GIẢM CHIỀU DỮ LIỆU

Giới thiệu

- Dữ liệu lớn đến từ rất nhiều nguồn và ba nguồn chính là:

  (1) các phương tiện truyền thông xã hội

  (2) các máy móc thu nhận dữ liệu, các thiết bị công nghiệp, các cảm biến (sensors), các dụng cụ giám sát...

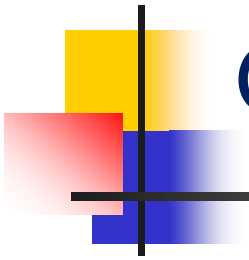
  (3) giao dịch kinh doanh, từ số liệu giá cả sản phẩm, thanh toán, dữ liệu chế tạo và phân bố...

Ba chìa khóa chính của khai thác dữ liệu lớn luôn được xem là:

(1) quản trị dữ liệu, tức lưu trữ, bảo trì và truy nhập các nguồn dữ liệu lớn;

(2) phân tích dữ liệu, tức tìm cách hiểu được dữ liệu và tìm ra các thông tin hoặc tri thức quý báu từ dữ liệu;

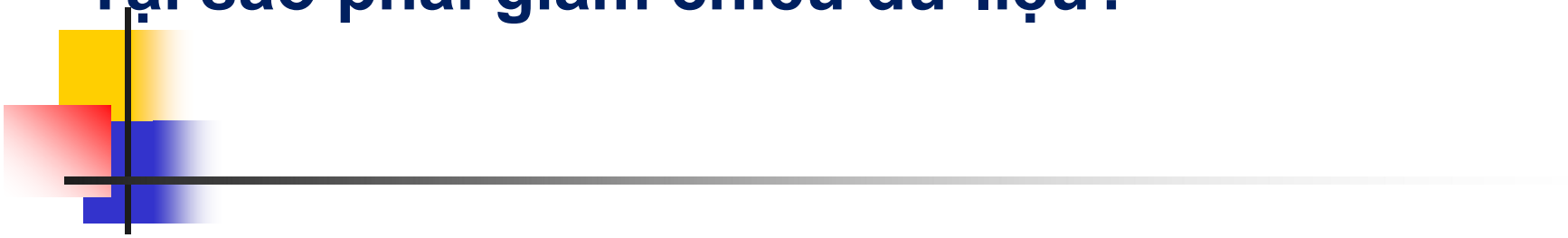
(3) hiển thị dữ liệu và kết quả phân tích dữ liệu. Phát triển công cụ quản trị dữ liệu lớn, nghiên cứu về các kỹ thuật hiển thị dữ liệu lớn, về mối quan hệ phức tạp trong chúng, là những thách thức không nhỏ, nhưng thách thức chính của dữ liệu lớn là các *phương pháp phân tích dữ liệu*.



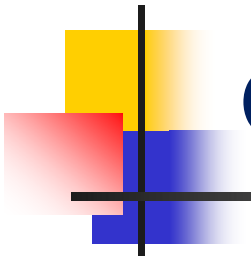
Giới thiệu

- Tuy nhiên, một kho (tập) dữ liệu lớn có thể chứa lượng dữ liệu lên đến terabytes
- Sự gia tăng của các tập dữ liệu lớn trong nhiều lĩnh vực đặt ra thách thức cho khai phá dữ liệu.
- Không chỉ tập dữ liệu lớn mà còn các kiểu dữ liệu mới:
  - Data stream trên web
  - Mạng xã hội
  - Hệ thống sinh học

Tại sao phải giảm chiều dữ liệu?



- Năm 1997 một số lĩnh vực sử dụng hơn 40 thuộc tính đặc trưng
- Năm 2003 hầu hết các bài báo cho thấy các lĩnh vực đã sử dụng 10^2 tới 10^4 biến (variable)
- Các kĩ thuật học máy và khai phá dữ liệu có thể không hiệu quả với dữ liệu có số chiều lớn
- Giảm chiều dữ liệu:
  - ✓ Là việc làm giảm chiều của không gian tìm kiếm dữ liệu
  - ✓ Giảm chi phí thu thập và lưu trữ dữ liệu
  - ✓ Nâng cao hiệu quả của việc khai phá dữ liệu
  - ✓ Làm đơn giản hóa các kết quả khai phá dữ liệu



Giảm chiều dữ liệu

- Giảm chiều dữ liệu là gì?
- Kỹ thuật lồi trong giảm chiều dữ liệu
- kỹ thuật không lồi cho giảm chiều

Giảm chiều dữ liệu

1. Giảm chiều dữ liệu là gì ?

- *Tại sao cần phải giảm bớt dữ liệu?*

Một kho (tập) dữ liệu lớn có thể chứa lượng dữ liệu lên đến terabytes. Do đó quá trình khai phá dữ liệu có thể sẽ chạy rất lâu (rất mất thời gian) đối với toàn bộ tập dữ liệu.

