Review

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

March 27, 2023

Ôn tập

Lư ơ ng Ngọc Hoàng

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

Ngày 27 tháng 3 năm 2023

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/duality.html`

Người 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:// allmodelsarerong.github.io/duality.html

Basic Notations

- Assume our data to be in a tabular format, which can be represented as a mathematical **matrix** object.
- An example of a data matrix $\boldsymbol{X}$ of size $n \times p$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

- We assume the rows of a data matrix correspond to the data items/individuals/objects.
- We assume the columns of a data matrix correspond to the variables/features observed on the individuals.
- x_{ij} represents the value observed for the j-th variable on the i-th individual.

Ký hiệu cơ bản

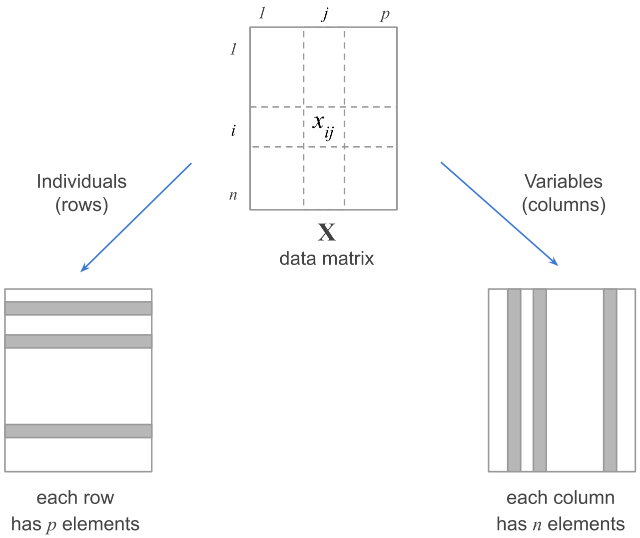
• Giả sử dữ liệu của chúng ta ở định dạng bảng, có thể được
biểu diễn dưới dạng đối tượng ma trận toán
học . • Một ví dụ về ma trận dữ liệu X có kích thước n × p:

x11 x12 . . . x1j . . . x1p
x21 x22xi1 xi2 x2j . . xij x2p . xip
xn1 xn2 . . . xnj . . .
xnp

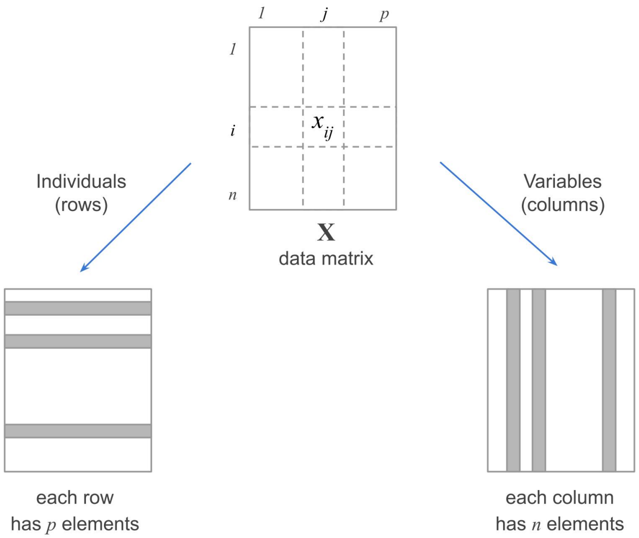
X =

• Chúng tôi giả sử các hàng của ma trận dữ liệu tương ứng với các
mục/cá nhân/đối tượng dữ
liệu. • Chúng tôi giả sử các cột của ma trận dữ liệu tương ứng với
các biến / tính năng được quan sát trên các cá
• nhân. xij biểu thị giá trị được quan sát cho biến thứ j trên biến thứ i
cá nhân.

