Principal Component Analysis

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

March 27, 2023

Phân tích thành phần chính

Lư ơ ng Ngọc Hoàng

Trư ờng Đại học Công nghệ Thông tin (UIT), ĐHQG-HCM

Ngày 27 tháng 3 năm 2023

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

References

The contents of the slides are from: Gaston Sanchez and Ethan Marzban: *All Models Are Wrong: Concepts of Statistical Learning* - `https://allmodelsarewrong.github.io/pca.html`

Ngư ời giới thiệu

Nội dung của các trang trình bày là từ: Gaston Sanchez và Ethan Marzban: Tất cả các mô hình đều sai: Các khái niệm về học tập thống kê - https://allmodelsarewrong.github.io/pca.html

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
2 / 26

Low-dimensional Representations

- Individuals form a cloud of points in a p-dim space. Variables form a cloud of arrows in an n-dim space.
- Suppose some data in which its cloud of points form a mug:

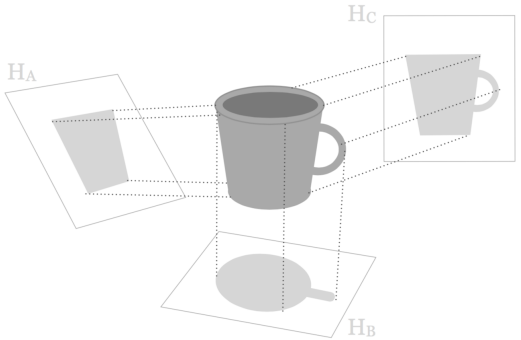
- Is there away to get a low-dimensional representation of this data?

Biểu diễn chiều thấp

- Các cá nhân tạo thành một đám mây điểm trong không gian p-dim. biến dạng một đám mây mũi tên trong không gian n-mờ.
- Giả sử một số dữ liệu trong đó đám mây điểm của nó tạo thành một cốc:

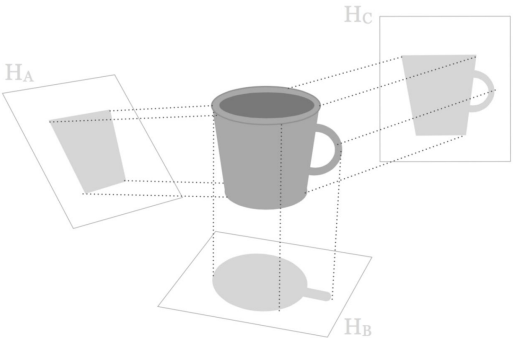
- Có cách nào để có đư ợc biểu diễn ít chiều của dữ liệu này không?

Low-dimensional Representations



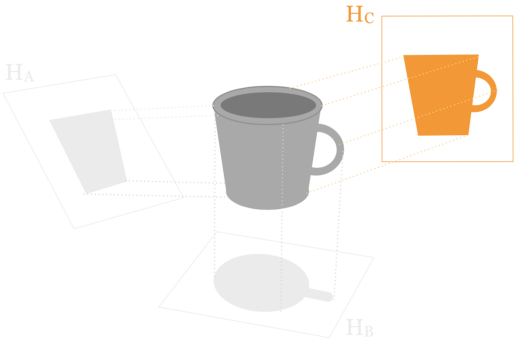
- We can look for projections of the data into sub-spaces of lower dimension.
- Assume we take a photo of the mug from different angles. What is the **best** angle to take a photo to get the images of the mug as similar as possible to the mug?

Biểu diễn chiều thấp



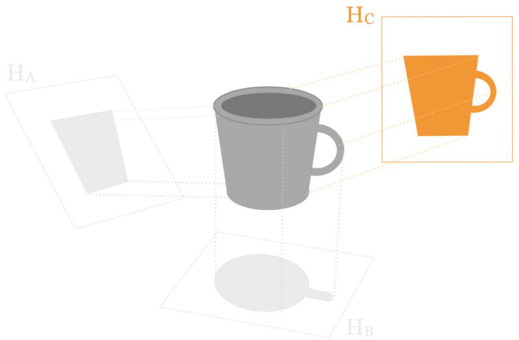
- Chúng ta có thể tìm các phép chiếu của dữ liệu vào các không gian con của cấp dư ới kích thư ớc.
- Giả sử chúng ta chụp ảnh chiếc cốc từ các góc độ khác nhau. Góc tốt nhất để chụp ảnh là gì để có đư ợc hình ảnh của chiếc cốc giống với chiếc cốc nhất có thể?

Low-dimensional Representations



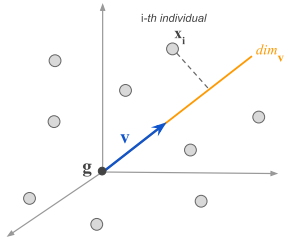
- Among 03 projections $\mathbb{H}_A, \mathbb{H}_B, \mathbb{H}_C$, the subspace $\mathbb{H}_C$ provides the best low-dimensional representation.
- The resulting image in low-dimensional space is not capturing the whole pattern: there is always some loss of information.
- Choosing the right projection, we try to minimize such loss.

