

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

Tên báo cáo:

**ÁP DỤNG PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH TRONG
MINH GIẢI DỮ LIỆU ĐỊA VẬT LÝ GIẾNG KHOAN**

GV thực hiện: ThS. Bùi Thị Ngân

Bộ môn: Địa chất dầu khí

Hà Nội, 01/2026

PCA (Principal Component Analysis) là một phương pháp thống kê dùng để giảm chiều dữ liệu (dimensionality reduction), bằng cách biến đổi tập hợp nhiều biến có tương quan thành một tập hợp biến mới không tương quan gọi là các thành phần chính (principal components).

Ý tưởng cốt lõi

Giả sử dữ liệu có hàng trăm đặc trưng (biến). PCA sẽ tìm cách xoay và chiếu không gian dữ liệu này lên các trục tọa độ mới gọi là các thành phần chính (Principal Components - PCs).

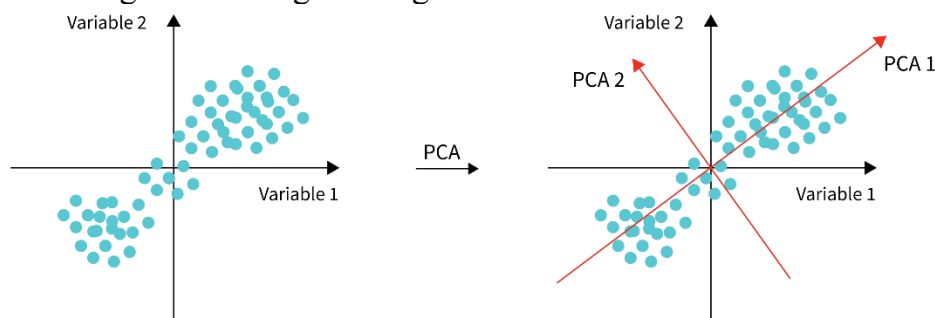
Các thành phần chính này vuông góc với nhau (trực giao) và không có sự tương quan.

Thành phần chính đầu tiên PC1 chứa phương sai (thông tin) lớn nhất, chứa phần phương sai lớn nhất tiếp theo, và cứ thế tiếp diễn.

Bằng cách loại bỏ các thành phần có phương sai rất nhỏ, bạn có thể giảm thiểu đáng kể số chiều dữ liệu mà không làm mất mát nhiều thông tin.

Việc giảm số lượng biến trong một tập dữ liệu đương nhiên sẽ dẫn đến giảm độ chính xác, nhưng thủ thuật trong việc giảm chiều dữ liệu là đánh đổi một chút độ chính xác để đổi lấy sự đơn giản. Bởi vì các tập dữ liệu nhỏ hơn dễ khám phá và trực quan hóa hơn, do đó giúp việc phân tích các điểm dữ liệu trở nên dễ dàng và nhanh chóng hơn đối với [các thuật toán học máy](#) mà không cần xử lý các biến không cần thiết.

Tóm lại, ý tưởng của PCA rất đơn giản: giảm số lượng biến trong một tập dữ liệu, đồng thời bảo toàn càng nhiều thông tin càng tốt.



Giảm số lượng biến trong một tập dữ liệu

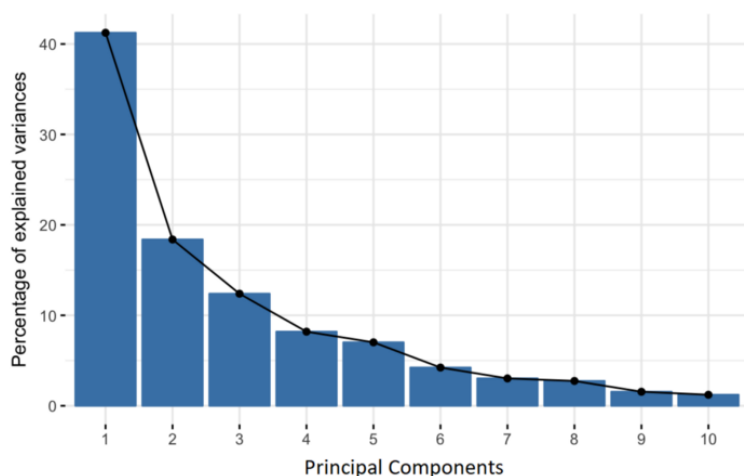
Thành phần chính là gì?

Các thành phần chính là các biến mới được xây dựng dưới dạng tổ hợp tuyến tính hoặc hỗn hợp của các biến ban đầu. Các tổ hợp này được thực hiện theo cách sao cho các biến mới (tức là các thành phần chính) không tương quan và hầu hết thông tin trong các biến ban đầu được nén lại vào thành phần đầu tiên. Ví dụ, dữ liệu 10 chiều cung cấp 10 thành phần chính, nhưng PCA cố gắng đặt tối đa thông tin có thể vào thành phần đầu tiên, sau đó là tối đa thông tin còn lại vào thành phần thứ hai, v.v., cho đến khi có được kết quả như được hiển thị trong biểu đồ Hình 1 bên dưới.

Việc sắp xếp thông tin theo các thành phần chính như vậy sẽ cho phép giảm chiều dữ liệu mà không làm mất nhiều thông tin, bằng cách loại bỏ các thành phần có ít thông tin và coi các thành phần còn lại là các biến mới.

Điều quan trọng cần nhận ra ở đây là các thành phần chính khó giải thích hơn và không có ý nghĩa thực sự nào vì chúng được xây dựng như là tổ hợp tuyến tính của các biến ban đầu.

