Clustering, principal component analysis, and autoencoders

June 25, 2025

Reminder

1. **Exercises**: solve 8 – 10 questions. Sumission deadline: **23.07.2025**
2. **Course project**: Select 1 topic to conduct numerical experiment, and write a short report. Submission deadline: **27.07.2025**
3. give a 10-15 min **presentation**. Date: 11.07 or 18.07.

Remark:

1. Two persons can work together on the course report. In this case, the project should be more comprehensive. Each should contribute equally to the experiment and to the writing.
2. Code of conduct for scientific writing, e.g. use of materials of this course, online materials, and ChatGPT!

Plan for the next 3 weeks

Lectures:

1. 02.07: transition pathways, string method.
2. 09.07: transition pathway theory on graphs, committor.
3. 16.07: summary

Practice:

1. 04.07: numerical examples
2. 11.07: numerical examples; presentation
3. 18.07: presentation; discussion on exercises

Machine learning tasks

1. Supervised learning

   Dataset contains features and labels: $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$.

   1. Classification (pattern recognition)
   2. Regression

Machine learning tasks

1. Supervised learning

   Dataset contains features and labels: $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$.

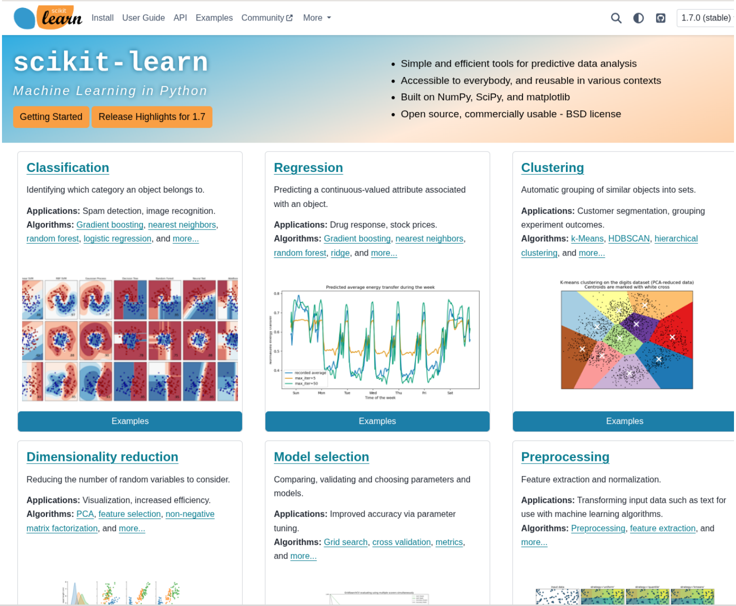
   1. Classification (pattern recognition)
   2. Regression

2. **Unsupervised learning**

   Data: $\mathcal{D} = \{x_n\}_{n=1}^{N}$.

   - clustering: k-means
   - discovering latent structure: PCA
   - generative modeling

ML package: scikit-learn



scikit learn | Install | User Guide | API | Examples | Community | More | 1.7.0 (stable)

scikit-learn

Machine Learning in Python

Getting Started | Release Highlights for 1.7

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

Predicted average energy transfer during the week

Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...

K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross

Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency.
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning.
Algorithms: Grid search, cross validation, metrics, and more...

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: Preprocessing, feature extraction, and more...

Part 1: K-means clustering

Clustering task

1. data points $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$

Clustering task

1. data points $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$
2. goal: partition data into k sets $S = \{S_1, S_2, \ldots, S_k\}$

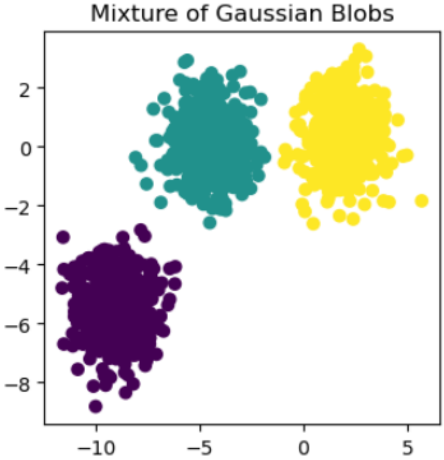
Clustering task

1. data points $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$
2. goal: partition data into k sets $S = \{S_1, S_2, \ldots, S_k\}$
3. *within-cluster sum-of-squares*:

$$\arg\min_S \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right) = \arg\min_S \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i)\,,$$

where $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$ is the mean (also called centroid) of points in S_i.

Clustering task



Clustering task

1. *within-cluster sum-of-squares*:

$$\arg\min_{S} \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right) = \arg\min_{S} \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i) \,,$$

where $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$.

