***PHƯƠNG PHÁP PCA (PRINCIPAL COMPONENT ANALYSIS)***

1. Mở đầu:

Đây là một phương pháp được sử dụng thường xuyên khi các nhà phân tích thống kê phải làm việc với những bộ số liệu với số chiều lớn (big data). Vậy làm thế nào để giảm thiểu chiều dữ liệu mà vẫn không mất đi thông tin và giữ lại được những thông tin cần thiết cho việc xây dựng các mô hình ?

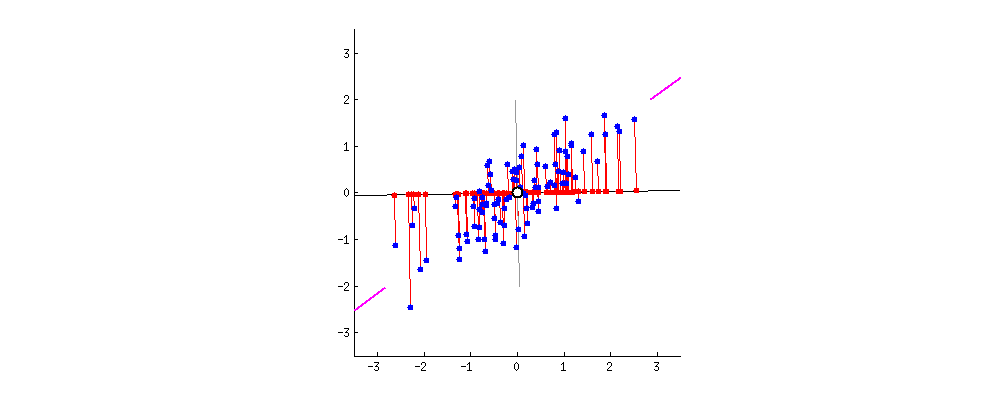


Cùng là 1 chú lạc đà, tuy nhiên với các cách nhìn khác nhau, chúng ta lại có những cách thu nhận thông tin khác nhau và cho ta những kết luận khác nhau.

1. Phân tích thành phần chính:
2. Khái niệm:

PCA là phương pháp biến đổi giúp giảm số lượng lớn các biến có tương quan với nhau thành tập ít các biến sao cho các biến mới tạo ra là tổ hợp tuyến tính của những biến cũ không có tương quan tuyến tính.

Ví dụ: chúng ta có 100 biến ban đầu có tương quan tuyến tính với nhau, khi đó chúng ta sử dụng phương pháp PCA xoay chiều không gian cũ thành chiều không gian mới mà ở đó chỉ còn 5 biến không có tương quan tuyến tính mà vẫn dữ được nhiều nhất lượng thông tin từ nhóm biến ban đầu.



1. Đặc tính PCA:

Một số đặc tính của PCA được kể đến như:

* Giúp giảm số chiều dữ liệu - Giúp visualization khi dữ liệu có quá nhiều chiều thông tin.
* Do dữ liệu ban đầu có số chiều lớn (nhiều biến) thì PCA giúp chúng ta xoay trục tọa độ để xây một trục tọa độ mới đảm bảo độ biến thiên của dữ liệu và giữ lại được nhiều thông tin nhất mà không ảnh hưởng tới chất lượng của các mô hình dự báo. (Maximize the variability).
* Do PCA giúp tạo 1 hệ trục tọa độ mới nên về mặt ý nghĩa toán học, PCA giúp chúng ta xây dựng những biến factor mới là tổ hợp tuyến tính của những biến ban đầu.
* Trong không gian mới, có thẻ giúp chúng ta khám phá thêm những thông tin quý giá mới khi mà tại chiều thông tin cũ những thông tin quý giá này bị che mất.

1. Mô hình PCA:

Xét tập không gian (dữ liệu) k biến, k biến này được biểu qua j thành phần chính sao cho (j < k). Xét thành phần chính đầu tiên có dạng:

= + + +... +

Thành phần chính đầu tiên chứa đựng hầu hết thông tin từ k biến ban đầu (được hình thành là 1 tổ hợp tuyến tính của các biến ban đầu) và lúc này tiếp tục xét thành phần chính thứ 2 được biểu diễn tuyến tính từ k biến ban đầu tuy nhiên thành phần chính thứ 2 phải không trực giao với thành phần chính ban đầu (hay thành phần chính thứ 2 không có mối tương quan tuyến tính với thành phần chính đầu tiên). Về lý thuyết chúng ta có thể xây dựng nhiều thành phần chính từ nhiều biến ban đầu. Tuy nhiên chúng ta cần tìm được trục không gian sao cho ít thành phần nhất mà có thể biểu diễn được hầu hết thông tin từ những biến ban đầu.

