Quiz: Dimensionality Reduction

Lecture series „Machine Learning“

Niels Landwehr

Research Group „Data Science“
Institute of Computer Science
University of Hildesheim

Quiz: Intrinsic Dimensionality of Data

- Assume a data set of 1000 points $\{\mathbf{x}_1, \ldots, \mathbf{x}_{1000}\}$ with $\mathbf{x}_n = (x_n, y_n, z_n) \in \mathbb{R}^3$ generated as follows:
  - Draw $a_n \sim uniform(0,1)$, $b_n \sim uniform(0,1)$
  - Set $x_n = a_n + b_n,\ \mathrm{y}_n = a_n^2 + b_n^2 + 2a_n b_n,\ z_n = -a_n - b_n$

- **Question**: What is the resulting intrinsic dimensionality of the data? In other words, which embedding dimension would a sufficiently large autoencoder model need to perfectly fit the data?
  - The intrinsic dimensionality of the data is 3, so three-dimensional embedding is needed
  - The intrinsic dimensionality of the data is 2, so a two-dimensional embedding is needed
  - The intrinsic dimensionality of the data is 1, so a one-dimensional is enough
  - Cannot say based on the data provided

Solution: Intrinsic Dimensionality of Data

- **Solution**: The intrinsic dimensionality of the data is one, so a one-dimensional embedding is enough

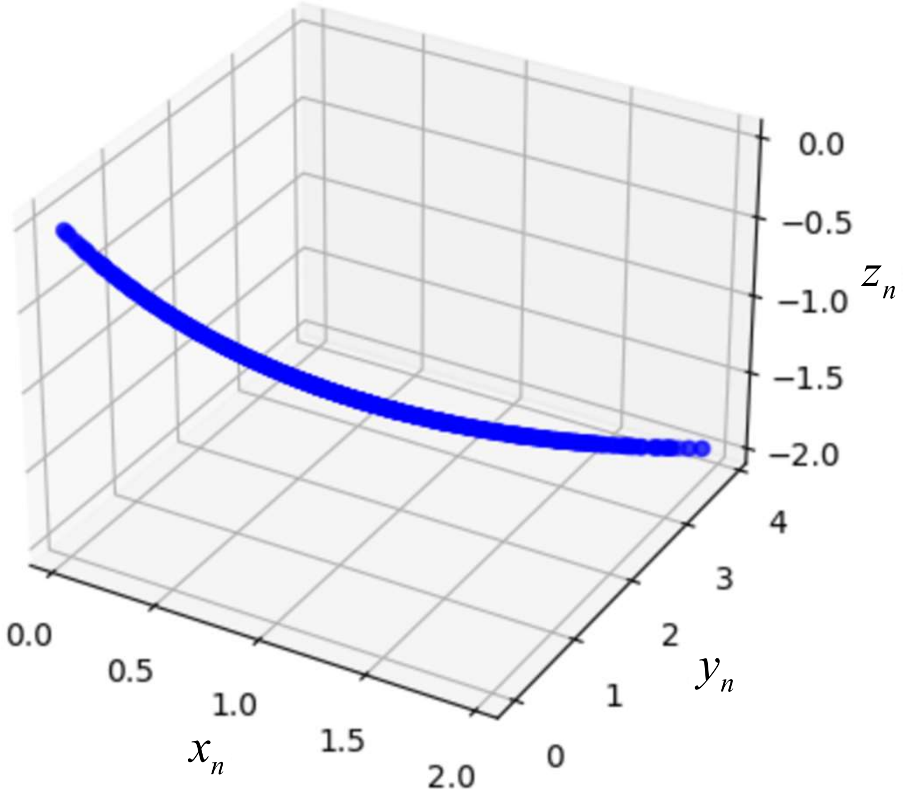
- We can rewrite the data generating process as follows:
  - Draw $a_n \sim uniform(0,1)$, $b_n \sim uniform(0,1)$
  - Set $c_n = a_n + b_n$
  - Set $x_n = c_n, \mathrm{y}_n = c_n^2, z_n = -c_n$

- Now it is clear that the data is one-dimensional, because running through all possible values of c_n will generate all possible data points

Solution: Intrinsic Dimensionality of Data

- **Solution**: The intrinsic dimensionality of the data is one, so a one-dimensional embedding is enough

- Here is a plot of the data:



Quiz: PCA

- Assume the same data set $\{\mathbf{x}_1,\ldots,\mathbf{x}_{1000}\}$ as above:
  - Draw $a_n \sim uniform(0,1)$, $b_n \sim uniform(0,1)$
  - Set $x_n = a_n + b_n,\ \mathrm{y}_n = a_n^2 + b_n^2 + 2a_n b_n,\ z_n = -a_n - b_n$
- **Question**: what is the needed number of components *K* in a PCA so that the PCA can perfectly reconstruct the data (without reconstruction error)?
  - *K*=3 is needed, data does not lie in any linear subspace (except the original space)
  - *K*=2 is needed, data lies on a two-dimensional linear subspace
  - *K*=1 is needed, data lies on a one-dimensional linear subspace

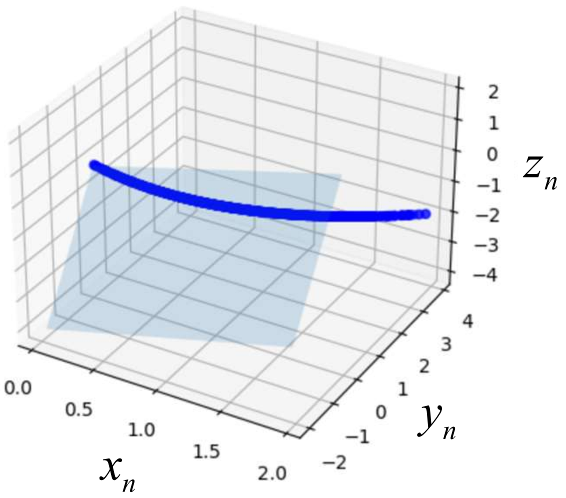
Solution: PCA

- **Solution**: *K*=2 is needed, data lies on a two-dimensional linear subspace

- A stated above, the data has one intrinsic dimension c, and all three coordinates in the original space can be computed from that intrinsic dimension:

$$x_n = c_n \qquad y_n = c_n^2 \qquad z_n = -c_n$$

- The first coordinate $x_n = c_n$ and the last coordinate $z_n = -c_n$ are linear functions of the intrinsic dimension c, so they lie within a one-dimensional linear subspace
- The second coordinate $y_n = c_n^2$ however depends non-linearly on c, thus one more dimension in the PCA for this coordinate is needed


2
1
0
−1
−2
−3
−4
z_n
4
3
2
1
0
−1
−2
y_n
0.0
0.5
1.0
1.5
2.0
x_n