Clustering task

1. *within-cluster sum-of-squares*:

$$\arg\min_S \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right) = \arg\min_S \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i)\,,$$

where $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$.

2. The identity

$$|S_i| \sum_{x \in S_i} |x - \mu_i|^2 = \frac{1}{2} \sum_{x,y \in S_i} |x - y|^2$$

Clustering task

1. *within-cluster sum-of-squares*:

$$\arg\min_S \sum_{i=1}^{k} \Big(\sum_{x \in S_i} |x - \mu_i|^2 \Big) = \arg\min_S \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i)\,,$$

where $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$.

2. The identity

$$|S_i| \sum_{x \in S_i} |x - \mu_i|^2 = \frac{1}{2} \sum_{x,y \in S_i} |x - y|^2 = \sum_{x \in S_i} |x|^2 - (\sum_{x \in S_i} x)^2\,.$$

Clustering task

1. *within-cluster sum-of-squares*:

$$\arg\min_S \sum_{i=1}^{k} \Big(\sum_{x \in S_i} |x - \mu_i|^2 \Big) = \arg\min_S \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i) ,$$

   where $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$.

2. The identity

$$|S_i| \sum_{x \in S_i} |x - \mu_i|^2 = \frac{1}{2} \sum_{x,y \in S_i} |x - y|^2 = \sum_{x \in S_i} |x|^2 - (\sum_{x \in S_i} x)^2 .$$

3. We have

$$\arg\min_S \sum_{i=1}^{k} \Big(\sum_{x \in S_i} |x - \mu_i|^2 \Big) \quad \Longleftrightarrow \quad \arg\min_S \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x,y \in S_i} |x - y|^2 .$$

Hence, finding partition to minimize the pairwise squared deviations of points within each cluster.

k-means algorithm

k-means algorithm

Input:

1. the number of clusters k
2. initial centers $m_1^{(1)}, \ldots, m_k^{(1)}$

k-means algorithm

Input:

1. the number of clusters k
2. initial centers $m_1^{(1)}, \ldots, m_k^{(1)}$

Algorithm:

1. Assignment step: assigning each point x to the cluster with the nearest center.

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \,\middle|\, |x - m_i^{(l)}| \leq |x - m_j^{(l)}|, \ \forall 1 \leq j \leq k \right\}. \tag{1}$$

k-means algorithm

Input:

1. the number of clusters k
2. initial centers $m_1^{(1)}, \ldots, m_k^{(1)}$

Algorithm:

1. Assignment step: assigning each point x to the cluster with the nearest center.

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \;\middle|\; |x - m_i^{(l)}| \leq |x - m_j^{(l)}|, \ \forall 1 \leq j \leq k \right\}. \tag{1}$$

2. Update step:

$$m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x\,. \tag{2}$$

k-means algorithm

Input:

1. the number of clusters k
2. initial centers $m_1^{(1)}, \ldots, m_k^{(1)}$

Algorithm:

1. Assignment step: assigning each point x to the cluster with the nearest center.

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \;\middle|\; |x - m_i^{(l)}| \leq |x - m_j^{(l)}|, \; \forall 1 \leq j \leq k \right\}. \tag{1}$$

2. Update step:

$$m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x \,. \tag{2}$$

Illustration: `https://en.wikipedia.org/wiki/K-means_clustering`

k-means algorithm

within-cluster sum-of-squares:

$$\arg\min_{S} \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right). \qquad (3)$$

k-means algorithm

within-cluster sum-of-squares:

$$\arg\min_{S} \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right). \tag{3}$$

Proposition

The objective (3) *is non-increasing in the k-means algorithm.*

k-means algorithm

within-cluster sum-of-squares:

$$\arg\min_{S} \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right). \tag{3}$$

Proposition

The objective (3) *is non-increasing in the k-means algorithm.*

Proof.

The update step and the assignment step imply

$$\sum_{i=1}^{k} \sum_{x \in S_i^{(l)}} |x - m_i^{(l)}|^2 \geq \sum_{i=1}^{k} \sum_{x \in S_i^{(l)}} |x - m_i^{(l+1)}|^2$$

k-means algorithm

within-cluster sum-of-squares:

$$\arg\min_{S} \sum_{i=1}^{k} \left(\sum_{x \in S_i} |x - \mu_i|^2 \right). \tag{3}$$

Proposition

The objective (3) *is non-increasing in the k-means algorithm.*

Proof.

