# **PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH** **(Principal Component Analysis – PCA)**

**MỘT SỐ KHÁI NIỆM CƠ BẢN**

* Phân tích thành phần chính là một trong những phương pháp phân tích thống kê đa biến cổ điển với mục đích thu gọn số chiều của tập dữ liệu quan sát mà không mất nhiều thông tin, cụ thể là thay các tập biến ban đầu bằng một tập biến nhỏ hơn gọi là thành phần chính và được sử dụng cho các quá trình phân tích kế tiếp.

Trong phương pháp này công cụ làm việc chủ yếu là ma trận hiệp phương sai (*ma trận hiệp phương sai là ma trận các momen tương quan* của các biến và các véctơ riêng, giá trị riêng tương ứng của ma trận

Với bảng số liệu cồng kềnh phức tạp khó tổng hợp và khó thấy các quan hệ thông tin chứa trong đó, vì vậy ta dùng phân tích thành phần chính để rút ra thông tin chủ yếu bằng cách xây dựng một biểu diễn đơn giản hơn sao cho trong biễu diễn đó tập dữ liệu thể hiện rõ nhất mà thông tin ít bị sai lạc nhất.

* **Trục chính và phương pháp tìm trục chính**

Với n cá thể thuộc , hai cá thể trong đó được gọi là gần nhau nếu p tọa độ của chúng gần nhau, về ý nghĩa hình học đơn giản là 2 cá thể gần nhau nếu p tọa độ của chúng gần nhau. Bài toán tìm trục chính tương đương với bài toán với các điều kiện sau:

* Tịnh tiến gốc về trọng tâm đám mây dữ liệu (tập dữ liệu).
* Quay hệ trục tọa độ sao cho trong hệ trục tọa độ mới đám mây dữ liệu thể hiện rõ nhất, nghĩa là tổng khoảng cách từ dữ liệu điểm đến chúng nhỏ nhất.
* Trong hệ trục mới chỉ cần giữ lại k trục chính (k < p), như vậy ta được siêu phẳng k chiều thể hiện rõ đám mây dữ liệu.

Cụ thể ta xét các biến ngẫu nhiên có trung bình là và ma trận hiệp phương sai Ta dùng thuật toán PCA thay thế các biến ban đầu (có tương quan) bởi một số ít các tổ hợp tuyến tính (không tương quan) với k < p.

Ta lập các tổ hợp tuyến tính đó như sau:

Trong đó véctơ hệ số với i=1,2,…,k được chọn sao cho:

* K thành phần chiếu đầu tiên của **X** (i=1,2,…,k) được xếp theo thứ tự có phương sai var( giảm dần:
* không tương quan với các , với j < i.

Ta gọi là các thành phần chính.

Các véctơ hệ số được chọn sao cho phương sai của các biến mới tạo ra xếp theo thứ tự giảm dần với ràng buộc: đồng thời chúng không tương quan với các biến tạo ra trước đó, nghĩa là ta sẽ có:

* Thành phần chính đầu tiên thu được từ véctơ hệ số, là tổ hợp tuyến tính làm cực đại phương sai của với Ta sẽ có:

.

Ta cần tìm các hệ số của sao cho đạt cực đại với điều kiện

Dùng phương pháp nhân tử Lagranger ta thu được:

Lấy đạo hàm theo và cho phương trình bằng 0, ta được:

⇒ .

Vậy nếu khác véctơ không thì phải thỏa mãn phương trình:

Điều đó có nghĩa là và là giá trị riêng và véctơ riêng của Mặt khác ta cần xác định giá trị cực đại của:

Nên ta phải chọn là trị riêng lớn nhất của ma trận **Σ** và là véctơ riêng tương ứng.

* Thành phần chính thứ hai thu được từ véctơ hệ số , là tổ hợp tuyến tính làm cực đại phương sai của và không tương quan với Tương tự như trên ta có:

với điều kiện: trong đó trực giao với

Lập hàm Lagrange như trên ta sẽ thu được:

Từ đó suy ra:

Nhân 2 vế của phương trình trên với ta sẽ có:

Do: nên hệ thức trên rút gọn thành:

Mặt khác vì các véctơ riêng của **Σ** đều trực giao với nhau nên ta có:

suy ra

Vậy . Tương tự như lập luận trên là giá trị riêng lớn thứ 2 của **Σ** vàlà véctơ riêng tương ứng.

