BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUY NHƠN

NGUYỄN NHẬT NAM

Một số phương pháp phân tích ma trận và ứng dụng

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

Ngành: Toán ứng dụng
Chuyên ngành: Khoa học dữ liệu

Người hướng dẫn: TS Lâm Thị Thanh Tâm

Bình Định, 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC QUY NHƠN

Một số phương pháp phân tích ma trận và ứng dụng

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

Ngành: Toán ứng dụng
Chuyên ngành: Khoa học dữ liệu

Sinh viên thực hiện: NGUYỄN NHẬT NAM
Mã số SV: 4251140002
Lớp, Khóa: Toán ứng dụng K42

Người hướng dẫn: TS Lâm Thị Thanh Tâm

Bình Định, 2023

Lời cảm ơn

Trước khi đi vào nội dung của luận văn, em xin bày tỏ lòng biết ơn chân thành tới toàn thể giáo viên của Trường Đại học Quy Nhơn nói chung và của Khoa Toán-Thống kê nói riêng vì đã tận tâm dạy bảo em trong suốt quá trình em theo học tại trường, và đặc biệt là sự định hướng dẫn dắt của TS Lâm Thị Thanh Tâm đối với đề tài khóa luận này.

Nhân dịp này em cũng xin được gửi những lời cảm ơn đến những người bạn, đặc biệt là những người thân trong gia đình đã luôn giúp đỡ em hết mình, luôn động viên, cổ vũ tinh thần và tạo điều kiện thuận lợi giúp em có thể hoàn thành được khoá luận tốt nghiệp này.

Trong quá trình viết báo cáo này mặc dù đã được chỉnh sửa nhiều lần nhưng không thể tránh khỏi việc thiếu sót và gây cho người đọc cảm giác khó hiểu. Em xin chân thành cảm ơn nếu nhận được sự góp ý từ các quý thầy cô, anh chị và bạn bè để có thể chỉnh sửa luận văn được tốt hơn.

Quy Nhơn, ngày ... tháng ... năm 2023

Sinh viên thực hiện

Nguyễn Nhật Nam

Mục lục

Lời mở đầu

Trong thời đại công nghệ thông tin phát triển mạnh mẽ ngày nay, việc thu thập và xử lý dữ liệu ngày càng trở nên quan trọng và cần thiết. Các phương pháp phân tích dữ liệu như phân rã giá trị suy biến (SVD) và phân tích thành phần chính (PCA) đóng vai trò quan trọng trong việc giải quyết các vấn đề liên quan đến dữ liệu lớn và đa chiều. Những phương pháp này không chỉ giúp giảm kích thước dữ liệu mà còn giúp khai thác thông tin hữu ích từ dữ liệu gốc. SVD và PCA đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực khoa học và kỹ thuật, chẳng hạn như xử lý hình ảnh, phân tích dữ liệu, nhận dạng mẫu, thống kê, v.v.

Trong luận văn này, chúng tôi tập trung nghiên cứu về các phương pháp phân tích dữ liệu như phân rã giá trị suy biến (SVD) và phân tích thành phần chính (PCA), cũng như một số ứng dụng của chúng trong các bài toán thực tế. Mục tiêu chính của luận văn là tìm hiểu về các tính chất cơ bản, cách tính và ứng dụng của SVD và PCA trong việc giải quyết các vấn đề liên quan đến dữ liệu.

Đối tượng nghiên cứu trong luận văn này là phương pháp phân rã giá trị suy biến (SVD) và phân tích thành phần chính (PCA). Ngoài lời cảm ơn, mục lục, lời mở đầu và một số ký hiệu, luận văn được chia làm bốn chương chính:

Chương 1: Kiến thức cơ bản. Trong chương này, chúng tôi trình bày một số kiến thức cơ sở về đại số tuyến tính, giải tích và thống kê, nhằm chuẩn bị cho việc tìm hiểu về SVD và PCA.

Chương 2: Phân rã giá trị suy biến (SVD). Trong chương này, chúng tôi giới thiệu về SVD, cách tính SVD và mối quan hệ giữa SVD và giá trị riêng của ma trận. Ngoài ra, chúng tôi cũng trình bày một số ứng dụng của SVD trong việc giải quyết các bài toán thực tế.

Chương 3: Phân tích thành phần chính (PCA). Trong chương này, chúng tôi giới thiệu về PCA, cách tính PCA và mối quan hệ giữa PCA và SVD. Chúng tôi cũng sẽ trình bày một số ứng dụng của PCA trong việc giảm kích thước dữ liệu, phân tích dữ liệu và các bài toán khác.

Chương 4: Ứng dụng của SVD và PCA trong các bài toán thực tế. Chương này sẽ trình bày một số ví dụ cụ thể về việc áp dụng SVD và PCA trong các bài toán thực tế như xử lý hình ảnh, nhận dạng mẫu, phân tích dữ liệu và thống kê.

