

SOME APPLICATIONS OF LINEAR ALGEBRA IN IMAGE PROCESSING AND STATISTICS

MỘT SỐ ỨNG DỤNG CỦA ĐẠI SỐ TUYẾN TÍNH TRONG XỬ LÝ ẢNH VÀ THỐNG KÊ

Nguyễn Thị Phương Dung
Học viện Ngân Hàng

ABSTRACT: *Describe in detail the principal component analysis method (PCA). This method is a very effective application of mathematics in data science - specifically in data dimensionality reduction. In this article, the method has been applied to solve some specific problems in order to gain a deeper understanding of PCA method.*

Key word: *Data dimensionality reduction, eigenvalue, private vector.*

TÓM TẮT: *Bài viết mô tả chi tiết phương pháp phân tích thành phần chính (PCA). PCA là phương pháp toán học có ứng dụng rất hiệu quả trong không khoa học dữ liệu - đặc biệt là trong phương pháp giảm chiều dữ liệu. Trong bài báo này, chúng tôi đưa ra một số ví dụ cụ thể, ứng dụng phương pháp PCA trong giảm chiều dữ liệu, đồng thời mô tả một cách tường minh phương pháp này thông qua bài toán cụ thể.*

Từ khóa: *Giảm chiều dữ liệu, giá trị riêng, vector riêng.*

1. GIỚI THIỆU

Các công nghệ xử lý hình ảnh khác nhau đã cách mạng hoá việc chụp ảnh, lưu trữ hoặc truy xuất một hình ảnh cụ thể từ một nguồn nhất định. Các tác vụ xử lý hình ảnh được sử dụng phổ biến nhất bao gồm: khôi phục hình ảnh; phân đoạn hình ảnh; nâng cao chất lượng hình ảnh; khử mờ; khử nhiễu... Những hình ảnh như vậy được xử lý chất lượng tốt lên bằng nhiều cách khác nhau bằng một số kỹ thuật hình ảnh như chụp MRI (chụp cộng hưởng từ), chụp cắt lớp. Trên thực tế, công cụ đắc lực nhất để xử lý chất lượng hình ảnh một cách tốt nhất chính là dựa vào công cụ then chốt - Ứng dụng của Toán học. Toán học đóng một vai trò quan trọng trong các tác vụ xử lý hình ảnh khác nhau. Bất chấp sự đổi mới và tiến bộ nhanh chóng trong các công nghệ

hình ảnh khác nhau, một điều vẫn quan trọng xuyên suốt, đó là sử dụng Toán học. Người ta quan sát thấy rằng xử lý hình ảnh có mối quan hệ vô cùng chặt chẽ với Toán học. Nhiều phương pháp xử lý ảnh dựa trên các kỹ thuật toán học cơ bản về cân bằng biểu đồ, xác suất và thống kê, tích phân, ma trận và đặc biệt là rất nhiều ứng dụng của lĩnh vực Đại số tuyến tính được sử dụng.

Phương pháp phân tích thành phần chính là một thuật toán sử dụng các nguyên lý toán học phức tạp để biến đổi một dãy số có tương quan với nhau thành một dãy có số biến nhỏ hơn - dãy đó được gọi là các thành phần chính. Trong phương pháp PCA, thông tin chứa trong một tập hợp dữ liệu, được lưu trữ với kích thước giảm dần dựa trên phép chiếu của tập dữ liệu lên một không gian con được

tạo bởi một cơ sở trực giao. Việc tính toán phải đảm bảo sao cho các đặc điểm quan trọng của dữ liệu phải được giữ lại và số thông tin bị mất phải ít và là phần không quan trọng của dữ liệu. Việc giảm chiều dữ liệu như vậy là một lợi thế trong một số lĩnh vực như nén ảnh, biểu diễn dữ liệu. Nó còn được sử dụng rộng rãi trong việc trích xuất đặc trưng, hợp nhất ảnh, nén ảnh, phân đoạn ảnh, đăng ký ảnh, khử nhiễu ... [1].

