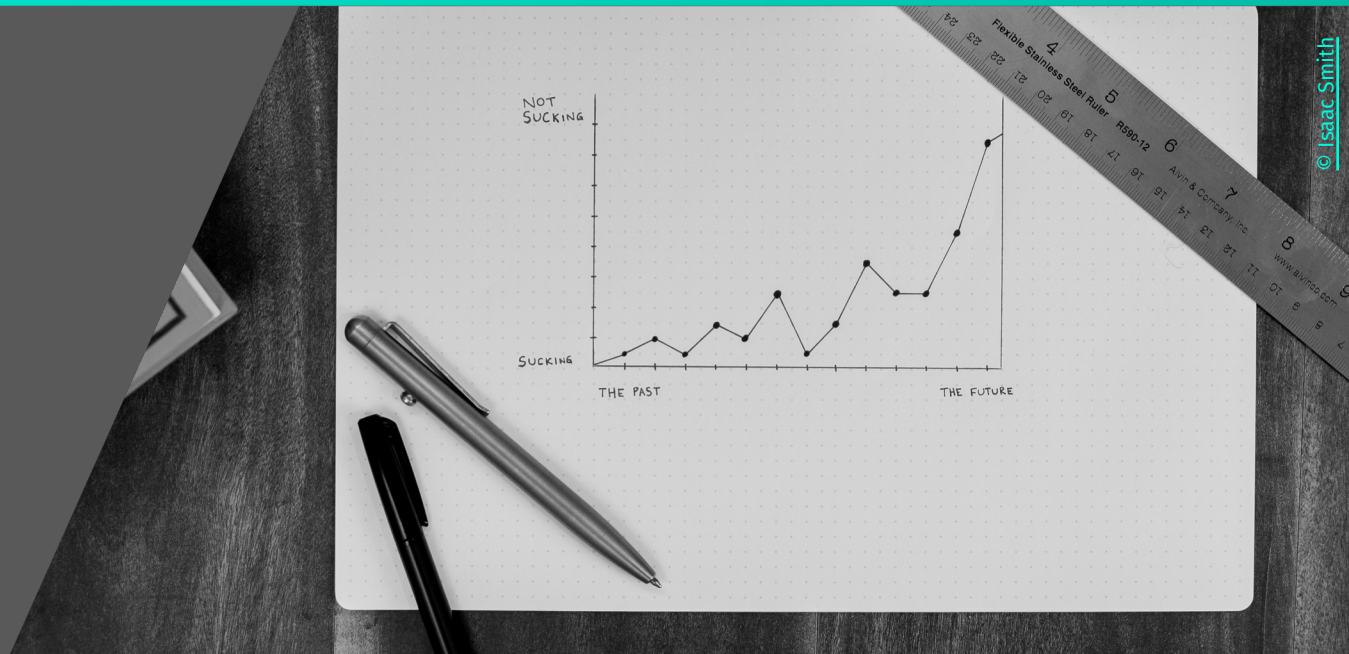
DATA VISUALIZATION

Part 8: Linear Projections

Prof. Dr. Heike Leitte

Visual Information Analysis

RPTU Kaiserslautern-Landau

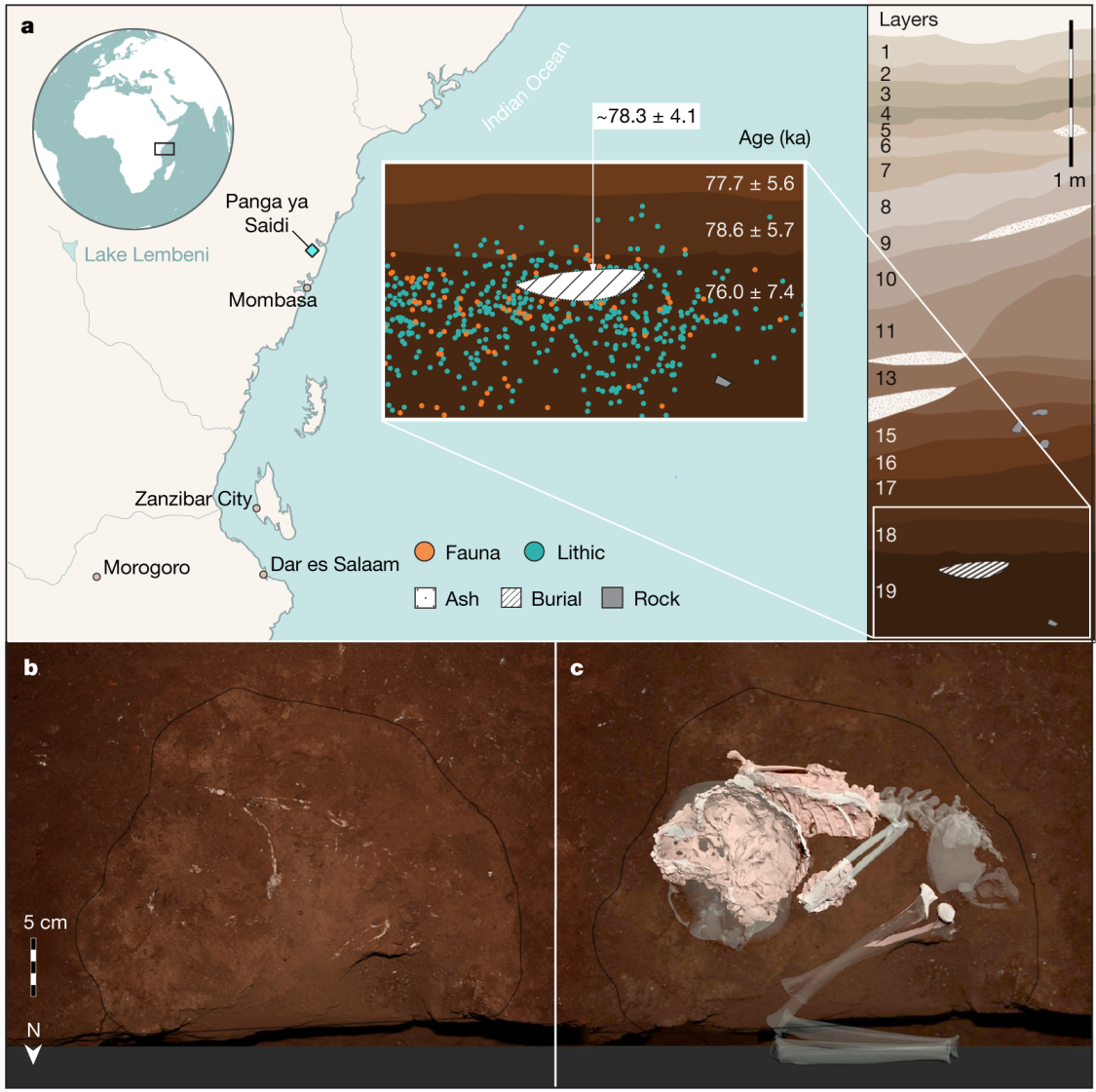


Lecture Outline

1. Introduction
2. Data visualization in the Jupyter Notebook
3. Charts
4. Perception and design
5. Visualization process
6. Vis of high-dimensional/multivariate data
7. Interaction with charts
8. **Linear Projections in visualization (PCA)**
9. Visualization of graphs
10. 2D scalar field visualization