Luận văn tốt nghiệp này được hoàn thành dưới sự hướng dẫn của TS. lâm Thị Thanh Tâm. Nhân dịp này, tôi xin bày tỏ lòng biết ơn sâu sắc đến thầy/cô hướng dẫn, đã không chỉ hỗ trợ tôi trong việc nghiên cứu khoa học mà còn tận tình giúp đỡ và tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình làm đề tài. Tôi cũng xin chân thành cảm ơn tập thể lớp Toán Ứng Dụng K42, khoa Toán và THống kê Trường Đại học Quy Nhơn đã giúp đỡ và tạo mọi điều kiện thuận lợi cho tôi hoàn thành khóa học cùng với luận văn này.

Cuối cùng, tôi xin chân thành cảm ơn gia đình, bạn bè, những người thân yêu đã luôn quan tâm, giúp đỡ và ủng hộ tôi trong

suốt quá trình học tập và thực hiện luận văn này. Dù đã cố gắng hết sức, nhưng do thời gian và năng lực có hạn, luận văn không tránh khỏi những thiếu sót và hạn chế. Rất mong nhận được sự góp ý và chỉ bảo của quý thầy cô, quý bạn đồng nghiệp để luận văn tốt nghiệp này được hoàn thiện hơn. Tôi xin chân thành cảm ơn.

Một số ký hiệu

$\mathbb{N}$	**tập các số tự nhiên**
$\mathbb{Z}$	**tập các số nguyên**
$\mathbb{Q}$	**tập các số hữu tỷ**
$\mathbb{R}$	**tập các số thực**
$\mathbb{C}$	**tập các số phức**
$\overline{\mathbb{R}}$	**tập các số thực mở rộng**
$\mathbb{R}^n$	**không gian vectơ thực n-chiều**
$\mathbb{K}$	**tập $\mathbb{R}$ hoặc $\mathbb{C}$**
$\mathcal{P}(X)$	**tập tất cả các tập con của X**
sign	**hàm dấu**
$A \triangle B$	**hiệu đối xứng của hai tập A và B**
$\bigsqcup_{i=1}^n A_i$	**hợp rời nhau các tập $A_1, \ldots, A_n$**
$f_n \uparrow$	**dãy f_n đơn điệu tăng**
$f_n \downarrow$	**dãy f_n đơn điệu giảm**
$f_n \nearrow f$	**dãy f_n đơn điệu tăng và hội tụ đến f**
$f_n \rightrightarrows f$	**dãy f_n hội tụ đều đến f**
$\mathcal{L}$	**σ-đại số Lebesgue các tập con của $\mathbb{R}$**
m	**độ đo Lebesgue**
$\sharp$	**độ đo đếm**

Chương 1

Một số kiến thức chuẩn bị

1.1 Ma trận

.

Định nghĩa 1.1.1. Ma trận cỡ $m \times n$ là một bảng gồm mn số thực được sắp xếp thành m dòng và n cột. Ma trận thường được kí hiệu như sau

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \vdots \\ a_{m1} & a_{m2} & a_{13} & \cdots & \cdots & \cdots & a_{mn} \end{bmatrix},$$

hoặc

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \vdots \\ a_{m1} & a_{m2} & a_{13} & \cdots & \cdots & \cdots & a_{mn} \end{pmatrix},$$

hoặc $\mathbf{A} = (a_{ij})_{m\times n}$, trong đó a_{ij} là phần tử của ma trận nằm trên dòng i, cột j, với $i = 1, 2, \ldots, m$, và $j = 1, 2, \ldots, n$.

Khi $m = n$, ta gọi ma trận cỡ $m \times m$ là ma trận vuông cấp m. Các phần tử a_{11}, $a_{22}, \ldots, a_{mm}$ nằm trên một đường thẳng được gọi đường chéo chính của ma trận.

Định nghĩa 1.1.2. Ma trận đơn vị cấp n là ma trận vuông cấp n có mọi phần tử nằm trên đường chéo chính bằng 1, các phần tử khác bằng 0 . Ta ký hiệu ma trận đơn vị cấp n bởi $\mathbf{I}_n$ và nó có dạng như sau

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Trong trường hợp không cần chú ý đến cấp của ma trận, ta ký hiệu ma trận đơn vị bởi $\mathbf{I}$.

Định nghĩa 1.1.3. Ma trận đường chéo là ma trận vuông có các phần tử nằm trên đường chéo chính khác 0, các phần tử nằm ngoài đường chéo chính bằng 0 . Ma trận đường chéo có dạng như sau

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{mm} \end{bmatrix}$$

Ma trận chỉ có một dòng được gọi là vector dòng. Ma trận chỉ có một cột được gọi là vector cột.

Định nghĩa 1.1.4. Ma trận vuông A cấp n được gọi là ma trận khả nghịch nếu tồn tại một ma trận $\mathbf{A}'$ vuông cấp n thỏa mãn $\mathbf{A}\mathbf{A}' = \mathbf{A}'\mathbf{A} = \mathbf{I}_n$. Ma trận $\mathbf{A}$ được gọi là ma trận nghịch đảo của ma trận $\mathbf{A}$, và được ký hiệu là $\mathbf{A}^{-1}$

Định nghĩa 1.1.5. Cho ma trận $\mathbf{A} = (a_{ij})_{m\times n}$. Ma trận chuyển vị của $\mathbf{A}$, ký hiệu là $\mathbf{A}^T$, có dạng $\mathbf{A}^T = (a_{ji})_{n\times m}$.

