PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH

Tran Van Long

AI Academy Vietnam

20/7/2024

Nội dung

Cơ sở, tọa độ [1]

Cho V là một không gian tuyến tính có $\dim V = n$. Giả sử $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$ là một cơ sở của V. Khi đó, mọi véc-tơ $x \in V$ có **biểu diễn duy nhất**

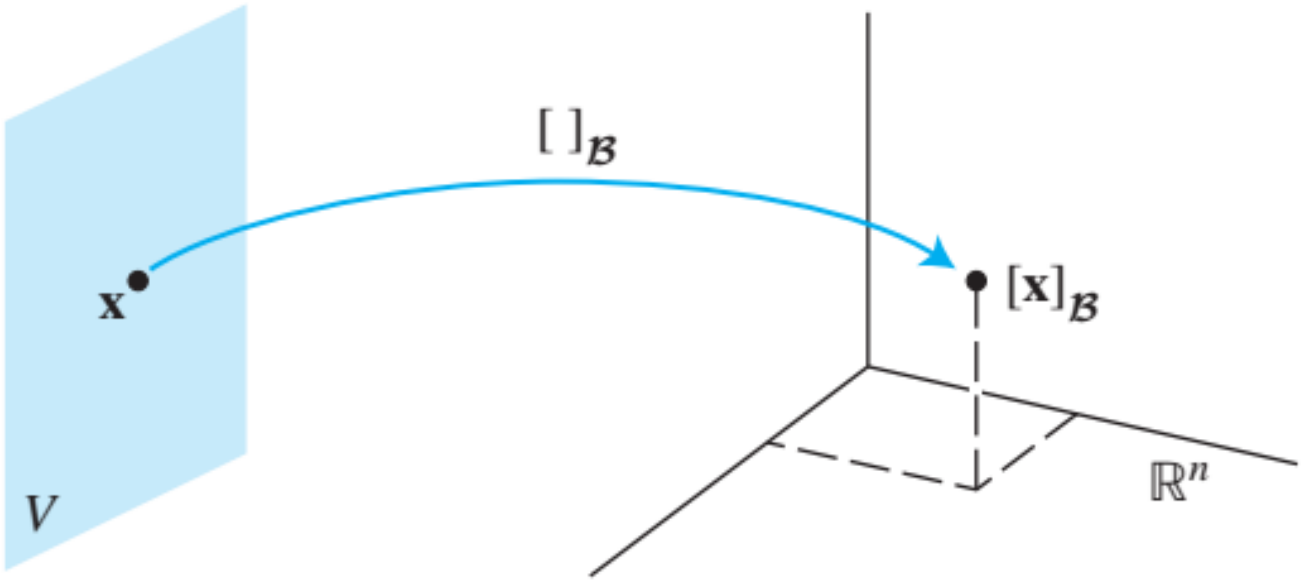
$$x = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n.$$

Ta gọi **tọa độ** của x đối với cơ sở $\mathcal{B}$ là $x_{\mathcal{B}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

[1] (Tài liệu 2)

Biến đổi từ không gian tuyến tính sang R^n

$$x = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n \Rightarrow x_{\mathcal{B}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$



Phép biến đổi tọa độ

Giả sử V có cơ sở mới $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ và véc-tơ x có tọa độ đối với cơ sở $\mathcal{C}$ là $x_{\mathcal{C}} = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix}$.

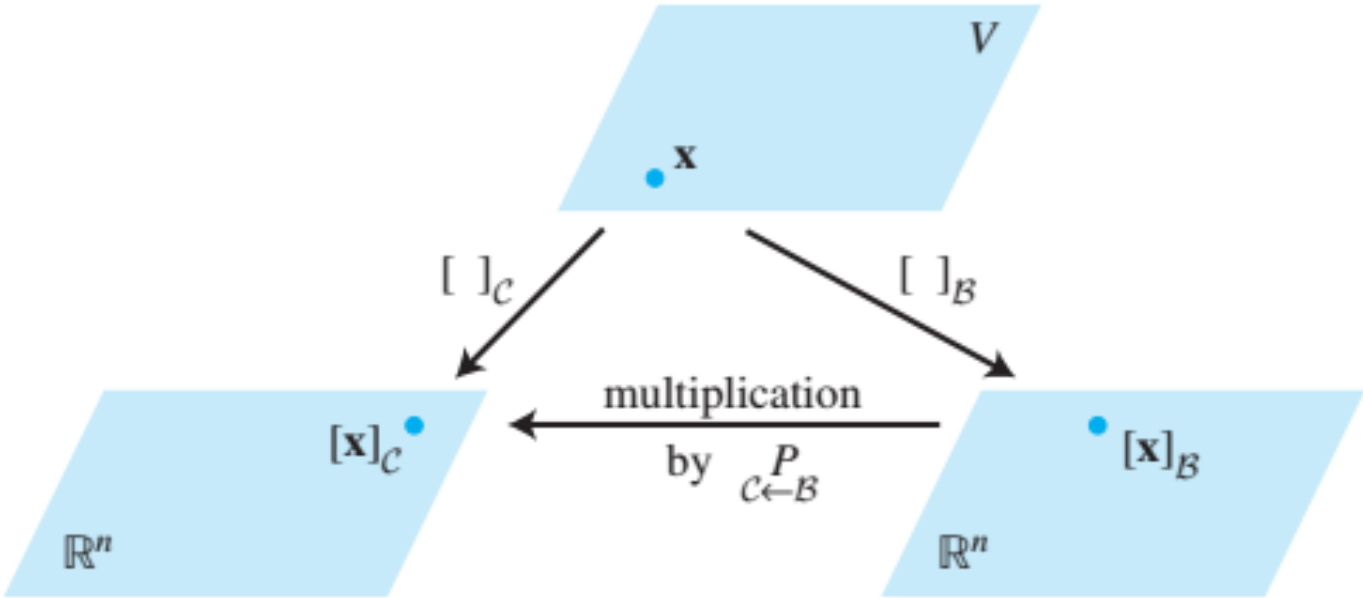
Ma trận T chuyển từ cơ sở $\mathcal{B}$ sang cơ sở $\mathcal{C}$

$$\{c_1, c_2, \ldots, c_n\} = \{b_1, b_2, \ldots, b_n\} T$$

Phép biến đổi tọa độ

$$x_{\mathcal{B}} = T x_{\mathcal{C}} \Leftrightarrow x_{\mathcal{C}} = T^{-1} x_{\mathcal{B}}.$$