Earliest known human burial in Africa

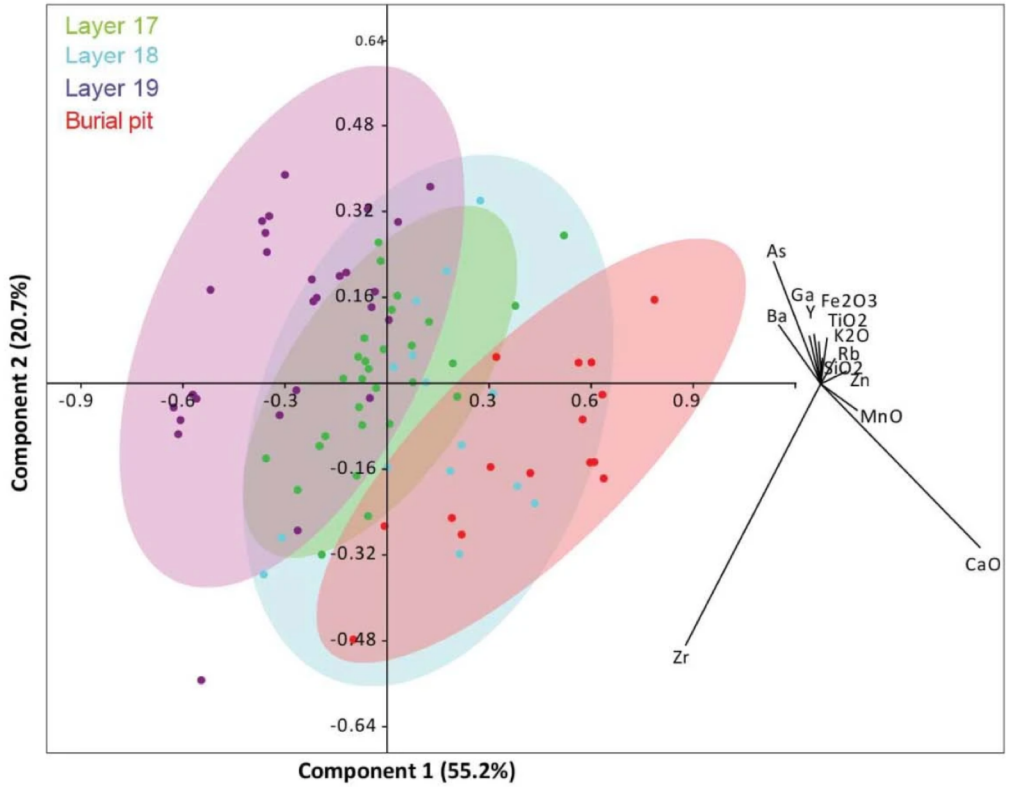
General view of the cave site of Panga ya Saidi where burial was unearthed.
A partial skeleton of a roughly 2.5- to 3.0-year-old child was found dating to 78.3 ± 4.1 thousand years ago.
© Mohammad Javad Shoaee



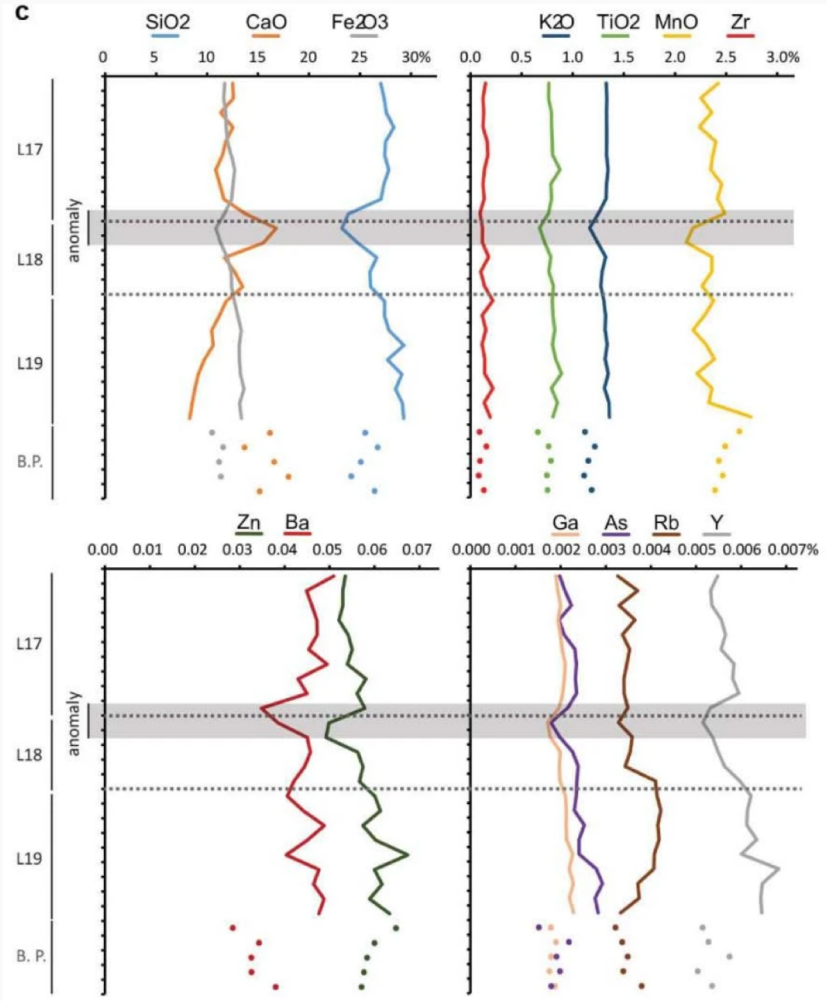
Martinón-Torres, M., d'Errico, F., Santos, E. et al. Earliest known human burial in Africa. Nature 593, 95–100 (2021). https://doi.org/10.1038/s41586-021-03457-8

Analysis of sediment samples

80 samples with 13 components



Results of PCA of the centre log ratio data (selected elements: SiO2, K2O, TiO2, MnO, Fe2O3, Zn, Ga, As, Rb, Y, Zr and Ba). Confidence ellipses at 95%. The burial pit samples markedly differ from layer 19.



Elemental profiles of sediment from layers 17, 18, 19 and the burial pit. Element concentrations are expressed in percentages. Data are presented as mean values. Sediment samples from the burial pit display an elemental composition notably similar to the three samples identified as an anomaly at the top of layer 18 and the base of layer 17.