Duality of a Data Matrix

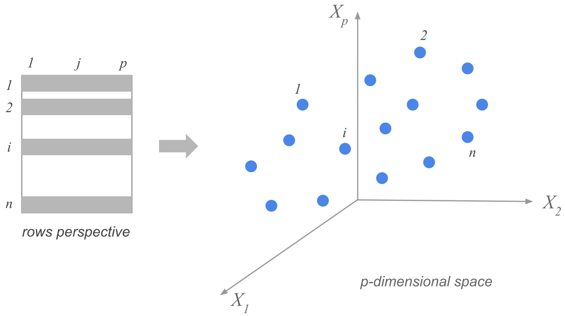


Tính hai mặt của ma trận dữ liệu



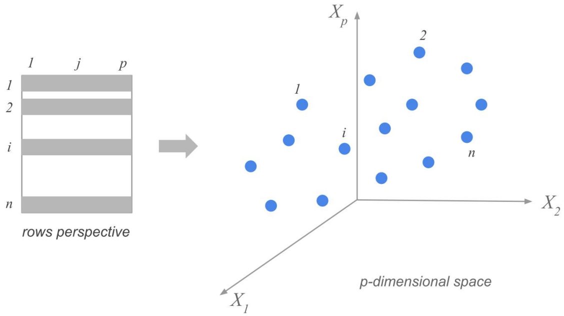
VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
4 / 24

Duality of a Data Matrix - Row Space



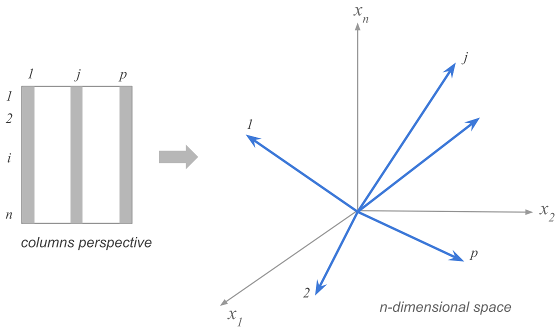
- Each row of the data matrix has p elements. We can regard each row as a single point in a p-dimensional space (with p axes).
- All together they form a cloud of points.

Tính hai mặt của ma trận dữ liệu - Row Space



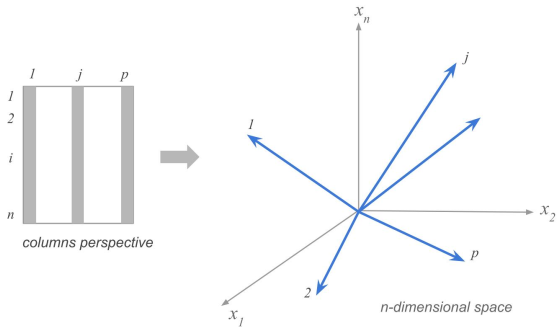
- Mỗi hàng của ma trận dữ liệu có p phần tử. Chúng ta có thể coi mỗi hàng là một điểm trong không gian p chiều (với p trục).
- Tất cả cùng nhau tạo thành một đám mây điểm.

Duality of a Data Matrix - Column Space



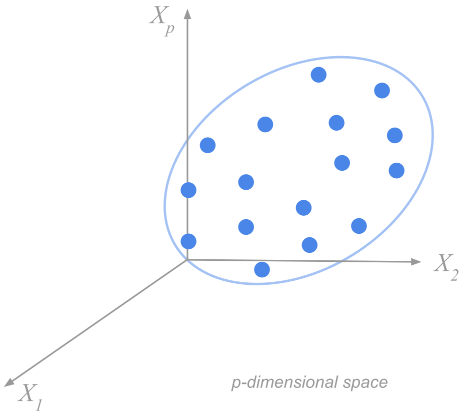
- Because each variable has n elements, we can regard the set of p variables as objects in an n-dimensional space (with n axes).
- Each variable is represented as an arrow (or a vector). In data pre-processing, we apply transformations on variables, that can change their scales (shrinking or stretching) without modifying their directions.

Tính hai mặt của ma trận dữ liệu - Không gian cột



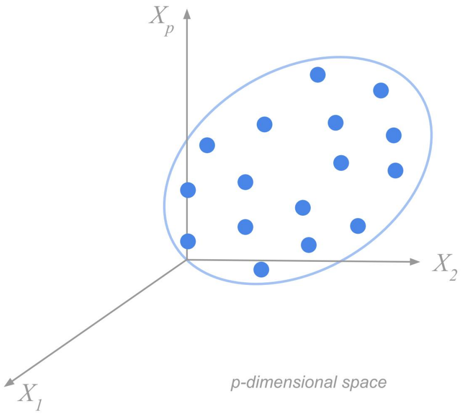
- Vì mỗi biến có n phần tử nên có thể coi tập p biến là các đối tư ợng trong không gian n chiều (có n trục). • Mỗi biến đư ợc biểu diễn dư ới dạng mũi tên (hoặc vectơ). Trong tiền xử lý dữ liệu, chúng tôi áp dụng các phép biến đổi trên các biến có thể thay đổi tỷ lệ của chúng (thu nhỏ hoặc kéo dài) mà không sửa đổi hư ớng của chúng.