Biểu diễn chiều thấp



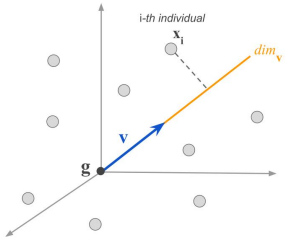
- Trong 03 phép chiếu HA,HB,HC, không gian con HC cung cấp đại diện chiều thấp tốt nhất.
- Hình ảnh thu đư ợc trong không gian ít chiều không thu đư ợc toàn bộ mẫu: luôn có một số thông tin bị mất. • Chọn phép chiếu phù hợp, chúng tôi cố gắng giảm thiểu tổn thất đó.

Projections



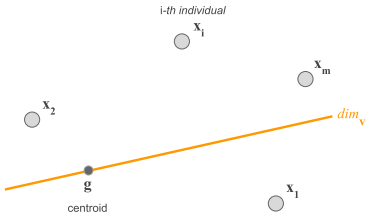
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.

dự đoán



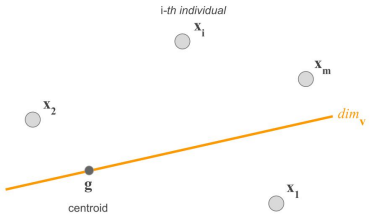
- Các điểm dữ liệu nằm trong không gian p chiều và đám mây có trọng tâm g.
- Trư ớc tiên, chúng tôi thử không gian ít chiều đơ n giản nhất: không gian 1D, trong đó có thể đư ợc hiển thị dư ới dạng một trục, đư ợc ký hiệu là dimv.

Projections



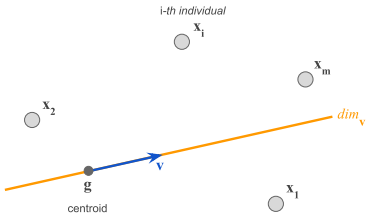
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.

dự đoán



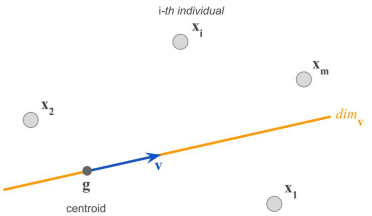
- Các điểm dữ liệu nằm trong không gian p chiều và đám mây có trọng tâm g.
- Trư ớc tiên, chúng tôi thử không gian ít chiều đơ n giản nhất: không gian 1D, trong đó có thể đư ợc hiển thị dư ới dạng một trục, đư ợc ký hiệu là dimv.

Projections



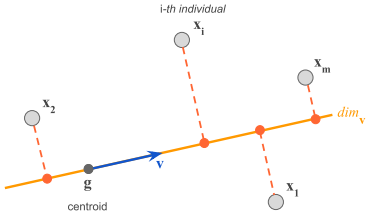
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_\mathbf{v}$.
- We manipulate $dim_\mathbf{v}$ via a vector $\mathbf{v}$ along this dimension.

dự đoán



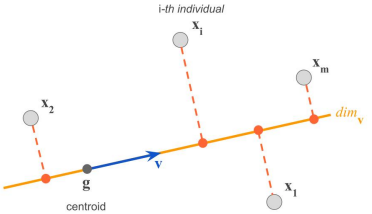
- Các điểm dữ liệu nằm trong không gian p chiều và đám mây có trọng tâm g.
- Trư ớc tiên, chúng tôi thử không gian ít chiều đơ n giản nhất: không gian 1D, trong đó có thể đư ợc hiển thị dư ới dạng một trục, đư ợc ký hiệu là dimv.
- Chúng ta thao tác dimv thông qua một vectơ v dọc theo chiều này.

Projections



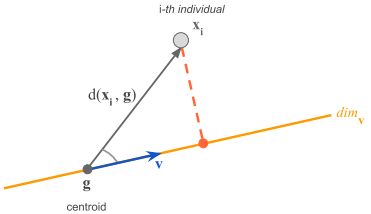
- Data points are in a p-dimensional space, and the cloud has its centroid $\boldsymbol{g}$.
- We first try the simplest low-dimensional space: a 1D space, which can be displayed as one axis, denoted as $dim_{\mathbf{v}}$.
- We manipulate $dim_{\mathbf{v}}$ via a vector $\mathbf{v}$ along this dimension.
- We want to project orthogonally the individuals onto this dimension.

dự đoán



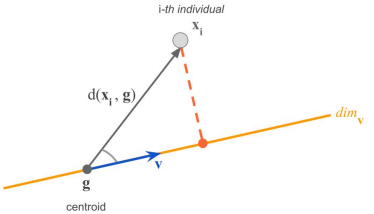
- Các điểm dữ liệu nằm trong không gian p chiều và đám mây có trọng tâm g.
- Trư ớc tiên, chúng tôi thử không gian ít chiều đơ n giản nhất: không gian 1D, trong đó có thể đư ợc hiển thị dư ới dạng một trục, đư ợc ký hiệu là dimv.
- Chúng ta thao tác dimv thông qua một vectơ v dọc theo chiều này. • Chúng tôi muốn chiếu trực giao các cá nhân lên điều này kích thư ớc.