Ma trận vuông $\mathbf{A}$ được gọi là ma trận đối xứng nếu $\mathbf{A}^T = \mathbf{A}$, và được gọi là ma trận phản đối xứng nếu $\mathbf{A}^T = -\mathbf{A}$.

Định nghĩa 1.1.6. Ma trận vuông $\mathbf{A}$ được gọi là ma trận trực giao nếu $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$

Nhận xét: (i) Ma trận $\mathbf{A} = [a_{ij}]_{n\times n}$ là ma trận trực giao khi và chỉ khi $\sum_{k=1}^{n} a_{ik}a_{jk} = \delta_{ij}$, với δ_{ij} là kí hiệu Kronecker.

(ii) Ma trận trực giao $\mathbf{A}$ là khả nghịch và $\mathbf{A}^T = \mathbf{A}^{-1}$.

(iii) Ma trận $\mathbf{A}$ trực giao khi và chỉ khi các vector cột và các vector hàng của $\mathbf{A}$ tạo thành các hệ trực chuẩn.

Định nghĩa 1.1.7. Cho $\mathbf{A}$ là một ma trận vuông cấp n. Định thức của ma trận $\mathbf{A}$, ký hiệu là $\det(\mathbf{A})$ hay $|\mathbf{A}|$ là một giá trị được xác định bằng công thức

$$\det(\mathbf{A}) = a_{11}\mathbf{A}_{11} + a_{12}\mathbf{A}_{12} + \cdots + a_{1n}\mathbf{A}_{1n}$$

trong đó $\mathbf{A}_{ik} = (-1)^{i+k} \det(\mathbf{M}_{ik})$, với $\mathbf{M}_{ik}$ là ma trận vuông cấp $n-1$ nhận được từ ma trận $\mathbf{A}$ bằng cách bỏ đi dòng thứ i và cột thứ k. Đại lượng $\mathbf{A}_{ik}$ được gọi là phần bù đại số của a_{ik}.

Nhận xét:

- Định thức cấp một: Nếu $\mathbf{A} = (a_{11})$ thì $\det(\mathbf{A}) = a_{11}$.
- Định thức cấp hai: Nếu $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ thì $\det(\mathbf{A}) = ad - bc$.

Định nghĩa 1.1.8. Cho $\mathbf{A}$ là ma trận cỡ $m \times n$ bất kì và s là số nguyên thỏa $1 \leqslant s \leqslant \min(m, n)$. Khi đó, các phần tử nằm trên giao của s dòng và s cột của ma trận $\mathbf{A}$ sẽ lập nên các ma trận vuông cấp s, ta gọi đó chính là các ma trận con cấp s của $\mathbf{A}$.

Định thức của các ma trận này được gọi là định thức con cấp s của ma trận $\mathbf{A}$.

Định nghĩa 1.1.9. Định thức con cấp cao nhất, khác 0 của ma trận $\mathbf{A}$ được gọi là định thức con cơ sở của ma trận $\mathbf{A}$. Một ma trận $\mathbf{A}$ có thể có nhiều định thức con cơ sở cùng cấp.

Hạng của ma trận $\mathbf{A}$ là cấp của định thức con cơ sở. Ký hiệu hạng của ma trận $\mathbf{A}$ là $\text{rank}(\mathbf{A})$.

Nhận xét: Cho $\mathbf{A}$ là ma trận cấp $m \times n, \mathbf{B}$ là ma trận cấp $n \times l$. (i) Nếu $\text{rank}(\mathbf{B}) = n$ thì $\text{rank}(\mathbf{A} \cdot \mathbf{B}) = \text{rank}(\mathbf{A})$.

(ii) Nếu $\text{rank}(\mathbf{A}) = n$ thì $\text{rank}(\mathbf{A}.\mathbf{B}) = \text{rank}(\mathbf{B})$.

Định nghĩa 1.1.10. Vết của ma trận vuông $\mathbf{A}$ cấp n được xác định bằng tổng các phần tử trên đường chéo chính của ma trận $\mathbf{A}$, và được ký hiệu là Tr(A).

1.2 Vector riêng- Giá trị riêng

Định nghĩa 1.2.1. Cho $\mathbf{A}$ là ma trận vuông cấp n. Khi đó đa thức bậc n của biến λ được xác định như sau

$$P_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix}$$

được gọi là đa thức đặc trưng của ma trận $\mathbf{A}$. Các nghiệm của đa thức $P_{\mathbf{A}}(\lambda)$ được gọi là các giá trị riêng của ma trận $\mathbf{A}$.

Vector $\mathbf{u} \in \mathbb{R}^n$ được gọi là vector riêng ứng với giá trị riêng λ của ma trận $\mathbf{A}$ nếu thỏa $\mathbf{Au} = \lambda\mathbf{u}$.

Nhận xét:

(i) Nếu λ là một giá trị riêng của $\mathbf{A}$ thì $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Khi đó hệ phương trình thuần nhất

$$(\mathbf{A} - \lambda\mathbf{I}) \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = 0$$

có vô số nghiệm.

(ii) Mỗi giá trị riêng có thể có nhiều vector riêng.

(iii) Mỗi vector riêng chỉ ứng với một giá trị riêng duy nhất.