Duality of a Data Matrix - Cloud of Individuals



- The rows of the data matrix correspond to n individuals/points in a p-dimensional space.
- We consider common operations we can apply on the individuals.

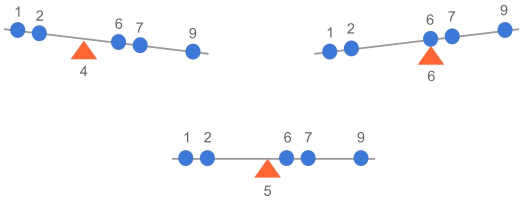
Tính hai mặt của ma trận dữ liệu - Đám mây cá nhân



- Các hàng của ma trận dữ liệu tư ơ ng ứng với n cá nhân/điểm trong một không gian p chiều.
- Chúng tôi xem xét các hoạt động chung mà chúng tôi có thể áp dụng cho các cá nhân.

Cloud of Individuals - Average Individual

- If we have only one variable, then all individual points lie in a one-dimensional space, which is a **line**.
- The **average individual** can be defined as the arithmetic average of the values, corresponding to the balancing point.



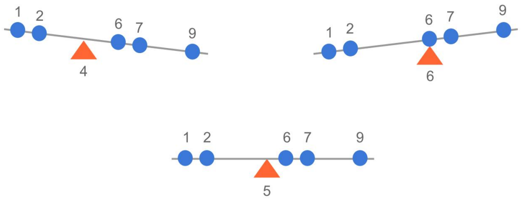
- The average of individuals $x_1, x_2, \dots, x_n$ is:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{x}^\top \mathbf{1}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{1} = (1, 1, \dots, 1)$.

Đám mây cá nhân - Cá nhân trung bình

- Nếu chúng ta chỉ có một biến, thì tất cả các điểm riêng biệt đều nằm trong không gian một chiều, đó là một đư ờng thẳng.
- Trung bình cá nhân có thể đư ợc định nghĩa là trung bình cộng của các giá trị, tư ơ ng ứng với điểm cân bằng.

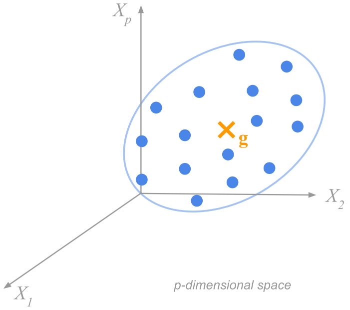


- Trung bình cộng của các cá nhân x1, x2, . . . , xn là:

x¯ = (x1 + x2 + · · · + xn) / N = 1/n N i=1 xi = 1/ 1 xn

trong đó x = (x1, x2, . . . , xn) và 1 = (1, 1, . . . , 1).

Cloud of Individuals - Average Individual

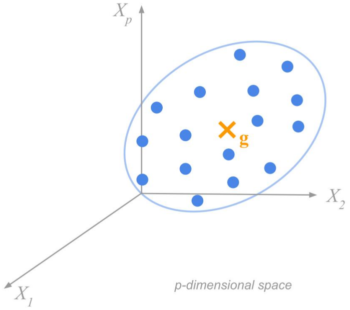


- If we have multiple variables (multivariate), the average individual is the point $\boldsymbol{g}$ with co-ordinates as the averages of all the variables:

$$\boldsymbol{g} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$$

  where $\bar{x}_j$ is the average of the j-th variable.
- $\boldsymbol{g}$ is also called the **centroid**, barycenter, or center of gravity of the cloud of points.

Đám mây cá nhân - Cá nhân trung bình

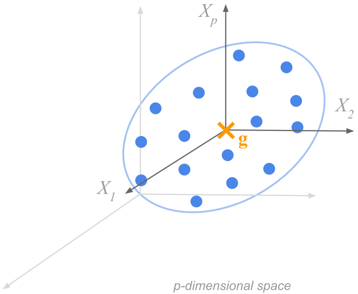


- Nếu chúng ta có nhiều biến (đa biến), trung bình cá nhân là điểm g với tọa độ là trung bình của tất cả các biến:

  g = (x¯1,

- x¯2, . . . , x¯p) trong đó x¯j là giá trị trung bình của biến thứ j. g còn đư ợc gọi là trọng tâm, tâm khối, hoặc trọng tâm của đám mây điểm.

