Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a **dimensionality reduction** technique used to simplify complex datasets by transforming them into a set of uncorrelated, lower-dimensional components while preserving as much variance (information) as possible. This makes PCA particularly useful for handling large datasets with many variables, where it's challenging to understand or visualize relationships between variables.

Key Concepts in PCA

1. **Principal Components**:
    - PCA transforms the original variables into new variables called *principal components*.
    - Each principal component is a linear combination of the original variables, capturing a certain proportion of the data's variance.
    - The components are uncorrelated with each other and are ranked by the amount of variance they explain, with the first component explaining the most variance, the second explaining the next highest amount, and so on.
2. **Purpose of PCA**:
    - **Dimensionality Reduction**: PCA reduces the number of variables in a dataset while preserving most of its variability. This is valuable for reducing data complexity and removing noise.
    - **Data Visualization**: By reducing high-dimensional data to two or three principal components, PCA allows for visualization of complex datasets.
    - **Feature Extraction**: PCA can be used to create new features that summarize the original data, especially when the data has redundant or highly correlated variables.
3. **How PCA Works**:
    - **Step 1**: **Standardization** (optional, but recommended): Center and scale the data so each variable has a mean of 0 and a standard deviation of 1. This step ensures that all variables contribute equally to the analysis, regardless of their original units or scales.
    - **Step 2**: **Covariance Matrix Calculation**: Compute the covariance (or correlation) matrix of the standardized data to understand relationships between variables.
    - **Step 3**: **Eigenvalue and Eigenvector Computation**: Compute the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues measure the amount of variance each principal component explains, and eigenvectors determine the direction of each component.
    - **Step 4**: **Selecting Components**: Choose the top components that capture the most variance. The sum of their eigenvalues will represent the total variance retained.
    - **Step 5**: **Transform Data**: Project the original data onto the selected components to obtain a new dataset in a lower-dimensional space.

4. **Explained Variance**:
    - Each principal component explains a proportion of the total variance in the dataset. The **explained variance ratio** for a component is the percentage of variance it explains relative to the total variance.
    - Often, a certain percentage of the variance (e.g., 90% or 95%) is used as a threshold to decide the number of components to retain.

Loadings

In PCA, **loadings** are coefficients that show the relationship between each original variable and each principal component. They indicate how much each variable "loads" onto each component and represent the correlation between the variable and the component. Loadings are essential in interpreting the components because they tell us which variables contribute the most to each component and help us understand what each component represents.

Key Points about Loadings in PCA

5. **Definition**:
    - The loading of a variable on a principal component is the *correlation* between the variable and the component.
    - It quantifies how strongly each variable is associated with the component.
6. **Interpretation**:
    - High absolute values of loadings indicate that the variable contributes significantly to the component.
    - If a variable has a high positive loading on a component, it means it positively influences that component.
    - If it has a high negative loading, it means it negatively influences the component.
7. **Loading Matrix**:
    - In PCA, we compute a matrix called the **loading matrix** (or **component loadings matrix**), often denoted by L. This matrix has dimensions $p \times k$, where p is the number of original variables and k is the number of components.
    - Each element l_{ij} in the loading matrix is the loading of variable i on component j.
8. **Connection to Eigenvectors**:
    - Mathematically, loadings are computed from the eigenvectors of the covariance (or correlation) matrix of the original data, scaled by the square root of the corresponding eigenvalues.
    - If we denote the eigenvectors by E and the eigenvalues by Λ, then the loading for each principal component can be derived as
      $L = E\sqrt{\Lambda}$
9. **Interpretation in Reduced Dimension**:
    - Loadings help us interpret the reduced-dimensional space created by PCA. By examining which variables have high loadings on each principal component, we can label the components based on the types of information they capture.
10. **Squaring Loadings for Variance Explanation**:

- Squaring each loading gives the amount of variance in the original variable explained by the component. Squared loadings are also referred to as **squared correlations**.