(iv) Nếu $\lambda = 0$ là một giá trị riêng của ma trận $\mathbf{A}$ thì $\mathbf{A}$ không khả nghịch. Ngược lại, nếu mọi giá trị riêng của $\mathbf{A}$ đều khác 0 thì ma trận $\mathbf{A}$ khả nghịch.

1.3 Định lí phổ của ma trận đối xứng

Định nghĩa 1.3.1. Cho $\mathbf{A}$ là ma trận đối xứng cấp n. Khi đó

i) Với mỗi giá trị riêng thực λ của $\mathbf{A}$, tồn tại một vector riêng tương ứng $\mathbf{u} \in \mathbb{R}^n$ sao cho $\mathbf{Au} = \lambda\mathbf{u}$.

ii) Tồn tại ma trận dường chéo $\mathbf{D}$ cấp n và ma trận trực giao $\mathbf{U}$ cấp n sao cho $\mathbf{A} = \mathbf{UDU}^T$, trong đó các phần tử nằm trên đường chéo chính của $\mathbf{D}$ là các giá trị

riêng của $\mathbf{A}$, và các vector cột của $\mathbf{U}$ là các vetor riêng của $\mathbf{A}$ tương ứng với các giá trị riêng đó. Tức là, nếu

$$\mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

và

$$\mathbf{U} = \left[\begin{array}{c|ccc} \mathbf{u}^{(1)} & \mathbf{u}^{(2)} & \ldots & \mathbf{u}^{(n)} \end{array} \right]$$

thì $\mathbf{A}\mathbf{u}^{(i)} = \lambda_i \cdot \mathbf{u}^{(i)}, \quad i = 1, 2, \ldots, n.$

Chứng minh. Ta sẽ chứng minh định lí này bằng phương pháp quy nạp toán học. Trường hợp $n = 1$, kết quả trên đúng. Giả sử kết quả trên đúng với mọi ma trận có cấp nhỏ hơn hoặc bằng $n - 1$, ta sẽ chứng minh kết quả trên đúng trong trường hợp ma trận $\mathbf{A}$ là ma trận đối xứng cấp n.

Xét hàm số $p(t) = \det(t\mathbf{I} - \mathbf{A})$. Ta có $p(t)$ là một đa thức bậc n và được gọi là đa thức đặc trưng của ma trận $\mathbf{A}$. Theo Định lí cơ bản của đại số, đa thức $p(t)$ sẽ có n nghiệm là $\lambda_1, \lambda_2, \ldots, \lambda n$, và ta gọi chúng là các giá trị riêng của ma trận $\mathbf{A}$.

Giả sử λ là một giá trị riêng của ma trận $\mathbf{A}$, ta có $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$, tức ma trận $(\lambda\mathbf{I} - \mathbf{A})$ không khả nghịch. Điều này có nghĩa là, tồn tại một vector thực, khác không $\mathbf{u}$ sao cho $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$. Ta có thể chuẩn hóa vector $\mathbf{u}$ sao cho $\mathbf{u}^T\mathbf{u} = 1$. Khi đó, $\lambda = \mathbf{u}^T\mathbf{A}\mathbf{u}$ là số thực.

Với λ_1 là một giá trị riêng của $\mathbf{A}$ và $\mathbf{u}_1$ là một vector riêng tương ứng. Sử dụng phép trực giao hóa Gramm-Schmidt, ta có thể tìm được ma trận $\mathbf{V}_1$ cấp $n \times (n-1)$ sao cho $[\mathbf{u}_1\mathbf{V}_1]$ là một ma trận trực giao. Ta có $\mathbf{V}_1^T\mathbf{A}\mathbf{V}_1$ là ma trận đối xứng cấp $n - 1$. Khi đó theo giả thiết quy nạp, ta có thể viết $\mathbf{V}_1^T\mathbf{A}\mathbf{V}_1 = \mathbf{Q}_1\mathbf{D}_1\mathbf{Q}_1^T$, trong đó $\mathbf{D}_1 = \text{diag}\,(\lambda_2, \lambda_3, \ldots, \lambda_n)$ là ma trận đường chéo với các phần tử nằm trên đường chéo chính là $n - 1$ giá trị riêng của $\mathbf{A}$ và $\mathbf{Q}_1$ là ma trận trực giao cấp $(n-1)$ gồm $(n-1)$ vector riêng của $\mathbf{V}_1^T\mathbf{A}\mathbf{V}_1$ tương ứng.

Ta định nghĩa ma trận $\mathbf{U}_1$ cấp $n \times (n-1)$ bởi $\mathbf{U}_1 = \mathbf{V}_1\mathbf{Q}_1$. Khi đó $\mathbf{U} = [\mathbf{u}_1, \mathbf{U}_1]$ là ma trận trực giao. Ta có

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{U}_1^T \end{pmatrix} \mathbf{A}\,(\mathbf{u}_1 - \mathbf{U}_1) = \begin{pmatrix} \mathbf{u}_1^T\mathbf{A}\mathbf{u}_1 & \mathbf{u}_1^T\mathbf{A}\mathbf{U}_1 \\ \mathbf{U}_1^T\mathbf{A}\mathbf{u}_1 & \mathbf{U}_1^T\mathbf{A}\mathbf{U}_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \mathbf{D}_1 \end{pmatrix} = \mathbf{D}.$$

Điều này chứng tỏ $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, với $\mathbf{U}$ là ma trận trực giao cấp n và $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Đây được gọi là phân tích giá trị riêng của ma trận $\mathbf{A}$ (Eigen value decomposition, EVD).