Cloud of Individuals - Centered Data

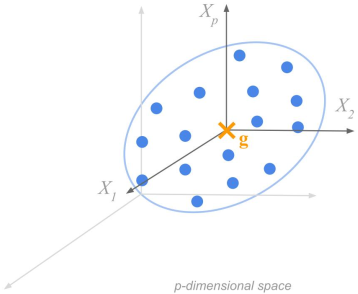


p-dimensional space

- It is convenient to make the centroid of a data set become the origin of the cloud of points.
- Geometrically, this transformation is a shift of the axes in the p-dimensional space.
- Algebraically, this transformation is expressing the value of each variable in terms of the deviations from their averages (means).

$$\mathbf{x}_1 - \mathbf{g}, \mathbf{x}_2 - \mathbf{g}, \ldots, \mathbf{x}_n - \mathbf{g}$$

Đám mây cá nhân - Dữ liệu tập trung

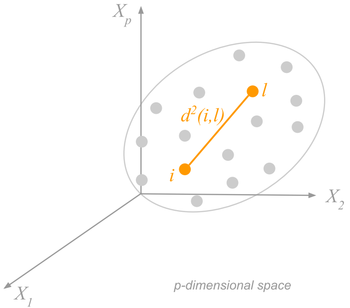


p-dimensional space

- Thuận tiện để làm cho trọng tâm của tập dữ liệu trở thành điểm gốc của đám mây điểm.
- Về mặt hình học, sự biến đổi này là sự dịch chuyển của các trục trong không gian p chiều.
- Về mặt đại số, phép biến đổi này biểu thị giá trị của từng biến theo độ lệch so với giá trị trung bình (trung bình) của chúng.

x1 g,x2 g, . . . , xn g

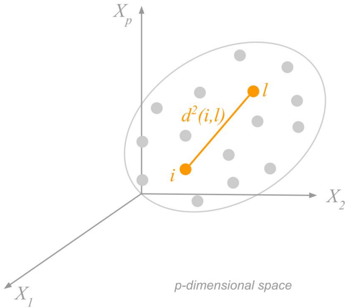
Cloud of Individuals - Distance between Individuals



- The most common type of distance is the (squared) Euclidean distance.
- With p variables, the squared distance between the i-th individual and the l-th individual is:

$$\begin{aligned} d^2(i,l) &= (x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \ldots + (x_{ip} - x_{lp})^2 \\ &= (\boldsymbol{x}_i - \boldsymbol{x}_l)^\top (\boldsymbol{x}_i - \boldsymbol{x}_l) \end{aligned}$$

Đám mây cá nhân - Khoảng cách giữa các cá nhân



- Loại khoảng cách phổ biến nhất là Euclidean (bình phư ơ ng) khoảng cách.
- Với p biến, bình phư ơ ng khoảng cách giữa cá thể thứ i và cá nhân thứ l là:

đ 2 (i,l) = (xi1 xl1)2 + (xi2 xl2)2 + · · · + (xip xlp)2

= (xi xl) (xi xl)

Cloud of Individuals - Overall Dispersion

- The centroid is a measure of center of individuals. The dispersion is a measure of spread/scatter among individuals.
- If we have three individuals, we can compute all pairwise distances and sum them up:

$$
\begin{aligned}
d^2(a,a) + d^2(b,b) + d^2(c,c) &+ \\
d^2(a,b) + d^2(b,a) &+ \\
d^2(a,c) + d^2(c,a) &+ \\
d^2(b,c) + d^2(c,b) &
\end{aligned}
$$

- The **overall dispersion** of n individuals is:

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

Đám mây cá nhân - Phân tán tổng thể

- Trọng tâm là thư ớc đo trung tâm của các cá nhân. Sự phân tán là thư ớc đo mức độ lây lan/phân tán giữa các cá nhân.
- Nếu chúng ta có ba cá nhân, chúng ta có thể tính toán tất cả các khoảng cách theo cặp và tóm tất chúng:

đ ^{2} (a, a) + d ^{2} (b, b) + d ^{2} (c,c)+

đ ^{2} (a, b) + d ^{2} (b, a)+

ngày ^{2} (a, c) + d ^{2} (c, a)+

ngày ^{2} (b, c) + d ^{2} (c, b)

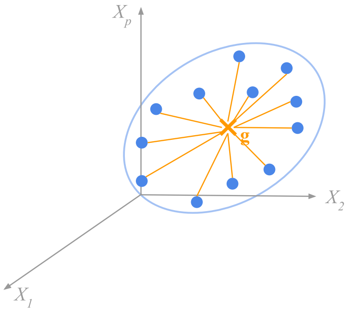
- Độ **phân tán chung** của n cá thể là:

N N

đ ^{2} (Il)

i=1 l=1

Cloud of Individuals - Inertia

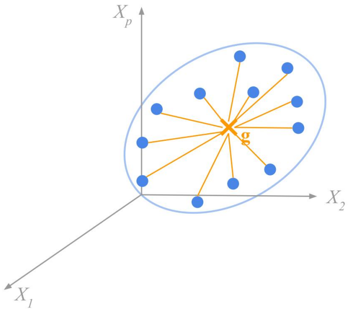


- The **inertia** can be computed by averaging the squared distances between all individuals and the centroid:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{g})^\top (\boldsymbol{x}_i - \boldsymbol{g})$$

- In uni-dimensional case, $p = 1$, we have: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Đám Mây Cá Nhân - Inertia



- Quán tính có thể đư ợc tính bằng cách lấy trung bình các khoảng cách bình phư ơ ng giữa tất cả các cá nhân và trung tâm:

$$\frac{1}{n}\sum_{i=1}^{N} đ^2 (tôi, g) = \frac{1}{n}\sum_{i=1}^{N} (xi \quad g) \quad (xi \quad g)$$

- Trong trư ờng hợp một chiều p = 1, ta có: $\frac{1}{N}\sum_{tôi=1}^{N}(xi \quad x^-)^2$

Cloud of Variables - Mean of a Variable

- The mean (or average) of an n-element variable $\bar{x}$ is computed by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= \frac{1}{n}x_1 + \frac{1}{n}x_2 + \ldots + \frac{1}{n}x_n$$

- We can generalize the concept of an average as a **weighted aggregation of information**.
- If we denote the weight of the i-th individual as w_i, then the average is:

$$\bar{x} = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$
$$= \sum_{i=1}^{n} w_ix_i$$
$$= \mathbf{w}^\top \mathbf{x}$$

Cloud of Variables - Mean of a Variable • Giá trị

trung bình (hoặc trung bình) của một biến n phần tử x¯ đư ợc tính bằng:

x¯ = (x1 + x2 + · · · + xn) / N = 1/n Σ (i=1 … N) xi

= 1/N x1 + 1/N x2 + · · · + 1/N lần

- Chúng ta có thể khái quát hóa khái niệm trung bình như một tập hợp thông tin có trọng số.
- Nếu chúng ta biểu thị trọng số của cá nhân thứ i là wi , thì giá trị trung bình là:

x¯ = w1x1 + w2x2 + · · · + wnxn

= Σ (tôi = 1 … N) wixi

= w x

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
14 / 24

Cloud of Variables - Variance of a Variable

- The variance is a measure of spread around the mean. We can take the average of the squared deviations from the mean.

$$Var(X) = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Because the variance has squared unit, we need to take the square root to recover the original unit in which X is expressed. This is the standard deviation.

$$sd(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Cloud of Variables - Phư ơ ng sai của một biến

- Phư ơ ng sai là thư ớc đo độ lan truyền xung quanh giá trị trung bình. chúng ta có thể lấy trung bình của các bình phư ơ ng độ lệch so với giá trị trung bình.

V a(X) = (x1 x¯) 2 + · · · + (xn x¯) 2 / N = 1/n Σ N i=1 (xi x¯) 2

- Vì phư ơ ng sai có bình phư ơ ng đơ n vị nên chúng ta cần lấy căn bậc hai để khôi phục lại đơ n vị ban đầu mà X đư ợc biểu thị. Đây là độ lệch chuẩn.

sd(X) = √(1/n Σ N i=1 (xi x¯) 2)

Cloud of Variables - Variance of a Variable - Vector Notation

- A variable $X = (x_1, x_2, \ldots, x_n)$ can be denoted as a vector $\mathbf{x}$. The variance of a vector $\mathbf{x}$ can be computed:

$$Var(\mathbf{x}) = \frac{1}{n}(\mathbf{x} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})$$

  where $\bar{\mathbf{x}}$ is an n-element vector of mean values $\bar{x}$.