Example

Suppose you have three variables (e.g., height, weight, and age) and two principal components (PC1 and PC2). If the loading matrix L looks like this:

$$L = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & -0.1 \\ 0.6 & 0.9 \end{bmatrix}$$

This means:

- **PC1** is strongly correlated with height (0.8), weight (0.7), and age (0.6), suggesting it might represent "overall size."
- **PC2** is mostly correlated with age (0.9), suggesting it may capture an "age-related" component.

In this way, the loading matrix is crucial for interpreting the structure uncovered by PCA.

Squared correlations

In Principal Component Analysis (PCA), squared correlations (also known as *squared loadings* or *communality*) reflect the proportion of variance in each original variable that is explained by the principal components.

Here's a breakdown of what squared correlations represent in PCA:

11. **Squared Loadings**: When we project the original variables onto the principal components, we get *loadings*, which are the correlation coefficients between each variable and each component. Squaring these loadings gives the proportion of variance in each variable explained by each component. If you have *p* original variables and *k* components, then for each variable, you'll have *k* squared loadings that add up to the *communality* of that variable.

12. **Communality**: For each original variable, the sum of its squared loadings across all components gives the *communality* of that variable. This communality measures how well the principal components, taken together, explain the variance in that specific variable. Communality values closer to 1 indicate that the components capture most of the variance in that variable, while values closer to 0 indicate a weaker representation.

13. **Interpretation**: These squared correlations are helpful to assess which variables are well-represented in the principal component space and to understand how much of each variable's information is retained in the reduced dimensionality.

Calculation of Squared Correlations

If the loading matrix L has elements l_{ij}, where l_{ij} is the loading of variable i on component j, then the squared correlations are l_{ij}^2. For the communality of variable i, you calculate:

$$\text{Communality}_i = \sum_{j=1}^{k} l_{ij}^2$$

where k is the total number of principal components retained.

Squared correlations are thus useful for understanding the explanatory power of each component for each variable and for determining how well the PCA captures the structure of the data.

PCA results in a geospatial context

In a geospatial context, Principal Component Analysis (PCA) can be highly effective for analyzing spatially distributed data with multiple variables (e.g., environmental data, socio-economic factors, or climate variables) across geographic regions. By using PCA, complex, high-dimensional geospatial data can be simplified and visualized in ways that reveal patterns, trends, or regions with similar characteristics. Here's how PCA can be applied and how it contributes to creating geographic maps:

14. **Identifying Patterns in Geospatial Data**

- **Reducing Complexity**: Geospatial datasets often contain numerous correlated variables collected for different regions (such as climate variables, soil characteristics, population density, etc.). PCA helps reduce these variables into a smaller set of uncorrelated principal components that capture the main patterns.

- **Finding Underlying Trends**: Each principal component can highlight a dominant spatial trend, such as regions with similar climate profiles, land use types, or socio-economic characteristics. This can be particularly useful in environmental studies, urban planning, and resource management.

15. **Data Preparation for Geospatial PCA**

- **Spatial Units**: Divide the area of interest into spatial units (e.g., grids, administrative boundaries, or custom regions).

- **Standardization**: Standardize the variables to make them comparable, as they might be on different scales (e.g., temperature, elevation, rainfall).

- **Applying PCA**: Run PCA on the standardized dataset to obtain the principal components. The results will give each spatial unit a set of component scores for each principal component.

16. **Interpreting Principal Components for Mapping**

- **Component Scores for Mapping**: Each principal component gives a new score for each spatial unit, reflecting how strongly that unit is associated with the component's trend. These scores can be mapped to visualize spatial patterns.

- **Thematic Mapping**: The component scores can be visualized as thematic maps, where each spatial unit is colored or shaded based on its score for a particular component. For example:
    - **PC1 Map**: If the first principal component (PC1) represents a climate gradient, a map of PC1 scores might show how climate varies across the region.
    - **PC2 Map**: If the second principal component captures population density, a map of PC2 scores might highlight areas of high vs. low population density.