Chương 2

PHÂN TÍCH GIÁ TRỊ KÌ DỊ

Trong chương này, chúng tôi trình bày cơ sở của phân tích giá trị kì dị của ma trận, thuật toán phân tích giá trị kì dị, và trình bày một số tính chất của ma trận liên quan đến phân tích giá trị kì dị của nó.

2.1 Giới thiệu

Phân tích giá trị kì dị (SVD) là một trong những phân tích rất quan trọng của ma trận, có nhiều ứng dụng trong khoa học và kĩ thuật. Dạng cổ điển nhất của phân tích giá trị kì dị được phát hiện trong lĩnh vực Hình học vi phân []. Vào năm 1873 và 1874 , hai nhà toán học Eugenio Beltrami và Camille Jordan đã độc lập đưa ra giá trị kì dị của dạng song tuyến tính. Năm 1889, James Joshept Sylvester đã đưa ra phân tích giá trị kì dị của ma trận vuông thưc. Đến năm 1915, nhà toán học Autonne phát hiện ra phân tích giá trị kì dị dưới dạng phân tích cực. Năm 1936, hai nhà toán học Carl Eckart và Gale Young đã lần đầu tiên chứng minh phân tích giá trị kì dị đối với ma trận hình chữ nhật thực [] và ma trận vuông phức []. Trong chương này, chúng tôi sẽ tìm hiểu về phân tích giá trị kì dị cho một ma trận hình chữ nhật thực $\mathbf{A}$ cỡ $m \times n$.

2.2 Phân tích giá trị kì dị

Để phát biểu và chứng minh định lý phân tích giá trị kì dị, chúng ta sẽ cần định lí sau đây.

Định lý 2.2.1. *Cho A là một ma trận có cấp $m \times n$. Khi đó, mọi giá trị riêng của ma trận $\mathbf{A}^T\mathbf{A}$ đều không âm.*

Chứng minh. Dễ thấy $\mathbf{A}^T\mathbf{A}$ là một ma trận đối xứng. Theo Định lí 1.3.1, mọi giá trị riêng của $\mathbf{A}^T\mathbf{A}$ đều là số thực. Với mỗi giá trị riêng λ của $\mathbf{A}^T\mathbf{A}$, giả sử $\mathbf{v}$ là véc tơ riêng tương ứng. Khi đó

$$\mathbf{A}^T\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

Để ý rằng ta có thể chọn $\mathbf{v}$ là véc tơ đơn vị, i.e. $\|\mathbf{v}\| = 1$. Với cách chọn như vậy, ta nhận được

$$\|\mathbf{A}\mathbf{v}\|^2 = \langle \mathbf{A}\mathbf{v}, \mathbf{A}\mathbf{v}\rangle = (\mathbf{A}\mathbf{v})^T(\mathbf{A}\mathbf{v}) = \mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda\|\mathbf{v}\|^2 = \lambda$$

Do vậy, $\lambda \geqslant 0$. Ta có định nghĩa và cách xác định các giá trị kì dị của một ma trận như sau

Định nghĩa 2.2.1. Cho ma trận $\mathbf{A}$ cỡ $m \times n$ bất kì ($m \leqslant n$) và $\lambda_1, \lambda_2 \ldots, \lambda_r$, $1 \leqslant r \leqslant \min(m, n)$, lần lượt là các giá trị riêng của ma trận $\mathbf{A}^T\mathbf{A}$. Các giá trị $\sigma_i = \sqrt{\lambda_i}, 1 \leqslant i \leqslant r$, được gọi là các giá trị kì dị của ma trận $\mathbf{A}$.

Định lí sau cho ta sự tồn tại của SVD của một ma trận bất kỳ

Định lý 2.2.2. *Mọi ma trận* A *cỡ $m \times n$ bất kì đều có phân tích SVD có dạng*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

trong đó $\mathbf{U}$ và $\mathbf{V}$ lần lượt là các ma trận trực giao cấp m và n,

D là ma trận đường chéo cỡ $m \times n$ có dạng

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & 0 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \sigma_2 & 0 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \sigma_r & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \ldots & \ldots & 0 \end{bmatrix}$$

Chứng minh. Vì $\mathbf{A}^T\mathbf{A}$ là ma trận đối xứng nên theo Định lí 2.2.1 n giá trị riêng $\lambda_1, \ldots, \lambda_n$ của nó đều không âm. Do đó tồn tại $r \leqslant n$ sao cho $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_r > 0$ và

$\lambda_j = 0$ với $j > r$. Khi đó ma trận $\mathbf{A}$ có các giá trị kì dị là $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r \geqslant 0$, và $\sigma_j = 0$ với $j > r$.

Chọn $\mathbf{v}_i \in \mathbb{R}^n, i = 1, 2, \ldots, n$, là các vetor riêng tương ứng với λ_i sao cho $\mathbf{v}_i$ là các vetor đơn vị. Đặt $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, với $\mathbf{v}_i$ là các vector cột. Khi đó $\mathbf{V}$ là ma trận trực giao.