Phương pháp phân tích thành phần chính PCA bao gồm các bước cơ bản sau đây:

- Thu thập (chuẩn bị) dữ liệu, chuẩn hoá dữ liệu.
- Tính toán ma trận hiệp phương sai của dữ liệu ban đầu.
- Tìm các thành phần chính của dữ liệu thông qua tìm các giá trị riêng và các vector riêng của ma trận hiệp phương sai.
- Chọn số lượng thành phần chính thích hợp để giữ lại nhằm mục đích giữ lại đáng kể các thông tin của dữ liệu ban đầu.
- Biến đổi dữ liệu ban đầu sang không gian mới được tạo nên bởi các thành phần chính mà ta giữ lại.

Bài viết này nêu chi tiết phương pháp - thuật toán phân tích thành phần chính (PCA) trong xử lý ảnh, đồng thời sẽ trình bày công cụ và kỹ thuật toán học cơ bản mà ta sẽ dùng trong phương pháp PCA.

2. THUẬT TOÁN PHÂN TÍCH THÀNH PHẦN CHÍNH

Xuất phát từ mỗi bức ảnh trên thực tế được mô tả bởi hàng triệu điểm (pixel), mỗi điểm ảnh là bộ gồm 3 tọa độ mô tả mật độ màu (Red, blue, Green). Vì vậy,

nếu đem xử lý bằng thuật toán trên ma trận, thì ta có ma trận có kích thước quá lớn, gây khó khăn cho việc xử lý ảnh. Mục tiêu, giảm chiều ma trận bằng phương pháp PCA. Trước hết, ta cần có một số khái niệm sau đây:

- Trung bình và Hiệp phương sai:

Cho $[X_1, X_2, \dots, X_N]$ là ma trận cấp $p \times N$ của các quan sát, ma trận trung bình được định nghĩa như sau:

$$M = \frac{1}{N} \sum_{i=1}^N X_i ; \widehat{X}_k = X_k - M.$$

Khi đó, ta được ma trận cấp $p \times N$: $B = [\widehat{X}_1, \widehat{X}_2, \dots, \widehat{X}_N]$. B có ma trận trung bình là ma trận 0. B được gọi là ma trận dạng độ lệch trung bình.

Ma trận hiệp phương sai, được định nghĩa: $S = \frac{1}{N-1} B \cdot B^T$, với B^T là ma trận chuyển vị của ma trận B. Dễ dàng chứng minh được S là ma trận đối xứng.

Trên quan điểm của xác suất ta coi $X = [x_1, x_2, \dots, x_N]$, trong đó các x_i là các biến ngẫu nhiên nhận các giá trị trên tập các tọa độ thứ i của các vector quan sát X_i . Khi đó, ta ký hiệu $s_{ij} = \text{var}(x_i)$. Tổng var của dữ liệu là:

Total var = trace (S) là tổng các phần tử trên đường chéo chính của S.

Ký hiệu $S_{ij} := \text{cov}(x_i, x_j)$ là hiệp phương sai của x_i và x_j , được hiểu bằng đo mối tương quan giữa chúng.

Tiếp theo, chúng tôi sẽ trình bày thuật toán PCA và ứng dụng của nó trong phương pháp giảm chiều dữ liệu.

- Thuật toán PCA

Giả sử ta có tập các vector quan sát cấu thành ma trận $[X_1, X_2, \dots, X_N]$, cấp $p \times N$, mục tiêu tìm một ma trận trực giao

$P = [u_1, u_2, \dots, u_p]$, sao cho

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix} = [u_1, u_2, \dots, u_p] \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{bmatrix}.$$

Với tính chất các $\text{cov}(y_i, y_j) = 0$ và chúng được sắp xếp theo thứ tự phương sai giảm dần. Để thấy ma trận hiệp phương sai của $[Y_1, Y_2, \dots, Y_N]$ là P^TSP mà P^TSP lại là ma trận chéo hoá được. Nên tồn tại ma trận đường chéo D để $P^TSP = D$. Các vector riêng (đơn vị) của S được gọi là các thành phần chính của dữ liệu. Thành phần chính thứ nhất là vector riêng ứng với giá trị riêng lớn nhất.