General idea

Problem statement

Direct mapping of high-dimensional data has multiple issues:

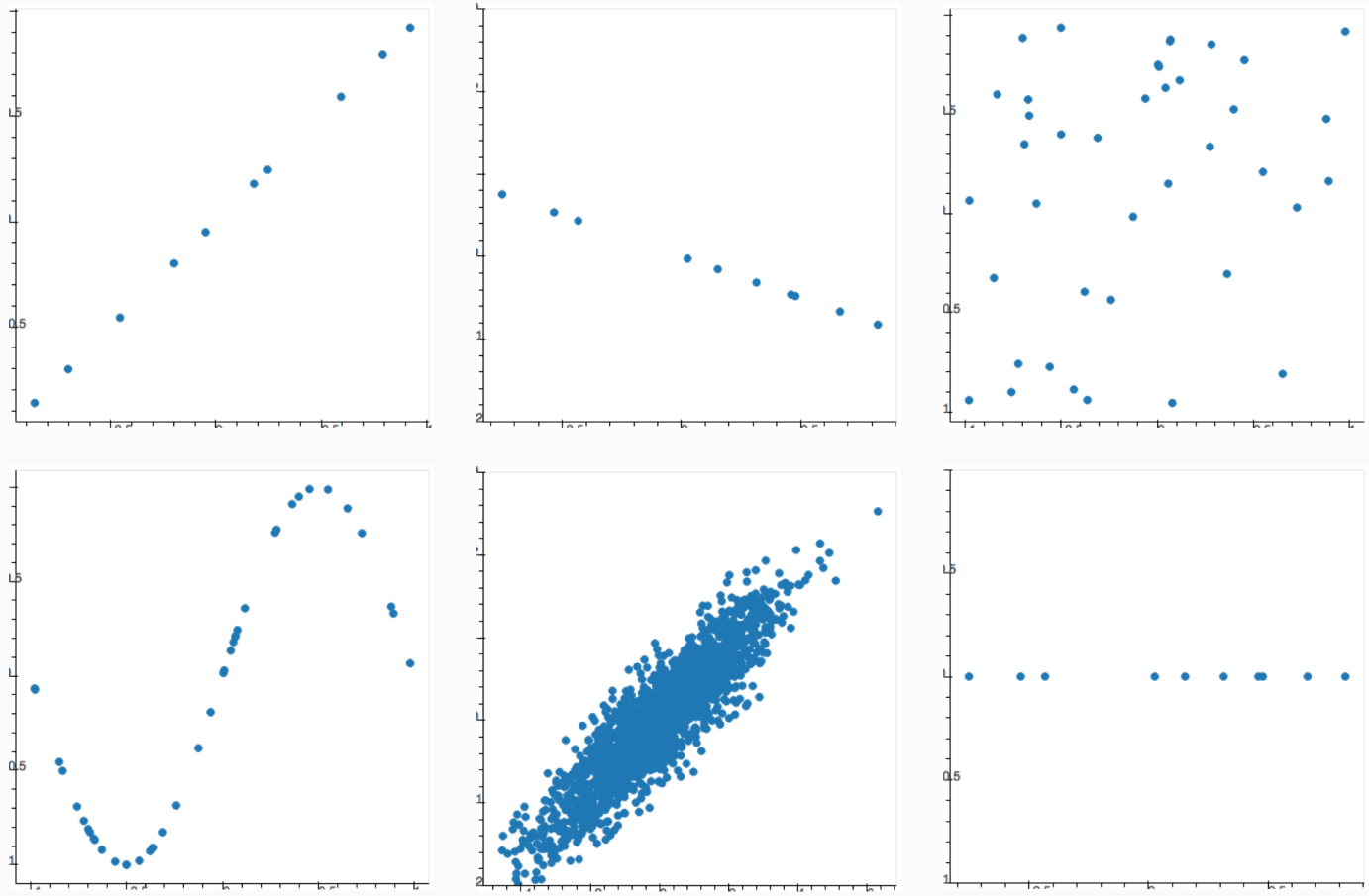
- Plots in one variable (bar charts/line charts) do not show correlations.
- Multivariate plots (scatter plot matrices, parallel coordinates) are not feasible for more than approx. 10-15 variables.
- Patterns become difficult to spot due to the huge amount of visual input.

Can we do some preprocessing that makes our charts more expressive?

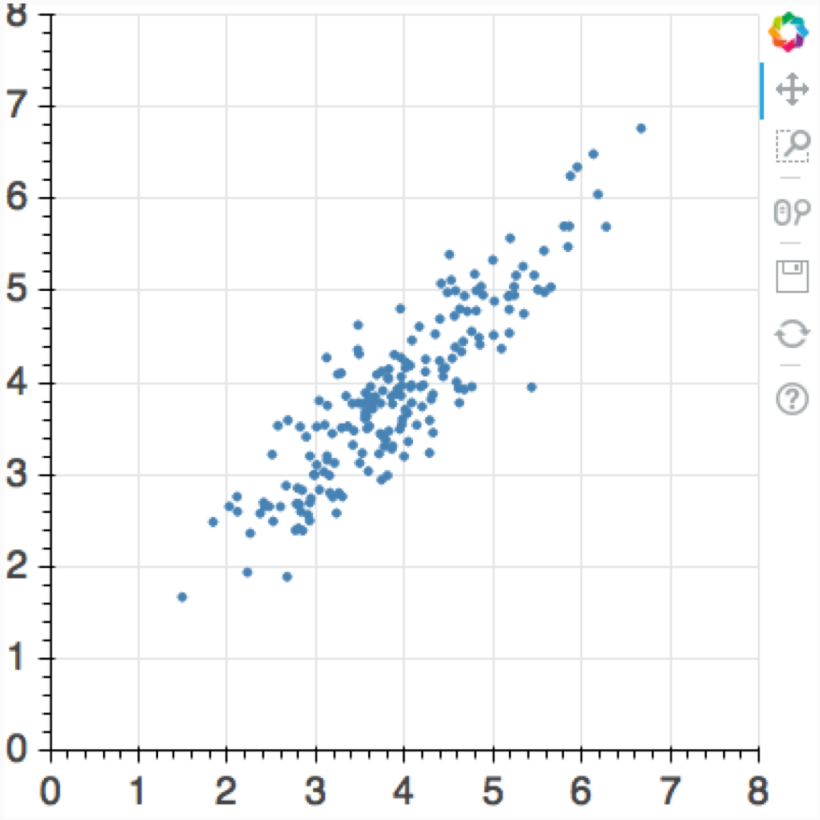
Goal: reduce the number of variables/features without loosing too much information

Intrinsic dimensionality

How many dimensions do you need to describe the data?