- ✓ Ảnh hưởng tiêu cực của số chiều (số thuộc tính) lớn:

  - Khi số chiều tăng, dữ liệu trở nên thưa hơn

  - Mật độ và khoảng cách giữa các điểm (quan trọng đối với việc phân cụm, phát hiện ngoại lai) trở nên ít có ý nghĩa;

Giảm chiều dữ liệu

- *Giảm bớt dữ liệu* (data reduction): để thu được một biểu diễn thu gọn (giảm bớt) nhưng vẫn sinh ra cùng (hoặc xấp xỉ) các kết quả phân tích (khai phá) như với tập dữ liệu ban đầu
- *Các chiến lược giảm bớt dữ liệu:*
  - Giảm số chiều (dimensionality reduction): loại bỏ bớt các thuộc tính không (ít) quan trọng;
  - Giảm lượng dữ liệu (data/numerosity reduction)
    - kết hợp khối dữ liệu (data cube aggregation)
    - nén dữ liệu (data compression)
    - hồi quy (regression)
    - rời rạc hóa (discretization)

Tiến trình

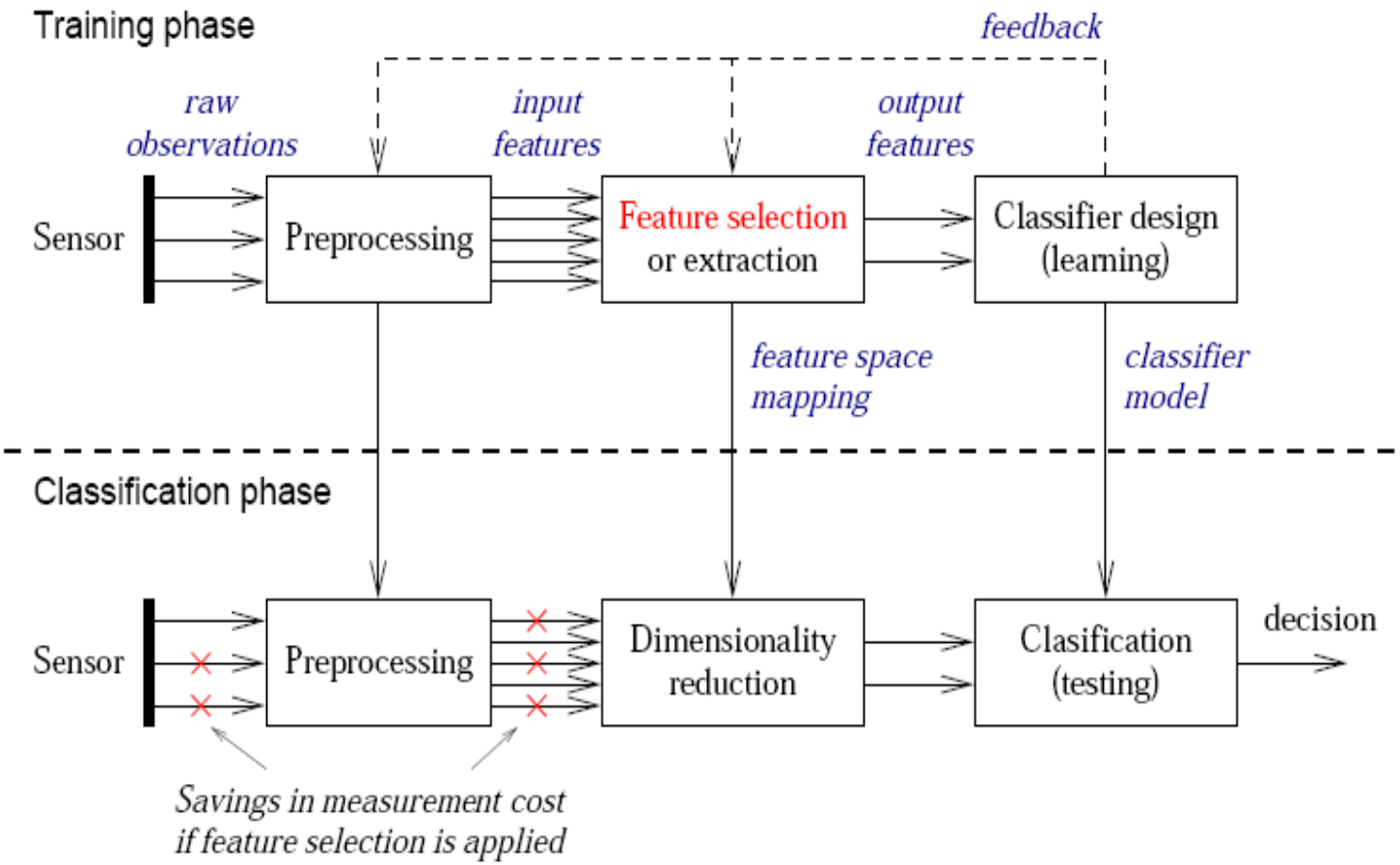
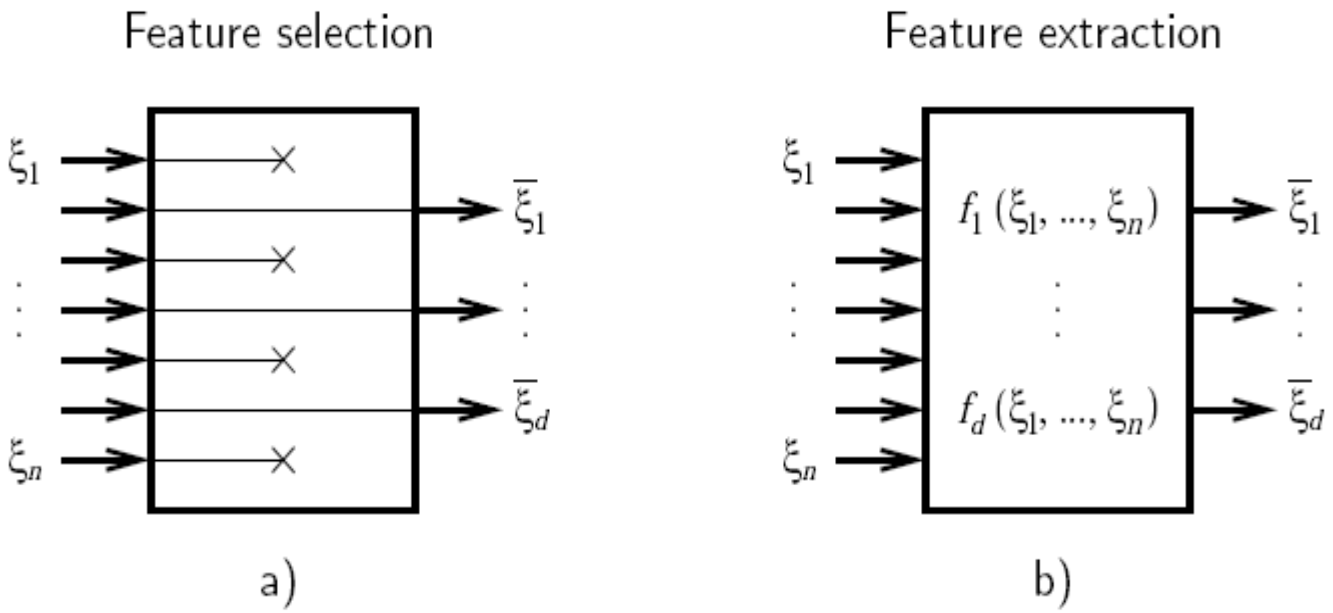