Thành phần chính u_1 thứ nhất xác định vector y_1 theo cách sau đây: nếu $u_1 = [c_1, c_2, \dots, c_p]$, u_1^T là hàng đầu tiên của ma trận P^T , khi đó:

$$y_1 = u_1^T X = c_1 x_1 + c_2 x_2 + \dots + c_p x_p.$$

• Ứng dụng của phương pháp PCA trong giảm chiều dữ liệu

PCA là phương pháp có giá trị tiềm năng trong hầu hết các tập dữ liệu lớn với mục tiêu giảm từ dữ liệu có chiều lớn về tập dữ liệu có chiều nhỏ hơn nhưng vẫn gần như là giữ lại hầu hết các nội dung bản chất của tập hợp dữ liệu ban đầu.

Từ cách xây dựng ma trận Y , ta có Tổng các phương sai (varian) của $x_i =$ Tổng các phương sai của $y_i = \text{trace}(D) = \sum_{i=1}^p \lambda_i$.

Trong đó $\text{var}(y_j) = \lambda_j$, với $\frac{\lambda_j}{\text{trace}(S)}$ là phần trăm giải thích dữ liệu ban đầu qua thành phần y_j .

Để đơn giản hoá dữ liệu, ta thường tính toán ma trận hiệp phương sai của hầu hết các vector quan sát ban đầu (ma trận

tọa độ quang phổ của bức ảnh có tới hàng triệu pixel), đưa tới một ma trận cấp $3 \times N$. Điều này làm đơn giản cho việc tính toán của chúng ta, giúp người đọc dễ dàng hiểu và tiếp cận phương pháp. [2]

Dưới đây, chúng tôi đưa ra một số ví dụ để minh họa phương pháp PCA.

Ví dụ mô tả chi tiết phương pháp phân tích thành phần chính:

Một hình ảnh vệ tinh với 3 thành phần quang phổ (red, blue, green) được cung cấp bởi căn cứ không quân Homestead ở Florida. Ma trận hiệp phương sai của hình ảnh đó được cho sau đây:

$$S = \begin{bmatrix} 164.12 & 32.73 & 81.04 \\ 32.73 & 539.44 & 249.13 \\ 81.04 & 249.13 & 189.11 \end{bmatrix}.$$

Hãy tìm thành phần chính mô tả hầu hết dữ liệu của bức ảnh?

• Bước 1: Tìm giá trị riêng của ma trận hiệp phương sai $\det(A - \lambda I_3) = 0$.

$$\det(A - \lambda I_3) = 0 \Leftrightarrow \begin{cases} \lambda_1 = 677.4971 \\ \lambda_2 = 181.5811 \\ \lambda_3 = 33.5918. \end{cases}$$

+ Vector riêng ứng với các giá trị riêng lớn nhất là:

$U_1(x, y, z)$ là vector riêng ứng với giá trị riêng λ_1

$$A u_1 = \lambda_1 u_1 \Leftrightarrow$$

$$\begin{cases} 164.12x + 32.73y + 81.04z = 677.4971x \\ 32.73x + 539.44y + 249.13z = 677.4971y \\ 81.04x + 249.13y + 189.11z = 677.4971z \end{cases}$$

Hệ có nghiệm $(x, y, z) = t(0.277; 1.8702; 1)$ với mọi số thực t khác 0.

• Bước 2: Tìm phần trăm lớn nhất biểu diễn ảnh:

Phần trăm biểu diễn ảnh lớn nhất là

$$\frac{\lambda_1}{\text{trace}(S)} = \frac{677.4971}{892.67} = 75.8956\%.$$

Vậy, không gian vector con sinh bởi vector riêng ứng với giá trị riêng thứ nhất giữ lại được gần 75.89% nội dung chính của bức ảnh sau khi dùng phương pháp phân tích thành phần chính PCA.