17. **Examples of PCA in Geospatial Applications**

- **Environmental Studies**: PCA can help identify regions with similar environmental profiles, such as areas with high precipitation and low temperatures or regions with high vegetation density and low soil moisture. Maps of these components can be used for ecosystem monitoring or conservation planning.
- **Urban Analysis**: In urban studies, PCA can reveal socio-economic and demographic patterns. For instance, one component may capture urbanization gradients, distinguishing urban from rural areas. Thematic maps of these components can aid in urban planning and resource allocation.
- **Climate Analysis**: PCA can simplify climate data, such as temperature, humidity, and precipitation across a region, into dominant patterns. For example, PCA might reveal major climate zones, and maps of these patterns can be used in climate change impact studies.

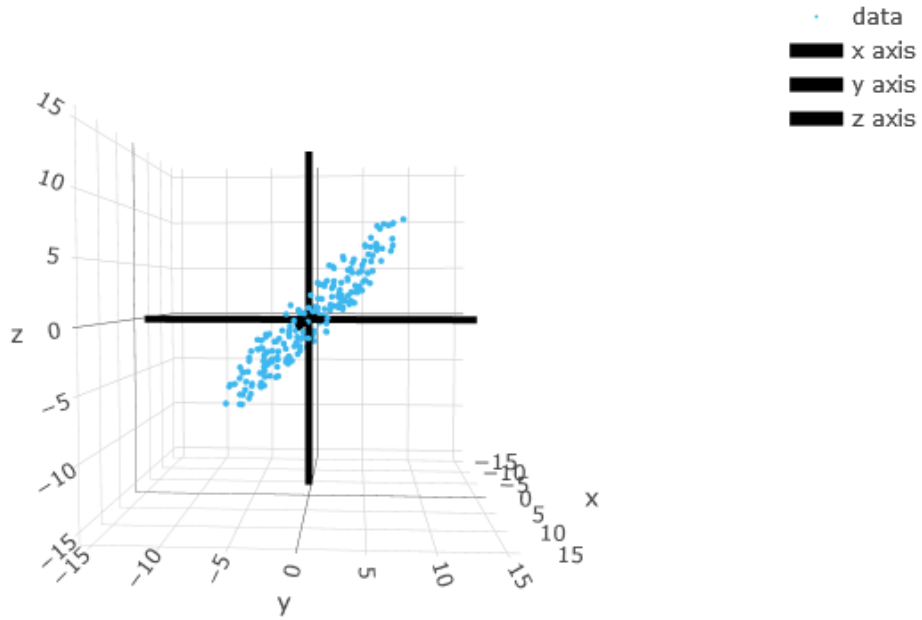
18. **Creating Geographic Maps from PCA Results**

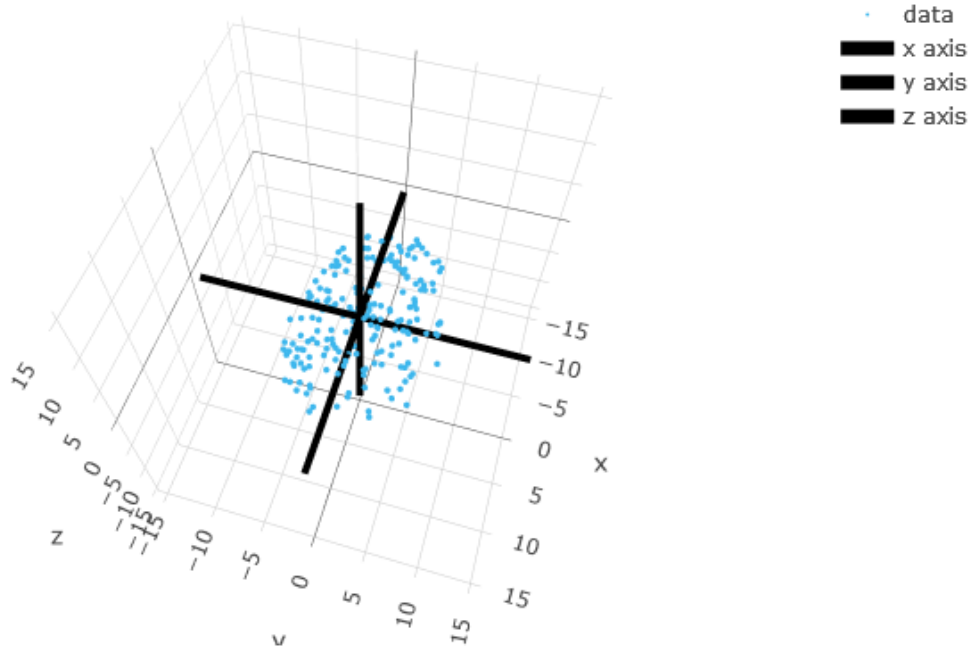
- **Visualizing Scores as Heatmaps or Choropleth Maps**: For each principal component, use the component scores of each spatial unit to create a map (usually a heatmap or choropleth). Regions with similar colors on the map share similar values for the component and thus exhibit similar characteristics according to the patterns found by PCA.
- **Multiple Component Maps**: If the first few components capture significant and distinct patterns, each can be mapped separately to highlight different aspects of the data. Alternatively, overlaying maps of the first few components can create a composite map showing more nuanced relationships.
- **Interpreting Color Schemes**: Choose color schemes that make it easy to differentiate high, medium, and low values for each component. For example, in a climate study, blue might indicate cool, humid areas, while red indicates hot, dry regions.

