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

- Geometrical Introduction
- Classical Derivation
- Dimensionality Reduction
- Statistical Point of View
- Non Linear PCA

Geometrical Introduction

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

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

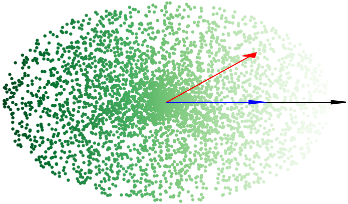
Notations:

- $x_i \in \mathbb{R}^n$ 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 $X^T \mathbb{1}_N = 0$

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.

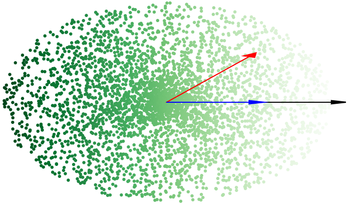
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



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.

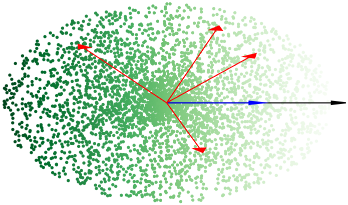
$$w_1 \in \operatorname*{argmax}_{\|w\|_2=1} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



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.

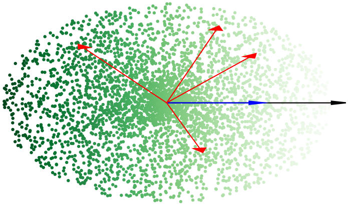
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



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.

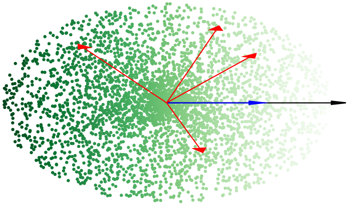
$$w_1 \in \underset{\|w\|_2=1}{\operatorname{argmax}} \sum_{j=1}^{N} \left(w \cdot x_j\right)^2$$



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.

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

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

$$\vdots$$

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

Example

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

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

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

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

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

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

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

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

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?

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

- 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: 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

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

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

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

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

Why switching to an eigen-pair problem?[1]

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

- $w_1, \cdots, w_n$ eigenvectors
- $w_i^T X^T X w_i = V(w_i)$ eigenvalues
- $V(w_1) > \cdots > V(w_n) \geq 0$

[1] Appendix for further details.

Classical Derivation: An Eigenvalue Problem

Why switching to an eigen-pair problem?[1]

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

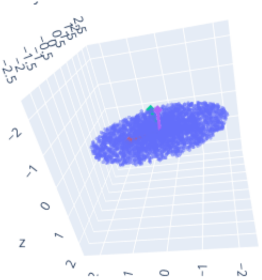
- $w_1, \cdots, w_n$ eigenvectors
- $w_i^T X^T X w_i = V(w_i)$ eigenvalues
- $V(w_1) > \cdots > V(w_n) \geq 0$

[1] Appendix for further details.

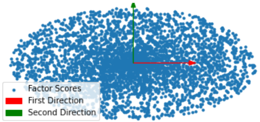
Dimensionality Reduction

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

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

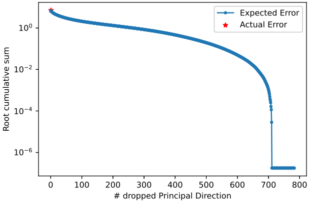


Feature space



Factor scores restricted to the first two principal directions.

Dimensionality Reduction: A concrete example

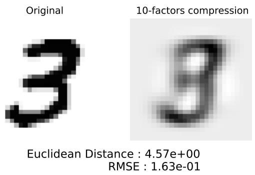


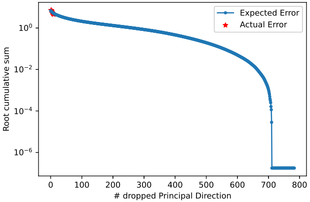
Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1|\cdots|w_{n-k}]\tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}}\|X - X_k\| = \frac{1}{\sqrt{N}}\sqrt{V(n-k+1) + \cdots + V(n)} \qquad \text{(EB)}$$

Dimensionality Reduction: A concrete example

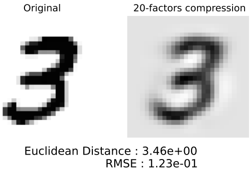


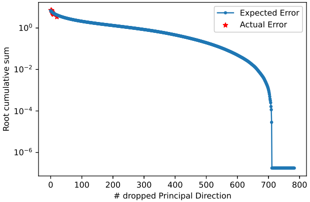


Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \qquad \text{(EB)}$$



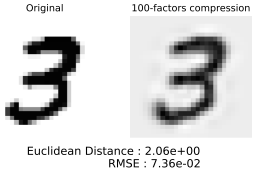


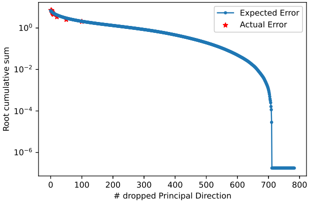
Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \qquad \text{(EB)}$$