Choosing a better coordinate system



PCA

Principal component analysis (PCA)

origin: center of points
axes: orthogonal + max variance

Given: high-dimensional data of any size

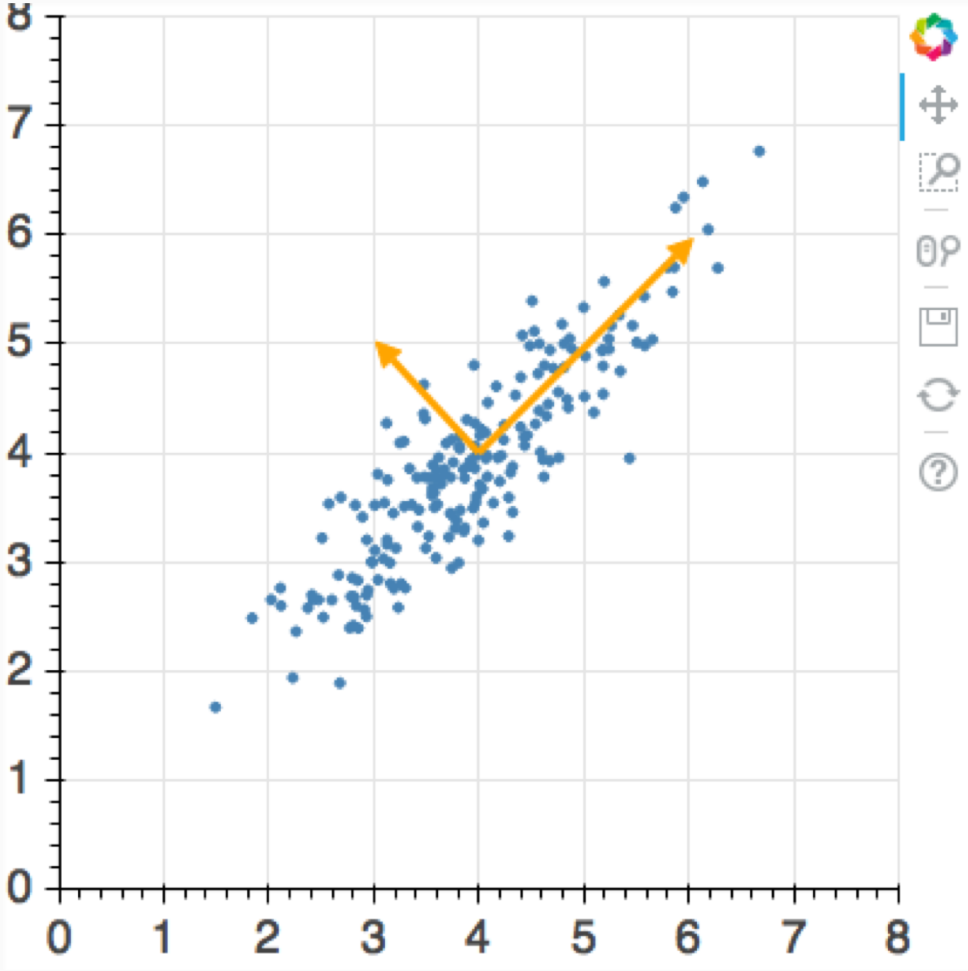
PCA finds a new coordinate system obtained from the old one by rotation and scaling only.

The origin of the new coordinate system is the center of the data.

The first axis is aligned with the principal direction of variation in the data.

All other axes are orthogonal to the previous ones and are sorted by decreasing amount of variation in the data.

So, we are trying to find a high-dimensional ellipsoid that is as small as possible and still contains most of the data points.



Definition Variance & Covariance

The variance of a random variable X is the expected value of the squared deviation from the mean of X ($\mu = E[X]$):

$$Var(X) = E[(X - \mu)^2].$$

The covariance between two jointly distributed real-valued random variables X and Y with finite second moments is defined as the expected product of their deviations from their individual expected values:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Variance in sampled data

How to compute variance for sampled data?

Definition: The sample mean, denoted $\bar{x}$, is the average of the n data points $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sample mean summarizes the "location" or "center" of the data.

Definition: The sample variance, denoted s^2 summarizes the "spread" or "variation" of the data:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Covariance matrix

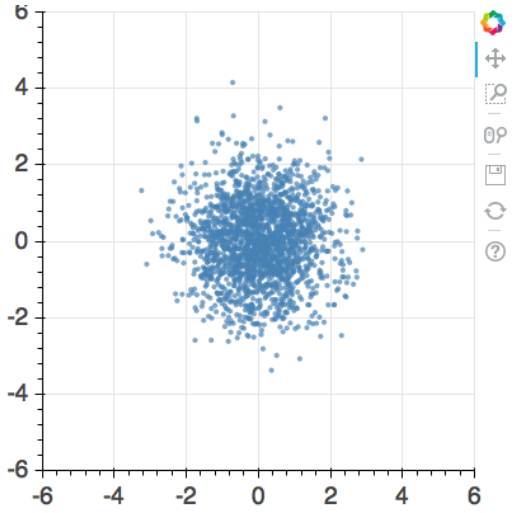
The covariance matrix Σ summarizes the covariance for all pairwise combinations of k features.

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & \dots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) & \dots & Cov(X_2, X_k) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) & \dots & Cov(X_3, X_k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & Cov(X_k, X_3) & \dots & Var(X_n) \end{bmatrix}$$

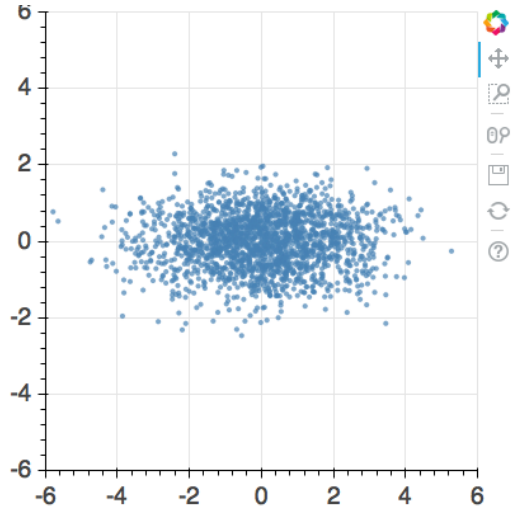
Note: $Var(X) = Cov(X, X)$

Covariance matrix – examples

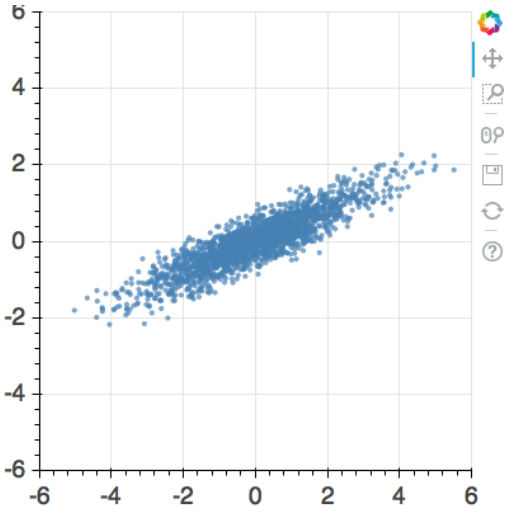
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 2.5 & 1 \\ 1 & 0.5 \end{bmatrix}$$