The update step and the assignment step imply

$$\sum_{i=1}^{k} \sum_{x \in S_i^{(l)}} |x - m_i^{(l)}|^2 \geq \sum_{i=1}^{k} \sum_{x \in S_i^{(l)}} |x - m_i^{(l+1)}|^2 \geq \sum_{i=1}^{k} \sum_{x \in S_i^{(l+1)}} |x - m_i^{(l+1)}|^2 .$$

□

Remark

Algorithm:

1. Assignment step:

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \;\middle|\; |x - m_i^{(l)}| \le |x - m_j^{(l)}|,\ \forall 1 \le j \le k \right\}. \tag{4}$$

2. Update step: $m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x$.

Remark

Algorithm:

1. Assignment step:

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \;\middle|\; |x - m_i^{(l)}| \le |x - m_j^{(l)}|,\ \forall 1 \le j \le k \right\}. \tag{4}$$

2. Update step: $m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x$.

Remark

1. *The k-means algorithm may not be able to find the optimal partition.*

Remark

Algorithm:

1. Assignment step:

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \,\middle|\, |x - m_i^{(l)}| \leq |x - m_j^{(l)}|,\ \forall 1 \leq j \leq k \right\}. \tag{4}$$

2. Update step: $m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x$.

Remark

1. *The k-means algorithm may not be able to find the optimal partition.*
2. *The number of clusters k has to be chosen beforehand. The optimal number of k needs to be determined, e.g. by comparing clustering results obtained with different k.*

Remark

Algorithm:

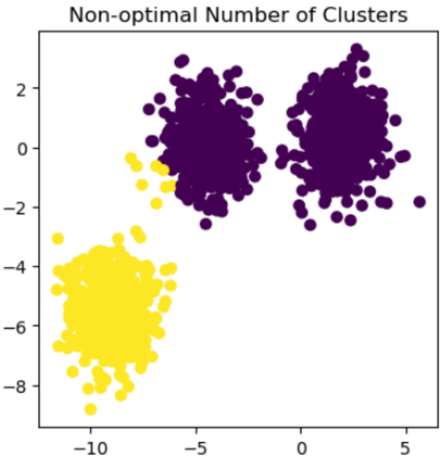
1. Assignment step:

$$S_i^{(l)} = \left\{ x \in \mathcal{D} \,\middle|\, |x - m_i^{(l)}| \leq |x - m_j^{(l)}|,\ \forall 1 \leq j \leq k \right\}. \tag{4}$$

2. Update step: $m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x \in S_i^{(l)}} x$.

Remark

1. *The k-means algorithm may not be able to find the optimal partition.*
2. *The number of clusters k has to be chosen beforehand. The optimal number of k needs to be determined, e.g. by comparing clustering results obtained with different k.*
3. *The computational complexity is $\mathcal{O}(n \times k \times N)$, where N is the total number of iterations.*


Non-optimal Number of Clusters
2
0
−2
−4
−6
−8
−10
−5
0
5

Part 2: Principal component analysis (PCA)

Problem setup

1. data $\mathcal{D} = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, where $d \gg 1$.

Problem setup

1. data $\mathcal{D} = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, where $d \gg 1$.
2. PCA aims at identifying a linear and orthogonal projection

$$f(x) = WW^\top x + b, \quad \forall\, x \in \mathbb{R}^d, \tag{5}$$

where $W \in \mathbb{R}^{d \times k}$ satisfies $W^\top W = I_k$ and $b \in \mathbb{R}^d$.

Problem setup

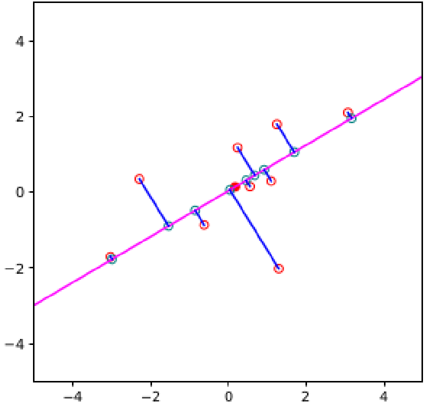
1. data $\mathcal{D} = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, where $d \gg 1$.
2. PCA aims at identifying a linear and orthogonal projection

$$f(x) = WW^\top x + b, \quad \forall\, x \in \mathbb{R}^d\,, \tag{5}$$

   where $W \in \mathbb{R}^{d\times k}$ satisfies $W^\top W = I_k$ and $b \in \mathbb{R}^d$.
3. criteria: *reconstruction error*

$$L(W, b) = \frac{1}{n}\sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2\,. \tag{6}$$

Illustration



Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n} \sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.

Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n}\sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.
2. Clearly, the optimal b is

$$b = \frac{1}{n}\sum_{i=1}^{n}(x_i - WW^\top x_i)$$

Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n} \sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.
2. Clearly, the optimal b is

$$b = \frac{1}{n} \sum_{i=1}^{n} (x_i - WW^\top x_i) = \bar{x} - WW^\top \bar{x}\,, \quad \text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\,. \tag{7}$$

Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n}\sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.
2. Clearly, the optimal b is

$$b = \frac{1}{n}\sum_{i=1}^{n}(x_i - WW^\top x_i) = \bar{x} - WW^\top \bar{x}\,, \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{7}$$

3. Hence

$$f(x) = WW^\top x + b = WW^\top (x - \bar{x}) + \bar{x}\,,$$

Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n}\sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.
2. Clearly, the optimal b is

$$b = \frac{1}{n}\sum_{i=1}^{n}(x_i - WW^\top x_i) = \bar{x} - WW^\top \bar{x}\,, \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{7}$$

3. Hence

$$\begin{aligned} f(x) =& WW^\top x + b = WW^\top(x - \bar{x}) + \bar{x}\,, \\ L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top(x_i - \bar{x})|^2\,, \end{aligned} \tag{8}$$

under the constraint $W^\top W = I_k$.

Simplification by removing b

1. Reconstruction error: $L(W, b) = \frac{1}{n}\sum_{i=1}^{n} |x_i - (WW^\top x_i + b)|^2$.
2. Clearly, the optimal b is

$$b = \frac{1}{n}\sum_{i=1}^{n}(x_i - WW^\top x_i) = \bar{x} - WW^\top \bar{x}\,, \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{7}$$

3. Hence

$$\begin{aligned} f(x) =& WW^\top x + b = WW^\top(x - \bar{x}) + \bar{x}\,, \\ L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top(x_i - \bar{x})|^2\,, \end{aligned} \tag{8}$$

under the constraint $W^\top W = I_k$.

Goal: find W that minimizes $L(W)$.

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top (x_i - \bar{x})|^2$$

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$\begin{aligned} L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top(x_i - \bar{x})|^2 \\ =& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - 2(x_i - \bar{x})^\top WW^\top(x_i - \bar{x}) + (x_i - \bar{x})^\top WW^\top(x_i - \bar{x}) \Big) \end{aligned}$$

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$
\begin{aligned}
L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top (x_i - \bar{x})|^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - 2(x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big)
\end{aligned}
$$

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$
\begin{aligned}
L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top (x_i - \bar{x})|^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - 2(x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\Big(W^\top (x_i - \bar{x})(x_i - \bar{x})^\top W \Big)
\end{aligned}
$$

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$
\begin{aligned}
L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top (x_i - \bar{x})|^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - 2(x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\Big(W^\top (x_i - \bar{x})(x_i - \bar{x})^\top W \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \mathrm{tr}(W^\top \hat{\Sigma} W)\,,
\end{aligned}
$$

Expression of $L(W)$

Using $W^\top W = I_k$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we can derive

$$
\begin{aligned}
L(W) =& \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top (x_i - \bar{x})|^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - 2(x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) + (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(|x_i - \bar{x}|^2 - (x_i - \bar{x})^\top WW^\top (x_i - \bar{x}) \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\Big(W^\top (x_i - \bar{x})(x_i - \bar{x})^\top W \Big) \\
=& \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \mathrm{tr}(W^\top \hat{\Sigma} W)\,,
\end{aligned}
$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$.

Expression of $L(W)$

To summarize, we have obtained the following result.

Proposition

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x}) - WW^\top(x_i - \bar{x})|^2 = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|^2 - \mathrm{tr}(W^\top \hat{\Sigma} W)\,.$$

where $\hat{\Sigma}$ the empirical covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top \in \mathbb{R}^{d\times d}\,. \tag{9}$$

Reformulation

1. Minimizing $L(W)$ is equivalent to solving

$$\max_{W \in \mathbb{R}^{d \times k}, W^\top W = I_k} \operatorname{tr}(W^\top \hat{\Sigma} W) \,. \tag{10}$$

Reformulation

1. Minimizing $L(W)$ is equivalent to solving

$$\max_{W \in \mathbb{R}^{d \times k}, W^\top W = I_k} \operatorname{tr}(W^\top \hat{\Sigma} W) \,. \tag{10}$$

2. Denote by $W_1, \ldots, W_k \in \mathbb{R}^d$ the column vectors of W. Direct computation shows that

$$\operatorname{tr}(W^\top \hat{\Sigma} W) = \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i \,,$$

$$W^\top W = I_k \quad \Longleftrightarrow \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \leq i, j \leq k \,.$$