Figure 1.1: A block diagram of a pattern recognition system.

Phương pháp giảm chiều dữ liệu

Để giảm chiều:

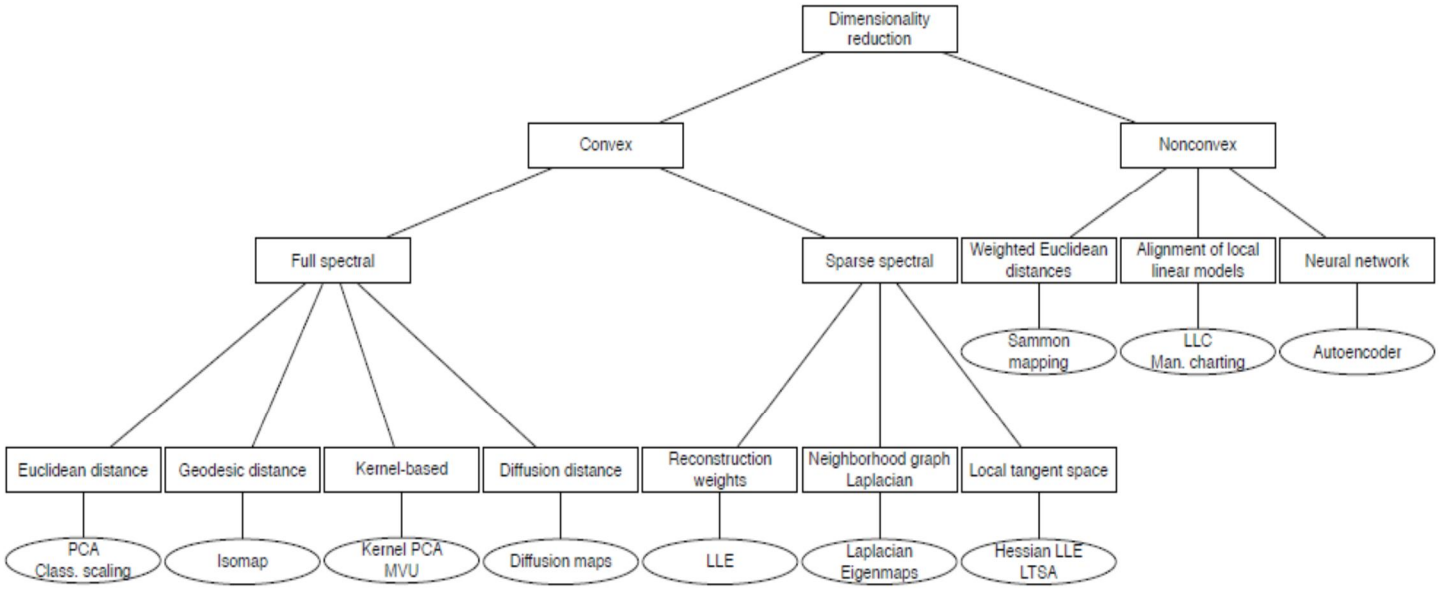
- Lựa chọn đặc trưng (Feature Selection)
- Trích chọn đặc trưng (Feature Extraction)