PCA computation

Goal: Find the principal components/factors of a distribution

1. Characterize the distribution by
   - covariance matrix (similar variable scales, normalized data)
   - correlation matrix (different variable scales)
2. Perform a Eigendecomposition of the matrix C to get

$$C = Q \Lambda Q^{-1} \quad [Eq.1]$$

   Q: matrix with Eigenvectors as columns
   Λ: diagonal matrix with Eigenvalues λ

3. Order the Eigenvectors in terms of their Eigenvalues λ

PCA as rotation and scaling

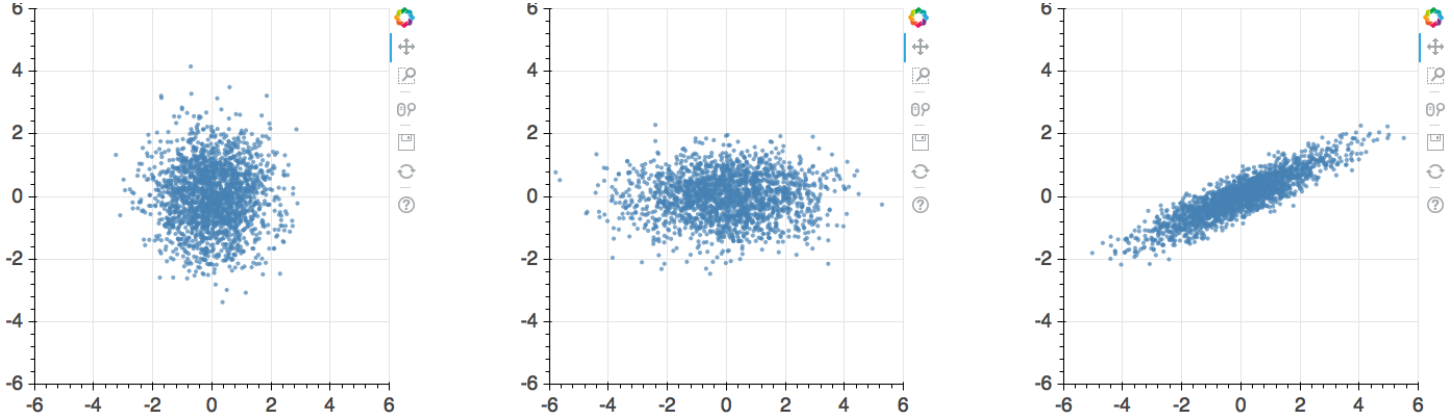
The eigendecomposition of the covariance matrix, as given in Eq. 1 on the last slide, extracts the rotation and scaling matrices of the linear transformation:

$$\begin{aligned}\Sigma &= Q\,\Lambda\,Q^{-1} \\ &= R\,S\,S\,R^{-1}\end{aligned}$$

where $R = Q$ is a rotation matrix and $S = \sqrt{\Lambda}$ is a scaling matrix.

PCA as rotation and scaling – examples

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2.5 & 1 \\ 1 & 0.5 \end{bmatrix}$$



$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad Q = \begin{bmatrix} 0.92 & 0.38 \\ 0.38 & -0.92 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Lambda = \begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \Lambda = \begin{bmatrix} 2.91 & 0 \\ 0 & 0.09 \end{bmatrix}$$

PCA interpretation

Outline for PCA analysis

1. Preprocess data
2. Check correlation
3. Compute PCA
4. Check explained variance → number of principal components
5. Analyze biplot → data groups + relationship with features
6. Analyze principal axes

Step 2: Correlation analysis

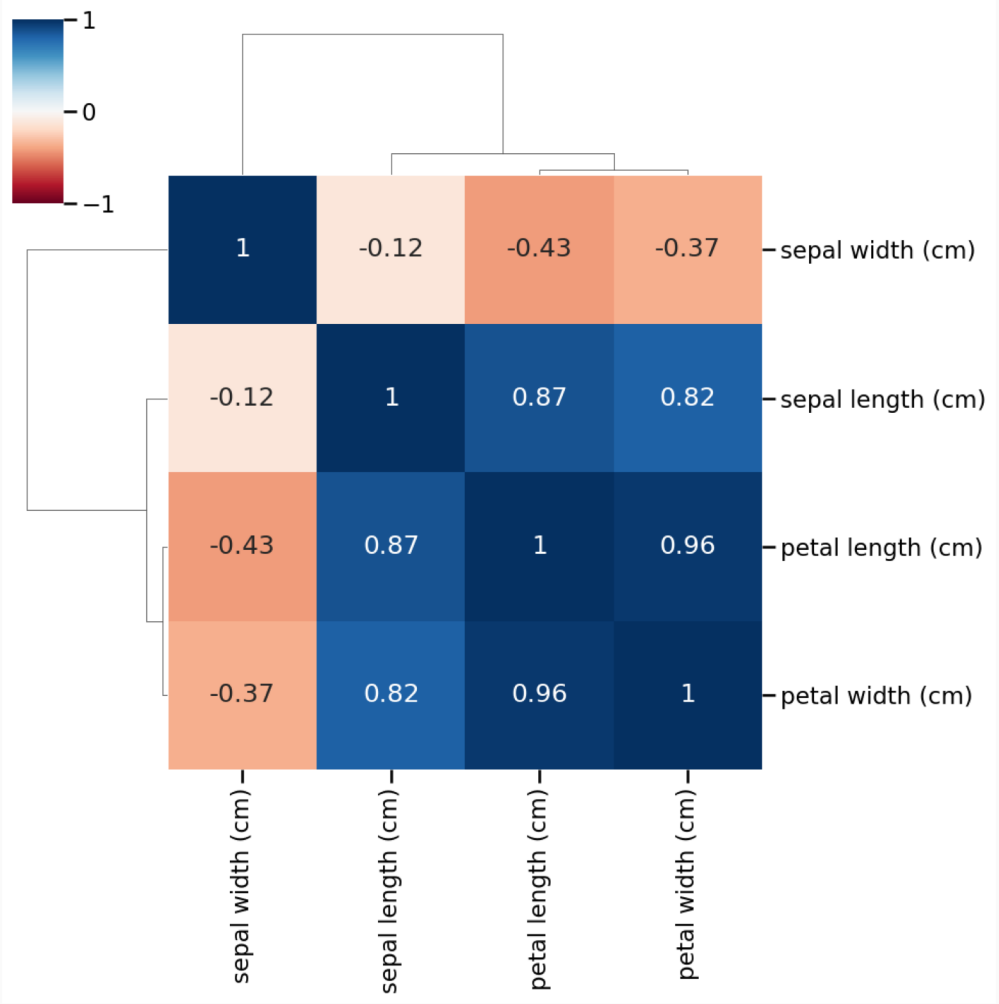
PCA only works if data is correlated.

Check that your data is correlated. PCA can only reduce the number of dimensions if the data features (linear) correlation. You can measure correlation using correlation coefficients and inspect it visually using heatmaps. The cluster map (right) groups attributes that have a similar correlation profile.

