Principal Component Analysis (PCA)

Shusen Wang

Some Statistical Concepts

Mean and Variance (Scalar)

- Let X be a random scalar variable and $p(\cdot)$ be the probability density function (PDF).
- **Mean**: $\mu = \mathbb{E}[X] = \int x \cdot p(x)\, dx.$
- **Variance**: $\sigma^2 = \mathbb{E}[(X - \mu)^2] = \int (x - \mu)^2 \cdot p(x)\, dx.$

Mean and Variance (Scalar)

- Let X be a random scalar variable and $p(\cdot)$ be the probability density function (PDF).
- **Mean**: $\mu = \mathbb{E}[X] = \int x \cdot p(x)\, dx.$
- **Variance**: $\sigma^2 = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 \cdot p(x)\, dx.$

- Let $x_1, \cdots, x_n$ be independently drawn observations of X.
- **Sample mean**: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- **Sample variance**: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$

Mean and Covariance (Vector)

- Let X be a random vector variable and $p(\cdot)$ be the probability density function (PDF).
- **Mean**: $\boldsymbol{\mu} = \mathbb{E}[X] = \int \mathbf{x} \cdot p(\mathbf{x})\, d\mathbf{x} \in \mathbb{R}^d$.
- **Covariance**: $\mathbf{C} = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \in \mathbb{R}^{d \times d}$.

- Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ be independently drawn observations of X.
- **Sample mean**: $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \in \mathbb{R}^d$.
- **Sample covariance**: $\hat{\mathbf{C}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \in \mathbb{R}^{d \times d}$.

Principal Component Analysis (PCA)

PCA Explained

- Transforms the data to a new coordinate system.
  - The greatest variance lie on the first coordinate (called the 1st principal component).
  - The 2nd greatest variance on the second coordinate, and so on…

PCA Explained

The original data.

PCA Explained

Step 1: subtract the mean

- **Sample mean:** $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$

PCA Explained

Step 1: subtract the mean

- Sample mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
- **Translation**: $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$

PCA Explained

Step 1: subtract the mean

- Sample mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
- **Translation**: $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$

PCA Explained

Step 2: rotation

PCA Explained

Step 2: rotation

- Find a rotation matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$
- **Rotation**: $\mathbf{x}_i'' = \mathbf{V}^T \mathbf{x}_i' \in \mathbb{R}^d$

PCA Explained

The result

- Find a rotation matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$
- Rotation: $\mathbf{x}_i'' = \mathbf{V}^T \mathbf{x}_i' \in \mathbb{R}^d$

the greatest variance direction

Question: How to Perform the Rotation?

- Find a rotation matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$
- **Rotation**: $\mathbf{x}_i'' = \mathbf{V}^T \mathbf{x}_i' \in \mathbb{R}^d$

Question: How to Perform the Rotation?

- $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.

Question: How to Perform the Rotation?

- $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}'_1, \cdots, \mathbf{x}'_n$.
- Singular value decomposition (SVD):

$$\mathbf{X}' = \sum_{j=1}^{d} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T .$$

Question: How to Perform the Rotation?

- $\mathbf{X}' \in \mathbb{R}^{n\times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.
- Singular value decomposition (SVD):

$$\mathbf{X}' = \sum_{j=1}^{d} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T.$$

- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_d] \in \mathbb{R}^{d\times d}$.
- Rotation: $\mathbf{x}_i'' = \mathbf{V}^T \, \mathbf{x}_i' \in \mathbb{R}^d$.

Question: How to Perform the Rotation?

- $\mathbf{X}' \in \mathbb{R}^{n\times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.
- Singular value decomposition (SVD):

$$\mathbf{X}' = \sum_{j=1}^{d} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T.$$

- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_d] \in \mathbb{R}^{d\times d}$.
- Rotation: $\mathbf{x}_i'' = \mathbf{V}^T \, \mathbf{x}_i{}' \in \mathbb{R}^d$.

- $\mathbf{X}'' \in \mathbb{R}^{n\times d}$: the stack of $\mathbf{x}_1'', \cdots, \mathbf{x}_n''$.
- Its right singular vectors are the standard basis: $\mathbf{e}_1, \cdots, \mathbf{e}_d$.

The Procedure of PCA (Recap)

1. Input: data $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and target rank k ($\leq d$).
2. Subtract the mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$.
   - Obtain matrix $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.