Vector and Scalar Projections



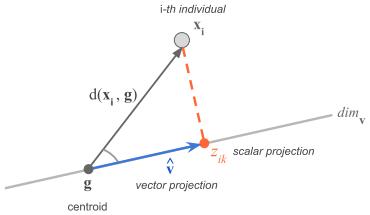
- Take the centroid $\boldsymbol{g}$ as the origin of the clouds of points.
- The dimension that we look for has to pass through the origin.
- Obtain the orthogonal projection of the i-th individual onto $dim_{\mathbf{v}}$ is projecting $\mathbf{x}_i$ onto any vector $\mathbf{v}$ along this dimension.

Phép chiếu vectơ và vô hướng



• Lấy trọng tâm g làm gốc của đám mây điểm. • Chiều ta tìm phải đi qua gốc tọa độ. • Có đư ợc hình chiếu trực giao của cá thể thứ i lên dimv là hình chiếu xi lên bất kỳ vectơ v nào dọc theo chiều này.

Vector and Scalar Projections



- The **vector projection** of $\mathbf{x}_i$ onto $\mathbf{v}$ is:

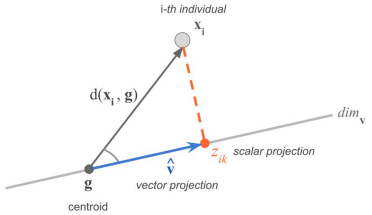
$$\hat{\mathbf{v}} = \frac{\mathbf{v}^\top \mathbf{x}_i}{\mathbf{v}^\top \mathbf{v}} \mathbf{v}$$

- The **scalar projection** of $\mathbf{x}_i$ onto $\mathbf{v}$ is:

$$z_{ik} = \frac{\mathbf{v}^\top \mathbf{x}_i}{\|\mathbf{v}\|}$$

- We would prefer the **scalar projection** to obtain the co-ordinate of $\mathbf{x}_i$ along this axis.

Phép chiếu vectơ và vô hướng



- Hình chiếu vectơ của xi lên v là:

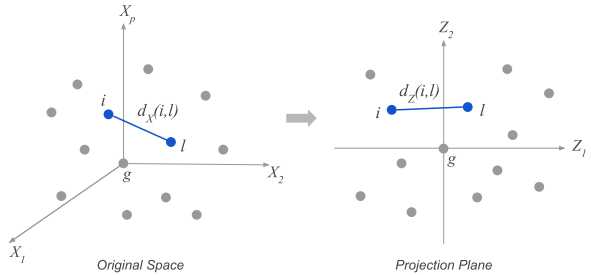
$$\hat{v} = \frac{v \quad xi}{v \quad v} v$$

- Hình chiếu vô hướng của xi lên v là: v

$$zik = \frac{xi}{v}$$

- Chúng tôi muốn phép chiếu vô hướng thu được tọa độ của xi dọc theo trục này.

Projected Inertia

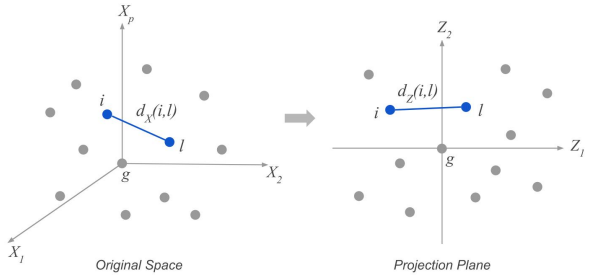


- Find the angle that give the best photo of the object $\Longleftrightarrow$ Find the subspace that the distances between the points are the most similar to the original points.
- The overall dispersion of the original data is: $\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l)$. We try to find a subspace $\mathbb{H}$ such that:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l) \approx \sum_{i=1}^{n}\sum_{l=1}^{n} d^2_{\mathbb{H}}(i,l)$$

Machine Translated by Google

quán tính dự kiến



- Tìm góc cho ảnh đẹp nhất của vật Tìm không gian con sao cho khoảng cách giữa các điểm bằng nhau nhất so với điểm ban đầu.
- Độ phân tán tổng thể của dữ

liệu gốc là: cố gắng tìm một không gian con H $\sum_{\text{tôi = 1}}^{N}\sum_{\text{2dl = 1}}^{N}$ (Il). Chúng tôi sao cho:

$$\sum_{i=1}^{N}\sum_{l=1}^{N} \text{ngày}^2(i,l) \approx \sum_{i=1}^{N}\sum_{l=1}^{N} \text{ngày}^2_{\text{Xin chào, tôi)}}$$