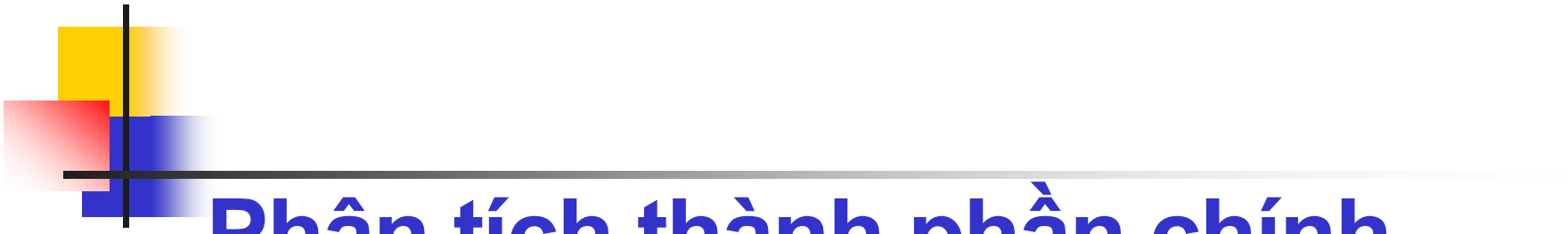


Kỹ thuật giảm chiều dữ liệu

- Kỹ thuật tuyến tính
- Kỹ thuật phi tuyến

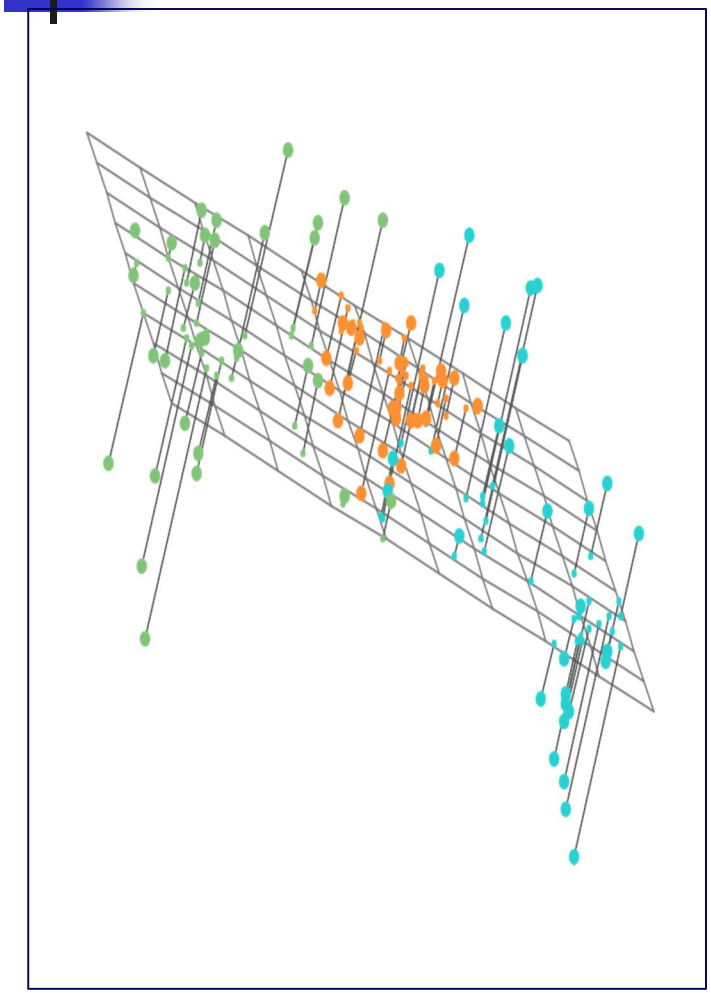
Có rất nhiều kỹ thuật giảm chiều dữ liệu, chia nhỏ kỹ thuật giảm chiều thành 2 nhóm chính: kỹ thuật lồi và kỹ thuật không lồi