- If $\mathbf{x}$ is already mean-centered, then

$$Var(\mathbf{x}) = \frac{1}{n}\mathbf{x}^\top\mathbf{x} = \frac{1}{n}\|\mathbf{x}\|^2$$

Đám mây biến - Phư ơ ng sai của một biến - Vector ký hiệu

- Một biến X = (x1, x2, . . , xn) có thể đư ợc ký hiệu là một vectơ x. Các phư ơ ng sai của một vectơ x có thể đư ợc tính:

$$V\ a(x) = \frac{1}{N}(x \quad \bar{x})\ (x \quad \bar{x})$$

  trong đó x¯ là một vectơ n phần tử của các giá trị trung bình x¯.

- Nếu x đã là trung tâm trung bình, thì

$$V\ a(x) = \frac{1}{N} x \quad x = \frac{1}{N} \quad x^2$$

Cloud of Variables - Covariance

- The covariance generalizes the concept of variance for two variables.

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x}$ is the mean value of $\mathbf{x}$:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and $\bar{y}$ is the mean value of $\mathbf{y}$:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- If the variables are mean-centered, we have

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n}(\mathbf{x}^\top \mathbf{y})$$

Đám mây biến số - Hiệp phư ơ ng sai

- Hiệp phư ơ ng sai tổng quát hóa khái niệm phư ơ ng sai cho hai biến.

cov(x,y) = 1/n N i=1 (xi x¯)(yi y¯)

trong đó x¯ là giá trị trung bình của x:

x¯ = 1/N (x1 + x2 + · · · + xn) = 1/n N i=1 xi

và y¯ là giá trị trung bình của y:

y¯ = 1/N (y1 + y2 + · · · + y) = 1/n N i=1 vâng

- Nếu các biến đư ợc lấy trung bình làm trung tâm, chúng ta có

cov(x,y) = 1/N (x y)

Cloud of Variables - Correlation

- Covariance indicates the direction (positive or negative) of a possible linear relation, but it does not tell how big the relation is.
- Use the standard deviations of the variables to normalize the covariance.
- The coefficient of linear correlation is defined as:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{\sqrt{var(\mathbf{x})}\sqrt{var(\mathbf{y})}}$$

- If the variables are mean-centered, we have:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^{\top}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

- If both $\mathbf{x}$ and $\mathbf{y}$ are standardized, the correlation is:

$$cor(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{y}$$

Đám mây biến - Tư ơ ng quan

- Hiệp phư ơ ng sai chỉ ra hư ớng (tích cực hoặc tiêu cực) của một mối quan hệ tuyến tính có thể, như ng nó không cho biết mức độ lớn của mối quan hệ.
- Sử dụng độ lệch chuẩn của các biến để chuẩn hóa hiệp phư ơ ng sai.
- Hệ số tư ơ ng quan tuyến tính đư ợc xác định như sau:

cor(x,y) = √ cov(x,y) / var(x) √ var(y)

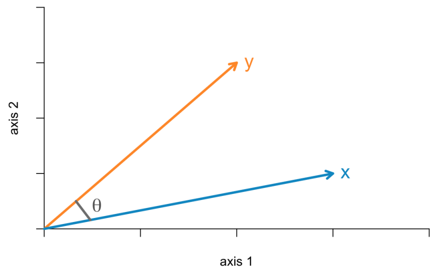
- Nếu các biến đư ợc lấy trung bình làm trung tâm, ta có:

y cor(x,y) = x / x y

- Nếu cả x và y đều đư ợc chuẩn hóa thì tư ơ ng quan là:

cor(x,y) = x y

Geometry of Correlation



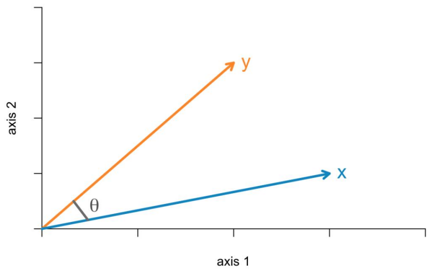
- The inner product of two mean-centered vectors:

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos(\theta_{\mathbf{x},\mathbf{y}})$$

- The correlation between mean-centered vectors $\mathbf{x}$ and $\mathbf{y}$ is the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$:

$$\cos(\theta_{\mathbf{x},\mathbf{y}}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = cor(\mathbf{x}, \mathbf{y})$$

Hình học tư ơ ng quan



- Tích trong của hai vectơ trung bình cộng:

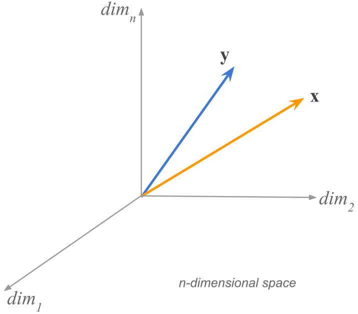
x y = x y cos (θx,y)

- Mối tư ơ ng quan giữa các vectơ trung bình x và y là cosin của góc giữa x và y:

x

y cos (θx,y) = = cor(x,y) x y

Cloud of Variables - Orthogonal Projections

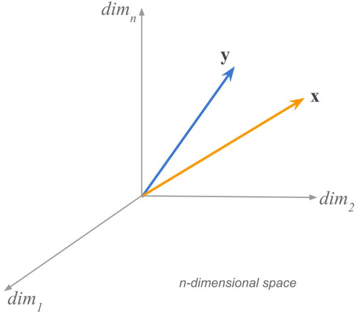


- Consider 2 variables $\mathbf{x}$ and $\mathbf{y}$. Can we approximate $\mathbf{y}$ in terms of $\mathbf{x}$?
- The approximation of $\mathbf{y}$, denoted by $\hat{\mathbf{y}}$, means finding a scalar b:

$$\hat{\mathbf{y}} = b\mathbf{x}$$

- To get $\hat{\mathbf{y}}$, we minimize the squared difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

Đám mây biến - Phép chiếu trực giao

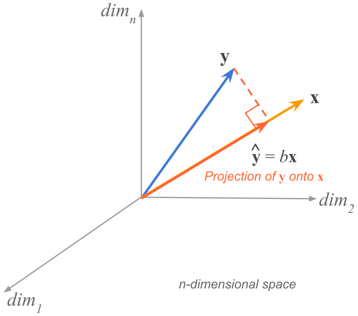


• Xét 2 biến x và y. Chúng ta có thể xấp xỉ y theo x không? • Xấp xỉ của y, ký hiệu là yˆ, có nghĩa là tìm một số vô hư ớng b:

yˆ = bx

• Để có đư ợc yˆ, chúng ta cực tiểu hóa bình phư ơ ng chênh lệch giữa y và yˆ.

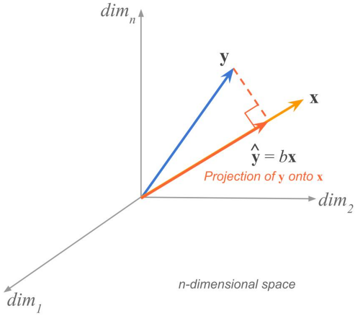
Cloud of Variables - Orthogonal Projection



- We project $\mathbf{y}$ orthogonally onto $\mathbf{x}$:

$$\hat{\mathbf{y}} = b\mathbf{x} = \mathbf{x}\left(\frac{\mathbf{y}^\top\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}\right) = \mathbf{x}\left(\frac{\mathbf{y}^\top\mathbf{x}}{\|\mathbf{x}\|^2}\right)$$
$$= \mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{y}$$

Đám mây biến - Phép chiếu trực giao

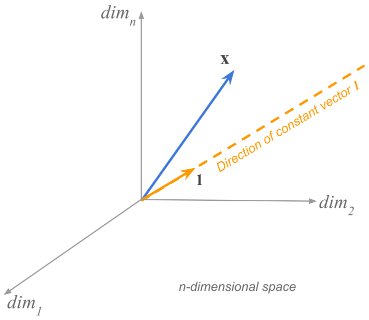


- Ta chiếu y trực giao lên x:

y^ = bx = x(yx x) = x(y x 2)

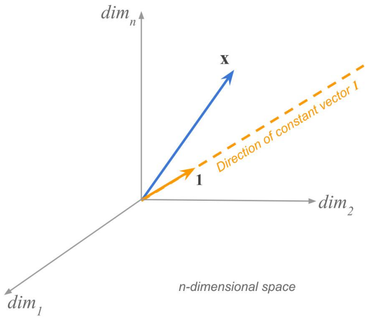
= x(x x) 1x y

Cloud of Variables - The Mean as an Orthogonal Projection



- A variable $X = (x_1, x_2, \ldots, x_n)$ can be denoted as a vector $\mathbf{x}$ in an n-dimensional space.
- Consider the constant vector $\mathbf{1} = (1, 1, \ldots, 1)$.
- What is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{1}$?