Dimensionality Reduction: A concrete example



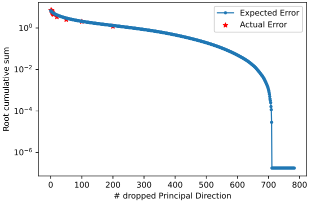


Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \tag{EB}$$

Dimensionality Reduction: A concrete example

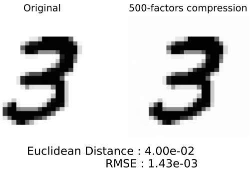


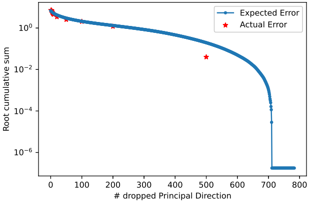
Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \tag{EB}$$

Dimensionality Reduction: A concrete example



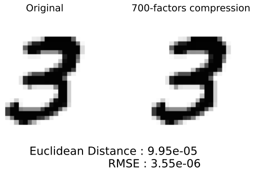


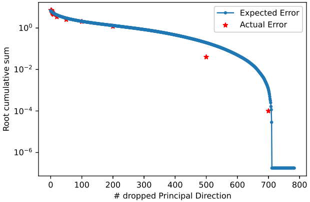
Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \tag{EB}$$

Dimensionality Reduction: A concrete example





Compression Error Estimation

1. $x \in \mathbb{R}^n$ original sample.
2. $f = W^T x \in \mathbb{R}^n$ coordinates in factor-scores space.
3. $\tilde{f} = [f_1, \cdots, f_{n-k}] \in \mathbb{R}^{n-k}$ dropping last k coordinates.
4. $\tilde{x} = [w_1 | \cdots | w_{n-k}] \tilde{f} \in \mathbb{R}^n$, approximation of x.

$$\|x - \tilde{x}\| \approx \frac{1}{\sqrt{N}} \|X - X_k\| = \frac{1}{\sqrt{N}} \sqrt{V(n-k+1) + \cdots + V(n)} \qquad \text{(EB)}$$

Where is statistic?

Statistical Point of View: 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: Notations

$\mathcal{V}$ random variable, $V = (v_1, \dots, v_N)$ 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: 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 $Cov(\mathcal{U}, \mathcal{V}) = \mathbb{E}[(\mathcal{U} - \mathbb{E}[\mathcal{U}])(\mathcal{V} - \mathbb{E}[\mathcal{V}])]$
- If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$Cov(\mathcal{U}) = \begin{bmatrix} Cov(\mathcal{U}_1, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_1, \mathcal{U}_m) \\ \vdots & \ddots & \vdots \\ Cov(\mathcal{U}_m, \mathcal{U}_1) & \cdots & Cov(\mathcal{U}_m, \mathcal{U}_m) \end{bmatrix}$$

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. $Cov(\mathcal{U}, \mathcal{V}) = \frac{1}{N}\sum_{j=1}^{N} u_j v_j$
3. If $\mathcal{U} = (\mathcal{U}_1, \cdots, \mathcal{U}_m)$, then

$$\mathrm{Var}(w \cdot \mathcal{U}) = w^T \mathcal{U} w$$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$
  momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N \operatorname{Cov}(X)$
- $N \operatorname{Var}(w \cdot X)$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$ momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N \, \mathrm{Cov}(X)$
- $N \, \mathrm{Var}(w \cdot X)$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$ momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N \, \mathrm{Cov}(X)$
- $N \, \mathrm{Var}(w \cdot X)$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$
  momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N \, \mathrm{Cov}(X)$
- $N \, \mathrm{Var}(w \cdot X)$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$
  momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N\, \mathrm{Cov}(X)$
- $N\, \mathrm{Var}(w \cdot X)$

Statistical Point of View: Language Translation

Geometrical

- $X^T \mathbb{1}_N = 0$
- $X^T X$
- $V(w) = \sum_j (w \cdot x_j)^2$
  momentum along w.

Statistical

- $\mathbb{E}[X^{(1)}], \cdots, \mathbb{E}[X^{(n)}] = 0.$
- $N\,\mathrm{Cov}(X)$
- $N\,\mathrm{Var}(w \cdot X)$

Compression Error Estimation

$$\|x_j - \tilde{x}_j\| \approx \sqrt{\mathbb{E}[\|x_j - \tilde{x}_j\|^2]} = \sqrt{Var(w_{k+1} \cdot X) + \cdots + Var(w_n \cdot X)}$$

a.k.a

$$w_1, \cdots, w_k \quad \text{explain} \quad 100 * \left(\frac{\sum_{i=1}^{k} \mathrm{Var}(w_i \cdot X)}{\sum_i \mathrm{Var}(w_i \cdot X)} \right) \% \quad \text{of the variance.}$$