Về mặt hình học, các thành phần chính biểu thị các hướng của dữ liệu giải thích được lượng phương sai tối đa, nghĩa là các đường thẳng nắm bắt được nhiều thông tin nhất của dữ liệu. Mối quan hệ giữa phương sai và thông tin ở đây là: phương sai càng lớn thì độ phân tán của các điểm dữ liệu dọc theo đường thẳng đó càng lớn, và độ phân tán dọc theo một đường thẳng càng lớn thì nó càng chứa nhiều thông tin. Nói một cách đơn giản, hãy coi các thành phần chính như những trục mới cung cấp góc nhìn tốt nhất để quan sát và đánh giá dữ liệu, giúp sự khác biệt giữa các quan sát được thể hiện rõ ràng hơn.



Hình 1. Tỷ lệ phương sai (Thông tin) cho mỗi PC với dữ liệu ban đầu là 10 chiều (10 biến đầu vào)

PCA xây dựng các thành phần chính như thế nào?

Vì số lượng thành phần chính bằng số lượng biến trong dữ liệu, nên các thành phần chính được xây dựng theo cách sao cho thành phần chính đầu tiên giải thích được phương sai lớn nhất có thể trong tập dữ liệu.

Ví dụ, giả sử biểu đồ phân tán của tập dữ liệu của chúng ta được hiển thị như bên dưới, liệu chúng ta có thể đoán được thành phần chính đầu tiên không? Có, đó gần đúng là đường thẳng trùng với các dấu màu tím vì nó đi qua gốc tọa độ và là đường thẳng mà hình chiếu của các điểm (chấm đỏ) trải rộng nhất. Hay nói theo cách toán học, đó là đường thẳng tối đa hóa phương sai (trung bình của bình phương khoảng cách từ các điểm được chiếu (chấm đỏ) đến gốc tọa độ).

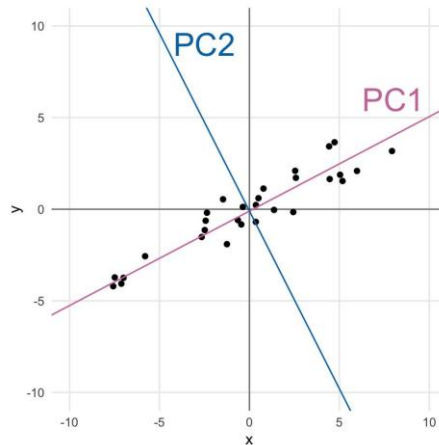
Thành phần chính thứ hai được tính toán theo cùng một cách, với điều kiện là nó không tương quan với (tức là vuông góc với) thành phần chính thứ nhất và nó giải thích cho phương sai cao thứ hai.

Quá trình này tiếp tục cho đến khi tổng cộng p thành phần chính được tính toán, bằng với số lượng biến ban đầu.

PC1 lấy thông tin lớn nhất, PC2 lấy phần lớn nhất còn sót lại nhưng không được trùng với PC1, PC3 lấy phần lớn nhất còn sót lại nhưng không được trùng với cả PC1 và PC2, cứ thế tiếp tục cho đến khi hết số chiều ban đầu.

Trong không gian 2 chiều

Giả sử dữ liệu có 2 biến: RHOB và NPHI.



1. PC1: tìm hướng có phương sai lớn nhất trong toàn bộ mặt phẳng.
2. PC2: phải vuông góc với PC1.

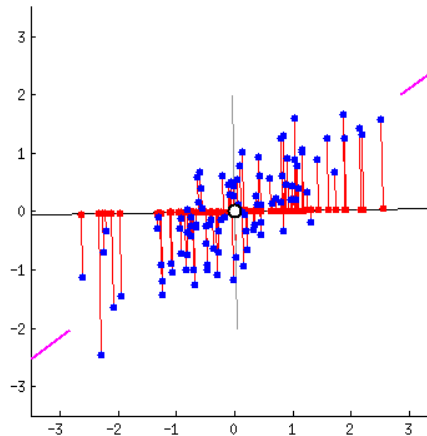
Không gian 3 chiều

Giả sử dữ liệu có 3 biến: RHOB, NPHI, DT.

1. PC1 được tìm trước.
2. Các hướng vuông góc với PC1 không còn là một đường nữa, mà tạo thành cả một mặt phẳng có vô số hướng (mặt phẳng trực giao với PC1). PCA sẽ chọn trong số đó hướng mà dữ liệu chiếu lên có phương sai lớn nhất, chính là PC2.

Sau đó:

3. PC3 là hướng vuông góc với cả PC1 và PC2. Trong 3D thì lúc này chỉ còn một hướng duy nhất.



Xây dựng các thành phần chính

2. Các bước thực hiện cơ bản

Về mặt toán học, PCA hoạt động dựa trên **đại số tuyến tính** và **ma trận hiệp phương sai (covariance matrix)**. Quá trình có thể được tóm tắt như sau:

Chuẩn hóa dữ liệu: Các đặc trưng cần được đưa về cùng một thang đo (thường có trung bình bằng 0 và độ lệch chuẩn bằng 1).

Mục đích của bước này là chuẩn hóa phạm vi của các biến ban đầu liên tục sao cho mỗi biến đều đóng góp như nhau vào quá trình phân tích.