Phép biến đổi tọa độ



Cơ sở trực giao, cơ sở trực chuẩn

Hệ véc-tơ $\{v_1, v_2, \ldots, v_n\}$ khác không trong R^n là hệ trực giao nếu

$$v_i \perp v_j, \quad \forall i \neq j.$$

Hệ véc-tơ $\{u_1, u_2, \ldots, u_n\}$ các véc-tơ đơn vị trong R^n là hệ trực chuẩn nếu

$$u_i \perp u_j, \quad \forall i \neq j.$$

Ma trận $U = [u_1, u_2, \ldots, u_p]$ gọi là ma trận có các cột trực giao nếu hệ véc-tơ cột $\{u_1, u_2, \ldots, u_p\}$ là hệ trực chuẩn hay

$$U^T U = I.$$

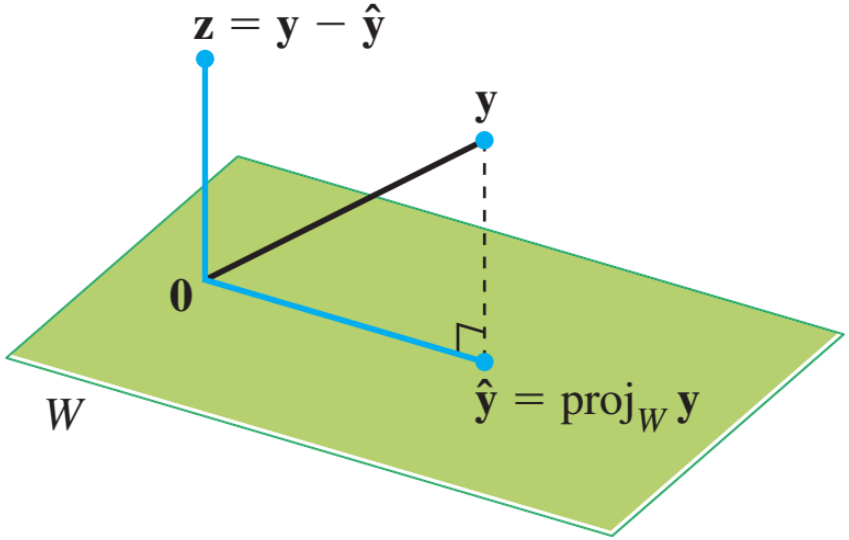
Phép chiếu trực giao

Giả sử không gian con W có cơ sở trực chuẩn $\{u_1, u_2, \ldots, u_p\}$. Thì mọi $y \in \mathrm{R}^n$ có khai triển trực giao

$$y = \hat{y} + z, \hat{y} \in W, z \perp W.$$

Hơn nữa,

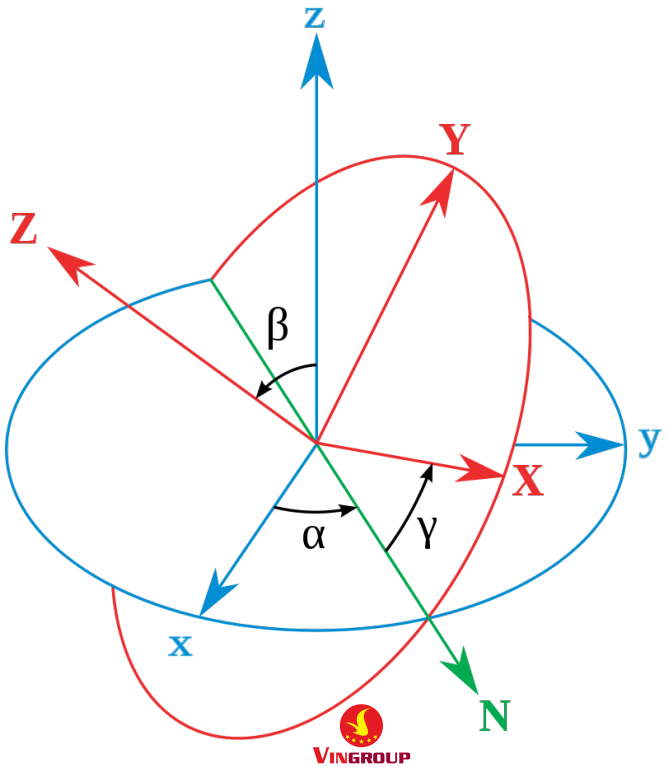
$$\hat{y} = proj_W(y) = UU^T y = (y.u_1)u_1 + (y.u_2)u_2 + \cdots + (y.u_p)u_p.$$