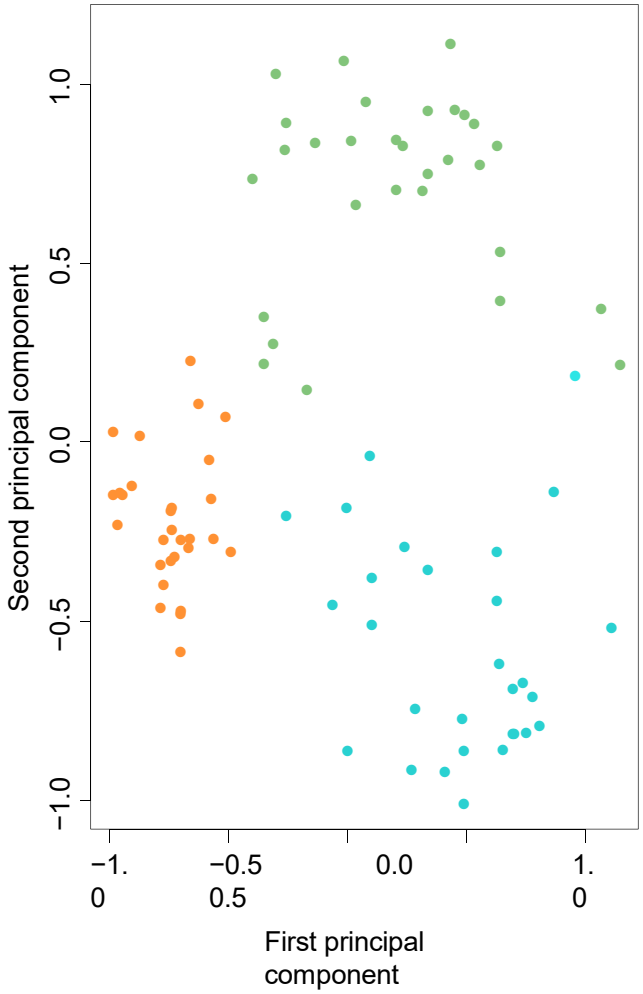




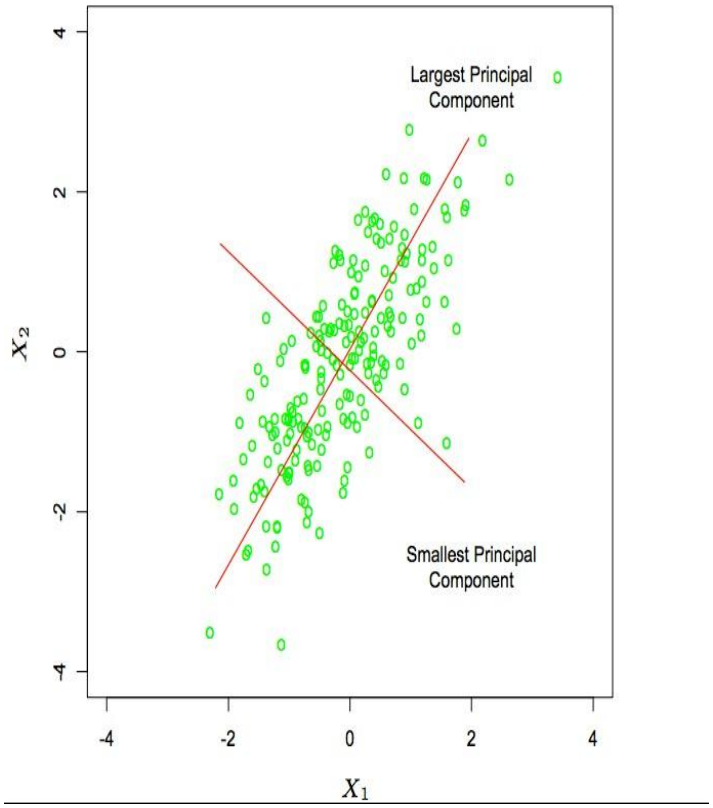
Phân tích thành phần chính
principal component analysis (PCA)

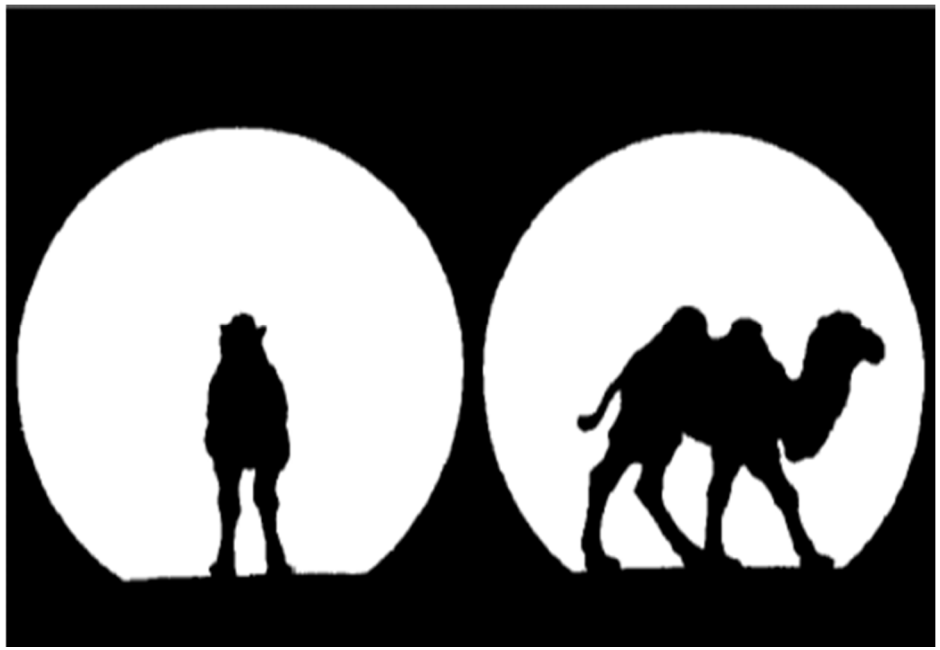
Giảm chiều dữ liệu

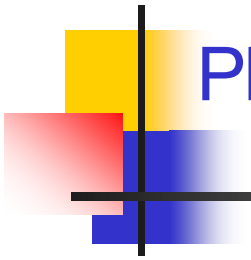




Phép chiếu


Largest Principal Component
Smallest Principal Component
X_1
X_2
-4
-2
0
2
4



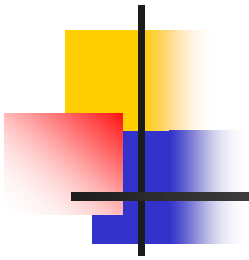


Phân tích thành phần chính là gì

- Đề xướng bởi Pearson (1901) và Hotelling (1933)
- "PCA là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn (2 hoặc 3 chiều) nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu"
- "Phân tích thành phần chính là một phương pháp trích xuất các biến quan trọng (dưới dạng các thành phần) từ một tập hợp lớn các biến có sẵn trong một tập dữ liệu với mục đích thu thập càng nhiều thông tin càng tốt"

Ưu điểm của PCA

- Giảm số chiều của không gian chứa dữ liệu khi nó có số chiều lớn, không thể thể hiện trong không gian 2 hay 3 chiều.
- Xây dựng những trục tọa độ mới, thay vì giữ lại các trục của không gian cũ, nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương, và đảm bảo độ biến thiên của dữ liệu trên mỗi chiều mới.
- Tạo điều kiện để các liên kết tiềm ẩn của dữ liệu có thể được khám phá trong không gian mới, mà nếu đặt trong không gian cũ thì khó phát hiện vì những liên kết này không thể hiện rõ.
- Đảm bảo các trục tọa độ trong không gian mới luôn trực giao đôi một với nhau, mặc dù trong không gian ban đầu các trục có thể không trực giao.