Cụ thể hơn, lý do tại sao việc chuẩn hóa dữ liệu trước khi thực hiện PCA lại rất quan trọng là vì phương pháp này khá nhạy cảm với sự khác biệt về phương sai của các biến ban đầu. Nghĩa là, nếu có sự khác biệt lớn giữa phạm vi giá trị của các biến ban đầu, thì các biến có phạm vi lớn hơn sẽ chiếm ưu thế hơn so với các biến có phạm vi nhỏ hơn (ví dụ, một biến có phạm vi từ 0 đến 100 sẽ chiếm ưu thế hơn so với một biến có phạm vi từ 0 đến 1), dẫn đến kết quả bị sai lệch. Vì vậy, việc chuyển đổi dữ liệu sang các thang đo tương đương có thể ngăn ngừa vấn đề này.

Về mặt toán học, điều này có thể được thực hiện bằng cách trừ đi giá trị trung bình và chia cho độ lệch chuẩn cho mỗi giá trị của mỗi biến. Sau khi quá trình chuẩn hóa hoàn tất, tất cả các biến số sẽ được chuyển đổi sang cùng một thang đo.

$$z = \frac{value - mean}{standard\ deviation}$$

Tính ma trận hiệp phương sai: Đo lường mức độ tương quan giữa các cặp đặc trưng.

Mục đích của bước này là để hiểu các biến số trong tập dữ liệu đầu vào đang biến thiên như thế nào so với giá trị trung bình, hay nói cách khác, để xem liệu có mối quan hệ nào giữa chúng hay không. Bởi vì đôi khi, các biến số có tương quan cao đến mức chúng chứa thông tin dư thừa. Vì vậy, để xác định các mối tương quan này, chúng ta tính toán **ma trận hiệp phương sai**.

Ma trận hiệp phương sai là một ma trận đối xứng $p \times p$ (trong đó p là số chiều) có các phần tử là các hiệp phương sai liên quan đến tất cả các cặp biến ban đầu có thể có. Ví dụ, đối với một tập dữ liệu 3 chiều với 3 biến x , y và z , ma trận hiệp phương sai là một ma trận dữ liệu 3×3 có dạng như sau:

$$\begin{bmatrix} Cov(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & Cov(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & Cov(z, z) \end{bmatrix}$$

Ma trận hiệp phương sai cho dữ liệu 3 chiều

Vì hiệp phương sai của một biến với chính nó chính là phương sai của nó ($Cov(a, a) = Var(a)$), nên trên đường chéo chính (từ trên cùng bên trái xuống dưới cùng bên phải) ta thực sự có phương sai của từng biến ban đầu. Và vì hiệp phương sai có tính chất giao hoán ($Cov(a, b) = Cov(b, a)$), nên các phần tử của ma trận hiệp phương sai đối xứng với đường chéo chính, có nghĩa là phần tam giác trên và phần tam giác dưới bằng nhau.

Các hệ số hiệp phương sai mà chúng ta có trong ma trận cho chúng ta biết điều gì về mối tương quan giữa các biến?

Thực ra, dấu của hiệp phương sai mới là điều quan trọng:

- Nếu giá trị dương thì: hai biến số cùng tăng hoặc cùng giảm (có tương quan).

- Nếu giá trị âm thì: biến này tăng khi biến kia giảm (Tương quan nghịch đảo)

Tức là, ma trận hiệp phương sai không hơn gì một bảng tóm tắt các mối tương quan giữa tất cả các cặp biến có thể có.

Tính toán trị riêng và vector riêng: Phân tích ma trận hiệp phương sai để tìm ra các thành phần chính (vector riêng) và tầm quan trọng của chúng (trị riêng).

Các vector riêng và giá trị riêng là những khái niệm đại số tuyến tính mà chúng ta cần tính toán từ ma trận hiệp phương sai để xác định *các thành phần chính* của dữ liệu.

Điều đầu tiên bạn cần biết về vector riêng và giá trị riêng là chúng luôn xuất hiện theo cặp, sao cho mỗi vector riêng đều có một giá trị riêng tương ứng. Ngoài ra, số lượng vector riêng bằng với số chiều của dữ liệu. Ví dụ, đối với tập dữ liệu 3 chiều, có 3 biến, do đó có 3 vector riêng với 3 giá trị riêng tương ứng.

Chính các vector riêng và giá trị riêng là yếu tố tạo nên sự kỳ diệu của phân tích thành phần chính, bởi vì các vector riêng của ma trận hiệp phương sai thực chất là *hướng của các trục có độ biến thiên lớn nhất* (nhiều thông tin nhất) và chúng ta gọi đó là các thành phần chính. Còn các giá trị riêng đơn giản là các hệ số gắn liền với các vector riêng, cho biết lượng *biến thiên được mang trong mỗi thành phần chính*.

Bằng cách xếp hạng các vector riêng theo thứ tự giá trị riêng của chúng, từ cao nhất đến thấp nhất, bạn sẽ nhận được các thành phần chính theo thứ tự mức độ quan trọng.

Ví dụ về phân tích thành phần chính:

Giả sử tập dữ liệu của chúng ta là không gian 2 chiều với 2 biến x, y và các vector riêng và giá trị riêng của ma trận hiệp phương sai như sau:

$$v_1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \quad \lambda_1 = 1.284028$$

$$v_2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \quad \lambda_2 = 0.04908323$$

Nếu ta xếp hạng các giá trị riêng theo thứ tự giảm dần, ta được $\lambda_1 > \lambda_2$, có nghĩa là vector riêng tương ứng với thành phần chính thứ nhất (PC1) là v_1 và vector riêng tương ứng với thành phần chính thứ hai (PC2) là v_2 .

Sau khi có được các thành phần chính, để tính toán phần trăm phương sai (thông tin) mà mỗi thành phần giải thích, ta chia giá trị riêng của mỗi thành phần cho tổng các giá trị riêng. Nếu áp dụng điều này vào ví dụ trên, ta thấy rằng PC1 và PC2 lần lượt giải thích 96% và 4% phương sai của dữ liệu.