The Procedure of PCA (Recap)

1. Input: data $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and target rank k ($\leq d$).
2. Subtract the mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$.
   - Obtain matrix $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.
3. Find the rotation matrix $\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k] \in \mathbb{R}^{d \times k}$
   - Truncated SVD: $\mathbf{X}_k' = \sum_{j=1}^{k} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T$.

Question: Why setting a target rank $k \leq d$? Why not using $k = d$?

The Procedure of PCA (Recap)

1. Input: data $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and target rank k ($\leq d$).
2. Subtract the mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$.
   - Obtain matrix $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.
3. Find the rotation matrix $\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k] \in \mathbb{R}^{d \times k}$
   - Truncated SVD: $\mathbf{X}_k' = \sum_{j=1}^{k} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T$.

- $\mathbf{v}_1$ is the greatest variance direction, $\mathbf{v}_2$ is the 2nd greatest, and so on.
- Only the top variance directions are interesting.
  - The top ones are features.
  - The bottom ones are noise.
- For visualization, we set $k = 2$ or 3.

The Procedure of PCA (Recap)

1. Input: data $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and target rank k ($\leq d$).
2. Subtract the mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$.
   - Obtain matrix $\mathbf{X}' \in \mathbb{R}^{n \times d}$: the stack of $\mathbf{x}_1', \cdots, \mathbf{x}_n'$.
3. Find the rotation matrix $\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k] \in \mathbb{R}^{d \times k}$
   - Truncated SVD: $\mathbf{X}_k' = \sum_{j=1}^{k} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T$.
4. Rotation (and dimensionality reduction):
   - $\mathbf{x}_i'' = \mathbf{V}_k^T \, \mathbf{x}_i' \in \mathbb{R}^k$, for $i = 1$ to n, are the outputs.

Dimensionality reduction: $\mathbf{x}_i \in \mathbb{R}^d \longrightarrow \mathbf{x}_i'' \in \mathbb{R}^k$

Why is $\mathbf{v}_1$ the Greatest Variance Direction?

- The rows of $\mathbf{X}' \in \mathbb{R}^{n \times d}$ have zero mean.
  - $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ is the result of translation (Step 1 of PCA).
- The sample covariance matrix of $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ is:
  - $\hat{\mathbf{C}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \in \mathbb{R}^{d \times d}$.

Why is $\mathbf{v}_1$ the Greatest Variance Direction?

- The rows of $\mathbf{X}' \in \mathbb{R}^{n\times d}$ have zero mean.
  - $\boxed{\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}}$ s the result of translation (Step 1 of PCA).
- The sample covariance matrix of $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ is:
  - $\boxed{\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T} \in \mathbb{R}^{d\times d}$.
- ➔ $\hat{\mathbf{C}} = \frac{1}{n-1}\mathbf{X}'^T\mathbf{X}'$.

Why is $\mathbf{v}_1$ the Greatest Variance Direction?

- The rows of $\mathbf{X}' \in \mathbb{R}^{n\times d}$ have zero mean.
  - $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ is the result of translation (Step 1 of PCA).
- The sample covariance matrix of $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ is:
  - $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \in \mathbb{R}^{d\times d}$.
- ➔ $\hat{\mathbf{C}} = \frac{1}{n-1}\mathbf{X}'^T\mathbf{X}'$.
- $\mathbf{X}' = \sum_{j=1}^{d} \sigma_j \, \mathbf{u}_j \, \mathbf{v}_j^T$
  - $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$.

Why is $\mathbf{v}_1$ the Greatest Variance Direction?

- The rows of $\mathbf{X}' \in \mathbb{R}^{n\times d}$ have zero mean.
  - $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ is the result of translation (Step 1 of PCA).
- The sample covariance matrix of $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ is:
  - $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \in \mathbb{R}^{d\times d}$.
- $\boxed{\rightarrow \quad \hat{\mathbf{C}} = \frac{1}{n-1}\mathbf{X}'^T\mathbf{X}'.}$
- $\boxed{\mathbf{X}' = \sum_{j=1}^{d}\sigma_j\,\mathbf{u}_j\,\mathbf{v}_j^T}$ $\Longrightarrow$ $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{j=1}^{d}\sigma_j^2\,\mathbf{v}_j\,\mathbf{v}_j^T$.
  - $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$.

Why is $\mathbf{v}_1$ the Greatest Variance Direction?