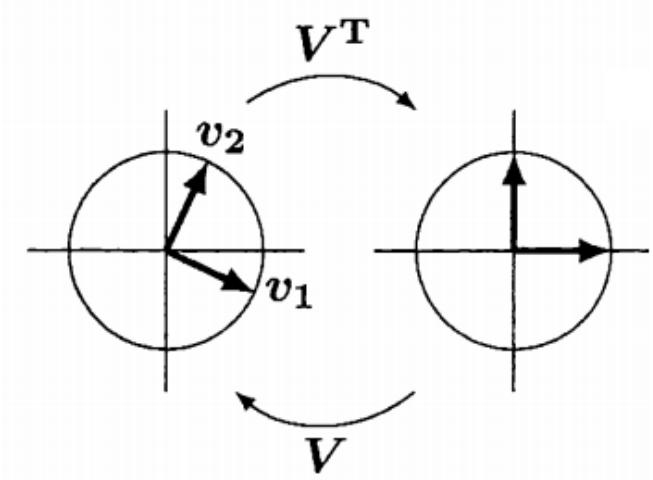
Biến đổi trực giao

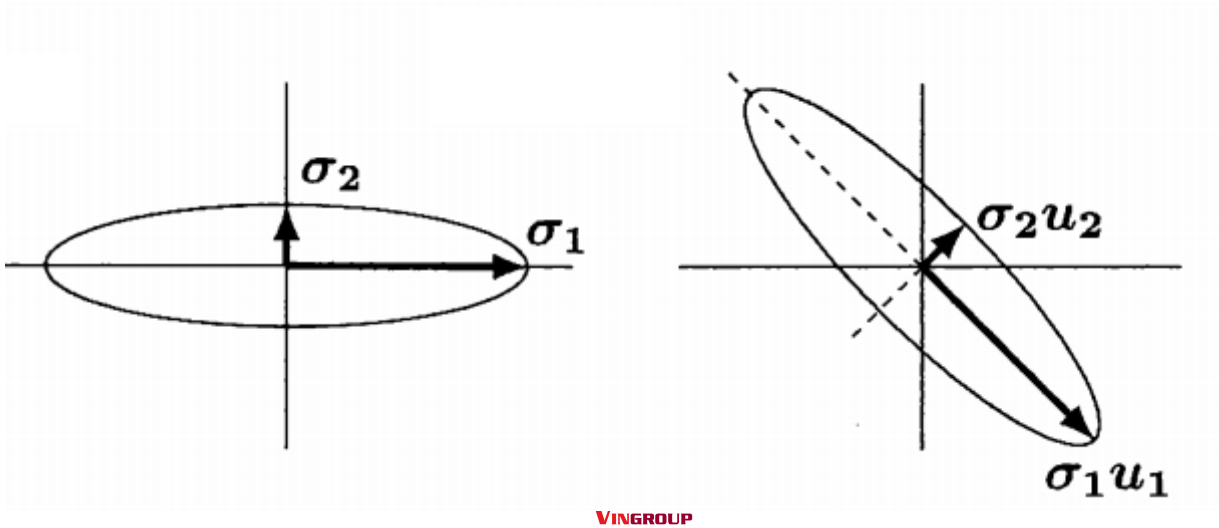
Trong R^n xét hai cơ sở trực chuẩn $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$ và $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ thì ma trận Q chuyển cơ sở từ $\mathcal{E}$ sang cơ sở $\mathcal{U}$ là ma trận trực giao

$$Q^{-1} = Q^T.$$



Biến đổi trực giao





Phương pháp PCA [2]

Phương pháp PCA: Tìm cơ sở mới biểu diễn lại dữ liệu từ đó lọc được nhiễu và khám phá cấu trúc dữ liệu.
Cơ sở mới là tổ hợp tuyến tính của cơ sở gốc.

Xét dữ liệu $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ có n quan sát trong không gian m chiều. Mỗi hàng là một véc-tơ trong R^m.
Gọi Y là ma trận biến đổi của X bởi phép biến đổi trực giao U xác định bởi

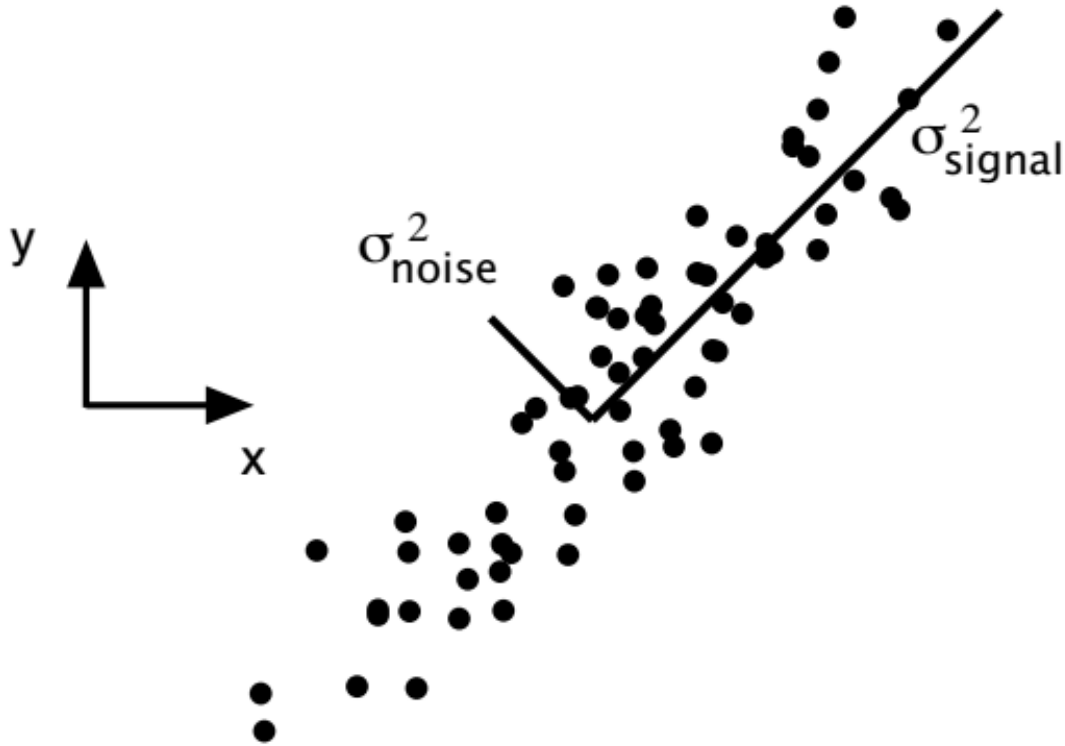
$$Y = XU,$$

với $U = [u_1, u_2, \ldots, u_m]$.
Các cột của U là hệ cơ sở mới để biểu diễn dữ liệu X.

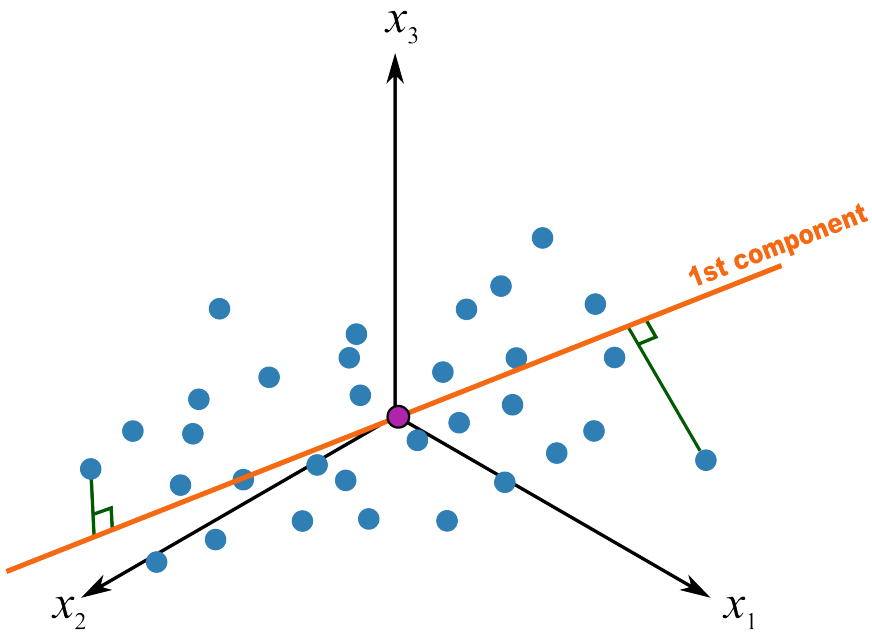
[2] (Tài liệu 1)