Ta xét tiếp một ví dụ thứ 2:

Giả sử ta có ma trận hiệp phương sai của bức ảnh ban đầu như sau:

$$S = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2533.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}.$$

Hãy tìm các thành phần chính của S, hãy đưa ra biến mới xác định bởi thành phần chính thứ nhất?

- Bước 1: Xác định các giá trị riêng của ma trận hiệp phương sai trên bằng cách giải hệ phương trình:

$$\det(S - \lambda I_3) = 0 \Leftrightarrow \lambda_1 = 7614.23; \lambda_2 = 427.63; \lambda_3 = 98.10.$$

- Bước 2: sắp xếp các giá trị riêng theo thứ tự giá trị giảm dần.

- Bước 3: Tìm các vector riêng ứng với các giá trị riêng trên, bằng cách giải đồng thời 3 hệ phương trình:

$$S.v_1 = \lambda_1 v_1; S.v_2 = \lambda_2 v_2; S.v_3 = \lambda_3 v_3.$$

Giải 3 hệ phương trình trên, ta thu được

$$v_1 = \begin{bmatrix} 0.5417 \\ 0.6295 \\ 0.5570 \end{bmatrix}; v_2 = \begin{bmatrix} -0.4894 \\ -0.3026 \\ 0.81790 \end{bmatrix}; v_3 = \begin{bmatrix} 0.6834 \\ -0.7157 \\ 0.1441 \end{bmatrix}.$$

Vậy biến mới, xác định bởi thành phần chính thứ nhất là:

$$Y_1 = 0.54 x_1 + 0.63 x_2 + 0.56 x_3.$$

Qua ví dụ trên ta thấy, từ dữ liệu ban đầu của bức ảnh, nếu mô tả dữ liệu dưới dạng ma trận, thì ma trận đó có kích thước vô cùng lớn và để xử lý được nó thật không dễ dàng, nhưng chúng ta dùng

phương pháp PCA, thì dữ liệu ban đầu chỉ cần tập trung vào thành phần chính thứ nhất của ma trận Hiệp phương sai có chiều thấp đã mô tả được dữ liệu ban đầu.

3. KẾT LUẬN

Giảm chiều dữ liệu là một trong những kỹ thuật quan trọng của machine learning. Các vector đặc trưng của dữ liệu trong các bài toán thực tế thường có số chiều rất lớn, ngoài ra còn có rất nhiều các vector ta cần quan tâm khi nghiên cứu dữ liệu. Nếu thực hiện lưu trữ và tính toán trên các dữ liệu có số chiều lớn như thế này thì thực sự là rất khó khăn với thuật toán và máy tính. Vì vậy, người ta đưa ra phương pháp PCA nhằm mục đích giảm chiều dữ liệu ban đầu về dữ liệu có số chiều nhỏ hơn rất là nhiều, điều đó giúp cho việc lưu trữ và tính toán trên dữ liệu một cách dễ dàng và hiệu quả hơn nhiều.

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.

Phương pháp PCA giúp ta đưa ra được thành phần dữ liệu quan trọng nhất, nó giữ lại hầu hết thông tin của dữ liệu ban đầu.

TÀI LIỆU THAM KHẢO

- [1] Jolliffe IT. (2002), *Principal Component Analysis*, New York: Springer.
- [2] Saporta G and Niang N. (2009), *Principal component analysis: application to statistical process control*, In: Govaert G, ed. *Data Analysis*. London: John Wiley & Sons. 1-23.
- [3] Jolliffe IT. 2002, *Principal Component Analysis*, New York: Springer.
- [4] Strang G. (2003) *Introduction to Linear Algebra*, Cambridge, MA: Wellesley-Cambridge Press.

Liên hệ:

TS. Nguyễn Thị Phương Dung

Bộ môn toán - Học viện Ngân hàng.

Địa chỉ: Số 12 Chùa Bộc, Quang Trung, Đống Đa, Hà Nội

Email: dungnp@hvn.edu.vn

Ngày nhận bài: 22/3/2023

Ngày gửi phản biện: 27/3/2023

Ngày duyệt đăng: 19/02/2024