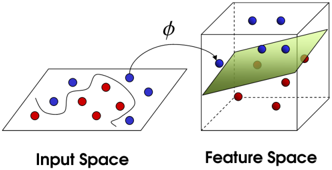
Non Linear PCA

Non Linear PCA: Kernel PCA

$$V_\kappa(w) = \sum_j \kappa(w, x_j)^2$$

where

$$\kappa(v, w) = \Phi(v) \cdot \Phi(w)$$



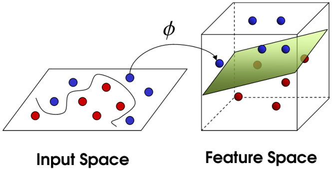
Learning with kernels - Bernhard Schölkopf, Alexander J. Smola

Non Linear PCA: Kernel PCA

$$V_\kappa(w) = \sum_j \kappa(w, x_j)^2$$

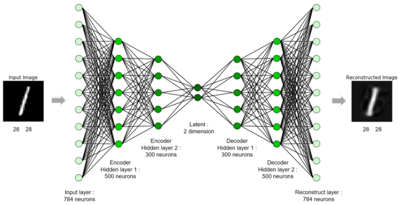
where

$$\kappa(v, w) = \Phi(v) \cdot \Phi(w)$$



Learning with kernels - Bernhard Schölkopf, Alexander J. Smola

Non Linear PCA: Autoencoders



Autoencoders Training

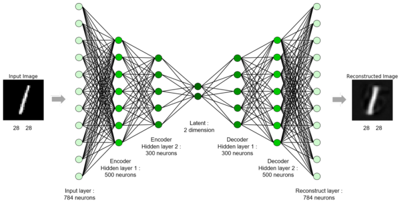
$$\min_{\theta} \frac{1}{N} \sum_j \|f_\theta(x_j) - x_j\|^2 \qquad \text{(mP)}$$

Claim[2]

- $f_\theta(x) = U V x$ is a 1-depth autoencoder with hidden space of dimension k.
- If $W = [w_1|\cdots|w_n]$ principal components of $X \in \mathbb{R}^{N \times n}$
- $V^* = [w_1|\cdots|w_k]$ and $U^* = (V^*)^T$ solves mP

[2] From Principal Subspaces to Principal Components with Linear Autoencoders

Non Linear PCA: Autoencoders



Autoencoders Training

$$\min_{\theta} \frac{1}{N} \sum_{j} \|f_\theta(x_j) - x_j\|^2 \qquad \text{(mP)}$$

Claim[2]

- $f_\theta(x) = U\,V\,x$ is a 1-depth autoencoder with hidden space of dimension k.
- If $W = [w_1|\cdots|w_n]$ principal components of $X \in \mathbb{R}^{N\times n}$
- $V^* = [w_1|\cdots|w_k]$ and $U^* = (V^*)^T$ solves mP

[2] From Principal Subspaces to Principal Components with Linear Autoencoders

appendix

Eigen-pairs of Simmetric def.positive matrices

A matrix $A \in M(n)$ is symmetric and def.positive if respectively

$$A^T A = AA^T, \quad v^T A v > 0 \, \forall v \in \mathbb{R}^n \tag{3}$$

From spectral theorem it's exists an isometry $V = [v_1 | \cdots | v_n]$ such that

$$V^T A V = D$$

where $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix.
Because of $V^T V = Id$ then

$$AV = \begin{bmatrix} Av_1 & | \cdots | & Av_n \end{bmatrix} = VD = \begin{bmatrix} \lambda_1 v_1 & | \cdots | & \lambda_n v_n \end{bmatrix} \tag{4}$$

This shows that **there exists an orthonormal bases of eigenvectors for** A. Because of A is def.positive then

$$\lambda_i = v_i^T A v_i > 0$$

and so A **has only positive eigenvalues**.

Approximation Error

For each $j = 1, \cdots, N$ we can write $x_j = f_{j1} v_i + \cdots + f_{jn} v_n$ where $f_{ij} = w_i \cdot x_j$. The approximated samples can be written as $\tilde{x}_j = f_{j1} v_i + \cdots + f_{j,n-k} v_{n-k}$. The main idea is to write the **expected value of the square euclidean distance** between the two samples (i.e. original end compressed).

$$\begin{aligned} \|x - \tilde{x}\|^2 &\approx \frac{1}{N} \sum_{j=1}^{N} \|f_{j,n-k+1} v_{n-k+1} + \cdots + f_n v_n\|^2 \\ &= \frac{1}{N} \sum_{j=1}^{N} f_{j,n-k+1}^2 + \cdots + f_{j,n}^2 \\ &= \frac{1}{N} \left(V(w_1) + \cdots + V(w_n) \right) \end{aligned} \tag{5}$$

By taking the root we obtain the approximation in EB. Moreover we can compute also the **Variance of the squared euclidean distance** to increase the accuracy of the error approximation.