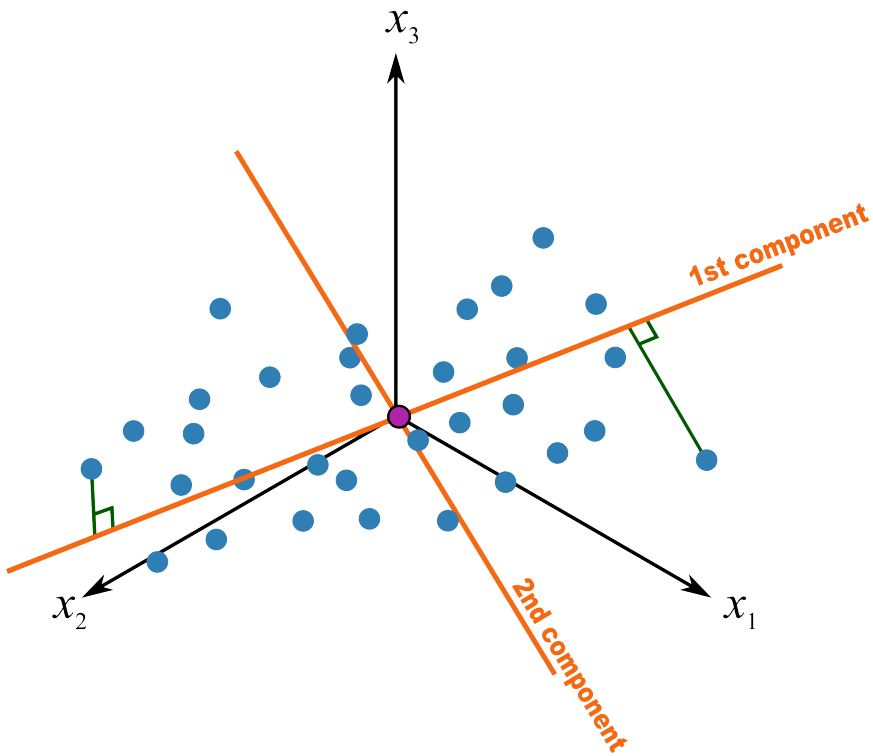
Thành phần chính - Principal components

Thành phần chính là véc-tơ đơn vị (hướng) có phương sai lớn nhất hay hướng của đường thẳng phù hợp nhất (best-fit line).