Đặc điểm của PCA

- Một số ứng dụng của PCA bao gồm nén dữ liệu (đặc biệt là dữ liệu ảnh), đơn giản hóa dữ liệu để dễ dàng học tập, hình dung.
- Lưu ý rằng kiến thức miền là rất quan trọng trong khi lựa chọn có nên tiếp tục với PCA hay không?
- PCA hữu ích hơn khi xử lý dữ liệu 3 chiều trở lên.
- PCA không phù hợp trong trường hợp dữ liệu bị nhiễu (tất cả các thành phần của PCA đều có độ biến thiên khá cao)

Phân tích thành phần chính

- Nếu hai biến (hay 2 items) có tương quan với nhau
  - Chúng có thể phản ảnh một hiện tượng tiềm ẩn (hay một yếu tố không quan sát được – latent factor)
  - Nếu chúng phản ảnh một latent variable, thì tổng hợp chúng thành 1 biến là hợp lí
- Các biến ẩn (latent variables) còn gọi là "**factors**" hay "**principal components**"

Phân tích thành phần chính

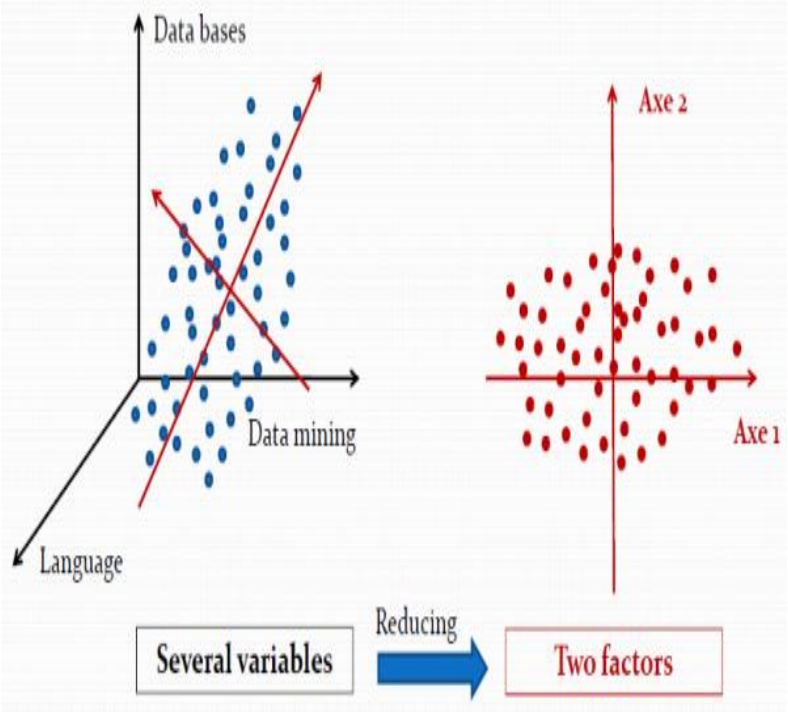
- Trong không gian mới, các liên kết tiềm ẩn của dữ liệu có thể được khám phá
- Ví dụ: Thị trường ta quan tâm có hàng ngàn mã cổ phiếu làm cách nào để khi quan sát dữ liệu từ hàng ngàn cổ phiếu này ta hình dung được xu hướng của toàn thị trường…

Phân tích thành phần chính

Minh họa PCA: phép chiếu lên các trục tọa độ khác nhau có thể cho cách nhìn rất khác nhau về cùng một dữ liệu.

Phân tích thành phần chính

- Giả sử tập dữ liệu ban đầu (tập điểm màu xanh) được quan sát trong không gian 3 chiều (trục màu đen) như hình bên trái.
- Rõ ràng 3 trục này không biểu diễn được tốt nhất mức độ biến thiên của dữ liệu.
- PCA do đó sẽ tìm hệ trục tọa độ mới (là hệ trục màu đỏ trong hình bên trái).
- Sau khi tìm được không gian mới, dữ liệu sẽ được chuyển sang không gian này để được biểu diễn như trong hình bên phải.
- Rõ ràng hình bên phải chỉ cần 2 trục tọa độ nhưng biểu diễn tốt hơn độ biến thiên của dữ liệu so với hệ trục 3 chiều ban đầu.



Bối cảnh và dữ liệu

- Bối cảnh: Có ma trận dữ liệu gồm *n* hàng và *p* biến số

$$[X_1, X_2, .., X_P]$$

- PCA tìm cách hoán chuyển các X_i thành p biến mới (y_i) nhưng không có liên quan với nhau!

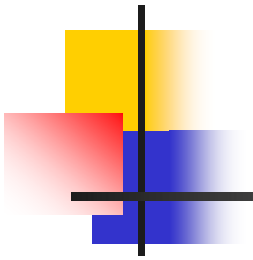
Các biến liên quan đến nhau?

- Cách đơn giản nhất là chỉ lưu lại 1 biến duy nhất (bỏ các biến còn lại) – không hợp lí!
- Cho trọng số mỗi biến. Trọng số nào?
- Tiêu chuẩn nào?
- Tìm phương pháp hoán chuyển ma trận **X** (n xp) sao cho

$$Y = \delta^T \mathbf{X} = \delta_1 X_1 + \delta_2 X_2 + .. + \delta_p X_p$$

trong đó: $\delta = (\delta_1, \delta_2, .., \delta_p)^T$ là cột vector trọng số sao cho:

$$\delta_1^2 + \delta_2^2 + ... + \delta_p^2 = 1$$



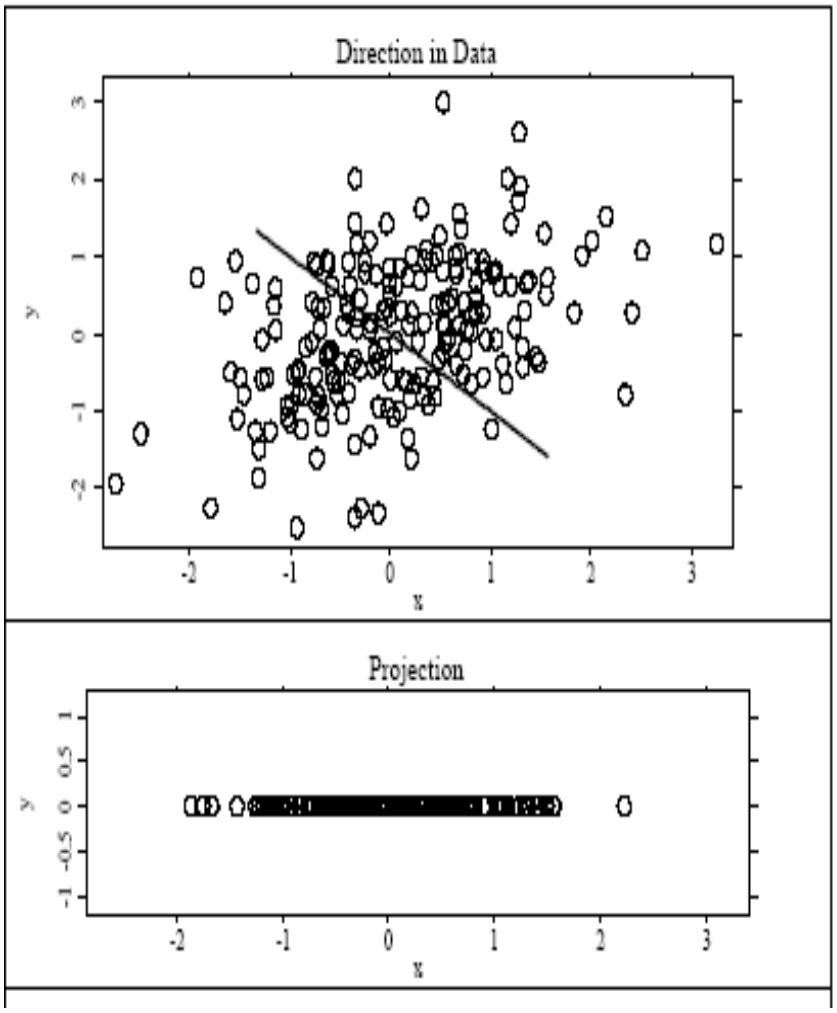
Tiêu chuẩn

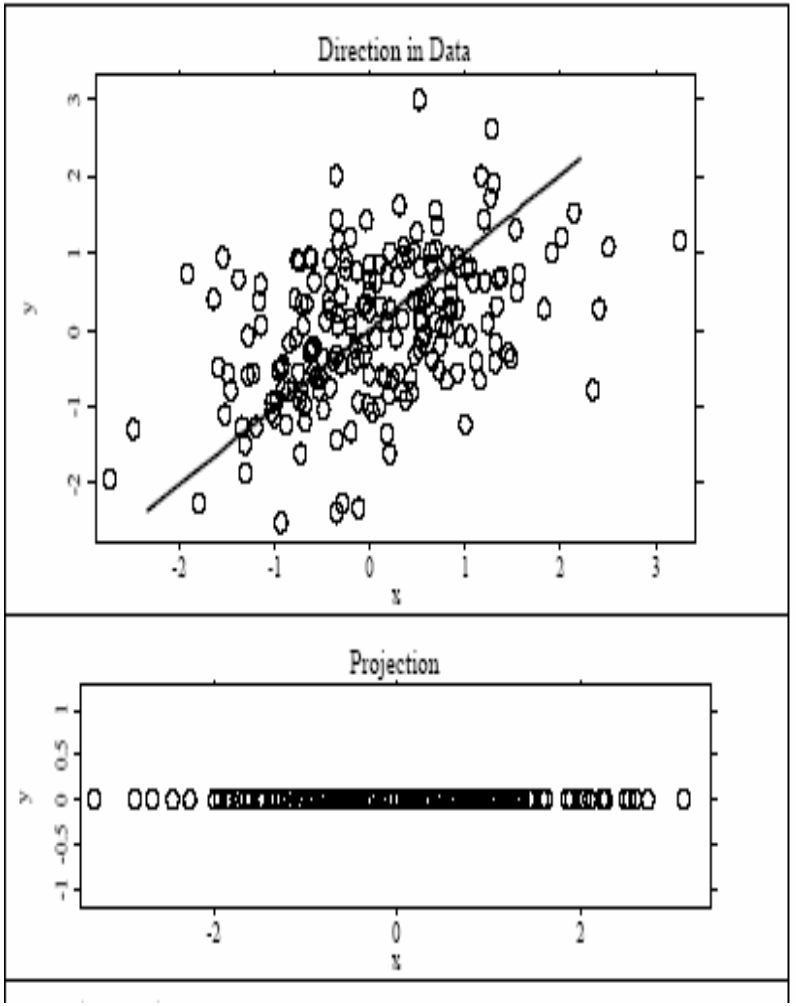
- Tối đa hoá phương sai của dữ liệu dựa trên các biến Y
- Tìm δ sao cho

  $Var(\delta^T X) = \delta^T Var(X)\, \delta$ tối đa

Ma trận C = Var(X) là hiệp biến (covariance) của các biến X_i

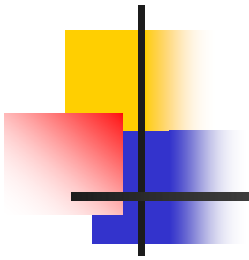
Thử tưởng tượng ...