Reformulation

1. Minimizing $L(W)$ is equivalent to solving
$$\max_{W\in\mathbb{R}^{d\times k}, W^\top W = I_k} \operatorname{tr}(W^\top \hat{\Sigma} W)\,. \tag{10}$$

2. Denote by $W_1, \ldots, W_k \in \mathbb{R}^d$ the column vectors of W. Direct computation shows that
$$\operatorname{tr}(W^\top \hat{\Sigma} W) = \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i\,,$$
$$W^\top W = I_k \quad \Longleftrightarrow \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \le i, j \le k\,.$$

3. Therefore, (10) is equivalent to
$$\max_{W_1,\ldots,W_k\in\mathbb{R}^d} \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i\,, \quad \text{subject to} \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \le i, j \le k\,.$$

Maximization task

$$\max_{W_1,\ldots,W_k \in \mathbb{R}^d} \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i \,, \quad \text{subject to} \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \leq i, j \leq k \,. \tag{11}$$

Maximization task

$$\max_{W_1,\ldots,W_k \in \mathbb{R}^d} \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i\,, \quad \text{subject to} \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \leq i,j \leq k\,. \tag{11}$$

1. $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$ is symmetric.

Maximization task

$$\max_{W_1,\dots,W_k \in \mathbb{R}^d} \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i \,, \quad \text{subject to} \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \le i,j \le k \,. \qquad (11)$$

1. $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$ is symmetric.
2. $\hat{\Sigma}$ is semi-positive definite, since

$$v^\top \hat{\Sigma} v = \frac{1}{n}\sum_{i=1}^{n} v^\top (x_i - \bar{x})(x_i - \bar{x})^\top v = \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x})^\top v|^2 \ge 0 \,, \quad \forall\, v \in \mathbb{R}^d \,.$$

Eigenvalues of $\hat{\Sigma}$ are real and non-negative.

Maximization task

$$\max_{W_1,\ldots,W_k \in \mathbb{R}^d} \sum_{i=1}^{k} W_i^\top \hat{\Sigma} W_i\,, \quad \text{subject to} \quad W_i^\top W_j = \delta_{ij}, \quad \forall 1 \le i,j \le k\,. \tag{11}$$

1. $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$ is symmetric.
2. $\hat{\Sigma}$ is semi-positive definite, since

   $$v^\top \hat{\Sigma} v = \frac{1}{n}\sum_{i=1}^{n} v^\top (x_i - \bar{x})(x_i - \bar{x})^\top v = \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x})^\top v|^2 \ge 0\,, \quad \forall\, v \in \mathbb{R}^d\,.$$

   Eigenvalues of $\hat{\Sigma}$ are real and non-negative.
3. The maximum of (11) is achieved when $W_1, \ldots, W_k$ are the k (pairwise orthogonal and normalized) eigenvectors of $\hat{\Sigma}$ corresponding to the largest k eigenvalues.

PCA algorithm

Based on previous analysis, we summarize the algorithm of PCA as follows.

PCA algorithm

Based on previous analysis, we summarize the algorithm of PCA as follows.

1. Compute $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

PCA algorithm

Based on previous analysis, we summarize the algorithm of PCA as follows.

1. Compute $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Compute eigenvectors $W_1, \ldots, W_k$ corresponding to the k largest eigenvalues of $\hat{\Sigma}$, so that $W_i^\top W_j = \delta_{ij}$ for $i, j = 1, \ldots, k$.

PCA algorithm

Based on previous analysis, we summarize the algorithm of PCA as follows.

1. Compute $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Compute eigenvectors $W_1, \ldots, W_k$ corresponding to the k largest eigenvalues of $\hat{\Sigma}$, so that $W_i^\top W_j = \delta_{ij}$ for $i, j = 1, \ldots, k$.
3. Let $W \in \mathbb{R}^{d\times k}$ be the matrix whose column vectors are $W_1, \ldots, W_k$.

PCA algorithm

Based on previous analysis, we summarize the algorithm of PCA as follows.

1. Compute $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Compute eigenvectors $W_1, \ldots, W_k$ corresponding to the k largest eigenvalues of $\hat{\Sigma}$, so that $W_i^\top W_j = \delta_{ij}$ for $i, j = 1, \ldots, k$.
3. Let $W \in \mathbb{R}^{d \times k}$ be the matrix whose column vectors are $W_1, \ldots, W_k$.

