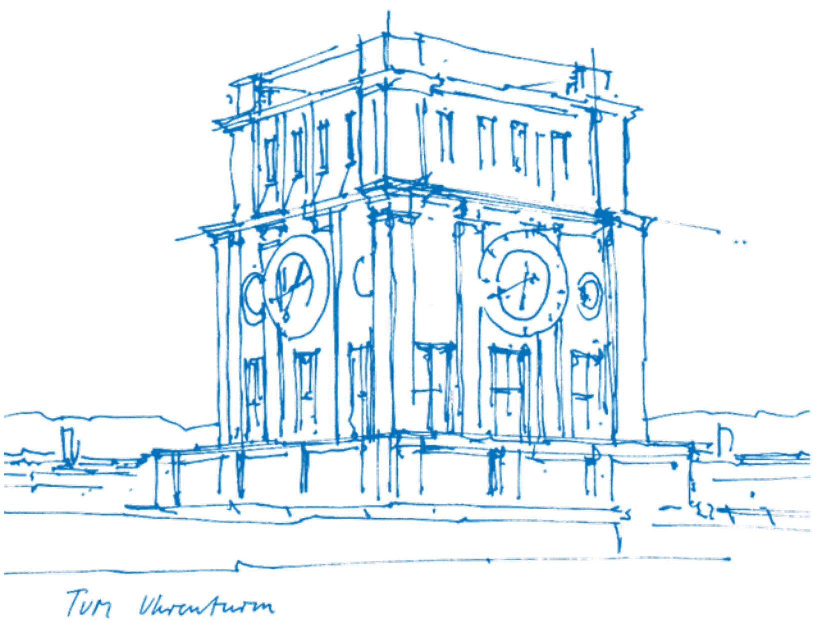
# Representation of data: Principal Components

Felix Dietrich



# Representation of data

## High-dimensional data with low-dimensional structure - general idea

1. Given input: data matrix $X \in \mathbb{R}^{N \times n}$ with $N$ data points in $n$-dimensional space.
2. ...algorithm...
3. Output: new representation of the data, e.g. as another coordinate matrix $U \in \mathbb{R}^{N \times p}$.

Ideally: $p \ll n$, so that the dimension of the data is reduced (manifold learning, compression).
For visualization, $p = 2, 3, (4)$ is necessary.
Example for a low-dimensional structure: $U \in \mathbb{R}^{1000 \times 3}$ with rows $u_i \in \mathbb{R}^3$, $\|u_i\| = 1$:

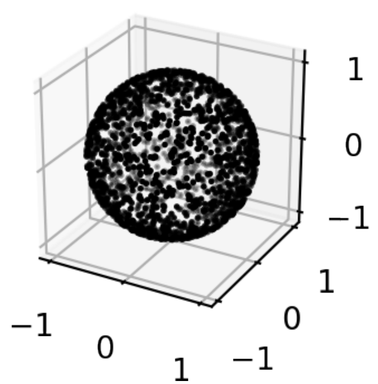



Figure: Data set where the points $u_i$ (black) are distributed on a sphere.

# Representation of data

## Challenges to be solved

1. High-dimensional ambient space (example: images, many sensors, brain wave detector, LHC, ...)
2. Complicated structure of data (fractal, sparse, noisy - instead of spherical, planar, periodic, ...)
3. Visualization of intrinsically high-dimensional data (social graphs, biological networks, economy)
4. Generation of new data from observations without running more experiments

# Representation of data

## Principal Component Analysis

1. Basic idea: approximate the “best” linear subspace in which the data $X$ lies [Hotelling, 1933, Hotelling, 1936].
2. Algorithm: iteratively find orthogonal directions with largest variance in the data.
3. Challenge: what if the data is highly nonlinear?

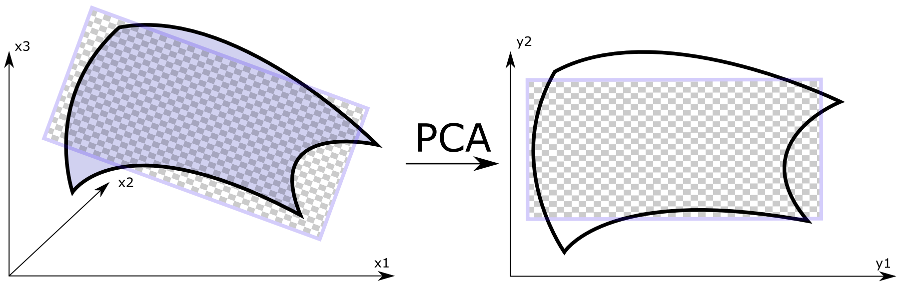



Figure: An embedding of a manifold in two-dimensional Euclidean space, using Principal Component Analysis. The manifold on the left is already embedded into three-dimensional Euclidean space, but PCA is able to find a two-dimensional embedding, because the manifold is almost planar.