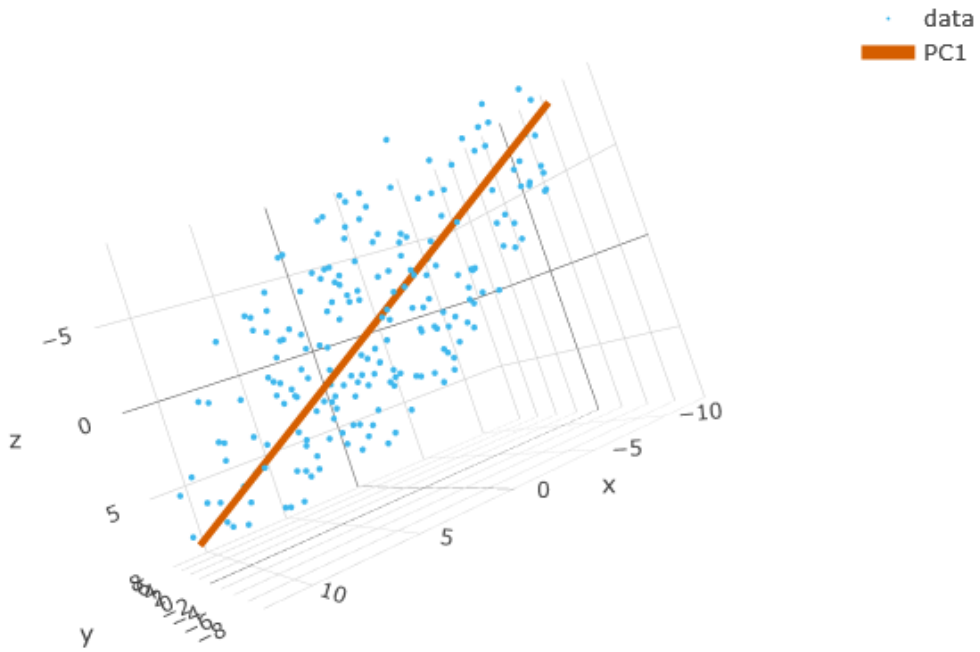
Visualizing PCA in 3D

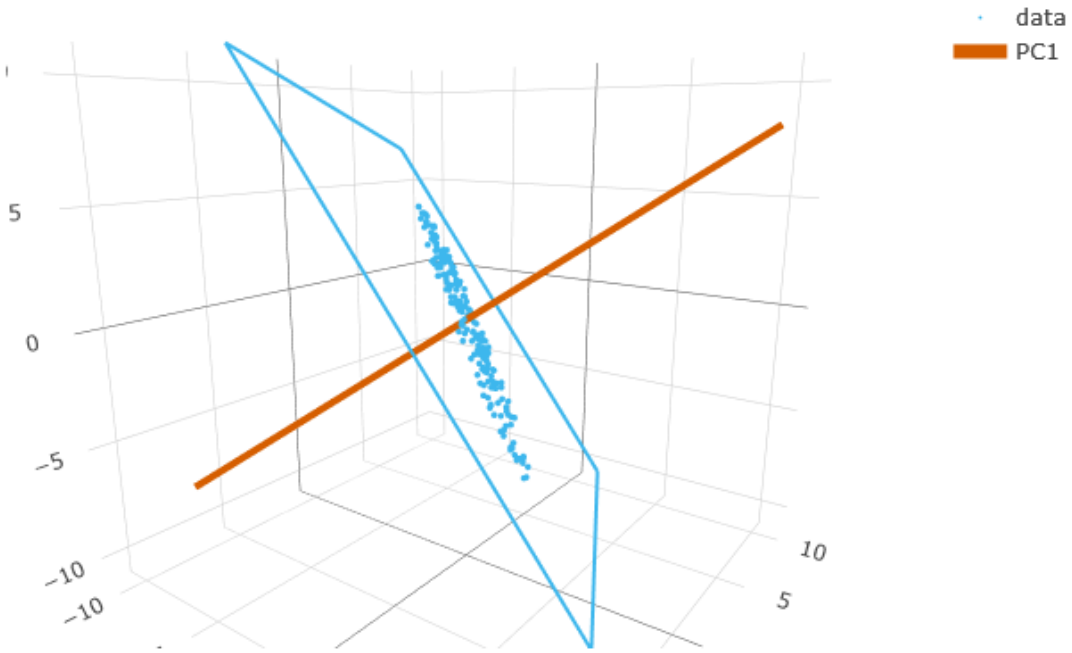
Use this link for interactive data manipulation:
https://bryanhanson.github.io/LearnPCA/articles/Vig_05_Visualizing_PCA_3D.html

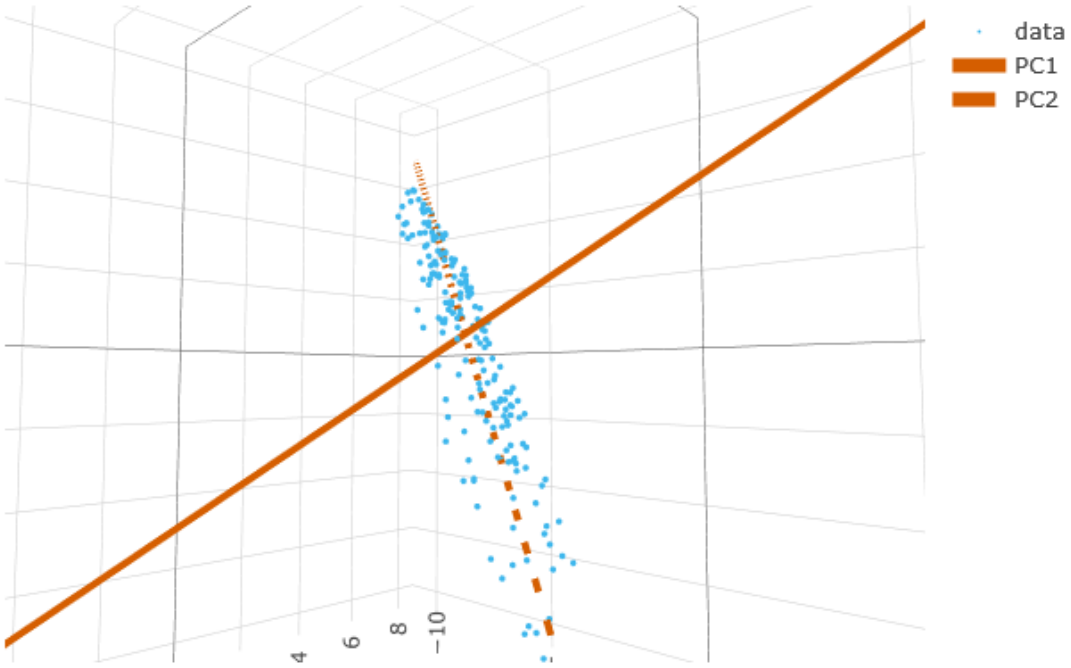
The images hereunder are extracted from these animations.