Với $\sigma_i, 1 \leqslant i \leqslant r$, là các giá trị kì dị dương của $\mathbf{A}$, đặt $\mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{\sigma_i}, i = 1, 2, \ldots, r$. Khi đó $\mathbf{u}_i$ là vector đơn vị trong $\mathbb{R}^m$.

Xây dựng ma trận $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$, với $\mathbf{u}_j = 0, r < j \leqslant m$. Ta có $\mathbf{U}$ là ma trận trực giao.

Ta chứng minh $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, hay $\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{D}$.

Ta có

$$\begin{aligned}\mathbf{A}\mathbf{V} &= \mathbf{A}\left[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\right] = \left[\mathbf{A}\mathbf{v}_1, \mathbf{A}\mathbf{v}_2, \ldots, \mathbf{A}\mathbf{v}_n\right] = \left[\mathbf{A}\mathbf{v}_1, \mathbf{A}\mathbf{v}_2, \ldots, \mathbf{A}\mathbf{v}_r, 0, \ldots, 0\right] \\ &= \left[\sigma_1\mathbf{u}_1, \sigma_2\mathbf{u}_2, \ldots, \sigma_r\mathbf{u}_r, 0, \ldots, 0\right] = \left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\right] \cdot \operatorname{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_r\right) \\ &= \mathbf{U}\mathbf{D}\end{aligned}$$

Vậy

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Các vector cột $\mathbf{u}_i$ trong ma trận $\mathbf{U}$ được gọi là các vector kì dị trái của $\mathbf{A}$. Các vetor cột $\mathbf{v}_i$ trong ma trận $\mathbf{V}$ được gọi là các vector kì dị phải của $\mathbf{A}$.

Định lý 2.2.3 (Về dạng khai triển của phân tích giá trị kì dị). *Cho $\mathbf{A}$ là ma trận cỡ $m \times n$ bất kì. Khi đó*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$$

trong đó σ_i là các giá trị kì dị dương của ma trận $\mathbf{A}, \mathbf{u}_i$ là các vector kì dị trái và $\mathbf{v}_i$ là các vector kì dị phải của ma trận $\mathbf{A}$.

Chứng minh. Ta có

$$\mathbf{UDV}^T = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m] \begin{bmatrix} \mathbf{D}_{r\times r} & \ldots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix}$$

$$= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_m] \begin{bmatrix} \mathbf{D}_{r\times r} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix}$$

$$= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \begin{bmatrix} \sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{r+1} & \mathbf{u}_{r+2} & \ldots & \mathbf{u}_m \end{bmatrix} [0] \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix}$$

$$= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \begin{bmatrix} \sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix}$$

$$= [\sigma_1\mathbf{u}_1, \ldots, \sigma_r\mathbf{u}_r] \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T.$$

2.3 Thuật toán tìm SVD của một ma trận

Cho $\mathbf{A}$ là ma trận cỡ $m \times n$, với $m \geq n$. Để tìm SVD của ma trận $\mathbf{A}$, chúng ta thực hiện các bước sau.

- Bước 1. Tính ma trận $\mathbf{A}^T\mathbf{A}$ và giải phương trình $\det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) = 0$ để tìm các giá trị riêng $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ của ma trận $\mathbf{A}^T\mathbf{A}$. Từ đó suy ra các giá trị kì dị

của $\mathbf{A}$ là $\sigma_i = \sqrt{\lambda_i}, i = \overline{1,n}$ và $\mathbf{D} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n)$

- Bước 2. Tương ứng với mỗi giá trị riêng λ_i, tìm véctơ riêng $v_i \in \mathbb{R}^n$ sao cho $\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) v_i = 0$. Từ đó tìm được ma trận trực giao $\mathbf{V}$ cấp n chứa các véctơ kì dị phải của $\mathbf{A}$.

- Bước 3. Xác định các véctơ kì dị trái của $\mathbf{A}$ theo công thức

$$u_i = \frac{1}{\sigma_i}\mathbf{A}v_i$$

Bổ sung $n-r$ véctơ $u_{r+1}, \ldots, u_n$ vào hệ $\{u_1, u_2, \ldots, u_r\}$ sao cho $\{u_1, u_2, \ldots, u_n\}$ lập thành một cơ sở trực chuẩn của $\mathbb{R}^n$. Từ đó nhận được ma trận trực giao $\mathbf{U}$ chứa các véctơ kì dị trái của $\mathbf{A}$, và

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

là phân tích SVD của ma trận $\mathbf{A}$.

Ví dụ

Tìm SVD của ma trận $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$. Lời giải.