* Thành phần chính thứ i,  (i =3,…,k) được xác định tiếp như sau, , làm cực đại phương sai của  với điều kiện và ,

Khi đó là giá trị riêng lớn thứ I của ma trận **Σ** và là véctơ riêng tương ứng.

Ta sẽ có:

.

Điều đó có nghĩa là phương sai của các thành phần chính sẽ là giá trị riêng của ma trận Σ. Tổng của chúng sẽ tạo nên vết của ma trận:

Theo biểu diễn phổ của ma trận ta viết được:

Trong đó là ma trận đường chéo có các phần tử trên đường chéo chính là và **C** là ma trận cột thành lập từ các véctơ riêng của **Σ**. Chú ý rằng **C** là ma trận trực giao nên ta sẽ có

Theo cách biến đổi trên ta thu được:

Nếu k thành phần chính đầu tiên biễu diễn cho một lượng lớn tổng phương sai của các dữ liệu thì sau này chúng được dùng trong phân tích với vai trò thay thế hoàn toàn các biến ban đầu. Các dữ liệu p-chiều đã cho đã được thu gọn lại thành k-chiều.

Tóm lại, cho **Σ** là ma trận hiệp phương sai của véctơ

Gọi với và là các giá trị riêng và véctơ riêng tương ứng của ma trận **Σ**. Khi đó thành phần chính thứ i, được xác định bởi:

Do đó để xác định các thành phần chính của tập dữ liệu ban đầu ta chỉ cần tìm các giá trị riêng và véctơ riêng của ma trận hiệp phương sai **Σ**.

* ***Ví dụ 1***

Cho các biến ngẫu nhiên có ma trận hiệp phương sai là:

.

Tìm các trị riêng và vectơ riêng của ma trận ta sẽ thu được:

Từ đó ta có các thành phần chính là và , trong đó

* ***Chọn ma trận tương quan hay chọn ma trận hiệp phương sai?***

Trong thực hành ta nên dùng ma trận tương quan sẽ cho kết quả tốt hơn. Vì nếu phương sai của các biến dữ liệu gốc chênh lệch nhiều thì sẽ có ảnh hưởng lớn đến việc xác định thành phần chính. Cách giải quyết thông thường là chuẩn hóa các biến đó hay nói cách khác là xác định ma trận các hệ số tương quan bởi:

Khi đó tập biến mới sẽ có:

Trong đó là *ma trận các hệ số tương quan* dữ liệu gốc và ta gọi tắt nó *là ma trận tương quan*. Lúc này nếu ta xử lý tập dữ liệu mới bằng cách dùng phương pháp PCA để phân tích ma trận hiệp phương sai thì đây cũng chính là phân tích với ma trận tương quan.

Khi đó phương sai của các biến chuẩn hóa bằng p, chính là tổng số phần tử trên đường chéo chính của ma trận hay là số biến ban đầu.

**CHỌN SỐ LƯỢNG CÁC THÀNH PHẦN CHÍNH**

Mục đích của phân tích thành phần chính là thu gọn số chiều, tức làm biến đổi nhiều biến tương quan về tập mới nhỏ hơn và không tương quan, từ đó các biến này sẽ dược dùng làm dữ liệu chính trong các phân tích tiếp theo. Vấn đề là cần chọn bao nhiêu thành phần chính trong số các tổ hợp tuyến tính đó để giữ lại phân tích. Người ta thường dùng 03 cách sau để chọn số lượng thành phần chính.

* Dựa vào tỷ lệ phương sai tích lũy.
* Dựa vào độ lớn phương sai của các thành phần chính.
* Dựa vào đồ thị độ dốc (Screeplot).

Ta sẽ lần lượt làm quen với chúng.