Python code (seaborn)

```
sns.clustermap(df.corr(), cmap='RdBu‘,
               vmin=-1, vmax=1, annot=True)
```

Correlation matrix of the iris dataset



Step 2: Correlation analysis

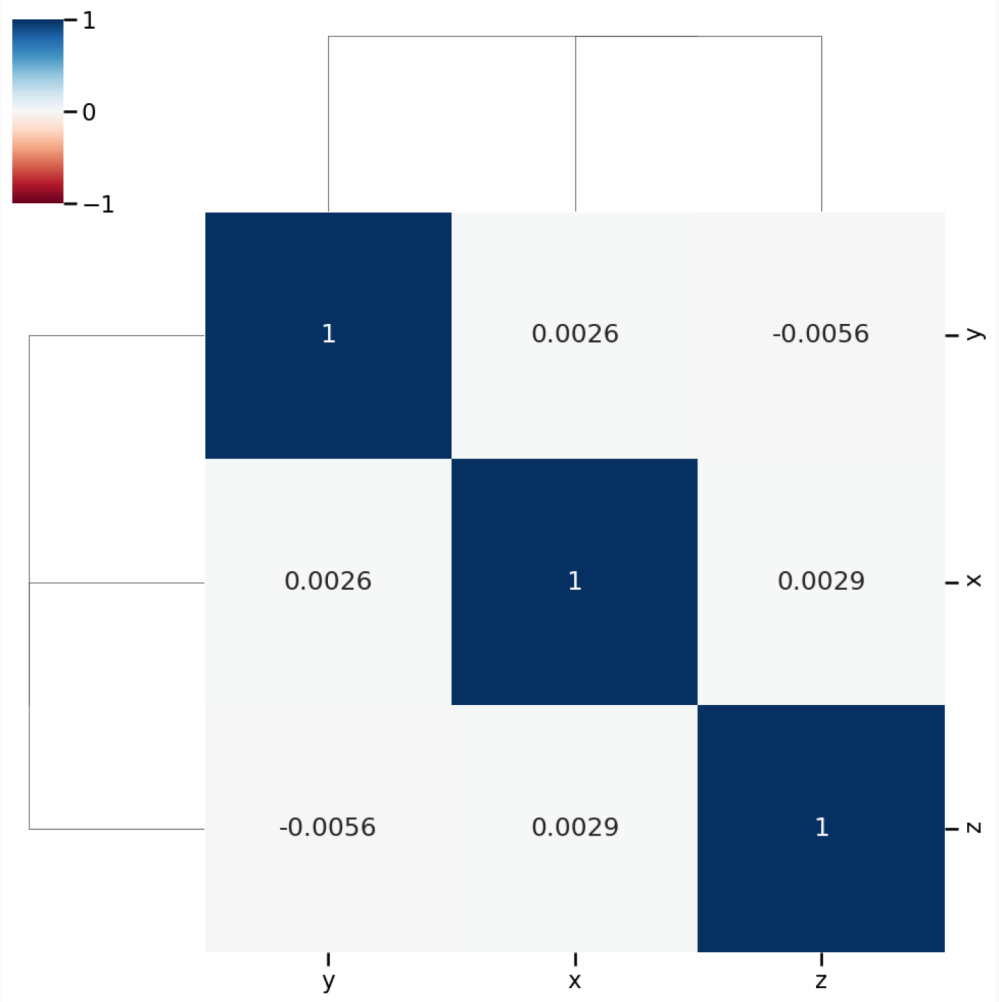
PCA only works if data is correlated.

Counter example on the right.

Data sampled from a 3D Gaussian distribution does not feature correlation between variables.

PCA will not be able to reduce this data.

Correlation matrix of 3D Gaussian data



Step 4: Explained variance

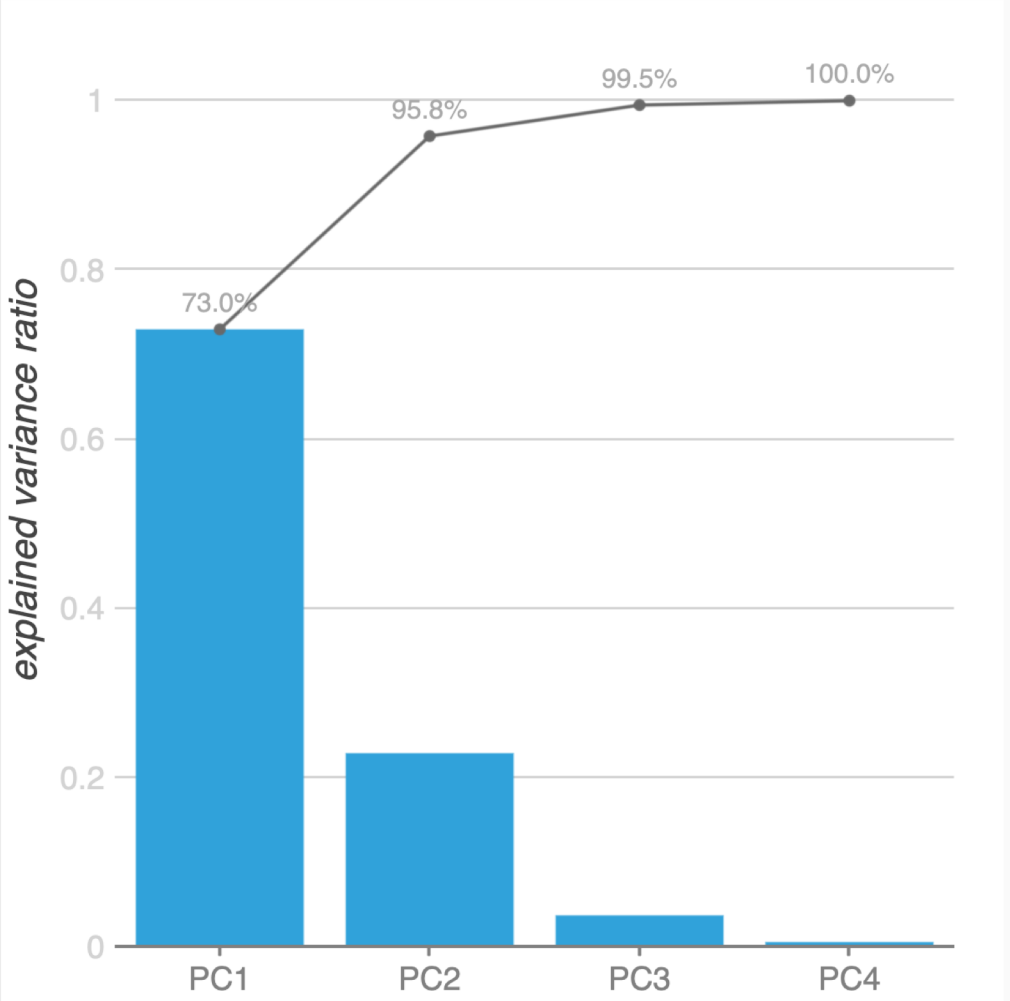
Find the number of relevant principal axes