Bước 1: Tìm các giá trị kì dị củ ma trận $\mathbf{A}$

Ta có

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

Giải phương trình $\det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) = 0$, ta tìm đượcác giá trị riêng λ của $\mathbf{A}^T\mathbf{A}$ là $\lambda_1 = 2, \lambda_2 = 1$. Do đó các giá trị kì dị của $\mathbf{A}$ là $\sigma_1 = \sqrt{2}, \sigma_2 = 1$. Suy ra

$$\mathbf{D} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}$$

Bước 2: Tìm ma trận $\mathbf{V}$

Giải phương trình $\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right)v = 0$ ta tìm được các véctơ riêng tương ứng là $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ và $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Từ đó suy ra

$$\mathbf{V}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Bước 3: Tìm ma trận **U**

Ta có

$$u_1 = \frac{1}{\sigma_1}Av_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

và

$$u_2 = \frac{1}{\sigma_2}Av_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Do đó

$$\mathbf{U} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix}$$

Vậy phân tích SVD của ma trận **A** là

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

2.4 Một số tính chất của ma trận liên quan đến SVD của nó

Định lý 2.4.1. *Cho ma trận* **A** *cỡ* $m \times n$ *bất kì, có phân tích SVD là*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Khi đó, hạng của $\mathbf{A}$ *đúng bằng số các giá trị kì dị dương của* $\mathbf{A}$.

Chứng minh. Giả sử r là số các giá trị kì dị dương của $\mathbf{A}$. Đặt $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$, $\mathbf{V}_r = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]$. Do tính chất của hạng của ma trận và tính trực giao của $\mathbf{U}, \mathbf{V}$ ta suy ra

$$\begin{aligned}
&\operatorname{rank}(\mathbf{U}_r) = \operatorname{rank}\left(\mathbf{U}_r\mathbf{U}_r^T\right) = \operatorname{rank}(\mathbf{I}_r) = r,\\
&\operatorname{rank}\left(\mathbf{V}_r^T\right) = \operatorname{rank}\left(\mathbf{V}_r^T\mathbf{V}_r\right) = \operatorname{rank}(\mathbf{I}_r) = r.
\end{aligned}$$

Do đó

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}\left(\mathbf{U}\mathbf{D}\mathbf{V}^T\right) = \operatorname{rank}\left(\mathbf{U}_r\mathbf{D}\mathbf{V}_r^T\right) = \operatorname{rank}\left(\mathbf{D}\mathbf{V}_r^T\right) = \operatorname{rank}(\mathbf{D}) = r$$

Định lý 2.4.2. *Cho* A *là một ma trận có cỡ* $m \times n$. *Giả sử* $\mathbf{A}$ *có phân tích* SVD *dưới dạng khai triển là*

$$\mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$$

Khi đó, với mỗi số nguyên dương $k < r$, *ma trận*

$$\mathbf{A}_k = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T$$

có hạng $k, \operatorname{rank}(\mathbf{A}_k) = k$.

Chứng minh. Dễ thấy rằng ma trận $\mathbf{A}_k$ có thể viết ở dạng

$$\mathbf{A}_k = \left(\sigma_1\mathbf{v}_1^T + \cdots + \sigma_k\mathbf{v}_k^T\right)\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \vdots\\ \mathbf{u}_k\end{bmatrix}$$

Ta có

$$\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix}^T = \mathbf{I}_k \text{ và rank } \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix} = \text{rank} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix}^T = k.$$

Xét rank $\left(\sigma_1\mathbf{v}_1^T, \sigma_2\mathbf{v}_2^T, \ldots, \sigma_k\mathbf{v}_k^T\right)$. Ta có

$$\left(\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_k^T\right)(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) = \mathbf{I}_k$$

Do đó

$$\left(\sigma_1\mathbf{v}_1^T, \ldots, \sigma_k\mathbf{v}_k^T\right)(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_k \end{bmatrix} \mathbf{I}_k = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_k \end{bmatrix} = \mathbf{B}$$

Vì rank$(\mathbf{B}) = k = \text{rank}\,(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ nên

$$\text{rank}\left(\sigma_1\mathbf{v}_1^T, \sigma_2\mathbf{v}_2^T, \ldots, \sigma_k\mathbf{v}_k^T\right) = k$$

Từ đó, ta suy ra $\text{rank}\,(\mathbf{A}_k) = k$.

Định lý 2.4.3. *Cho* $\mathbf{A}$ *là ma trận cỡ* $m \times n$ *bất kì và* $\sigma_1, \sigma_2, \ldots, \sigma_r$ *là các giá trị kì dị dương của* $\mathbf{A}$. *Khi đó* $\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}$.

Chứng minh. Giả sử $\mathbf{Q}$ là ma trận trực giao cấp m. Khi đó

$\|\mathbf{QA}\|_F^2 = \|[\mathbf{Qa}_1, \ldots, \mathbf{Qa}_n]\|_F^2 = \|\mathbf{Q_1}\|_F^2 + \cdots + \|\mathbf{Qa}_n\|_F^2 = \|\mathbf{a}_1\|_F^2 + \cdots + \|\mathbf{a}_n\|_F^2 = \|\mathbf{A}\|_F^2$.

Giả sử $\mathbf{A}$ có phân tích SVD là $\mathbf{A} = \mathbf{UDV}^T$. Vì $\mathbf{U}$ và $\mathbf{V}$ là các ma trận trực giao nên

$$\|\mathbf{A}\|_F^2 = \left\|\mathbf{UDV}^T\right\|_F^2 = \left\|\mathbf{DV}^T\right\}\|_F^2 = \left\|\left(\mathbf{DV}^T\right)^T\right\|_F^2 = \left\|\mathbf{VD}^T\right\|_F^2 = \left\|\mathbf{D}^T\right\|_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$$

Vậy

$$\|\mathbf{A}\|_F^2 = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$

với $\sigma_i, i = 1, 2, \ldots, r$ là các giá trị kì dị của ma trận $\mathbf{A}$.