1. ***Dựa vào tỷ lệ phương sai tích lũy***

Nếu là các giá trị riêng thì tỷ lệ phương sai tích lũy của k giá trị riêng đầu tiên ứng với k thành phần chính đầu tiên là:

Khi áp dụng thuật toán PCA, các giá trị riêng có phương sai nhỏ sẽ bị loại bỏ, chỉ giữ lại k giá trị riêng (tức k thành phần chính) nắm giữ hầu như toàn bộ thông tin của dữ liệu ban đầu. Thông thường người ta ưu tiên chọn các thành phần chính chiếm từ 70% - 90% trở đi của tổng phương sai.

Khi phân tích với ma trận tương quan thì tỷ lệ phương sai tích lũy là:

với .

1. ***Dựa vào độ lớn phương sai của các thành phần chính***

Phương pháp này thường được áp dụng trong trường hợp dùng ma trận tương quan để phân tích. Khi dùng ma trận tương quan các biến được chuẩn hóa và có phương sai bằng 1, nên cho phép giữ lại tất cả các thành phần chính mà có phương sai từ 1 trở lên. Cách này được giới thiệu bởi Henry Kaiser nên được gọi là tiêu chuẩn Kaiser. Dựa vào điều này nếu phân tích bằng **Σ** thay cho (ma trận tương quan) thì người ta giữ lại những véctơ riêng nào có giá trị riêng không kém hơn giá trị riêng trung bình của các biến ban đầu.

1. ***Dựa vào đồ thị độ dốc (Screeplot)***

Đồ thị độ dốc là đồ thị biểu diễn độ lớn của các giá trị tương ứng với mỗi số i, nên được xem là đồ thị của các phương sai. Các véctơ riêng được xắp xếp tương ứng với các trị riêng theo thứ tự giảm dần. Để xác định số lượng các thành phần thích hợp giữ lại, cần tìm một điểm uốn khúc (elbow) trong đồ thị dốc. Điểm uốn khúc là điểm mà bắt đầu từ nó các giá trị riêng còn lại tương ứng là rất nhỏ và chúng hầu như xấp xỉ nhau. Điều đó có nghĩa là nếu thêm vào bất kỳ véctơ riêng nào khác thì cũng không làm tăng thêm lượng phương sai nào đáng kể. Phương pháp này được đề xuất bởi Cattell (1966).

* ***Ví dụ 2***

Sử dụng phương pháp phân tích thành phần chính để xét các biến với ma trận tương quan sau:

Trước hết ta tìm các giá trị riêngta tìm được 5 giá trị riêng:

3,263; 1,538; 0,168; 0,031; 0.

Hai giá trị riêng lớn nhất ta chọn dựa vào tỷ lệ phương sai tích lũy:

Các véctơ riêng tương ứng với 3,263 và 1,538 sẽ là:

Trong ví dụ này ta có và Mặt khác ta biết rằng:

Ta xác định

trong đó là các giá trị riêng lớn tương ứng đã chọn.

Khi đó, nếu ký hiệu ta có ước lượng:

.

Trong ví dụ này ta sẽ có:

**.**

***MỘT SỐ KIẾN THỨC TOÁN CẦN ÔN LẠI KHI ĐỌC PHẦN NÀY***

* *Momen tương quan và ma trận hiệp phương sai*

*Hệ số tương quan và ma trận tương quan (Trong LT Xác suất)*

* *Cực trị có điều kiện của hàm số nhiều biến số và phương pháp nhân tử Lagranger*

*(Trong Toán giải tích )*

* *Phân tích phổ và khái niệm về vec tơ riêng và trị riêng của ma trận (Trong Đại số TT)*
* *Với* ***Σ***  *là ma trận hiệp phương sai*
* *là ma trận có các cột là các vec tơ riêng của* ***Σ,*** *là ma trận chuyển vị của là ma trận*  ***Λ*** *là ma trận đường chéo tạo bởi các trị riêng tương ứng,**phân tích:*

***Λ (\*)***

*được gọi là phân tích phổ của ma trận* ***Σ.*** *Từ**hệ thức* ***(\*)*** *sẽ suy ra*

***Λ (\*\*)***

*do tính chất của ma trận trực giao,*

* *Ma trận trực giao (orthogonal matrix) là ma trận vuông có các dòng và các cột là hệ trực chuẩn.*

*Cho**là ma trận không suy biến* ( *Nếu* *trực giao thì* *cũng trực giao. Ma trận trực giao có chuyển vị là ma trận nghịch đảo*