Python code (seaborn)

```
from sklearn.decomposition import PCA

pca = PCA()
pca.fit(X_scaled)
pca.explained_variance_ratio_
```

Explained variance of the iris dataset

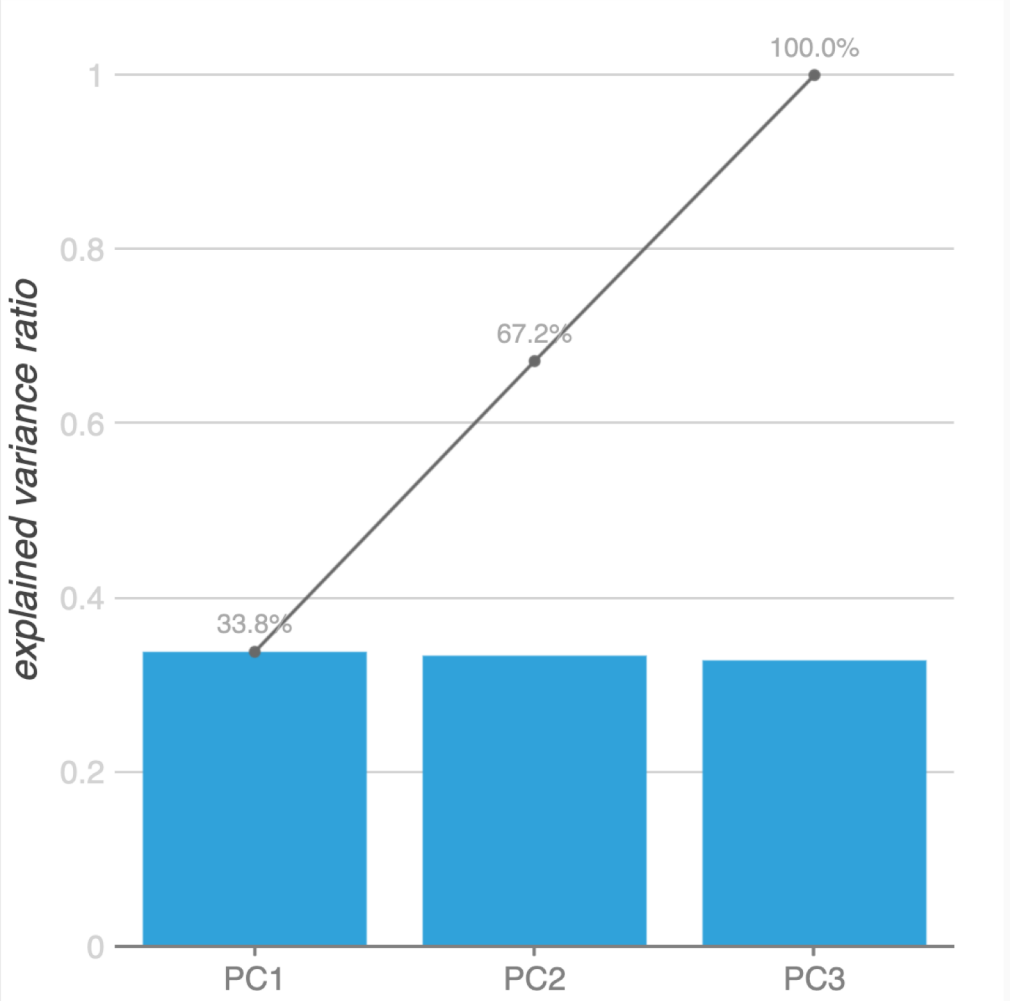


Step 4: Explained variance

Find the number of relevant principal axes

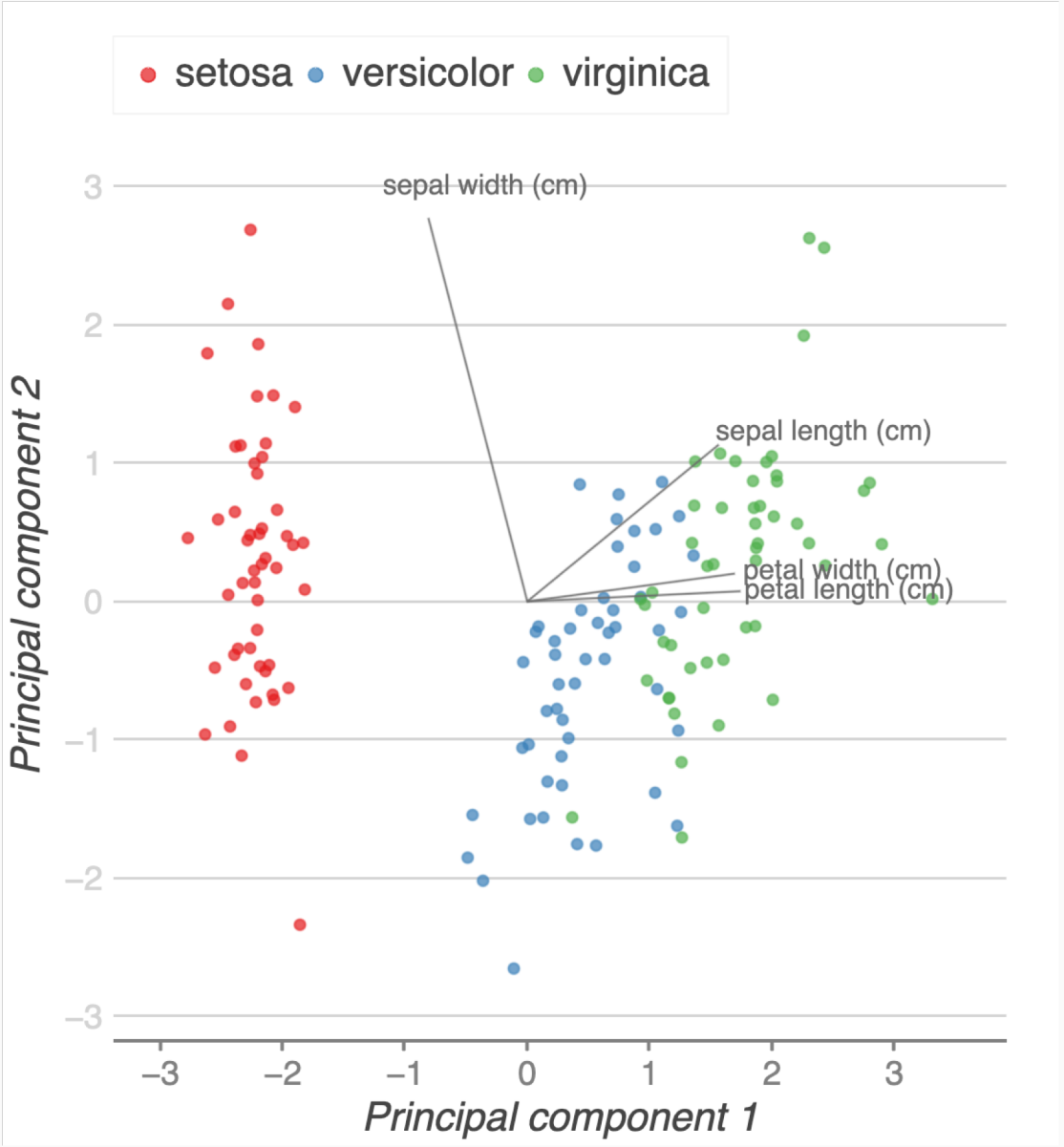
Counter example for PCA. We observe that all variables are (equally) important.