Chọn số chiều mới: Sắp xếp các trị riêng theo thứ tự giảm dần, giữ lại k thành phần chính tương ứng với mức phương sai tích lũy mong muốn (thường là 85% - 95%).

Việc tính toán các vector riêng và sắp xếp chúng theo giá trị riêng giảm dần cho phép chúng ta tìm ra các thành phần chính theo thứ tự mức độ quan trọng. Trong bước này, chúng ta sẽ lựa chọn xem có giữ lại tất cả các thành phần này hay loại bỏ những thành phần ít quan trọng hơn (có giá trị riêng thấp), và tạo thành một ma trận vector với những thành phần còn lại, mà chúng ta gọi là *vector đặc trưng*.

Như vậy, vector đặc trưng chỉ đơn giản là một ma trận có các cột là các vector riêng của các thành phần mà chúng ta quyết định giữ lại. Điều này tạo nên bước đầu tiên hướng tới việc giảm chiều dữ liệu, bởi vì nếu chúng ta chọn chỉ giữ lại p vector riêng (thành phần) trong $số\ n\ vector$ riêng, thì tập dữ liệu cuối cùng sẽ chỉ có p chiều.

Ví dụ về phân tích thành phần chính :

Tiếp tục với ví dụ từ bước trước, chúng ta có thể tạo một vector đặc trưng bằng cả hai vector riêng $v1$ và $v2$:

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Hoặc loại bỏ vector riêng $v2$, là vector có tầm quan trọng thấp hơn, và chỉ tạo một vector đặc trưng với $v1$:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

Việc loại bỏ vector riêng $v2$ sẽ làm giảm số chiều đi 1, và do đó sẽ gây ra mất mát thông tin trong tập dữ liệu cuối cùng. Nhưng vì $v2$ chỉ mang 4% thông tin, nên sự mất mát này sẽ không đáng kể và chúng ta vẫn sẽ có 96% thông tin mà $v1$ mang theo.

Như chúng ta đã thấy trong ví dụ, việc giữ lại tất cả các thành phần hay loại bỏ những thành phần ít quan trọng hơn là tùy thuộc vào mục đích của bạn. Bởi vì nếu bạn chỉ muốn mô tả dữ liệu của mình bằng các biến mới (thành phần chính) không tương quan mà không cần giảm chiều dữ liệu, thì việc loại bỏ các thành phần ít quan trọng hơn là không cần thiết.

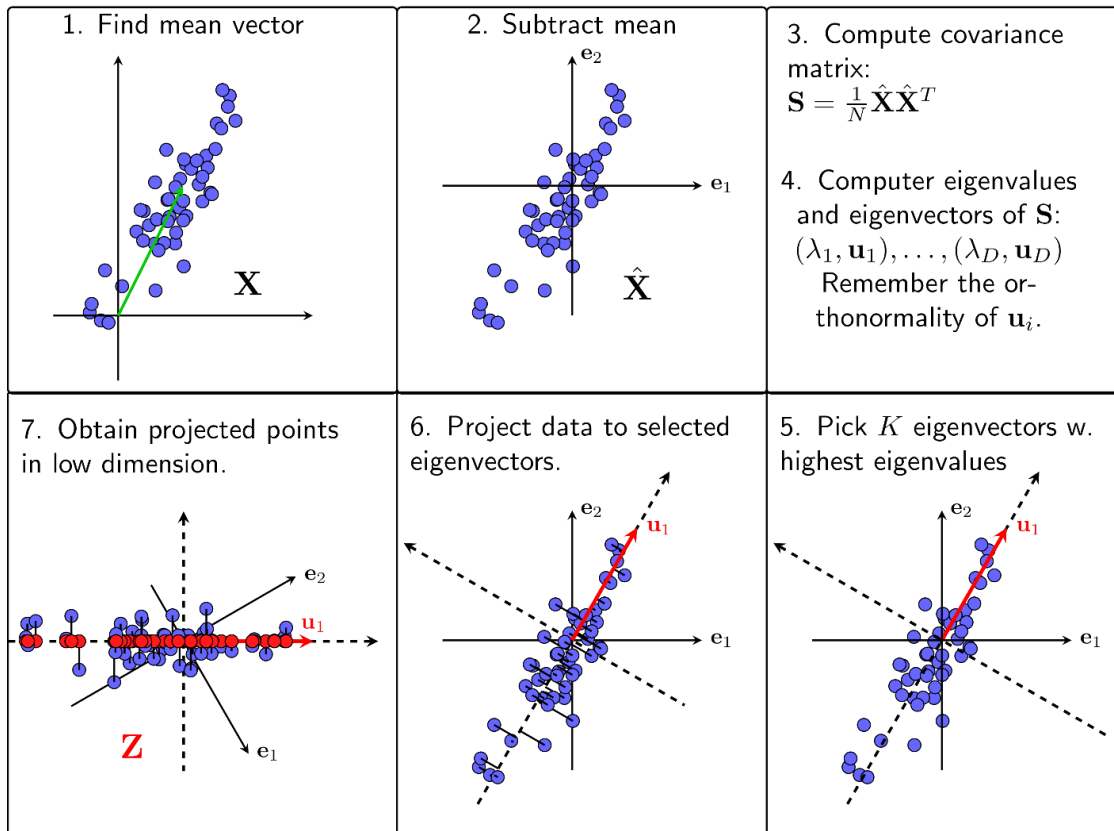
Chiều dữ liệu: Nhân tập dữ liệu gốc với các vector riêng đã chọn để tạo ra bộ dữ liệu mới có số chiều ít hơn.