Projected Inertia

- The overall dispersion is related to the inertia as:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l) = 2n\sum_{i=1}^{n} d^2(i,g) = 2n^2\frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = 2n^2\texttt{Inertia}$$

- Finding the subspace $\mathbb{H}$ that yields similar distances to the original subspace corresponds to maximize the projected inertia:

$$\max_{\mathbb{H}}\left\{\frac{1}{n}\sum_{i=1}^{n} d^2_{\mathbb{H}}(i,g)\right\}$$

quán tính dự kiến

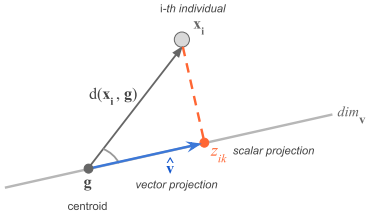
- Độ phân tán tổng thể liên quan đến quán tính như :

$$\sum_{i=1}^{N}\sum_{l=1}^{N} d^2(i,l) = 2n\sum_{i=1}^{N} d^2(i,g) = 2n^2\frac{1}{n}\sum_{i=1}^{N} d^2(i,g) = 2n^2\text{quán tính}$$

- Tìm không gian con H có khoảng cách tương tự với không gian con ban đầu tương ứng để tối đa hóa quán tính dự kiến:

$$\text{tối đa}_{H}\left\{\frac{1}{N}\sum_{i=1}^{N} d^2_{H}(i,g)\right\}$$

Projected Inertia



- We are consider $1D$ case, $\mathbb{H} \subseteq \mathbb{R}^1$, the projected inertia becomes:

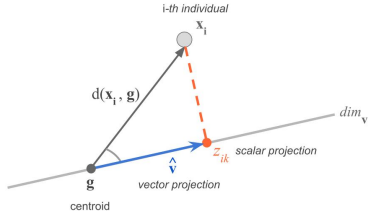
$$\frac{1}{n}\sum_{i=1}^{n} d^2_{\mathbb{H}}(i,g) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{v})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2$$

- Our maximization problem becomes:

$$\max_{\mathbf{v}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{v})^2 \right\} \quad \texttt{s.t.} \quad \mathbf{v}^\top \mathbf{v} = 1$$

- We constraint $\mathbf{v}$ to be a unit vector; otherwise, the maximization objective is unbounded.

quán tính dự kiến



- Ta xét trư ờng hợp 1D, H R¹, quán tính dự kiến trở thành:

$$\frac{1}{n}\sum_{i=1}^{N} đ^2_{H(i,\ g)} = \frac{1}{n}\sum_{i=1}^{N} (x \cdot v)^2 = \frac{1}{n}\sum_{i=1}^{N} {}^{2}{}_{t\acute{o}i}$$

- Bài toán tối đa hóa của chúng ta trở thành:

$$\text{tối đa}_{v\ \{\ 1} \frac{1}{n}\sum_{i=1}^{N} (x \cdot v)\ 2\}\ \text{st}\ v \quad v = 1$$

- Ta ràng buộc v là véc tơ đơ n vị; mặt khác, việc tối đa hóa mục tiêu là không giới hạn.

Maximization Problem

- Assume mean-centered data, the centroid $\mathbf{g}$ of the cloud of points is the origin $\mathbf{g} = \mathbf{0}$.
- We are projecting onto a line spanned by a unit-vector $\mathbf{v}$, the projected inertia $I_{\mathbb{H}}$ is the variance of the projected data points:

$$I_{\mathbb{H}} = \frac{1}{n}\sum_{i=1}^{n} d^2_{\mathbb{H}}(i,0) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{v})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\mathbf{z}^\top\mathbf{z} = \frac{1}{n}\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}$$

where

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \mathbf{X}\mathbf{v} = \begin{bmatrix} - - - \mathbf{x}_1^\top - - - \\ - - - \mathbf{x}_2^\top - - - \\ - - - - - - - \\ - - - \mathbf{x}_n^\top - - - \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{pmatrix}$$

Vấn đề tối đa hóa

- Giả sử dữ liệu lấy trung bình làm trung tâm, tâm g của đám mây điểm là gốc tọa độ g = 0. •

Chúng ta đang chiếu lên một đư ờng đư ợc kéo dài bởi một vectơ đơ n vị v, quán tính dự kiến IH là phư ơ ng sai của các điểm dữ liệu dự kiến:

$$IH = \frac{1}{n}\sum_{i=1}^{N} {}^{2}H(i,0) \quad \frac{1}{n}\sum_{i=1}^{N}(x_{\cdot} v)^2$$
$$= \frac{1}{n}\sum_{i=1}^{N} {}^{2} = \frac{1}{} zz = n \quad \frac{1}{N} v \quad X \quad Xv$$