- The rows of $\mathbf{X}' \in \mathbb{R}^{n\times d}$ have zero mean.
  - $\mathbf{x}_i' = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ is the result of translation (Step 1 of PCA).
- The sample covariance matrix of $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ is:
  - $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\boldsymbol{\mu}})(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \in \mathbb{R}^{d\times d}$.
- ➔ $\hat{\mathbf{C}} = \frac{1}{n-1}\mathbf{X}'^T\mathbf{X}'$.
- $\mathbf{X}' = \sum_{j=1}^{d}\sigma_j\,\mathbf{u}_j\,\mathbf{v}_j^T$ ⟹ $\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{j=1}^{d}\sigma_j^2\,\mathbf{v}_j\,\mathbf{v}_j^T$.
  - $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$.
- $\mathbf{v}_1 = \operatorname{argmax}_{\mathbf{v}}\ \mathbf{v}^T\,\hat{\mathbf{C}}\,\mathbf{v}$.

PCA: A Real-World Example

Ocean Temperature Data

ocean temperatures taken 1979—2011 at 6 hour intervals

Ocean Temperature Data

- Just keep the surface temperature
- Shape: $n \times d_1 \times d_2$ (oder-3 tensor)
  - $n = 4 \times 365 \times 32 = 46{,}720$ (4 measurements per day)
  - $d_1 = 2 \times 360 = 720$ (2 measurement per degree of longitude)
  - $d_2 = 2 \times 180 = 360$ (2 measurement per degree of latitude)

Ocean Temperature Data

- Just keep the surface temperature
- Shape: $n \times d_1 \times d_2$ (oder-3 tensor)
  - $n = 4 \times 365 \times 32 = 46{,}720$ (4 measurements per day)
  - $d_1 = 2 \times 360 = 720$ (2 measurement per degree of longitude)
  - $d_2 = 2 \times 180 = 360$ (2 measurement per degree of latitude)

- Reshape the tensor to a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Temporal dimension: $n = 46{,}720$
  - Spatial dimension: $d = d_1 d_2 = 259{,}200$

Compute the Top $k = 4$ Principal Components

- Ocean temperature data: $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Temporal dimension: $n = 46{,}720$
  - Spatial dimension: $d = d_1 d_2 = 259{,}200$

- Treat each row as a sample.
- The result of PCA is $n \times 4$ matrix.

- Treat each column as a sample.
- The result of PCA is $d \times 4$ matrix.

Compute the Top $k = 4$ Principal Components

- Ocean temperature data: $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Temporal dimension: $n = 46{,}720$
  - Spatial dimension: $d = d_1 d_2 = 259{,}200$

- Treat each row as a sample.
- The result of PCA is $n \times 4$ matrix.

- Treat each column as a sample.
- The result of PCA is $d \times 4$ matrix.

The first principal component $(n \times 1)$

One of the n time points

Compute the Top $k = 4$ Principal Components

- Ocean temperature data: $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Temporal dimension: $n = 46{,}720$
  - Spatial dimension: $d = d_1 d_2 = 259{,}200$

- Treat each row as a sample.
- The result of PCA is $n \times 4$ matrix.

- Treat each column as a sample.
- The result of PCA is $d \times 4$ matrix.

Compute the Top $k = 4$ Principal Components

- Ocean temperature data: $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Temporal dimension: $n = 46{,}720$
  - Spatial dimension: $d = d_1 d_2 = 259{,}200$

- Treat each row as a sample.
- The result of PCA is $n \times 4$ matrix.

- Treat each column as a sample.
- The result of PCA is $d \times 4$ matrix.

The first principal component $(d \times 1)$

One of the d locations (color indicates value)

The 1st Principal Components

Annual cycle

The 2nd Principal Components

Annual cycle

The 3rd Principal Components

El Nino

The 4th Principal Components

La Nina

Summary

Dimensionality Reduction

- Applications
  - Visualization and analysis
  - Data processing (to make downstream ML more efficient or accurate)

Dimensionality Reduction

- Applications
  - Visualization and analysis
  - Data processing (to make downstream ML more efficient or accurate)

- Methods
  - Principal component analysis (PCA)
  - Kernel PCA
  - Manifold learning
  - Autoencoder
  - Linear discriminant analysis (LDA)

3 Steps of PCA

1. Subtract mean (translation/shift).
2. Find the rotation matrix by (truncated) SVD.
   - The top right singular vectors are the maximum variance directions.
3. Rotate the zero-mean vectors.