Trong các bước trước, ngoài việc chuẩn hóa, bạn không thực hiện bất kỳ thay đổi nào khác trên dữ liệu, bạn chỉ chọn các thành phần chính và tạo thành vector đặc trưng, nhưng tập dữ liệu đầu vào luôn được giữ nguyên theo các trục ban đầu (tức là theo các biến ban đầu).

Ở bước này, cũng là bước cuối cùng, mục đích là sử dụng vector đặc trưng được tạo thành từ các vector riêng của ma trận hiệp phương sai để định hướng lại dữ liệu từ các trục ban đầu sang các trục được biểu diễn bởi các thành phần chính (do đó có tên là Phân tích Thành phần Chính). Điều này có thể được thực hiện bằng cách nhân ma trận chuyển vị của tập dữ liệu gốc với ma trận chuyển vị của vector đặc trưng.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

PCA procedure



Hình 5: Các bước thực hiện PCA.

Tại sao nên dùng PCA?

Khi làm việc với các bộ dữ liệu lớn, việc xử lý quá nhiều biến có thể dẫn đến:

- Mô hình phức tạp, dễ overfitting
- Thời gian huấn luyện lâu
- Khó trực quan hóa dữ liệu
- Hiện tượng đa cộng tuyến (multicollinearity)

PCA giúp khắc phục các vấn đề này thông qua những lợi ích sau:

- ✓ **Thu gọn không gian đặc trưng:** Giảm số biến, đơn giản hóa mô hình mà vẫn giữ được độ chính xác cao.
- ✓ **Loại bỏ thông tin nhiễu:** PCA tập trung vào các hướng có phương sai lớn nhất – tức là nơi dữ liệu thay đổi mạnh mẽ và có ý nghĩa nhất.
- ✓ **Tăng tốc độ và hiệu suất mô hình:** Giảm chiều giúp rút ngắn thời gian huấn luyện và tăng khả năng khái quát.
- ✓ **Phát hiện dữ liệu bất thường (outlier detection):** PCA có thể chỉ ra những điểm dữ liệu lệch khỏi cấu trúc chính của dữ liệu.

Nhược điểm

Các đặc trưng mới sau khi giảm chiều mang tính toán học thuần túy, mất đi ý nghĩa diễn giải vật lý ban đầu của từng cột dữ liệu. Ngoài ra, PCA chỉ phù hợp tốt với các mô hình tuyến tính.

✗ **Khó giải thích ý nghĩa các thành phần mới**, vì chúng là tổ hợp tuyến tính của nhiều biến gốc.

✗ **Nhạy cảm với tỷ lệ dữ liệu** – nếu không chuẩn hóa đúng, kết quả có thể sai lệch.

✗ **Mất một phần thông tin** khi giảm chiều quá mạnh.

✗ **Giả định mối quan hệ tuyến tính**, không phù hợp cho dữ liệu phi tuyến tính.

4. Ứng dụng thực tế

PCA giúp giảm số lượng biến hoặc đặc trưng trong tập dữ liệu trong khi vẫn bảo toàn những thông tin quan trọng nhất như xu hướng hoặc mô hình chính. Việc giảm số lượng này có thể rút ngắn thời gian huấn luyện mô hình máy học và giúp tránh hiện tượng quá khớp (overfitting) trong mô hình.

- **Trực quan hóa:** PCA giảm dữ liệu đa chiều xuống không gian hai chiều hoặc ba chiều để dễ hình dung hơn, giúp phát hiện các xu hướng dữ liệu, các cụm hoặc các điểm bất thường.

- **Tiền xử lý cho Machine Learning:** PCA thường được sử dụng trước khi áp dụng các thuật toán học máy nhằm giảm chiều dữ liệu, tăng tốc quá trình huấn luyện và cải thiện khả năng khái quát hóa của mô hình. PCA chuyển đổi các đặc điểm dữ liệu tương quan thành một tập hợp nhỏ hơn các thành phần không tương quan, điều này có thể giúp cải thiện hiệu suất và khả năng giải thích của mô hình.

Khi nào nên sử dụng PCA?

- Bộ dữ liệu có **rất nhiều biến và mối tương quan cao**
- Cần **trực quan hóa dữ liệu nhiều chiều**
- Muốn **giảm nhiều và cải thiện hiệu suất mô hình**
- Muốn **chuẩn bị dữ liệu trước khi huấn luyện mô hình học máy**

Ví dụ trong dữ liệu địa vật lý giếng khoan với 2 đường: RHOB và NPHI. PCA tìm hướng mà dữ liệu biến thiên mạnh nhất. PCA tìm ra trục chính, ví dụ: 95% dữ liệu nằm dọc theo PC1. X có thể: RHOB bình thường, NPHI bình thường nhưng lại cách rất xa trục PC1.

PCA phát hiện được vì điểm đó không tuân theo cấu trúc tương quan chung của dữ liệu.

Với nhiều logs hơn, chúng ta không thể đánh giá dễ dàng được không gian nhiều chiều.

- Mỗi giá trị riêng lẻ đều nằm trong phạm vi bình thường.
- Nhưng tổ hợp đó lại phi lý về mặt địa chất.
- Boxplot không phát hiện được.
- PCA có khả năng phát hiện vì điểm này không nằm trên "cấu trúc đa chiều" của tập dữ liệu.

