An introduction to PCA

Weekly AI pills

Fabio Brau.

2020-10-16

SSSA, Emerging Digital Technologies, Pisa.

ISTITUTO
DI TECNOLOGIE DELLA
COMUNICAZIONE,
DELL'INFORMAZIONE
E DELLA
PERCEZIONE

Scuola Superiore
Sant'Anna

Retis
Real-Time Systems Laboratory

Summary

- The aim of Principal Component Analysis
- Derivation
    1. A Geometrical idea
    2. A statistical Derivation
    3. Singolar Value Decomposition
- PCA from Encoder Decoder NN
- Dummy examples

Geometrical Introduction

Geometrical Introduction

Let $X \in \mathbb{R}^{N \times n}$ be a dataset of N **observation** within n **variables**.

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} = \left[x^{(1)} \;\middle|\; \cdots \;\middle|\; x^{(n)} \right] \tag{1}$$

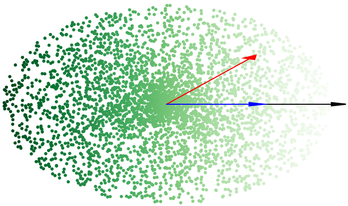
Notations:

- $x_i \in \mathbb{R}^n$ represents a single **observation**, i.e a **sample** in the feature space.
- $x^{(i)} \in \mathbb{R}^N$ represents the single **variable**, i.e a **column** of the dataset.
- The object $\mathbb{1}_n \in \mathbb{R}^n$ is the unitary columnar vector of length n $\mathbb{1}_n = [1, \cdots, 1]^T$.
- X is centered if $X^T \mathbb{1} = 0$

Geometrical Introduction: Finding a principal direction.

1. Scalar product measures the projection of x_j along the direction w.
2. We are only interested on module.
3. Summation over samples to get the global projection's contribute.
4. Searching for w which maximizes projection.
5. Adding constraint to avoid $w \to \infty$ solution.

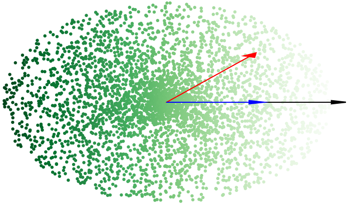
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



Geometrical Introduction: Finding a principal direction.

1. Scalar product measures the projection of x_j along the direction w.
2. We are only interested on module.
3. Summation over samples to get the global projection's contribute.
4. Searching for w which maximizes projection.
5. Adding constraint to avoid $w \to \infty$ solution.

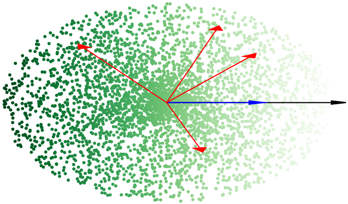
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



Geometrical Introduction: Finding a principal direction.

1. Scalar product measures the projection of x_j along the direction w.
2. We are only interested on module.
3. Summation over samples to get the global projection's contribute.
4. Searching for w which maximizes projection.
5. Adding constraint to avoid $w \to \infty$ solution.

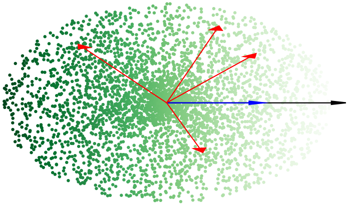
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} (w \cdot x_j)^2$$



Geometrical Introduction: Finding a principal direction.

1. Scalar product measures the projection of x_j along the direction w.
2. We are only interested on module.
3. Summation over samples to get the global projection's contribute.
4. Searching for w which maximizes projection.
5. Adding constraint to avoid $w \to \infty$ solution.

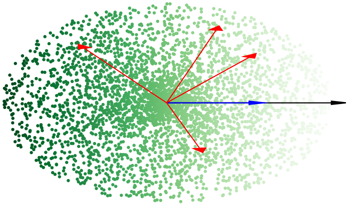
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



Geometrical Introduction: Finding a principal direction.

1. Scalar product measures the projection of x_j along the direction w.
2. We are only interested on module.
3. Summation over samples to get the global projection's contribute.
4. Searching for w which maximizes projection.
5. Adding constraint to avoid $w \to \infty$ solution.

$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



Geometrical Introduction: Finding other directions

We search for other orthogonal directions which maximize projections.