Explained variance of the 3D Gauss dataset



Step 5: Biplot

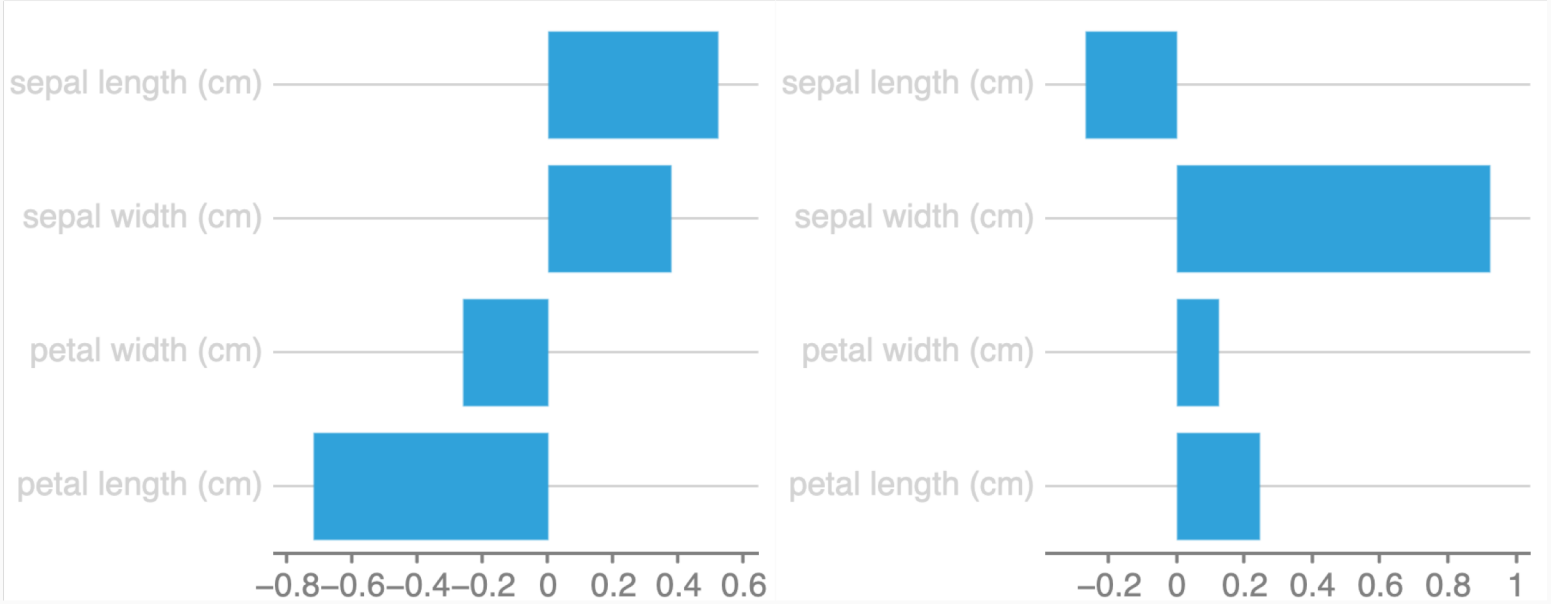
Explore data wrt the first two principal axes



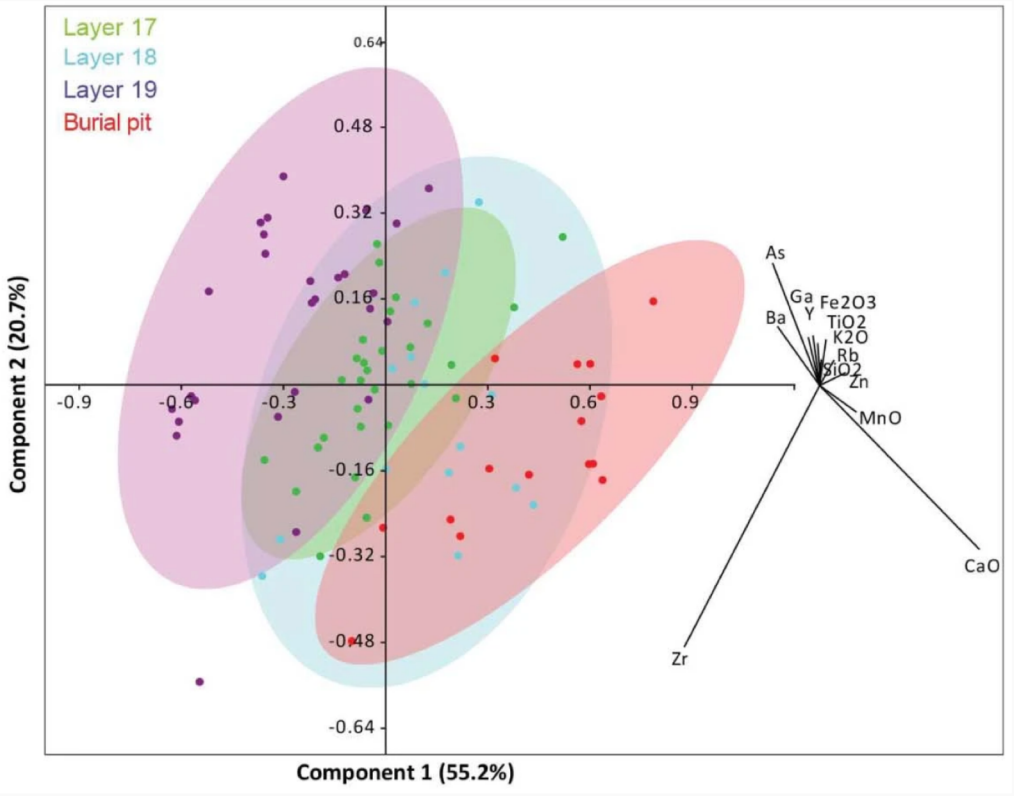
Biplot of the iris dataset

Step 6: Principal components

Understand how features and PC relate



PCA analysis of burial data



Results of PCA of the centre log ratio data (selected elements: SiO2, K2O, TiO2, MnO, Fe2O3, Zn, Ga, As, Rb, Y, Zr and Ba). Confidence ellipses at 95%. The burial pit samples markedly differ from layer 19.

Summary