1. Ý tưởng chính của PCA:

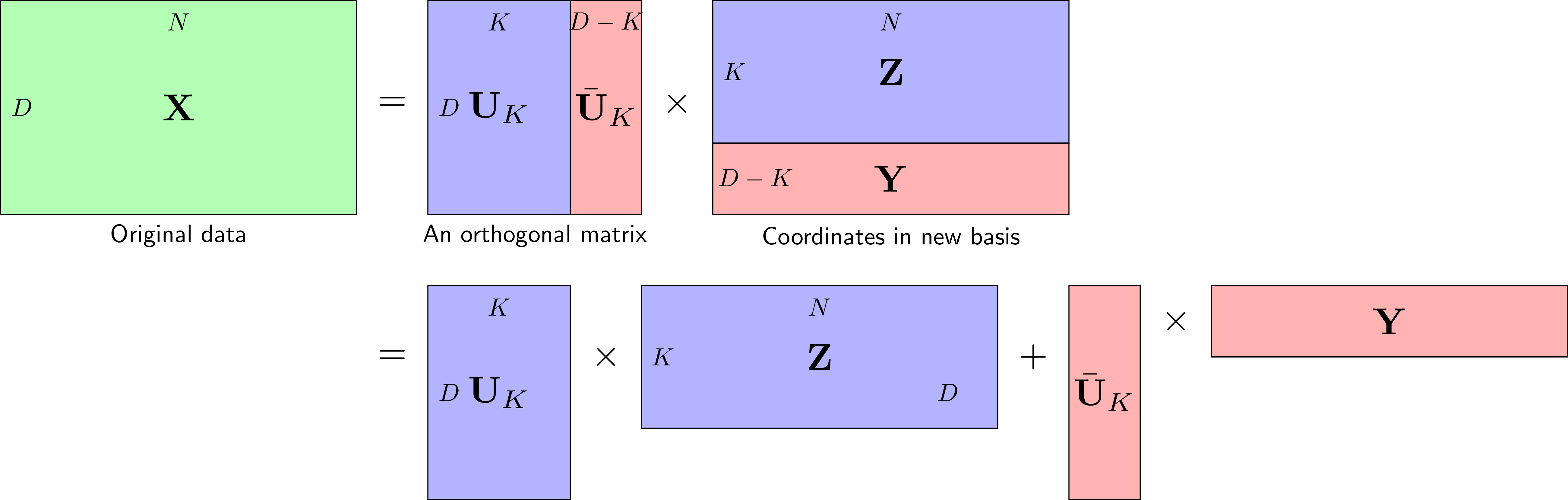
Cách đơn giản nhất để giảm chiều dữ liệu từ D về K<D là chỉ giữ lại K phần tử quan trọng nhất. Tuy nhiên, việc làm này chắc chắn chưa phải tốt nhất vì chúng ta chưa biết xác định thành phần nào là quan trọng hơn. Hoặc trong trường hợp xấu nhất, lượng thông tin mà mỗi thành phần mang là như nhau, bỏ đi thành phần nào cũng dẫn đến việc mất một lượng thông tin lớn.

Tuy nhiên, nếu chúng ta có thể biểu diễn các vector dữ liệu ban đầu trong một hệ cơ sở mới mà trong hệ cơ sở mới đó, tầm quan trọng giữa các thành phần là khác nhau rõ rệt, thì chúng ta có thể bỏ qua những thành phần ít quan trọng nhất.

Lấy một ví dụ về việc có hai camera đặt dùng để chụp một con người, một camera đặt phía trước người và một camera đặt trên đầu. Rõ ràng là hình ảnh thu được từ camera đặt phía trước người mang nhiều thông tin hơn so với hình ảnh nhìn từ phía trên đầu. Vì vậy, bức ảnh chụp từ phía trên đầu có thể được bỏ qua mà không có quá nhiều thông tin về hình dáng của người đó bị mất.

PCA chính là phương pháp đi tìm một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin. Và để cho đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới.

Giả sử hệ cơ sở trực chuẩn mới là U và chúng ta muốn giữ lại K toạ độ trong hệ cơ sở mới này. Không mất tính tổng quát, giả sử đó là K thành phần đầu tiên. Quan sát hình dưới đây:



U = [] là một hệ trực chuẩn với là ma trận con tạo bởi K cột đầu tiên của U. Với cơ sở mới này, ma trận dữ liệu có thể được viết thành: X =

Từ đây ta cũng suy ra:

= => Z =

Mục đích của PCA là đi tìm ma trận trực giao U sao cho phần lớn thông tin được giữ lại ở phần màu xanh  và phần màu đỏ  sẽ được lược bỏ và thay bằng một ma trận không phụ thuộc vào từng điểm dữ liệu.