PCA sẽ:

- nén dữ liệu xuống PC1, PC2
- giữ lại phần lớn thông tin
- làm nổi bật các điểm nằm xa cụm dữ liệu

Đây là lý do PCA được dùng nhiều trong:

- Well log QC
- Reservoir characterization
- Production surveillance
- Anomaly detection

Phương pháp	Ứng dụng	Độ phức tạp
Boxplot	Outlier của từng biến	Thấp
Z-score	Outlier của từng biến	Thấp
PCA	Outlier trong quan hệ nhiều biến	Trung bình
Isolation Forest	Outlier phi tuyến	Cao

Một quy trình thực tế trong xử lý dữ liệu địa vật lý giếng khoan thường là:

1. Dùng Boxplot hoặc Z-score để loại bỏ các giá trị sai rõ ràng (GR = -50 API, RHOB = 10 g/cm³,...).
2. Chuẩn hóa dữ liệu.
3. Dùng PCA hoặc Mahalanobis Distance để phát hiện các điểm bất thường về mặt tương quan giữa các log.
4. Kiểm tra lại bằng kiến thức địa chất và minh giải giếng khoan.

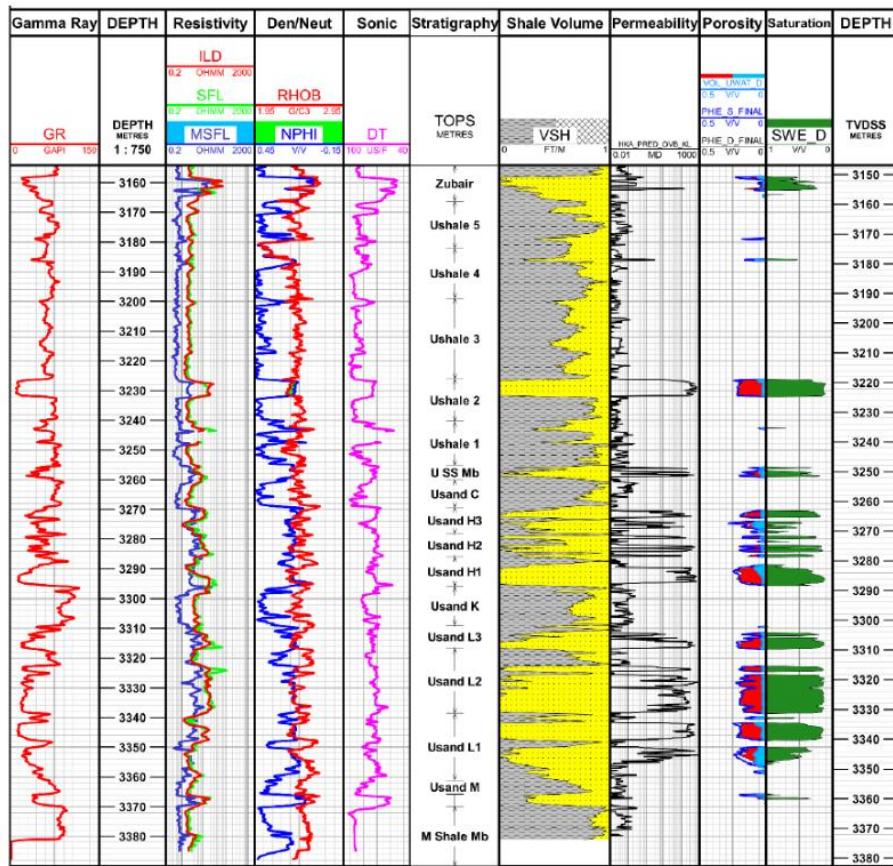
Vì vậy, PCA không phải để thay thế Boxplot hay Z-score, mà để phát hiện một loại outlier khác: **những điểm có vẻ bình thường trên từng biến riêng lẻ nhưng bất thường khi xét đồng thời nhiều biến địa vật lý**. Đây mới là giá trị lớn nhất của PCA trong dữ liệu giếng khoan và dữ liệu dầu khí nói chung.

Ứng dụng PCA trong địa vật lý giếng khoan

Kỹ thuật PCA có thể được áp dụng cho nhiều thông số đo lường khác nhau như well log bởi vì chúng phản ứng với các yếu tố địa chất liên quan. Ví dụ, các phép đo độ rỗng GR và NPHI thường có tương quan thống kê vì cả hai đều phản ứng với hàm lượng sét trong đá. Do đó PCA cung cấp thông tin chi tiết bằng cách chuyển đổi dữ liệu well log thành một hệ tọa độ mới, trực giao và đơn giản hóa.

PCA được thực hiện trên khoảng độ sâu 3160m-3280m ở giếng X. Phân tích sử dụng 7 well log, bao gồm: GR, NPHI, RHOB, DT, ILD, MSFL, SFL làm biến đầu vào, đại diện cho X_k với $p = 6$. Mỗi well log chứa 4500 phép đo về đặc tính đất đá dưới lòng đất. đại diện cho các quan sát x_{ik} với $n = 4500$.

Kết quả PCA được biểu diễn dưới dạng số liệu và đồ họa.



Ma trận tương quan R của toàn bộ dữ liệu well log được thể hiện trong bảng 1.

Mà trận tương quan cho thấy:

- Có mối tương quan dương cao và trung bình giữa GR và NPHI, DT
- Có mối tương quan dương cao giữa NPHI và DT
- Tồn tại mối tương quan âm ở mức trung bình giữa nhóm nhất ký (GR

NPHI, DT) và các nhật ký điện trở suất, bởi vì GR, NPHI, DT thường phản ánh hàm lượng sét, độ rỗng, còn Resistivity logs phản ánh độ dẫn điện của đá và độ bão hòa chất lưu. Thông thường, đá sét có GR, NPHI, DT cao, R thấp, còn cát sạch hoặc có chứa hydrocarbon thì R cao.