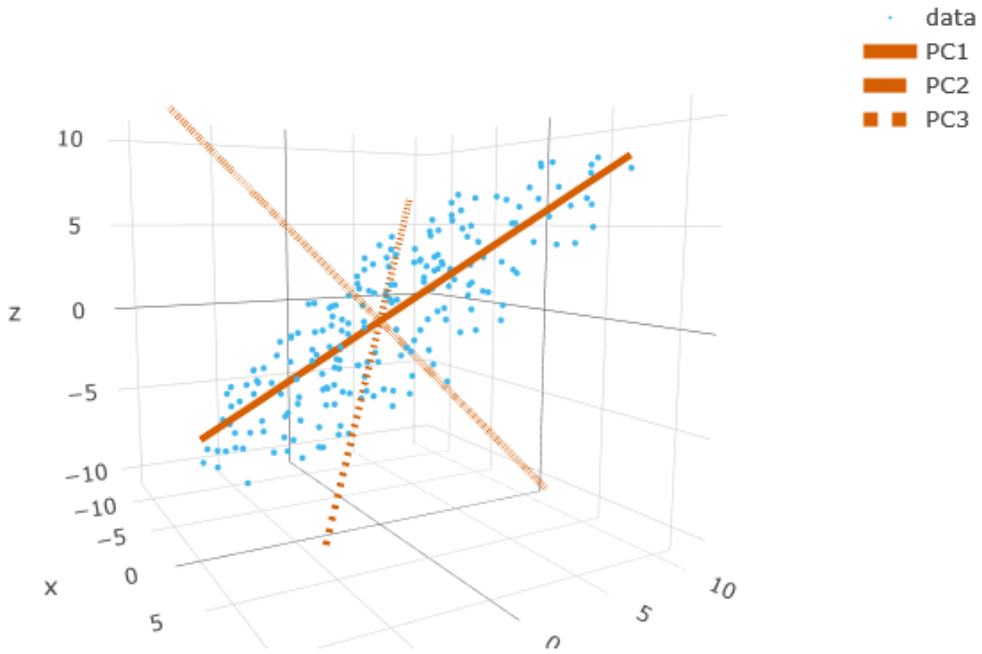





data
PC1
z
−5
0
5
y
10
5
0
−5
−10
x


data
PC1
5
0
−5
−10
−10
5
10


data
PC1
PC2
4
6
8
−10


data
PC1
PC2
PC3
10
5
z
0
−5
−10
−10
−5
x
0
5
10
5
0

References

1. **Jolliffe, I. T. (2002)**. *Principal Component Analysis* (2nd ed.). Springer Series in Statistics. Springer.

2. **Hastie, T., Tibshirani, R., & Friedman, J. (2009)**. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

3. **Shlens, J. (2014)**. *A Tutorial on Principal Component Analysis*. Available online through arXiv.

4. **Abdi, H., & Williams, L. J. (2010)**. *Principal component analysis*. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.

5. **Izenman, A. J. (2008)**. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

6. **Lever, J., Krzywinski, M., & Altman, N. (2017)**. *Points of Significance: Principal Component Analysis*. *Nature Methods*, 14(7), 641-642.

7. **Wold, S., Esbensen, K., & Geladi, P. (1987)**. *Principal Component Analysis*. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.

8. **Jackson, J. E. (1991)**. *A User's Guide to Principal Components*. John Wiley & Sons.

9. **Bishop, C. M. (2006)**. *Pattern Recognition and Machine Learning*. Springer.

10. **Pearson, K. (1901)**. *On Lines and Planes of Closest Fit to Systems of Points in Space*. *Philosophical Magazine*, 2(11), 559-572.

The last one is the original paper where Karl Pearson introduced the concept of PCA as a way of fitting lines and planes to points in multi-dimensional space)