Ở đâu

$$z = Xv = \begin{matrix} z1 \\ z2 \\ \\ zn \end{matrix} \quad \begin{matrix} x\ 1 \\ x\ 2 \\ \\ x \end{matrix} \quad \begin{matrix} v1 \\ v2 \end{matrix}$$

Maximization Problem

- The maximization problem becomes:

$$\max_{\mathbf{v}} \left\{ \frac{1}{n} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \right\} \quad \text{s.t.} \quad \mathbf{v}^\top \mathbf{v} = 1$$

- To solve this maximization, problem, we use Lagrange multipliers.

$$\mathcal{L} = \frac{1}{n} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} - \lambda (\mathbf{v}^\top \mathbf{v} - 1)$$

- Set the derivative of the Lagrangian $\mathcal{L}$ wrt $\mathbf{v}$ to $\mathbf{0}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = \frac{2}{n} \mathbf{X}^\top \mathbf{X} \mathbf{v} - 2\lambda \mathbf{v} = \mathbf{0} \Rightarrow \underbrace{\frac{1}{n} \mathbf{X}^\top \mathbf{X}}_{\mathbf{S} \in \mathbb{R}^{p \times p}} \mathbf{v} = \lambda \mathbf{v} \Rightarrow \mathbf{S} \mathbf{v} = \lambda \mathbf{v}$$

- This means that $\mathbf{v}$ is an eigenvector (with eigenvalue λ) of $\mathbf{S}$.
- λ is the value of the projected inertia $I_{\mathbb{H}}$ that we want to maximize.

Vấn đề tối đa hóa

- Bài toán cực đại hóa trở thành:

tối đa
v { 1 N v X Xv} v v = 1

- Để giải bài toán cực đại hóa này, ta sử dụng hệ số nhân Lagrange.

L = 1 N v X Xv λ(v v 1)

- Đặt đạo hàm của Lagrangian L wrt v thành 0:

L v = 2 N X Xv 2λv = 0 1 N X X v = λv Sv = λv

S Rp×p

- Điều này có nghĩa là v là một vectơ riêng (với giá trị riêng λ) của S. • λ là giá trị của IH quán tính dự kiến mà chúng ta muốn cực đại hóa.

Eigenvectors of S

- Assume $\mathbf{X}$ is full rank ($rank(\mathbf{X}) = p$). We have p eigenvectors:

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_k & \dots & \mathbf{v}_p \end{bmatrix}$$

- We also have the matrix of eigenvalues $\mathbf{\Lambda} = \texttt{diag}\{\lambda_i\}_{i=1}^{n}$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

- We then have the matrix of projected points $\mathbf{Z}$ (also known as **the matrix of principal components (PC's)**):

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_k & \dots & \mathbf{z}_p \end{bmatrix}$$

where the k-th principal component $\mathbf{z}_k$ is:

$$\mathbf{z}_k = \mathbf{X}\mathbf{v}_k = v_{1k}\mathbf{x}_1 + v_{2k}\mathbf{x}_2 + \ldots + v_{pk}\mathbf{x}_p$$

with $\mathbf{x}_k$ denotes columns of $\mathbf{X}$.

Véc tơ riêng của S •

Giả sử X là hạng đầy đủ (xếp hạng(X) = p). Ta có p vectơ riêng:

V = [v1 v2 . . . vk . . . phó chủ tịch]

- Ta cũng có ma trận các giá trị riêng Λ = diag{λi} n tôi = 1

$$= \begin{matrix} \lambda 1\ 0\ .\ .\ .\ 0\ 0 \\ \lambda 2\ .\ .\ .\ 0 \\ \\ 0\ 0\ .\ .\ .\ \lambda p \end{matrix}$$

- Khi đó ta có ma trận các điểm hình chiếu Z (còn được gọi là ma trận các thành phần chính (PC's)):

Z = [z1 z2 . . . zk . . . zp]

trong đó thành phần chính thứ k zk là:

zk = Xvk = v1kx1 + v2kx2 + · · · + vpkxp

với xk biểu thị các cột của X.

Eigenvalues of S

- Because the data is mean-centered, we have $\texttt{mean}(\mathbf{x}_i) = 0$. Then, $\texttt{mean}(\mathbf{z}_k) = 0$.
- How about the variance of $\mathbf{z}_k$?

$$Var(\mathbf{z}_k) = \frac{1}{n}\mathbf{z}^\top\mathbf{z} = \frac{1}{n}(\mathbf{X}\mathbf{v}_k)^\top(\mathbf{X}\mathbf{v}_k) = \frac{1}{n}\mathbf{v}_k^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}_k$$
$$= \mathbf{v}_k^\top\mathbf{S}\mathbf{v}_k = \mathbf{v}_k^\top(\lambda_k\mathbf{v}_k) = \lambda_k(\mathbf{v}_k^\top\mathbf{v}_k) = \lambda_k$$

- The k-th eigenvalue of $\mathbf{S}$ is the variance of the k-th principal component.
- If $\mathbf{X}$ is mean centered, $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is the covariance matrix of data.
- If $\mathbf{X}$ is standardized (mean-centered and scaled by the variance), then $\mathbf{S}$ is the correlation matrix.