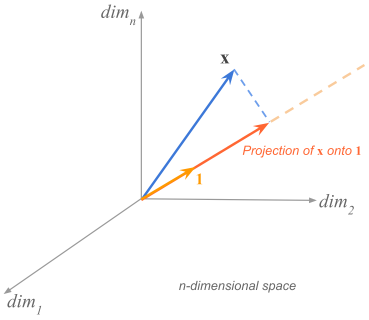
Đám mây biến - Giá trị trung bình như một phép chiếu trực giao



- Một biến X = (x1, x2, . . . , xn) có thể đư ợc ký hiệu là một vectơ x trong không gian n chiều.
- Xét vectơ hằng số 1 = (1, 1, . . . , 1). • Hình

chiếu trực giao của x lên 1 là gì?

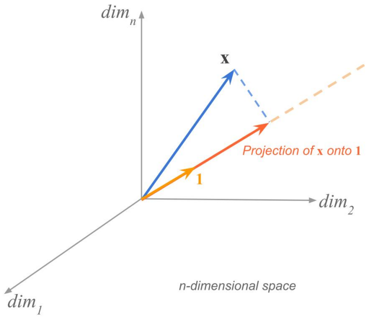
Cloud of Variables - The Mean as an Orthogonal Projection



- The projection is:

$$\hat{\mathbf{x}} = b\mathbf{1} = \mathbf{1}\left(\frac{\mathbf{x}^\top\mathbf{1}}{\mathbf{1}^\top\mathbf{1}}\right) = \mathbf{1}\left(\frac{\mathbf{x}^\top\mathbf{1}}{\|\mathbf{1}\|^2}\right)$$
$$= \left(\frac{x_1\cdot 1 + x_2\cdot 1 + \ldots + x_n \cdot 1}{1\cdot 1 + 1\cdot 1 + \ldots + 1\cdot 1}\right)\mathbf{1} = \frac{x_1 + x_2 + \ldots + x_n}{n}\mathbf{1}$$
$$= \bar{x}\mathbf{1}$$

Đám mây biến - Giá trị trung bình như một phép chiếu trực giao



- Hình chiếu là:

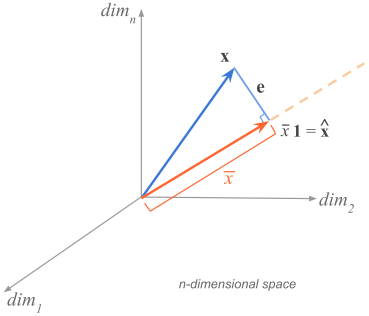
x ˆ = b1 = 1 (x 1 1 1) = 1 (x 11

· · · + xn 1 x1 + x2 + · · · + xn 1

2) = (1x1 1 +1 1+ x21 + .1. + + 1 1) 1 = N

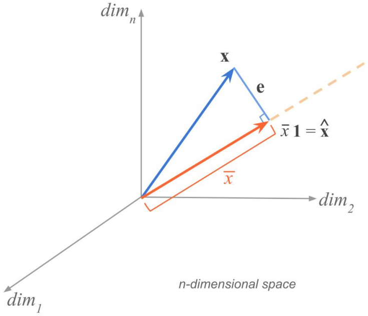
= x ¯1

Cloud of Variables - The Mean as an Orthogonal Projection



- The mean of the variable X, denoted by $\bar{x}$, is the scalar we multiply with $\mathbf{1}$ to obtain the vector projection $\hat{\mathbf{x}}$.

Đám mây biến - Giá trị trung bình như một phép chiếu trực giao



- Giá trị trung bình của biến X, ký hiệu là x¯, là đại lư ợng vô hư ớng ta nhân với 1 để đư ợc phép chiếu véc tơ x^.

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
24 / 24