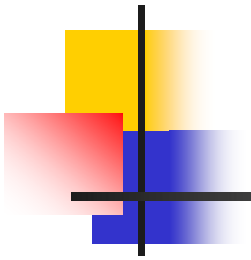
Variance-covariance matrix

$$\begin{pmatrix} v(x_1)\, c(x_1,x_2) \ldots\ldots c(x_1,x_p) \\ c(x_1,x_2)\, v(x_2) \ldots\ldots c(x_2,x_p) \\ \\ c(x_1,x_p)\, c(x_2,x_p) \ldots\ldots v(x_p) \end{pmatrix}$$



Có nghĩa là chúng ta tìm ...

- Hướng của δ được xác định bởi véc tơ riêng (vector eigen) γ_1 tương đương với giá trị riêng (eigen) lớn nhất của ma trận C.
- Vector thứ 2 cũng trực giao (orthogonal, tức không liên quan) với vector thứ 1.
- v.v.



PCA cung cấp

- Một nhóm biến mới (Y_i) là hàm số tuyến tính của các biến X_i:

$$Y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots a_{ip}x_p \; ; \; i = 1,2, \ldots p$$

- Biến mới Y_i được tạo ra theo mức độ quan trọng nhưng suy giảm theo độ quan trọng.
- Chúng được gọi là "principal components"

Tính toán eigenvalue và eigenvector

- Giá trị riêng λ_i được xác định bằng cách giải phương trình

$$Det(C-\lambda i) = 0$$

- Véc tơ riêng là những cột của ma trận A với đặc điểm

$$C = A\ D\ A^T$$

Trong đó

$$D = \begin{pmatrix} \lambda 1 & 0 & \ldots\ldots & 0 \\ 0 & \lambda_2 & \ldots\ldots & 0 \\ 0 & & & \\ 0 & \ldots\ldots & \ldots\ldots & \lambda_p \end{pmatrix}$$

Diễn giải PCA

- Những biến mới (Principal Components) có phương sai bằng giá trị riêng:

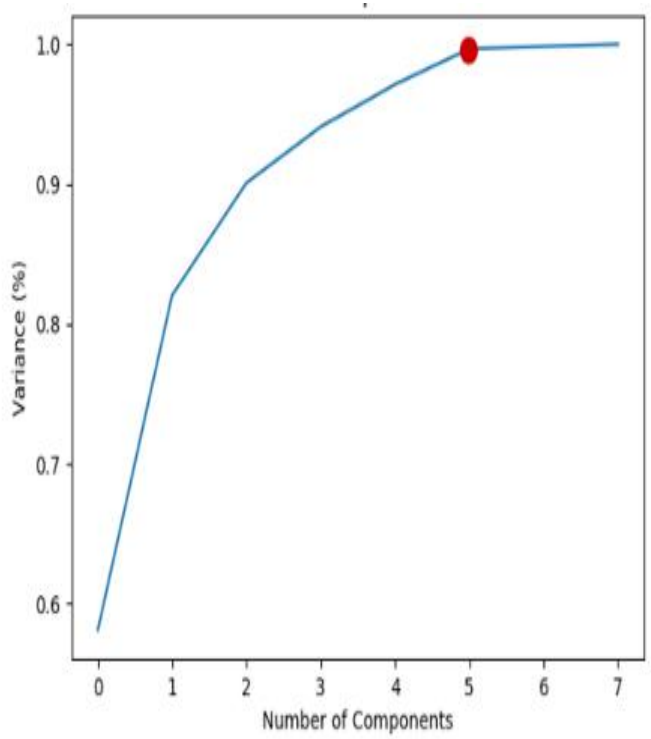
$$Var(Y_i) = \lambda_i \text{ cho tất cả } i = 1,2,\ldots,p$$

- Giá trị λ_i nhỏ $\Leftrightarrow$ phương sai thấp $\Leftrightarrow$ dữ liệu thay đổi nhỏ về hướng của Y_i

- Mức độ quan trọng của mỗi PC cho bởi $\lambda_i / \Sigma \lambda_i$

Cần bao nhiêu PC?

- Số PCs sao cho tỉ lệ phương sai giải thích được >90%
- Kaiser criterion: giữ PCs với eigenvalues >1
- Scree plot: thể hiện khả năng PC giải thích phương sai của dữ liệu



Số mẫu: 17898
Số biến: 9
Số PC = 5, khả năng giải thích phương sai xấp xỉ 99%

Diễn giải ý nghĩa của PC

- Trọng số của biến số trong mỗi PC:
- Nếu $Y_1 = 0.89\,X_1 + 0.15X_2 - 0.77X_3 + 0.51X_4$
- Thì X_1 và X_3 có trọng số cao nhất và là biến quan trọng nhất
- Xem mối tương quan giữa các biến X_i và PC

Các bước phân tích PCA

1. Chuẩn bị dữ liệu (chuẩn hoá dữ liệu)
2. Tính ma trận covariance hoặc correlation
3. Tính eigenvalues của ma trận covariance
4. Chọn components

Tổng kết

- PCA là một phương pháp giảm độ liên quan đa chiều của dữ liệu
- Ý tưởng: tìm các biến số mới (PC) là hàm số của các biến số gốc sao cho các PC không tương quan với nhau (orthogonal)
- PC đầu tiên giải thích nhiều phương sai nhất; PC 2 giải thích tỉ lệ phương sai ít hơn PC 1, v.v.