Giá trị riêng của S

- Vì dữ liệu lấy trung bình làm trung tâm nên chúng ta có mean(xi) = 0. Khi đó, mean(zk) =

0. • Còn phư ơ ng sai của zk thì sao ?

$$\text{V ar(zk)} = \frac{1}{n}\ \text{zz} = \frac{1}{N}\text{(Xvk)}\quad\text{(Xvk)} = \frac{1}{}\text{vn}\quad\text{kX Xvk}$$
$$= \text{kSvk} = \text{v}\quad\text{k}\,(\lambda\text{kvk}) = \lambda\text{k(v}\quad\text{kvk)} = \lambda\text{k}$$

- Giá trị riêng thứ k của S là phư ơ ng sai của giá trị gốc thứ k thành phần.

liệu. • Nếu X là trung bình, S = $\frac{1}{n}$X X là ma trận hiệp phư ơ ng sai của dữ

- Nếu X đư ợc chuẩn hóa (trung bình là trung tâm và đư ợc chia tỷ lệ theo phư ơ ng sai), thì S là ma trận tư ơ ng quan.

Eigenvalues of S

$$\texttt{Inertia} = \frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = \sum_k \lambda_k = \texttt{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)$$

- $\sum_{k=1}^{p} \lambda_k$ relates to the total amount of variability in the data.
- The principal components capture different parts of the variability in the data.

Giá trị riêng của S

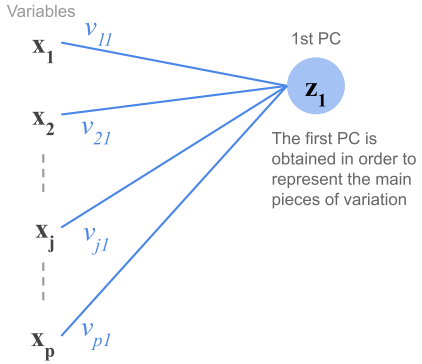
$$\text{quán tính} = \frac{1}{n}\sum_{i=1}^{N} đ^2(i, g) = {}_k\ \lambda k = \texttt{tr}\left(\frac{1}{n}\ X\ \ X\right)$$

- $\sum_{k=1}^{P} \lambda k$ liên quan đến tổng lượng biến đổi trong dữ liệu.
- Các thành phần chính nắm bắt các phần khác nhau của sự thay đổi trong dữ liệu.

VNUHCM - UIT
19 / 26

Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- The first PC is a linear combination:



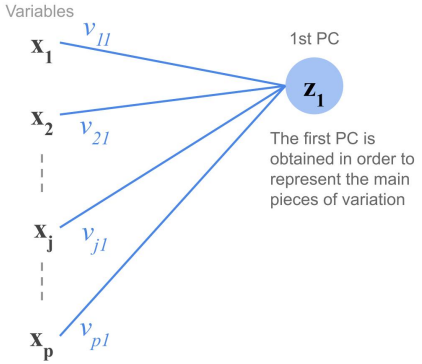
Phân tích thành phần chính (PCA)

- Cho tập p biến x1,x2, . . . , xp, chúng tôi muốn lấy k zk các biến z1, mới, đư ợc gọi là Thành phần chính (PC).

z2, . . . ,

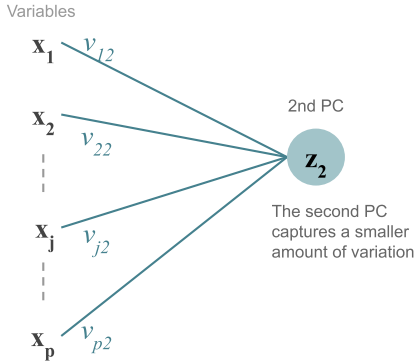
- Thành phần chính là sự kết hợp tuyến tính của p biến: z = Xv.