$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} (w \cdot x)^2$$

$$w_2 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} (w \cdot x)^2 \quad \text{and} \quad w_2 \perp w_1$$

$$\vdots$$

$$w_n \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} (w \cdot x)^2 \quad \text{and} \quad w_2 \perp \{w_1, \dots, w_{n-1}\}$$

Example

$$V(w) = \sum_j (w \cdot x_j)^2 \qquad \text{momentum along } w$$

If w_1, w_2, w_3 orthogonal that maximizes V in the 3D example, then

1. $V(w_1) = 3181.20$ ≈82.5%
2. $V(w_2) = 646.25$ ≈17.0%
3. $V(w_3) = 19.23$ ≈0.5 %

What if we forget the last direction?

Observation

- $x_j = \alpha_{1j} w_1 + \alpha_{2j} w_2 + \alpha_{3j} w_3$ (where $\alpha_{ij} = w_i \cdot x_j$).
- $\tilde{x}_j = \alpha_{1j} w_1 + \alpha_{2j} w_2$.

$$\frac{1}{N} \sum_j \|x_j - \tilde{x}_j\|^2 = \frac{V(w_3)}{N} \approx 4.8\,10^{-3} \qquad \text{(MSE)}$$

$$V(w) = \sum_j (w \cdot x_j)^2 \qquad \text{momentum along } w$$

If w_1, w_2, w_3 orthogonal that maximizes V in the 3D example, then

1. $V(w_1) = 3181.20$ $\approx$82.5%
2. $V(w_2) = 646.25$ $\approx$17.0%
3. $V(w_3) = 19.23$ $\approx$0.5 %

What if we forget the last direction?

Observation

- $x_j = \alpha_{1j} w_1 + \alpha_{2j} w_2 + \alpha_{3j} w_3$ (where $\alpha_{ij} = w_i \cdot x_j$).
- $\tilde{x}_j = \alpha_{1j} w_1 + \alpha_{2j} w_2$.

$$\frac{1}{N} \sum_j \|x_j - \tilde{x}_j\|^2 = \frac{V(w_3)}{N} \approx 4.8\,10^{-3} \qquad \text{(MSE)}$$

Geometrical Introduction: Direction Selection

$$V(w) = \sum_j (w \cdot x_j)^2 \qquad \text{momentum along } w$$

If w_1, w_2, w_3 orthogonal that maximizes V in the 3D example, then

1. $V(w_1) = 3181.20$ ≈82.5%
2. $V(w_2) = 646.25$ ≈17.0%
3. $V(w_3) = 19.23$ ≈0.5 %

What if we forget the last direction?

Observation

- $x_j = \alpha_{1j} w_1 + \alpha_{2j} w_2 + \alpha_{3j} w_3$ (where $\alpha_{ij} = w_i \cdot x_j$).
- $\tilde{x}_j = \alpha_{1j} w_1 + \alpha_{2j} w_2$.

$$\frac{1}{N} \sum_j \|x_j - \tilde{x}_j\|^2 = \frac{V(w_3)}{N} \approx 4.8\,10^{-3} \qquad \text{(MSE)}$$

Geometrical Introduction: Direction Selection

$$V(w) = \sum_j (w \cdot x_j)^2 \qquad \textbf{momentum along } w$$

If w_1, w_2, w_3 orthogonal that maximizes V in the 3D example, then

1. $V(w_1) = 3181.20$ ≈82.5%
2. $V(w_2) = 646.25$ ≈17.0%
3. $V(w_3) = 19.23$ ≈0.5 %

What if we forget the last direction?

Observation

- $x_j = \alpha_{1j} w_1 + \alpha_{2j} w_2 + \alpha_{3j} w_3$ (where $\alpha_{ij} = w_i \cdot x_j$).
- $\tilde{x}_j = \alpha_{1j} w_1 + \alpha_{2j} w_2$.

$$\frac{1}{N} \sum_j \|x_j - \tilde{x}_j\|^2 = \frac{V(w_3)}{N} \approx 4.8\,10^{-3} \qquad \text{(MSE)}$$

Geometrical Introduction: Direction Selection

$$V(w) = \sum_j (w \cdot x_j)^2 \qquad \text{momentum along } w$$

If w_1, w_2, w_3 orthogonal that maximizes V in the 3D example, then

1. $V(w_1) = 3181.20$ ≈82.5%
2. $V(w_2) = 646.25$ ≈17.0%
3. $V(w_3) = 19.23$ ≈0.5 %

What if we forget the last direction?