Once W is computed, we obtain the linear projection

$$f(x) = WW^\top(x - \bar{x}) + \bar{x}, \quad x \in \mathbb{R}^d.$$

PCA algorithm by SVD decomposition

1. Data matrix $X \in \mathbb{R}^{n \times d}$, whose ith row is $x_i - \bar{x}$. Then,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n} X^\top X.$$

PCA algorithm by SVD decomposition

1. Data matrix $X \in \mathbb{R}^{n \times d}$, whose ith row is $x_i - \bar{x}$. Then,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n} X^\top X.$$

2. SVD decomposition of $X = U \Lambda V^\top$, where
   1. $U \in \mathbb{R}^{n \times n}$ satisfies $U^\top U = I_n$
   2. $\Lambda \in \mathbb{R}^{n \times d}$ is a rectangular diagonal matrix with positive numbers
   3. $V \in \mathbb{R}^{d \times d}$ satisfies $V^\top V = I_d$.

PCA algorithm by SVD decomposition

1. Data matrix $X \in \mathbb{R}^{n\times d}$, whose ith row is $x_i - \bar{x}$. Then,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n}X^\top X.$$

2. SVD decomposition of $X = U\Lambda V^\top$, where
   1. $U \in \mathbb{R}^{n\times n}$ satisfies $U^\top U = I_n$
   2. $\Lambda \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix with positive numbers
   3. $V \in \mathbb{R}^{d\times d}$ satisfies $V^\top V = I_d$.

3. $\hat{\Sigma} = \frac{1}{n}X^\top X = \frac{1}{n}V(\Lambda^\top\Lambda)V^\top$, where $\Lambda^\top\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix.

PCA algorithm by SVD decomposition

1. Data matrix $X \in \mathbb{R}^{n\times d}$, whose ith row is $x_i - \bar{x}$. Then,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top = \frac{1}{n}X^\top X.$$

2. SVD decomposition of $X = U\Lambda V^\top$, where
    1. $U \in \mathbb{R}^{n\times n}$ satisfies $U^\top U = I_n$
    2. $\Lambda \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix with positive numbers
    3. $V \in \mathbb{R}^{d\times d}$ satisfies $V^\top V = I_d$.
3. $\hat{\Sigma} = \frac{1}{n}X^\top X = \frac{1}{n}V(\Lambda^\top\Lambda)V^\top$, where $\Lambda^\top\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix.

Therefore, we can construct W by selecting column vectors of V corresponding to the k largest singular values.

Part 3: Autoencoders

Autoencoders

1. Autoencoders find a low-dimensional representation of data using a **nonlinear** projection $\xi : \mathbb{R}^d \to \mathbb{R}^k$.

Autoencoders

1. Autoencoders find a low-dimensional representation of data using a **nonlinear** projection $\xi : \mathbb{R}^d \to \mathbb{R}^k$.

2. Dimension k: latent dimension or bottleneck dimension.

Autoencoders

1. Autoencoders find a low-dimensional representation of data using a **nonlinear** projection $\xi : \mathbb{R}^d \to \mathbb{R}^k$.

2. Dimension k: latent dimension or bottleneck dimension.

3. Autoencoder is a function of the form $f = \varphi \circ \xi$, where $\xi : \mathbb{R}^d \to \mathbb{R}^k$ is an encoder and $\varphi : \mathbb{R}^k \to \mathbb{R}^d$ is a decoder.

Autoencoders

1. Autoencoders find a low-dimensional representation of data using a **nonlinear** projection $\xi : \mathbb{R}^d \to \mathbb{R}^k$.

2. Dimension k: latent dimension or bottleneck dimension.

3. Autoencoder is a function of the form $f = \varphi \circ \xi$, where $\xi : \mathbb{R}^d \to \mathbb{R}^k$ is an encoder and $\varphi : \mathbb{R}^k \to \mathbb{R}^d$ is a decoder.

4. The autoencoder is optimized by minimizing the *reconstruction loss*

$$\mathcal{L}(\xi, \varphi) = \int_{\mathbb{R}^d} |\varphi(\xi(x)) - x|^2 \, d\mu = \mathbb{E}_{x \sim \mu} |\varphi(\xi(x)) - x|^2 \,, \qquad (12)$$

where μ is the data distribution.

Autoencoders and PCA

1. Given the data set $\mathcal{D} = \{x_1, \ldots, x_n\}$, the empirical loss function is

$$\hat{\mathcal{L}}(\xi, \varphi) = \frac{1}{n} \sum_{i=1}^{n} |\varphi(\xi(x_i)) - x_i|^2 .$$

Autoencoders and PCA

1. Given the data set $\mathcal{D} = \{x_1, \ldots, x_n\}$, the empirical loss function is

$$\hat{\mathcal{L}}(\xi, \varphi) = \frac{1}{n} \sum_{i=1}^{n} |\varphi(\xi(x_i)) - x_i|^2 \,.$$