1. Các bước thực hiện PCA

Từ các suy luận phía trên, ta có thể tóm tắt lại các bước trong PCA như sau:

1. Tính vector kỳ vọng của toàn bộ dữ liệu:

=

1. Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu:

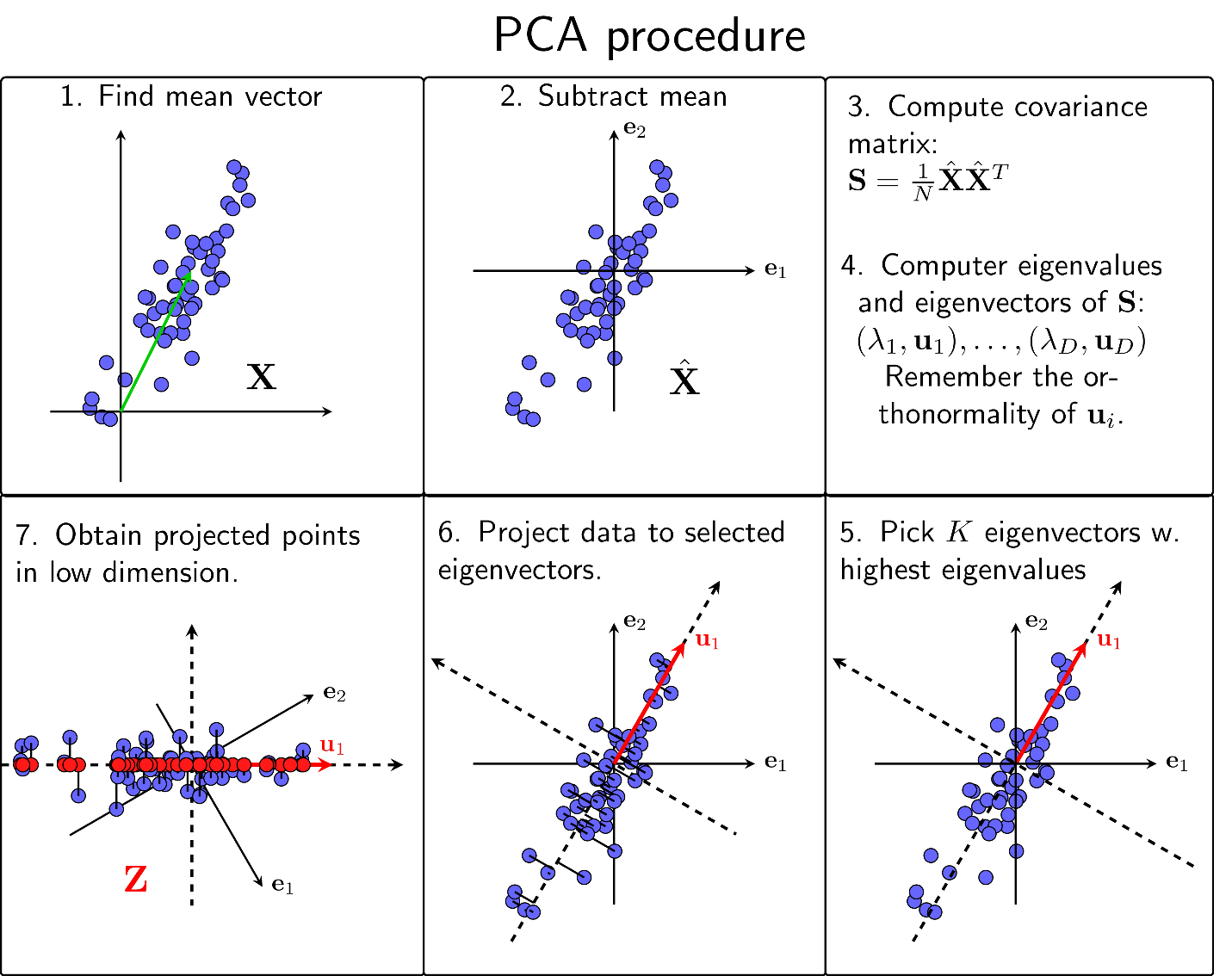
= -

1. Tính ma trận hiệp phương sai:

S =

1. Tính các trị riêng và vector riêng có norm bằng 1 của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
2. Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận  có các cột tạo thành một hệ trực giao. K vectors này, còn được gọi là các thành phần chính, tạo thành một không gian con *gần* với phân bố của dữ liệu ban đầu đã chuẩn hoá.
3. Chiếu dữ liệu ban đầu đã chuẩn hoá  xuống không gian con tìm được.
4. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới.

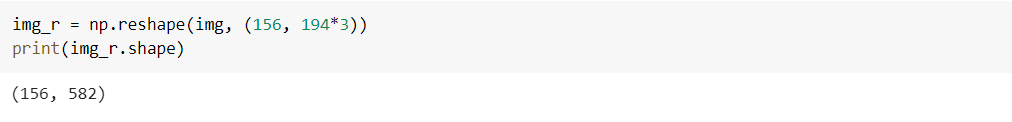
Z =

0

1. Code demo và ứng dụng:
2. Import thư viện và đọc dữ liệu:



1. Thay đổi kích thước ảnh:



1. Thực hiện PCA nén ảnh:

|  |  |  |
| --- | --- | --- |
| PCA(12) | PCA(64) | PCA(156) |
|  |  |  |