Observation

- $x_j = \alpha_{1j} w_1 + \alpha_{2j} w_2 + \alpha_{3j} w_3$ (where $\alpha_{ij} = w_i \cdot x_j$).
- $\tilde{x}_j = \alpha_{1j} w_1 + \alpha_{2j} w_2$.

$$\frac{1}{N} \sum_j \|x_j - \tilde{x}_j\|^2 = \frac{V(w_3)}{N} \approx 4.8\,10^{-3} \qquad \text{(MSE)}$$

Geometrical Introduction: Conclusion

- Given a set of data $X \in \mathbb{R}^{N \times n}$
- We can find $w_1, \cdots, w_n$ principal (orthonormal) directions the maximize their momentum.
- $V(w_1) > V(w_2) > \cdots > V(w_n)$
- Approximating X with $\tilde{X}$ by taking only the first k directions we are getting an error that is $V(w_{k+1})/N$

What's the catch?

$$\begin{aligned} \max_{w \in \mathbb{R}^n} \quad & \sum_{j=1}^{N} (w \cdot x_j)^2 \\ \text{s.t} \quad & w_i \cdot w = 0, \ \forall i < k \\ & w \cdot w = 1 \end{aligned} \tag{MP}$$

Geometrical Introduction: Conclusion

- Given a set of data $X \in \mathbb{R}^{N \times n}$
- We can find $w_1, \cdots, w_n$ principal (orthonormal) directions the maximize their momentum.
- $V(w_1) > V(w_2) > \cdots > V(w_n)$
- Approximating X with $\tilde{X}$ by taking only the first k directions we are getting an error that is $V(w_{k+1})/N$

What's the catch?

$$
\begin{aligned}
\max_{w \in \mathbb{R}^n} \quad & \sum_{j=1}^{N} (w \cdot x_j)^2 \\
\text{s.t} \quad & w_i \cdot w = 0, \ \forall i < k \\
& w \cdot w = 1
\end{aligned}
\qquad \text{(MP)}
$$

Geometrical Introduction: Conclusion

- Given a set of data $X \in \mathbb{R}^{N \times n}$
- We can find $w_1, \cdots, w_n$ principal (orthonormal) directions the maximize their momentum.
- $V(w_1) > V(w_2) > \cdots > V(w_n)$
- Approximating X with $\tilde{X}$ by taking only the first k directions we are getting an error that is $V(w_{k+1})/N$

What's the catch?

$$
\begin{aligned}
\max_{w \in \mathbb{R}^n} \quad & \sum_{j=1}^{N} (w \cdot x_j)^2 \\
\text{s.t} \quad & w_i \cdot w = 0, \ \forall i < k \\
& w \cdot w = 1
\end{aligned}
\tag{MP}
$$

Classical Derivation

Classical Derivation: Notations

$\mathcal{V}$ random variable, $V = (v_1, \ldots, v_N)$ N observations of the variable.

- Expected Value $\mathbb{E}[\mathcal{V}] = \frac{1}{N}\sum_{j=1}^{N} v_j$
- Variance $Var(\mathcal{V}) = \mathbb{E}[(\mathcal{V} - \mathbb{E}[\mathcal{V}])^2]$
- Covariance $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$

Observations

Under the assumption $\mathbb{E}[\mathcal{U}] = \mathbb{E}[\mathcal{V}] = 0$

1. $Var(\mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} v_j^2$
2. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$

Classical Derivation: Notations

$\mathcal{V}$ random variable, $V = (v_1, \ldots, v_N)$ N observations of the variable.

- Expected Value $\mathbb{E}[\mathcal{V}] = \frac{1}{N}\sum_{j=1}^{N} v_j$
- Variance $Var(\mathcal{V}) = \mathbb{E}[(\mathcal{V} - \mathbb{E}[\mathcal{V}])^2]$
- Covariance $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$

Observations

Under the assumption $\mathbb{E}[\mathcal{U}] = \mathbb{E}[\mathcal{V}] = 0$

1. $Var(\mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} v_j^2$
2. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$

Classical Derivation: Notations

$\mathcal{V}$ random variable, $V = (v_1, \ldots, v_N)$ N observations of the variable.

- Expected Value $\mathbb{E}[\mathcal{V}] = \frac{1}{N}\sum_{j=1}^{N} v_j$
- Variance $Var(\mathcal{V}) = \mathbb{E}[(\mathcal{V} - \mathbb{E}[\mathcal{V}])^2]$
- Covariance $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$

Observations

Under the assumption $\mathbb{E}[\mathcal{U}] = \mathbb{E}[\mathcal{V}] = 0$

1. $Var(\mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} v_j^2$
2. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$

Classical Derivation: Notations

$\mathcal{V}$ random variable, $V = (v_1, \ldots, v_N)$ N observations of the variable.