2. PCA is recovered, with a linear encoder and a linear decoder chosen as

$$\xi(x) = W^\top x, \quad \varphi(z) = Wz + b,$$
$$\text{where } W \in \mathbb{R}^{d \times k},\; b \in \mathbb{R}^d,\; \text{and } W^\top W = I_k \,.$$

Autoencoders and PCA

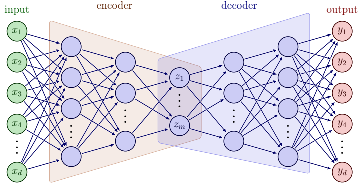
1. Given the data set $\mathcal{D} = \{x_1, \ldots, x_n\}$, the empirical loss function is

$$\hat{\mathcal{L}}(\xi, \varphi) = \frac{1}{n} \sum_{i=1}^{n} |\varphi(\xi(x_i)) - x_i|^2 .$$

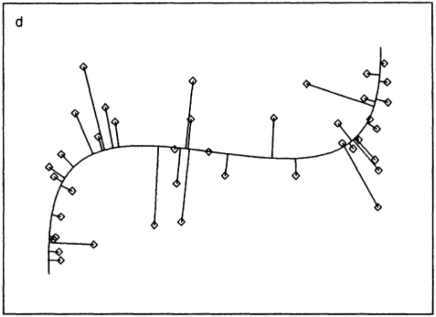
2. PCA is recovered, with a linear encoder and a linear decoder chosen as

$$\xi(x) = W^\top x, \quad \varphi(z) = Wz + b,$$
$$\text{where } W \in \mathbb{R}^{d\times k},\ b \in \mathbb{R}^d,\ \text{and } W^\top W = I_k .$$

3. In general, learning autoencoders can be viewed as a nonlinear generalization of PCA.



Illustration



$d = 2$ and $k = 1$. In this case, $\xi : \mathbb{R}^2 \to \mathbb{R}$ and $\varphi : \mathbb{R} \to \mathbb{R}^2$.

Characterization

We want to study the minimization task with loss

$$\min_{\xi} \min_{\varphi} \mathbb{E}_{x\sim\mu}|\varphi(\xi(x)) - x|^2 \,. \tag{13}$$

Characterization

We want to study the minimization task with loss

$$\min_{\xi} \min_{\varphi} \mathbb{E}_{x\sim\mu} |\varphi(\xi(x)) - x|^2 \,. \tag{13}$$

We use the following two facts.

1. Law of total expectation:

$$\mathbb{E}_{x\sim\mu}\big(|\varphi(\xi(x)) - x|^2\big) = \mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x)) - x|^2 \,\middle|\, \xi(x) = z\right)\right],$$

where $\widetilde{\mu}$ is the distribution of $z = \xi(x)$.

Characterization

We want to study the minimization task with loss

$$\min_{\xi} \min_{\varphi} \mathbb{E}_{x\sim\mu} |\varphi(\xi(x)) - x|^2 . \tag{13}$$

We use the following two facts.

1. Law of total expectation:

$$\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x)) - x|^2\right) = \mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x)) - x|^2 \,\middle|\, \xi(x) = z\right)\right],$$

   where $\widetilde{\mu}$ is the distribution of $z = \xi(x)$.

2. For a fixed $z \in \mathbb{R}^k$, we have

$$\min_{x' \in \mathbb{R}^d} \mathbb{E}_{x\sim\mu}\left(|x' - x|^2 \,\middle|\, \xi(x) = z\right) = \mathrm{Var}_{x\sim\mu}(x \,|\, \xi(x) = z) ,$$

   and the minimum is attained when $x' = \mathbb{E}_{x\sim\mu}(x \,|\, \xi(x) = z)$.

Characterization (1)

By first optimizing φ, we derive

$$\min_{\xi} \min_{\varphi} \mathbb{E}_{x\sim\mu} |\varphi(\xi(x)) - x|^2$$

Characterization (1)

By first optimizing φ, we derive

$$
\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\Big[\mathbb{E}_{x\sim\mu}\Big(|\varphi(\xi(x))-x|^2\,\Big|\,\xi(x)=z\Big)\Big]
\end{aligned}
$$

Characterization (1)

By first optimizing φ, we derive

$$
\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x))-x|^2\,\middle|\,\xi(x)=z\right)\right]\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(z)-x|^2\,\middle|\,\xi(x)=z\right)\right]
\end{aligned}
$$

Characterization (1)

By first optimizing φ, we derive

$$\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x))-x|^2\,\middle|\,\xi(x)=z\right)\right]\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(z)-x|^2\,\middle|\,\xi(x)=z\right)\right]\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left\{\min_{\varphi(z)}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(z)-x|^2\,\middle|\,\xi(x)=z\right)\right]\right\}
\end{aligned}$$