- Ngược lại, RHOB gần như không tương quan với resistivity, bởi vì type gần như không làm thay đổi RHOB nhiều, vì mật độ nước và mật độ dầu có sự khác biệt lớn, còn R nước thấp hơn R dầu. Cũng có thể do sự khác biệt nhỏ về giá trị mật độ đo được bằng RHOB so với các logs khác (giá trị nằm từ 2 – 2.6 g/cm³), do sự không đồng nhất về địa chất, các lớp đất đá có khác nhau nằm xen kẽ nên sự biến đổi mật độ giữa các tầng đá không đủ để liên quan hệ rõ ràng với các logs khác.

Mối tương quan giữa các nhật ký giếng khác nhau cho thấy các công cụ này rất cần thiết để giải thích well logs và đánh giá các tầng địa chất. Mối tương quan mạnh giữa nhiều phép đo có thể liên quan đến thực tế là các công cụ này phản ứng với các đặc điểm chung của tầng địa chất.

Bảng 1. Ma trận tương quan R của toàn bộ dữ liệu well log

	GR	NPHI	DT	ILD	SFL	MSFL	RHOB
GR	1.00						
NPHI	0.71	1.00					
DT	0.52		1.00				
ILD	-0.52	-0.61	-0.65	1.00			
SFL	-0.44	-0.53	-0.57	0.75	1.00		
MSFL	-0.38	-0.59	-0.61	0.71	0.58	1.00	
RHOB	0.12	-0.21	-0.24	0.09	0.07	0.19	1.00

Giá trị riêng và phương sai

Cung cấp thông tin chi tiết về cách diễn giải các PCA.

PC1 mô tả tỉ lệ phần trăm lớn nhất, 58% tổng biến thiên trong tập dữ liệu, trong khi PC2 chiếm 16% tổng biến thiên. Tổng PC1 và PC2 giải thích được 74% tổng biến thiên trong tập dữ liệu. Như vậy, cách diễn giải dữ liệu hiệu quả nhất được cung cấp bởi thành phần chính thứ nhất và thứ hai.

Bảng 2. Giá trị riêng và phương sai được giải thích bởi PCA

No.	PC	Eigenvalue	Variance (%)	Cum variance (%)
1	PC1	4.05	58	58
2	PC2	1.11	16	74
3	PC3	0.76	11	85
4	PC4	0.41	6	91
5	PC5	0.35	5	96
6	PC ₆	0.21	3	99
7	PC ₇	0.10	1	100

3. PC loading

Table 3. The PCA Components for the oil well X

	PC ₁	PC ₂	PC ₃	PC ₄	PC ₅	PC ₆	PC ₇
GR	-0.70	0.43	0.42	-0.15	-0.33	0.02	0.11
NPHI	-0.88	-0.02	0.39	0.05	0.11	0.08	-0.23
DT	-0.87	-0.14	0.21	0.12	0.37	-0.08	0.17
ILD	0.87	-0.07	0.32	0.05	0.09	0.36	0.06
MSFL	0.80	0.13	0.31	-0.43	0.21	-0.16	-0.03
SFL	0.78	-0.08	0.41	0.39	-0.11	-0.21	-0.01
RHOB	0.19	0.94	-0.14	0.19	0.17	0.01	-0.02

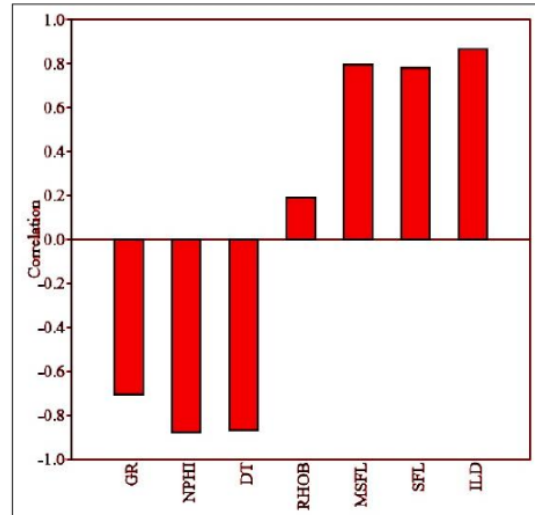


Fig. 3. Bar plot for loading coefficient of PC1

Hệ số tải thành phần (mối tương quan giữa nhật ký giếng khoan và các thành phần chính) The component loadings được trình bày trong Bảng 5 và được vẽ trong Hình 3. cho PC1, trong đó sự đóng góp của sáu nhật ký giếng khoan trong thành phần chính thứ nhất là GR, NPHI, DT, ILD, MSFL và SFL, và RHOB trong thành phần chính thứ hai chiếm ưu thế (trọng số rất cao và cao)

4. Tương quan vòng tròn (Circle Correlations)

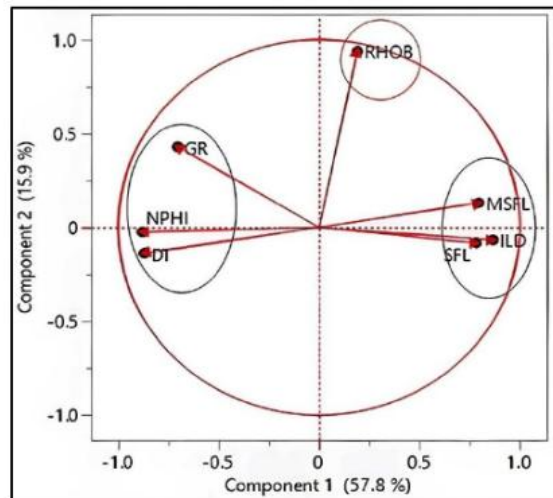


Fig. 4 . Circle correlation shows the correlation of components PC1 and PC2 in Table 4

Dấu chấm của các nhật ký nằm gần vòng tròn chu vi trong Hình cho thấy tất cả các nhật ký đã được diễn giải chính xác. Nó cung cấp một bức tranh rõ ràng về thành tạo của tầng chứa. Nó đã chia tập hợp các nhật ký giếng thành các nhóm theo góc giữa các nhật ký và các thành phần chính của chúng, vì góc giữa các nhật ký tỷ lệ thuận với tương quan mạnh (góc nhỏ có nghĩa là tương quan cao, và ngược lại).