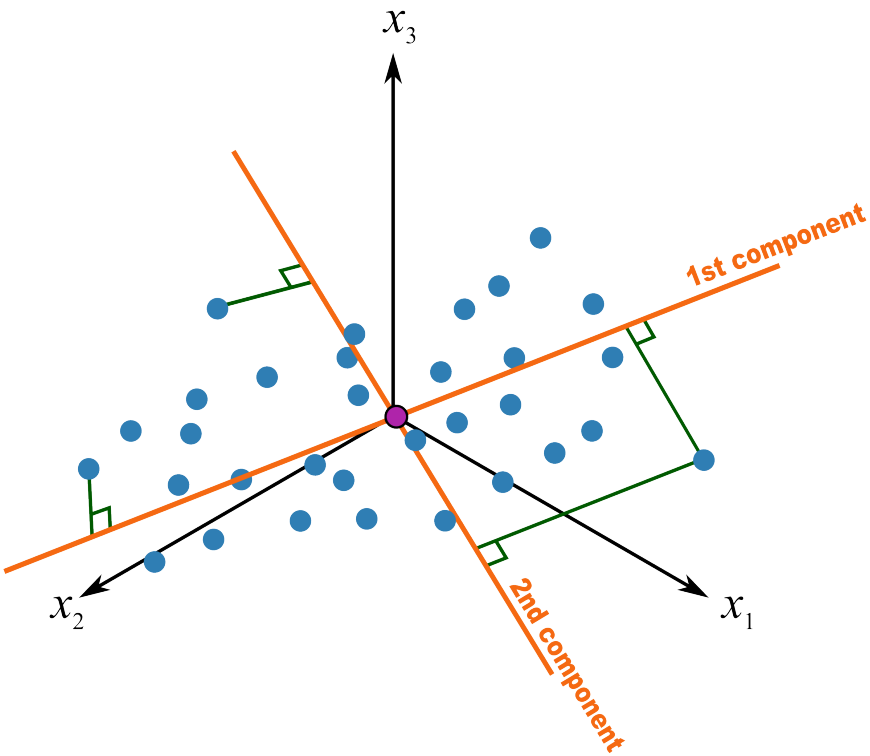


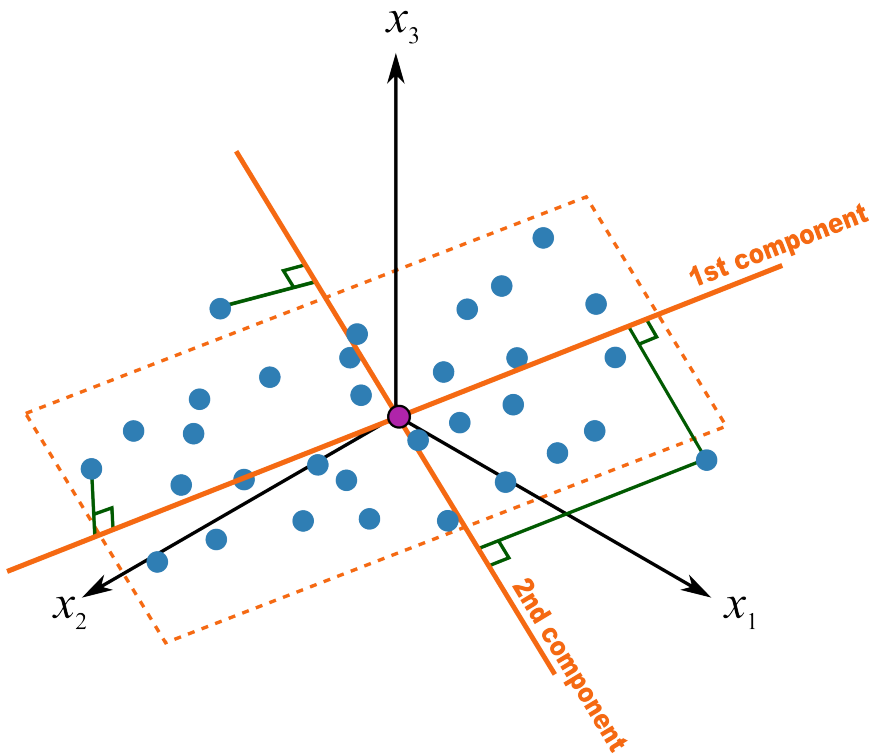
Các thành phần chính



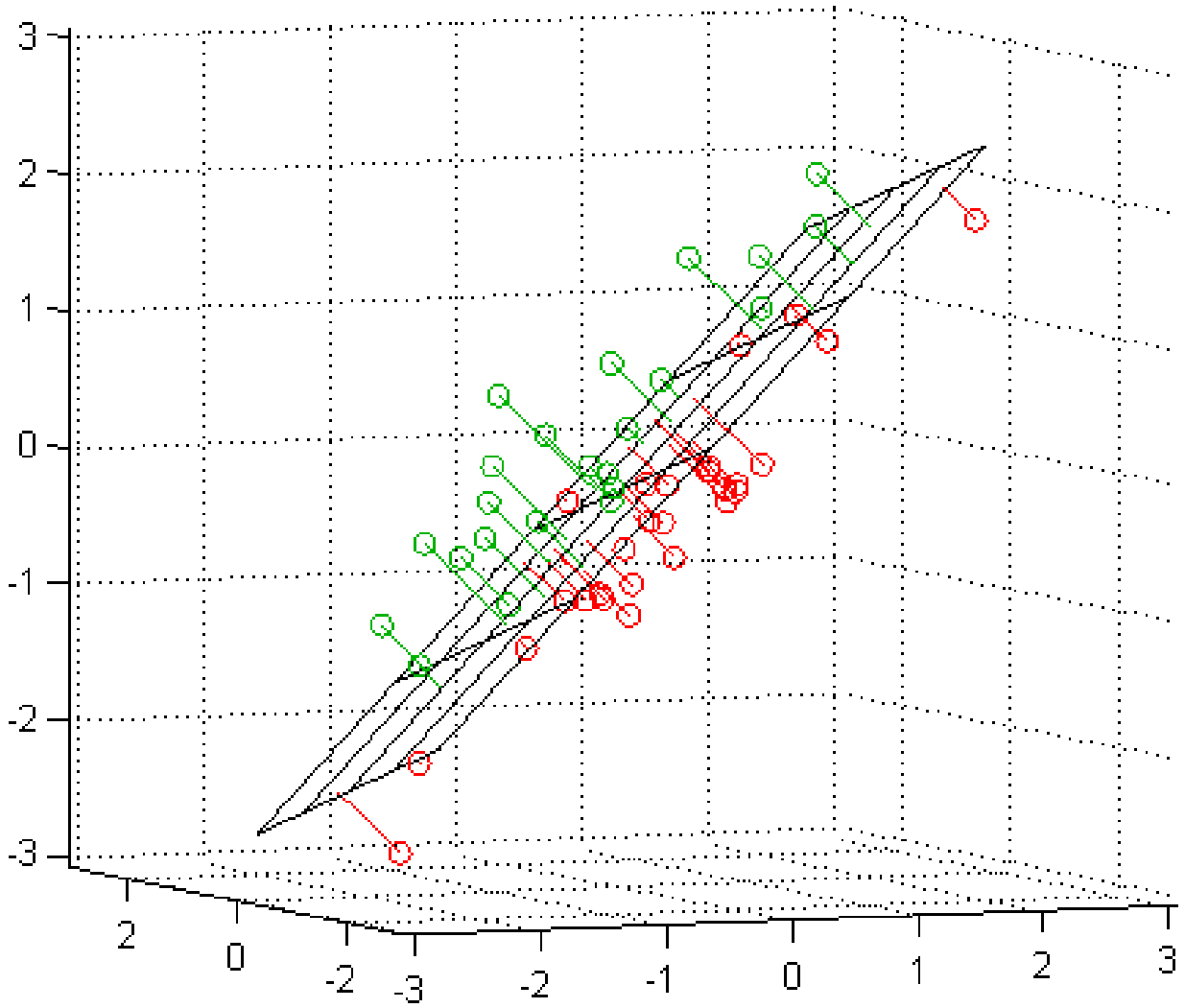


Các thành phần chính





Mặt phẳng chính



Thành phần chính thứ nhất

Giả sử dữ liệu được quy tâm (trung bình bằng 0), bằng cách sử dụng biến đổi

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$x_i \leftarrow x_i - \bar{x}.$$

Thành phần chính thứ nhất là véc-tơ u_1, hình chiếu của dữ liệu trên không gian con sinh với $\{u_1\}$ có các tọa độ là $y_i = u_1^T x_i$ có phương sai là

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \to \max$$

Thành phần chính thứ nhất

Ta biến đổi

$$\begin{aligned}\sigma_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(u_1^T x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} u_1^T x_i^T x_i u_1 \\ &= u_1^T\left(\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i\right)u_1 \\ &= u_1^T\left(\frac{1}{n}X^T X\right)u_1\end{aligned}$$

Ma trận $S = \frac{1}{n}X^T X$ gọi là ma trận hiệp phương sai.