Characterization (1)

By first optimizing φ, we derive

$$
\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(\xi(x))-x|^2 \,\middle|\, \xi(x)=z\right)\right]\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(z)-x|^2 \,\middle|\, \xi(x)=z\right)\right]\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left\{\min_{\varphi(z)}\left[\mathbb{E}_{x\sim\mu}\left(|\varphi(z)-x|^2 \,\middle|\, \xi(x)=z\right)\right]\right\}\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathrm{Var}_{x\sim\mu}(x|\,\xi(x)=z)\right].
\end{aligned}
$$

Characterization (1)

By first optimizing φ, we derive

$$
\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x)) - x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\Big[\mathbb{E}_{x\sim\mu}\Big(|\varphi(\xi(x)) - x|^2 \,\Big|\, \xi(x) = z\Big)\Big]\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\Big[\mathbb{E}_{x\sim\mu}\Big(|\varphi(z) - x|^2 \,\Big|\, \xi(x) = z\Big)\Big]\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\Big\{\min_{\varphi(z)}\Big[\mathbb{E}_{x\sim\mu}\Big(|\varphi(z) - x|^2 \,\Big|\, \xi(x) = z\Big)\Big]\Big\}\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\big[\mathrm{Var}_{x\sim\mu}(x|\,\xi(x) = z)\big]\,.
\end{aligned}
$$

Proposition

$$
\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x)) - x|^2 = \min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\big[\mathrm{Var}(x|\,\xi(x) = z)\big]\,.
$$

Characterization (1)

By first optimizing φ, we derive

$$
\begin{aligned}
&\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\Big(|\varphi(\xi(x))-x|^2\,\big|\,\xi(x)=z\Big)\right]\\
=&\min_{\xi}\min_{\varphi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathbb{E}_{x\sim\mu}\Big(|\varphi(z)-x|^2\,\big|\,\xi(x)=z\Big)\right]\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left\{\min_{\varphi(z)}\left[\mathbb{E}_{x\sim\mu}\Big(|\varphi(z)-x|^2\,\big|\,\xi(x)=z\Big)\right]\right\}\\
=&\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathrm{Var}_{x\sim\mu}(x|\,\xi(x)=z)\right].
\end{aligned}
$$

Proposition

$$
\min_{\xi}\min_{\varphi}\mathbb{E}_{x\sim\mu}|\varphi(\xi(x))-x|^2=\min_{\xi}\mathbb{E}_{z\sim\widetilde{\mu}}\left[\mathrm{Var}(x|\,\xi(x)=z)\right].
$$

Moreover, for a fixed encoder ξ, the optimal decoder is given by

$$
\varphi_{\xi}(z)=\mathbb{E}_{x\sim\mu}(x\,|\,\xi(x)=z)\,,\quad\forall\, z\in\mathbb{R}^k\,. \tag{14}
$$

Characterization (2)

By first optimizing ξ, we directly get the following result.

Proposition

$$\min_{\varphi} \min_{\xi} \mathbb{E}_{x\sim\mu} |\varphi(\xi(x)) - x|^2$$

Characterization (2)

By first optimizing ξ, we directly get the following result.

Proposition

$$\min_{\varphi} \min_{\xi} \mathbb{E}_{x\sim\mu}|\varphi(\xi(x)) - x|^2 = \min_{\varphi} \mathbb{E}_{x\sim\mu}|\varphi(\xi_\varphi(x)) - x|^2 \,,$$

where

$$\xi_\varphi(x) = \arg\min_{z\in\mathbb{R}^k} |\varphi(z) - x|, \quad \forall\, x \in \mathbb{R}^d \,. \tag{15}$$

Characterization (2)

By first optimizing ξ, we directly get the following result.

Proposition

$$\min_{\varphi} \min_{\xi} \mathbb{E}_{x\sim\mu} |\varphi(\xi(x)) - x|^2 = \min_{\varphi} \mathbb{E}_{x\sim\mu} |\varphi(\xi_{\varphi}(x)) - x|^2 \,,$$

where

$$\xi_{\varphi}(x) = \arg\min_{z\in\mathbb{R}^k} |\varphi(z) - x|, \quad \forall\, x \in \mathbb{R}^d \,. \tag{15}$$

To summarize, the optimal autoencoder satisfies the self-consistent condition:

1. $\varphi(z) = \mathbb{E}_{x\sim\mu}(x \,|\, \xi(x) = z), \quad z \in \mathbb{R}^k$
2. $\xi(x) = \arg\min_{z\in\mathbb{R}^k} |\varphi(z) - x| \,, \quad x \in \mathbb{R}^d$.