- Expected Value $\mathbb{E}[\mathcal{V}] = \frac{1}{N}\sum_{j=1}^{N} v_j$
- Variance $Var(\mathcal{V}) = \mathbb{E}[(\mathcal{V} - \mathbb{E}[\mathcal{V}])^2]$
- Covariance $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$

Observations

Under the assumption $\mathbb{E}[\mathcal{U}] = \mathbb{E}[\mathcal{V}] = 0$

1. $Var(\mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} v_j^2$
2. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$

Classical Derivation: Notations

$\mathcal{V}$ random variable, $V = (v_1, \ldots, v_N)$ N observations of the variable.

- Expected Value $\mathbb{E}[\mathcal{V}] = \frac{1}{N}\sum_{j=1}^{N} v_j$
- Variance $Var(\mathcal{V}) = \mathbb{E}[(\mathcal{V} - \mathbb{E}[\mathcal{V}])^2]$
- Covariance $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$

Observations

Under the assumption $\mathbb{E}[\mathcal{U}] = \mathbb{E}[\mathcal{V}] = 0$

1. $Var(\mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} v_j^2$
2. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$

Classical Derivation: An Eigenvalue Problem

$$\max_{\|w\|=1} V(w) = \max_{\|w\|=1} \sum_j (w^T x_j)^2 = \max_{w^T w=1} w^T (X^T X) w \qquad \text{(MP)}$$

Lagrange Multipliers Technique

Let consider the Lagrangian Function of MP

$$\mathcal{L}(w, \lambda) = V(w) - \lambda(w^T w - 1), \quad \forall w \in \mathbb{R}^n, \lambda \in \mathbb{R}$$

Claim

If w^* is a solution of MP then there exists λ^* such that

$$\nabla \mathcal{L}(w^*, \lambda^*) = 0, \quad i.e \quad (X^T X) w^* - \lambda^* w^* = 0 \qquad (2)$$

w Principal Direction $\implies$ w Eigenvector of $X^T X$

Classical Derivation: An Eigenvalue Problem

$$\max_{\|w\|=1} V(w) = \max_{\|w\|=1} \sum_j (w^T x_j)^2 = \max_{w^T w=1} w^T (X^T X) w \qquad \text{(MP)}$$

Lagrange Multipliers Technique

Let consider the Lagrangian Function of MP

$$\mathcal{L}(w, \lambda) = V(w) - \lambda(w^T w - 1), \quad \forall w \in \mathbb{R}^n, \lambda \in \mathbb{R}$$

Claim

If w^* is a solution of MP then there exists λ^* such that

$$\nabla \mathcal{L}(w^*, \lambda^*) = 0, \quad i.e \quad (X^T X) w^* - \lambda^* w^* = 0 \qquad (2)$$

w Principal Direction $\Longrightarrow$ w Eigenvector of $X^T X$

Classical Derivation: An Eigenvalue Problem

$$\max_{||w||=1} V(w) = \max_{||w||=1} \sum_j (w^T x_j)^2 = \max_{w^T w=1} w^T (X^T X) w \qquad \text{(MP)}$$

Lagrange Multipliers Technique

Let consider the Lagrangian Function of MP

$$\mathcal{L}(w, \lambda) = V(w) - \lambda(w^T w - 1), \quad \forall w \in \mathbb{R}^n, \lambda \in \mathbb{R}$$

Claim

If w^* is a solution of MP then there exists λ^* such that

$$\nabla \mathcal{L}(w^*, \lambda^*) = 0, \quad i.e \quad (X^T X) w^* - \lambda^* w^* = 0 \qquad (2)$$

w Principal Direction $\implies$ w Eigenvector of $X^T X$

Classical Derivation: An Eigenvalue Problem

Why switching to an eigen-pair problem?

$X^T X$ $+$ $\longrightarrow$

- $w_1, \cdots, w_n$ eigenvectors
- $w_i^T X^T X w_i = V(w_i)$ eigenvalues
- $V(w_1) > \cdots > V(w_n)$

Classical Derivation: An Eigenvalue Problem

Why switching to an eigen-pair problem?

$X^T X$ + MATLAB $\longrightarrow$

- $w_1, \cdots, w_n$ eigenvectors
- $w_i^T X^T X w_i = V(w_i)$ eigenvalues
- $V(w_1) > \cdots > V(w_n)$

Classical Derivation:Data Compression

The matrix $W = [w_1 | \cdots | w_n]$ can be used to reduce the dimensionality

$$F = \left[f^{(1)} \quad | \cdots | \quad f^{(n)} \right] = X W \quad \text{(factors)}$$