Phương pháp PCA

Ta giải bài toán

$$u_1^T S u_1 \to \max, \quad ||u_1|| = 1.$$

Vậy

$$u_1 = \arg\max\{w^T S w : \ ||w|| = 1\}$$

Thành phần chính thứ hai:

$$u_2 = \arg\max\{w^T S w : \ ||w|| = 1, w \perp u_1\}$$

Giả sử có các thành phần chính $u_1, \ldots, u_{k-1}$, thành phần chính thứ k xác định bởi:

$$u_k = \arg\max\{w^T S w : \ ||w|| = 1, w \perp u_1, \ldots, w \perp u_{k-1}\}$$

Phương pháp PCA

Các thành phành phần chính $u_1, u_2, \ldots, u_k$ là các véc-tơ riêng đơn vị ứng với các giá trị riêng lớn thứ i của ma trận hiệp phương sai S.

Thuật toán PCA

- Tính $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Tính $\hat{x}_i = x_i - \bar{x}$
- Tính ma trận hiệp phương sai $S = \frac{1}{n}\hat{X}^T\hat{X}$
- Tính các giá trị riêng và véc-tơ riêng đơn vị của S, sắp xếp các giá trị riêng theo thứ tự giảm dần.
- Các thành phần chính là $u_1, u_2, \ldots, u_n$

Phương pháp PCA

Biểu diễn dữ liệu [3]

Chọn các véc-tơ riêng $U = [u_1, u_2, \ldots, u_k]$ ứng với k giá trị riêng lớn nhất. Chiếu dữ liệu đã chuẩn hóa bởi phép chiếu trực giao xuống không gian con sinh bởi $\{u_1, u_2, \ldots, u_k\}$ bởi công thức

$$Z = \hat{X}U.$$

Dữ liệu ban đầu được xấp xỉ theo công thức

$$x_i \approx \hat{x}_i UU^T + \bar{x}.$$

[3](Tài liệu 1)

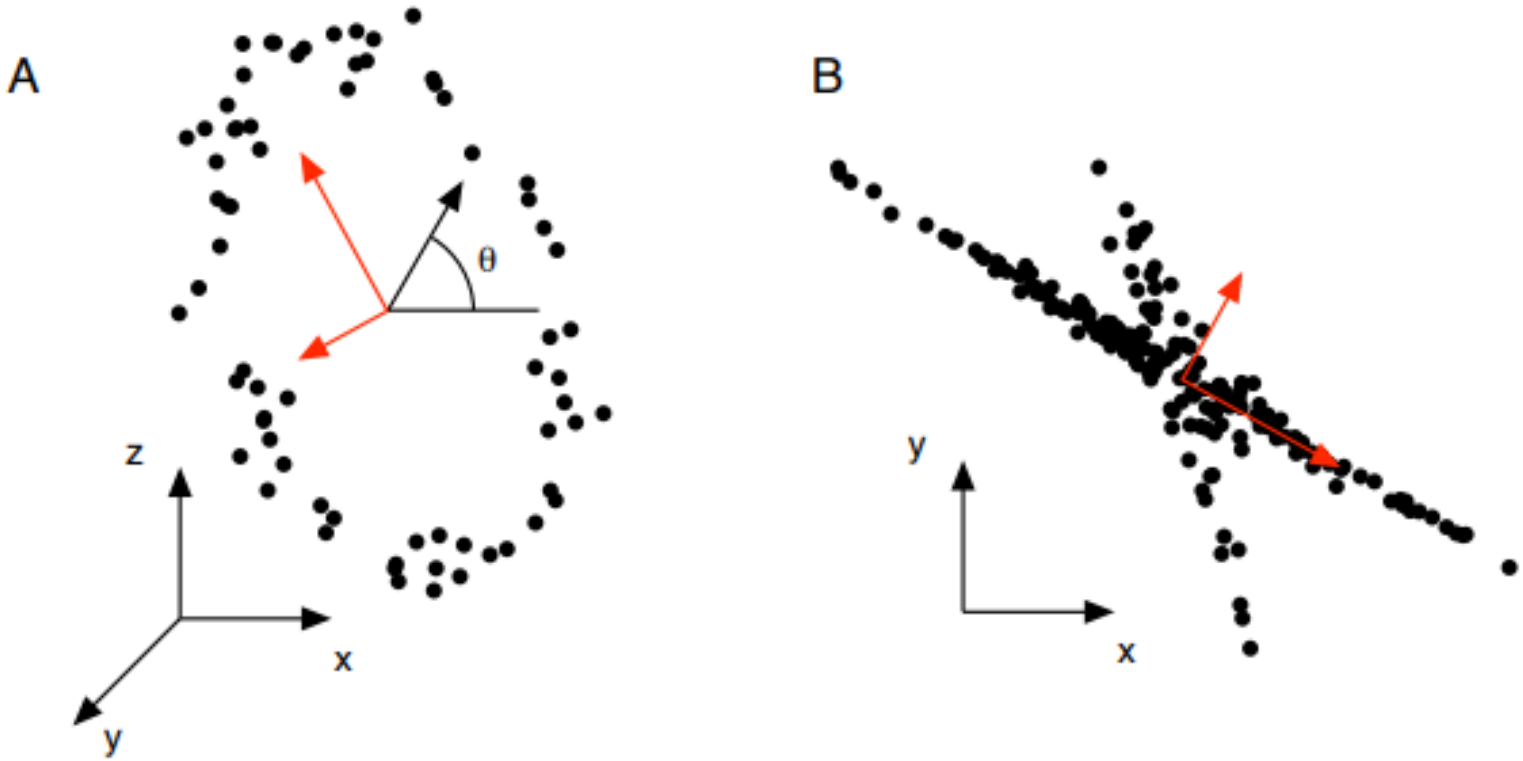
Phân tích sai số xấp xỉ

Hàm tổn thất do xấp xỉ dữ liệu từ không gian gốc xuống không gian k chiều xác định bởi

$$
\begin{aligned}
E &= \frac{1}{n}\sum_{i=1}^{n} ||x_i - \hat{x}_i UU^T - \bar{x}||^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} ||\hat{x}_i - \hat{x}_i UU^T||^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\hat{x}_i \hat{x}_i^T - \hat{x}_i UU^T \hat{x}_i^T\right) \\
&= trace(S) - trace(U^T SU) \\
&= \sum_{i=k+1}^{m} \sigma_i^2.
\end{aligned}
$$