- PC đầu tiên là tổ hợp tuyến tính:



Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- The second PC is another linear combination:

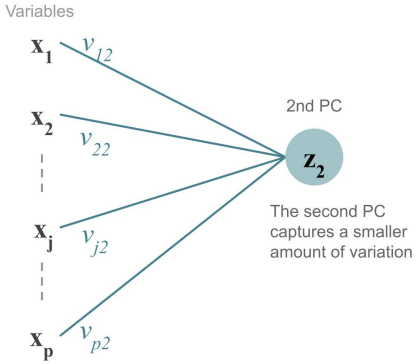


Phân tích thành phần chính (PCA)

- Cho tập p biến x1,x2, . . . , xp, chúng tôi muốn lấy k zk các biến z1, mới, đư ợc gọi là Thành phần chính (PC).

z2, . . . ,

- Thành phần chính là sự kết hợp tuyến tính của p biến: z = Xv.
- PC thứ hai là một tổ hợp tuyến tính khác:



Principal Component Analysis (PCA)

- Given a set of p variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, we want to obtain new k variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$, called the **Principal Components (PCs)**.
- A principal component is a **linear combination** of the p variables: $\mathbf{z} = \mathbf{X}\mathbf{v}$.
- We compute PCs as linear combinations of original variables:

$$\begin{aligned}
\mathbf{z}_1 &= v_{11}\mathbf{x}_1 + v_{21}\mathbf{x}_2 + \ldots + v_{p1}\mathbf{x}_p \\
\mathbf{z}_2 &= v_{12}\mathbf{x}_1 + v_{22}\mathbf{x}_2 + \ldots + v_{p2}\mathbf{x}_p \\
\vdots &= \vdots \\
\mathbf{z}_k &= v_{1k}\mathbf{x}_1 + v_{2k}\mathbf{x}_2 + \ldots + v_{pk}\mathbf{x}_p
\end{aligned}$$

Or:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

where $\mathbf{Z}$ is an $n \times k$ matrix of principal components, and $\mathbf{V}$ is a $p \times k$ matrix of weights (directional vectors of the principal axes).

Phân tích thành phần chính (PCA)

- Cho tập p biến x1,x2, . . . , xp, chúng tôi muốn lấy k mới các biến z1, z2, . . . ,zk, được gọi là Thành phần chính (PC).
- Thành phần chính là sự kết hợp tuyến tính của p biến: z = Xv.
- Chúng tôi tính toán PC dưới dạng tổ hợp tuyến tính của các biến ban đầu:

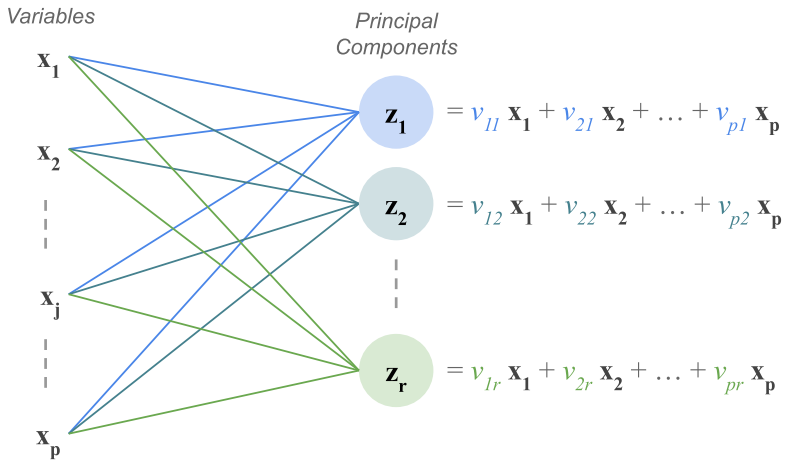
z1 = v11x1 + v21x2 + · · · + vp1xp
z2 = v12x1 + v22x2 + · · · + vp2xp
=
zk = v1kx1 + v2kx2 + · · · + vpkxp

Hoặc:

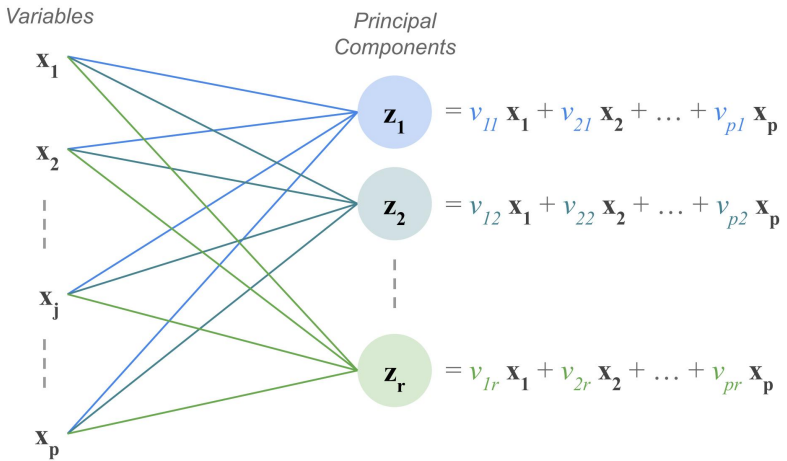
Z = XV

trong đó Z là ma trận n × k của các thành phần chính và V là một ma trận trọng số p × k (vectơ chỉ phương của các trục chính).