# Representation of data

## Principal Component Analysis

Given a data set $\{x_i \in \mathbb{R}^n\}_{i=1}^N$:

1. Form the data matrix $X \in \mathbb{R}^{N \times n}$ with rows $x_i$ from points (observations) in the data set.
2. Center the matrix by removing the data mean $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$ from every row (every data point):
$$\overline{X}_{ij} = X_{ij} - \overline{x}_j.$$
3. Decompose the centered data matrix into singular vectors $U$, $V$ and values $S$, such that
$$\overline{X} = USV^T,$$
where $U \in \mathbb{R}^{N \times N}$, $S \in \mathbb{R}^{N \times n}$, and $V \in \mathbb{R}^{n \times n}$.
4. The “energy” (explained variance) of the $i$-th principal component is contained in the singular value $\sigma_i$ on the diagonal of the matrix $S$. The percentage of the total energy explained by using a certain number $L$ of principal components to describe the data can be computed through
$$\frac{1}{\text{trace}(S^2)} \sum_{i=1}^{L} \sigma_i^2,$$
where trace$(S^2)$ is the sum over all squared singular values (not just $L$).

# Representation of data

## Principal Component Analysis

Important things to consider:

1. The functions from and to the principal components are linear.
2. The representation of the data is exact if all singular values are kept. However, ignoring small singular values does not change the reconstructed data too much:
$$\|\overline{X} - U\hat{S}V^T\| \leq N\varepsilon$$
if $\hat{S}_{ii} = S_{ii}$ for $S_{ii} > \varepsilon$ and zero otherwise.
3. PCA can be applied to data on nonlinear manifolds, but there may be better (nonlinear) representations with fewer “components”.

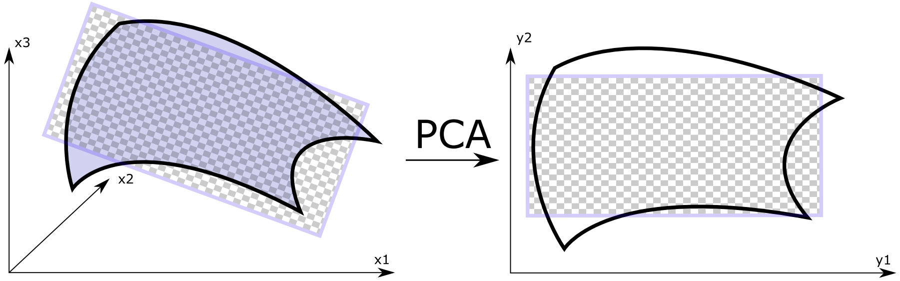



Figure: An embedding of a manifold in two-dimensional Euclidean space, using Principal Component Analysis. The manifold on the left is already embedded into three-dimensional Euclidean space, but PCA is able to find a two-dimensional embedding, because the manifold is almost planar.

# Representation of data

## Nonlinear manifold learning: Diffusion Maps (separate video!)

Important things to consider [Coifman and Lafon, 2006]:

1. The functions from and to the eigenfunction space are nonlinear.
2. The eigenvalues of the operator are not interpretable in the same terms of “energy” as in PCA.
3. Diffusion Maps works well if applied to densely sampled, nonlinear manifolds.

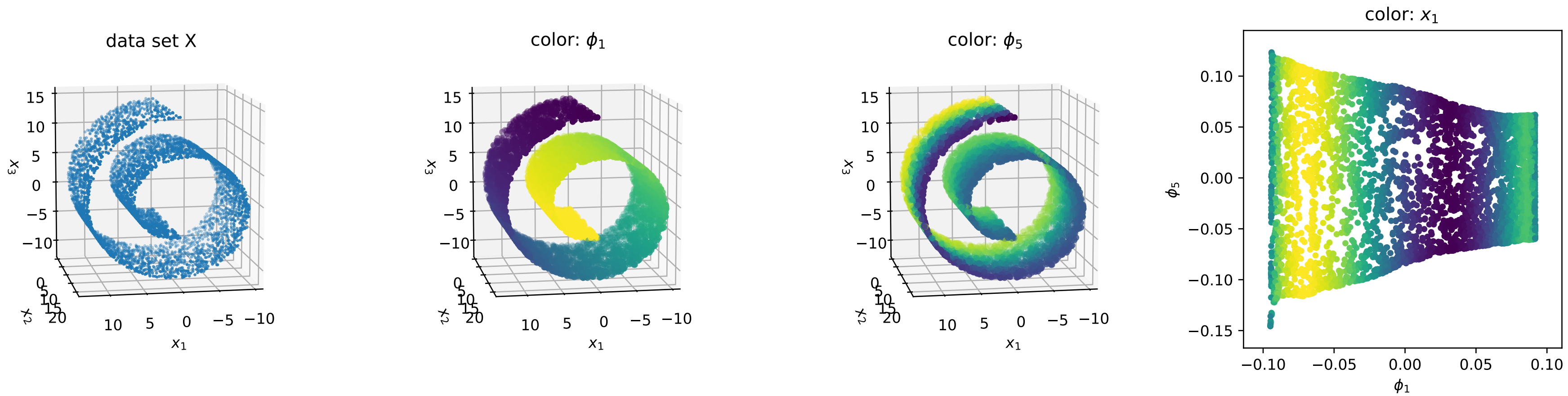



Figure: An embedding of a manifold in two-dimensional Euclidean space, using Diffusion Maps. The manifold on the left is already embedded into three-dimensional Euclidean space, but Diffusion Maps is able to find a two-dimensional embedding, even though the manifold is nonlinearly embedded.

# Literature I

Coifman, R. R. and Lafon, S. (2006).
Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions.
*Applied and Computational Harmonic Analysis*, 21(1):31–52.

Hotelling, H. (1933).
Analysis of a complex of statistical variables into principal components.
*Journal of Educational Psychology*, pages 417–441.

Hotelling, H. (1936).
Simplified calculation of principal components.
*Psychometrika*, 1(1):27–35.