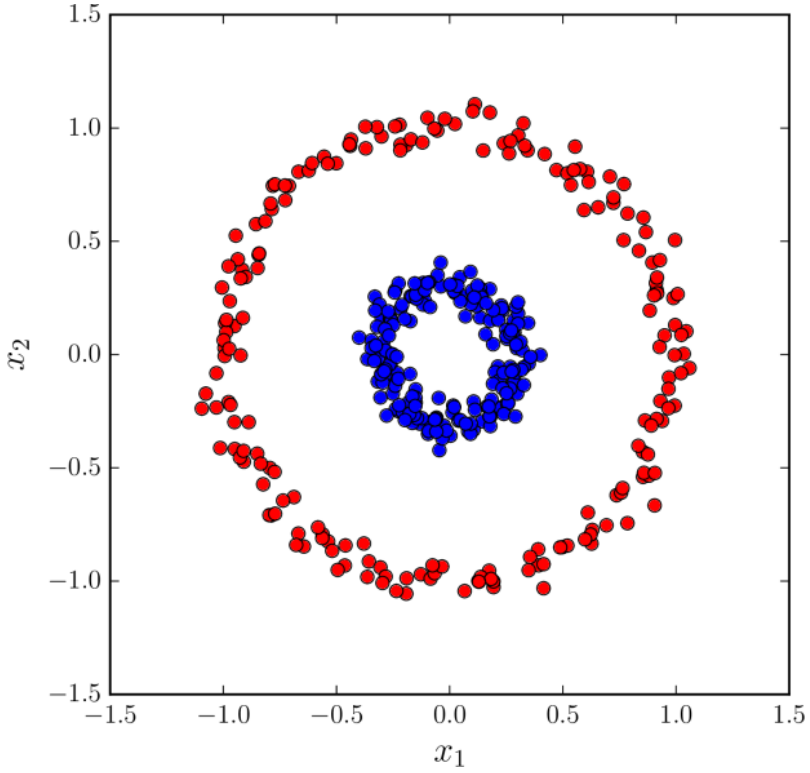
Hạn chế của phương pháp PCA

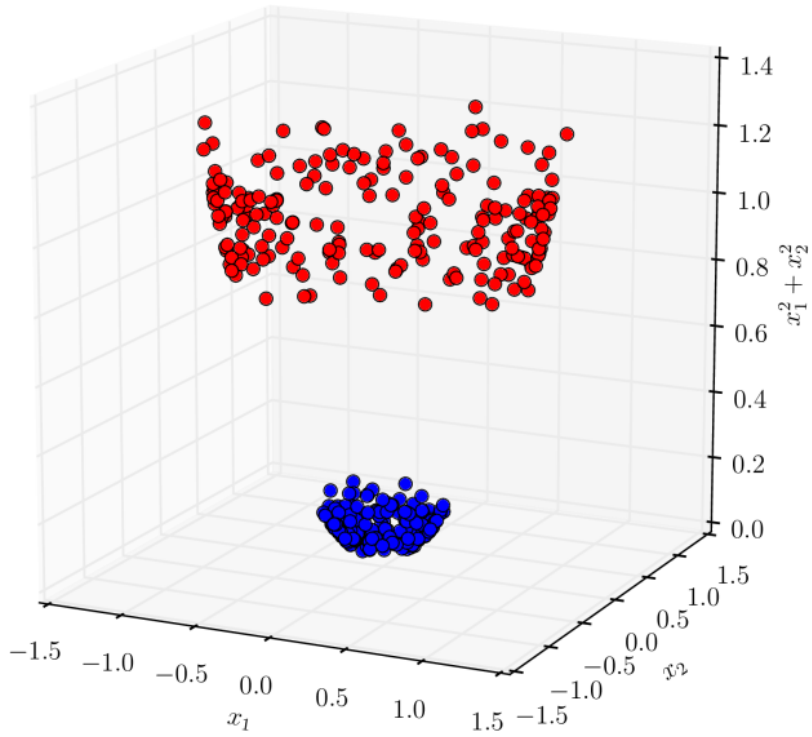
Phương pháp PCA không hiệu quả đối với dữ liệu phi tuyến



Phương pháp KPCA

Biến đổi dữ liệu





Phương pháp KPCA [4]

Ta sử dụng phép biến đổi phi tuyến chuyển dữ liệu lên không gian có số chiều lớn, sau đó áp dụng PCA đối với không gian mới.
Giả sử phép biến đổi phi tuyến

$$\varphi : R^m \to \mathrm{F}$$

không gian các đặc trưng F có số chiều $M > m$ và dữ liệu mới là $\varphi(x_i) \in \mathrm{F}, i = 1, 2, \ldots, n$.
Ta đặt $K = (K_{ij} = K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle)_{n \times n}$

[4] (Tài liệu 1)

Phương pháp KPCA

Tiếp theo ta cần tính ma trận hiệp phương sai C đối với dữ liệu mới $\varphi(x_i), i = 1, 2, \ldots, n$. Giả sử rằng $\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)$,

$$C = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\varphi(x_i)^T$$

Tính các véc-tơ thành phần chính v với giá trị riêng λ thỏa mãn

$$Cv = \lambda v.$$

Phương pháp KPCA

Véc-tơ v có biểu diễn tuyến tính

$$v = \sum_{j=1}^{n} \alpha_j \varphi(x_j)$$

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\varphi(x_i)^T \sum_{j=1}^{n} \alpha_j \varphi(x_j) = \lambda \sum_{j=1}^{n} \alpha_j \varphi(x_j)$$

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \sum_{j=1}^{n} \alpha_j K(x_i, x_j) = \lambda \sum_{j=1}^{n} \alpha_j \varphi(x_j)$$

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(x_k)^T \varphi(x_i) \sum_{j=1}^{n} \alpha_j K(x_i, x_j) = \lambda \sum_{j=1}^{n} \alpha_j \varphi(x_k)^T \varphi(x_j)$$

Phương pháp KPCA

Từ đó ta có

$$K^2\alpha = n\lambda K\alpha.$$

Ta chỉ cần tính

$$K\alpha = n\lambda\alpha, \quad \lambda \neq 0.$$

Điều kiện chuẩn hóa $v^T v = 1 \Rightarrow \alpha^T K\alpha = 1$

Từ $K\alpha = n\lambda\alpha \Rightarrow n\lambda\alpha^T\alpha = 1 \Rightarrow \alpha^T\alpha = \dfrac{1}{n\lambda}$

Phương pháp KPCA

Chuẩn hoá ma trận hiệp phương sai (khi $\bar{\varphi} \neq 0$)

$$C = \frac{1}{n} \sum_{i=1}^{n} \Big(\varphi(x_i) - \bar{\varphi}\Big)\Big(\varphi(x_i) - \bar{\varphi}\Big)^T$$

Ma trận hiệp phương sai được biểu diễn dạng

$$C = K - \frac{1}{n}1K - \frac{1}{n}K1 + \frac{1}{n^2}1K1$$

với 1 là ma trận vuông cấp n có các phần tử bằng 1.