Principal Component Analysis (PCA)



Phân tích thành phần chính (PCA)



Finding Principal Components

- The components $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$ are required to capture most of the variation in data $\mathbf{X}$.
- We look for a vector $\mathbf{v}_h$ such that a component $\mathbf{z}_h = \mathbf{X}\mathbf{v}_h$ has maximum variance:

$$\max_{\mathbf{v}_h} var(\mathbf{z}_h) \Rightarrow \max_{\mathbf{v}_h} var(\mathbf{X}\mathbf{v}_h) \Rightarrow \max_{\mathbf{v}_h} \frac{1}{n}\mathbf{v}_h^\top \mathbf{X}^\top \mathbf{X}\mathbf{v}_h$$

- If $\mathbf{v}_h$ can be arbitrarily big, the problem is unbounded. We need to restrict $\mathbf{v}_h$ to be of unit norm:

$$\|\mathbf{v}_h\| = 1 \Rightarrow \mathbf{v}_h^\top \mathbf{v}_h = 1$$

- If we denote the covariance matrix $\mathbf{S} = (1/n)\mathbf{X}^\top \mathbf{X}$, then

$$\max_{\mathbf{v}_h} \mathbf{v}_h^\top \mathbf{S}\mathbf{v}_h \quad \text{s.t.} \quad \mathbf{v}_h^\top \mathbf{v}_h = 1$$

- To avoid redundancy, we require $\mathbf{z}_h^\top \mathbf{z}_l = 0$ mutually orthogonal if $h \neq l$.

Tìm các thành phần chính • Các thành phần z1,

z2, . . . , zk đư ợc yêu cầu để nắm bắt hầu hết các biến thể trong dữ liệu X.

- Ta tìm véc tơ vh sao cho thành phần zh = Xvh có phư ơ ng sai lớn nhất:

tối đa var(zh) max var(Xvh) max 1vn hX Xvh
vh vh vh

- Nếu vh có thể lớn tùy ý thì bài toán là vô giới hạn. Chúng ta cần hạn chế vh ở dạng chuẩn đơ n vị:

vh = 1 v hvh = 1

- Nếu chúng ta ký hiệu ma trận hiệp phư ơ ng sai S = (1/n)X X, thì

cực đại hSvh st v hvh = 1
v vh

- Để tránh dư thừa, chúng tôi yêu cầu zh zl = 0 trực giao với nhau nếu h l.

Finding Principal Components

All PCs can be found by **diagonalizing** $\mathbf{S} = (1/n)\mathbf{X}^\top\mathbf{X}$.

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

- Λ is a diagonal matrix. The diagonal elements of Λ are the eigenvalues of $\mathbf{S}$.
- The columns of $\mathbf{V}$ are orthonormal: $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$
- The columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{S}$.
- $\mathbf{V}^\top = \mathbf{V}^{-1}$

Because $\mathbf{S}$ is a $p \times p$ symmetric matrix, we have:

- $\mathbf{S}$ has p real eigenvalues.
- The eigenvectors corresponding to different eigenvalues are orthogonal. $\mathbf{S}$ is orthogonally diagonalizable ($\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$).
- The set of eigenvalues of $\mathbf{S}$ is called the **spectrum of $\mathbf{S}$**.
- The PCA is obtained via an Eigenvalue Decomposition of $\mathbf{S}$.

Tìm các thành phần chính Tất cả các PC

có thể được tìm thấy bằng cách chéo hóa S = (1/n)X X.

S = VΛV

- Λ là ma trận đường chéo. Các phần tử đường chéo của Λ là các giá trị riêng của S.
- Các cột của V là trực giao: V V = I • Các cột của V là vectơ riêng của S. • V = V 1

Vì S là ma trận đối xứng ap × p nên ta có:

- S có p giá trị riêng thực.
- Các vectơ riêng tư ơ ng ứng với các giá trị riêng khác nhau là trực giao. S có thể chéo hóa trực giao (S = VΛV).
- Tập hợp các giá trị riêng của S được gọi là phổ của S. • PCA thu được thông qua Phân tách giá trị riêng của S.

Examples

- Principal Component Analysis - Intuitions: https://fmin.xyz/docs/applications/pca/
- Principal Component Analysis - Explained Visually: https://setosa.io/ev/principal-component-analysis/
- Principal Component Analysis (PCA): Iris data: https://www.math.umd.edu/~petersd/666/html/iris_pca.html
- Face Recognition using Principal Component Analysis: https://machinelearningmastery.com/face-recognition-using-principal-component-analysis/

ví dụ

- Phân tích thành phần chính - Trực giác: https://fmin.xyz/docs/applications/pca/
- Phân tích thành phần chính - Giải thích trực quan: https://setosa.io/ev/principal-component-analysis/
- Phân tích thành phần chính (PCA): Dữ liệu mống mắt: https://www.math.umd.edu/~petersd/666/html/iris_pca.html
- Nhận dạng khuôn mặt bằng cách sử dụng Phân tích thành phần chính : https://machinelearningmastery.com/face-recognition-using-principal-component-analysis/

VNUHCM - UIT
26 / 26