da resolução de autovetores e autovalores da matriz de covariância dos dados.