Phương pháp KPCA

Thuật toán KPCA:

- Tính ma trận K
- Tính ma trận C
- Tìm các giá trị riêng λ_j và véc-tơ riêng α_j và chuẩn hóa

$$a_j = \frac{1}{\sqrt{n\lambda_j}}\alpha_j$$

- Sắp xếp các giá trị riêng, véc-tơ riêng theo thứ tự giảm dần
- Chiếu dữ liệu $x \to y = (y_1, \ldots, y_d)$ xuống không gian con ứng với một số véc-tơ riêng

$$y_j = \sum_{i=1}^{n} a_{ji} K(x, x_i), j = 1, 2, \ldots, d.$$

Phương pháp KPCA

Một số hàm nhân thường dùng

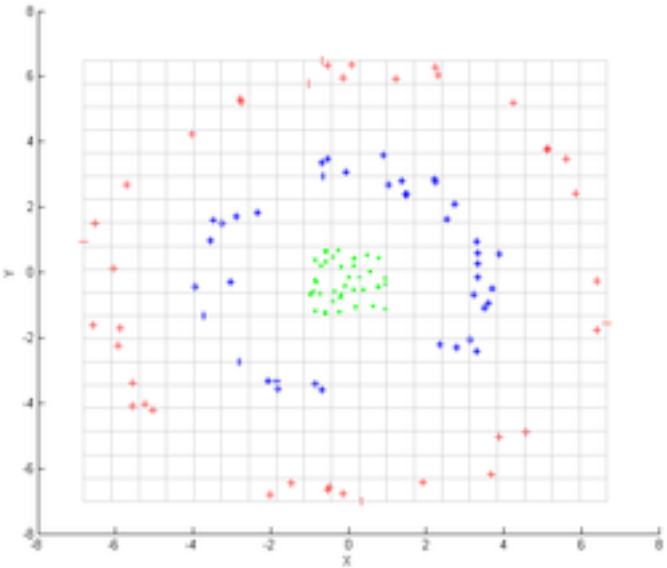
- **Nhân đa thức**

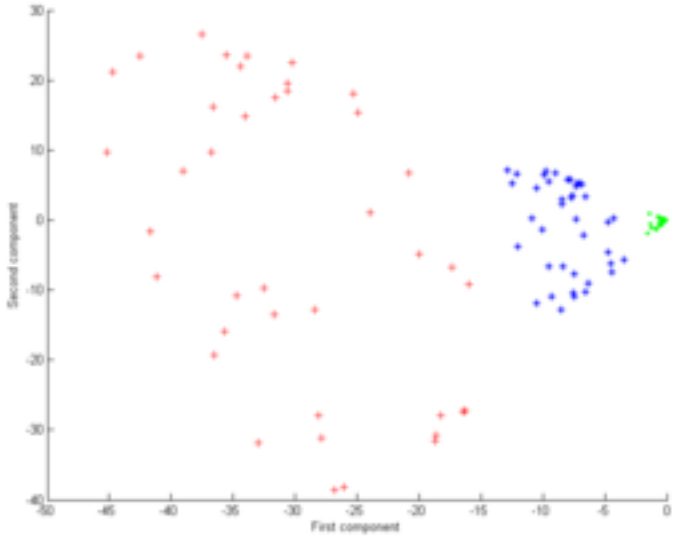
$$K(x, y) = (x^T y + c)^d$$

- **Nhân Gauss**

$$K(x, y) = \exp\left(- \frac{||x - y||^2}{2\sigma^2} \right)$$

Phương pháp KPCA





Thực hành 1

Cho đường bậc hai xác định bởi

$$x^2 - 4xy + 5y^2 = 1.$$

Tìm phép biến đổi trực giao để đưa về dạng chính tắc. Xác định các véc-tơ thành phần chính và giá trị các thành phần chính.

Thực hành 2

Cho mẫu quan sát

$$x_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, x_2 = \begin{pmatrix} 4 \\ 2 \\ 13 \end{pmatrix}, x_3 = \begin{pmatrix} 7 \\ 8 \\ 1 \end{pmatrix}, x_4 = \begin{pmatrix} 8 \\ 4 \\ 5 \end{pmatrix}$$

a) Tính trung bình $\bar{x}$.
b) Tính ma trận hiệp phương sai.
c) Tìm phép biến đổi trực giao V để $y_i = Vx_i, i = 1, 2, 3, 4$ có ma trận hiệp phương sai dạng đường chéo.
d) Biểu diễn dữ liệu trên không gian 2 chiều tạo bởi hai thành phần chính.

Bài tập 1
PCA: Dữ liệu iris

Thực hành phương pháp phân tích thành phần chính đối với dữ liệu iris và biểu diễn dữ liệu đối với hai thành phần chính.

Bài tập 2
KPCA: Dữ liệu circles

Thực hành phương pháp KPCA đối với dữ liệu circles và biểu diễn dữ liệu đối với 2 thành phần chính, so sánh với phương pháp PCA.

Tài liệu tham khảo

1. Charu C. Aggarwal; Linear Algebra and Optimization for Machine Learning, Springer, 2020.
2. David C. Lay, Steven R. Lay, Judi J. McDonald; Linear Algebra and Its Applications, Fifth edition, Pearson, 2016
3. Gilbert Strang; Linear Algebra and Learning from Data, Wellesley- Cambridge Press, 2019.
4. Stephen Boyd, Lieven Vandenberghe; Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares, Cambridge University Press, 2018.
5. Tom Lyche; Numerical Linear Algebra and Matrix Factorizations, Springer, 2020.
6. Hyun-Seok Son; Linear Algebra Coding with Python, 2020.