Nhóm đầu tiên: PC1 bao gồm ba đường biểu diễn điện trở suất theo hướng dương. Sự hiện diện đáng kể của các đường biểu diễn điện trở suất trong PC1 cho thấy sự thay đổi về độ bão hòa trong tầng địa chất.

Nhóm thứ hai: Trái ngược với các bản ghi điện trở suất, PC1 bao gồm các bản ghi tia gamma. Mỗi tương quan mạnh giữa bản ghi tia gamma và PC1 có thể cho thấy nồng độ đá phiến cao trong khoảng độ sâu này.

Nhóm thứ ba: Ngược hướng với các đường ghi điện trở suất và cùng hướng với đường ghi tia gamma, bao gồm các đường ghi độ rỗng (NPHI, DT) trên PC1. Sự quan trọng của NPHI và DT trong PC1 phản ánh vai trò độ rỗng của tầng địa chất. Có thể nhóm nhóm 2 và nhóm 3 (GR, NPHI và DT) thành nhóm phân biệt thạch học. Có thể thấy sự tương ứng chặt chẽ giữa kết quả PCA và các đánh giá địa vật lý của giếng X, như minh họa trong hình 2. Có thể phân biệt hiệu quả các tập sét và các khoảng không phải là sét (shale and non-shale) bằng cách sử dụng nhóm phân biệt thạch học. Góc nhỏ giữa các bản ghi DT và NPHI cho thấy mối tương quan mạnh mẽ, chỉ ra rằng các phép đo độ rỗng từ DT phù hợp chặt chẽ với các giá trị độ rỗng tổng cộng thu được từ NPHI. Sự vắng mặt của độ rỗng thứ cấp cho thấy địa tầng cát, vì độ rỗng thứ cấp chủ yếu liên quan đến các tầng đá vôi. Kết luận này phù hợp với các kết quả địa vật lý đối với giếng dầu X, như thể hiện trong Hình 2.

Thành phần chính 1 (PC1) chủ yếu bao gồm các nhật ký giếng ILD, MSFL, SFL, GR, DT và NPHI. Trong PC1, các nhật ký ILD, MSFL và SFL, có trọng số dương, tương quan nghịch với các nhật ký GR, NPHI và DT, có trọng số âm. Kết quả là, điểm PC1 dương tương ứng với các tầng có giá trị ILD, MSFL và SFL cao và giá trị GR, NPHI và DT thấp, trong khi điểm PC1 âm cho thấy điều ngược lại. Do đó, PC1 đóng vai trò như một nhật ký kép, phân biệt đá phiến sét với các vùng bão hòa. Tín hiệu PC1 có thể được sử dụng để xác định các vùng đá phiến sét (tín hiệu âm) và các vùng bão hòa (tín hiệu dương). Ngoài ra, PC1 phân biệt hiệu quả giữa các khoảng đá phiến sét và không phải đá phiến sét, cũng như giữa các trạng thái bão hòa khác nhau (nước hoặc dầu).

Nhóm thứ tư: Bao gồm một nhật ký giếng duy nhất thể hiện nhật ký mật độ trên PC2, gần như thẳng đứng so với PC1, cho thấy mối tương quan yếu. Thành phần chính thứ hai (PC2) chủ yếu bị ảnh hưởng bởi nhật ký ROHB. PC2 rất quan trọng để phân chia tầng địa chất thành hai vùng: đá phiến sét và vùng bão hòa, như thể hiện trong Hình 6. Hình này minh họa sự phân bố độ sâu trên các nhật ký và thành phần.

Kết luận: Các mối tương quan quan sát được giữa các nhật ký giếng khác nhau nhấn mạnh tầm quan trọng của các công cụ này trong việc đánh giá thành tạo và giải thích nhật ký. Nhóm nhật ký giếng đầu tiên, bao gồm nhật ký điện trở suất trên PC1, tương quan với sự biến đổi độ bão hòa trong thành tạo. Nhóm thứ hai, được đặc trưng bởi nhật ký tia gamma trên PC1, cho thấy mối tương quan mạnh mẽ, cho thấy nồng độ đá phiến cao trong khoảng độ sâu được phân tích. Nhóm thứ ba, bao gồm nhật ký độ rỗng (NPHI và DT) trên PC1, làm nổi bật vai trò thiết yếu của độ rỗng trong thành tạo, như được chỉ ra bởi hệ số tải trọng đáng kể của các nhật ký này. Nhóm thứ tư, chỉ được đại diện bởi nhật ký mật độ trên PC2, góp phần hơn nữa vào phân tích. Các mối tương quan mạnh mẽ giữa các nhật ký giếng này có thể là do tính nhạy cảm của chúng đối với các đặc điểm thành tạo chung, ngụ ý rằng các thông số vật lý cơ bản ảnh hưởng đến các phép đo này có liên quan chặt chẽ với nhau.