Chương 3

Phân tích Thành phần Chính

3.1 Giới thiệu

Phép phân tích thành phần chính (Principal Components Analysis - PCA) là một phương pháp phân tích ma trận đa biến, giúp giảm số chiều của dữ liệu bằng cách tìm ra các thành phần chính của ma trận đó. Ý tưởng của PCA là chuyển đổi dữ liệu ban đầu từ không gian có số chiều cao sang không gian có số chiều thấp hơn, giúp cho việc xử lý và phân tích dữ liệu dễ dàng hơn. Phép phân tích thành phần chính được đưa ra vào năm 1901 bởi Karl Pearson và đã được phát triển và ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau như nhận dạng hình ảnh, phân tích tín hiệu, dự báo kinh tế và thị trường tài chính.

Một trong những ứng dụng quan trọng nhất của phép phân tích thành phần chính là giảm số chiều của dữ liệu. Khi dữ liệu có số chiều lớn, việc phân tích và xử lý trở nên phức tạp và tốn nhiều thời gian. Đồng thời cho phép giảm số chiều của dữ liệu bằng cách xác định các thành phần chính của ma trận, giúp cho việc xử lý và phân tích dữ liệu trở nên đơn giản và nhanh chóng hơn.

Ngoài ra, phép phân tích này còn được sử dụng để giảm tác động của nhiễu trong dữ liệu, phát hiện các mối quan hệ giữa các biến đầu vào và giúp tối ưu hóa các thuật toán máy học.

Với những ứng dụng và lợi ích của mình, hép phân tích thành phần chính đã trở thành một công cụ quan trọng trong phân tích dữ liệu và được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau.

3.2 Ý tưởng

Giả sử dữ liệu ban đầu là $x \in \mathbb{R}^m$ và dữ liệu đã được giảm chiều là $z \in \mathbb{R}^r$ với $r < m$. Cách đơn giản nhất để giảm chiều dữ liệu từ m về $r < m$ là chỉ cần giữ lại r phần tử quan trọng nhất. Có hai câu hỏi được đặt ra ở đây. Câu hỏi thứ nhất, làm thế nào để xác định tầm quan trọng của mỗi chiều dữ liệu? Câu hỏi thứ hai, nếu tầm quan trọng của các chiều dữ liệu là như nhau, ta cần bỏ đi chiều nào?

Để trả lời câu hỏi thứ nhất, ta quan sát Hình 2.2a. Giả sử các điểm dữ liệu có thành phần thứ hai (phương đứng) giống hệt nhau hoặc sai khác nhau rất ít (phương sai nhỏ). Như vậy thành phần này hoàn toàn có thể được lược bỏ đi, và ta ngầm hiểu rằng nó sẽ được xấp xỉ bằng kỳ vọng của thành phần đó trên toàn bộ các điểm dữ liệu. Ngược lại, việc làm này nếu được áp dụng lên thành phần thứ nhất (phương ngang) sẽ khiến lượng thông tin bị mất đi rất nhiều do sai số xấp xỉ quá lớn. Vì vậy, lượng thông tin theo mỗi thành phần có thể được đo bằng phương sai của dữ liệu trên thành phần đó. Tổng lượng thông tin có thể được coi là tổng phương sai trên toàn bộ các thành phần.

Câu hỏi thứ hai tương ứng với trường hợp Hình 2.2b. Trong cả hai chiều, phương sai của dữ liệu đều lớn, việc bỏ đi một trong hai chiều đều dẫn đến việc lượng thông tin bị mất đi là rất lớn. Tuy nhiên, quan sát ban đầu của chúng ta là nếu xoay trục tọa độ đi một góc phù hợp, một trong hai chiều dữ liệu có thể được giảm đi vì dữ liệu có xu hướng phân bố xung quanh một đường thẳng.

thêm 2 hình zô Hình 2.2: Ví dụ về phương sai của dữ liệu trong không gian hai chiều. (a) Chiều thứ hai có phương sai (tỉ lệ với độ rộng của đường hình chuông) nhỏ hơn chiều thứ nhất. (b) Cả hai chiều có phương sai đáng kể. Phương sai của mỗi chiều là phương sai của thành phần tương ứng được lấy trên toàn bộ dữ liệu. Phương sai tỉ lệ thuận với độ phân tán của dữ liệu. Ý tưởng chính của PCA: Tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong r thành phần đầu tiên.

PCA là một phương pháp đi tìm một phép xoay trục tọa độ để được một hệ trục tọa độ mới sao cho trong hệ mới này, thông tin của dữ liệu chủ yếu tập trung ở vài thành phần. Phần còn lại chứa ít thông tin hơn có thể được lược bỏ.

3.3 Phân tích thành phần chính

Về mặt hình thức, cho ma trận dữ liệu $X \in \mathbb{R}^{m\times n}$ với mỗi giá trị đã được chuẩn hóa sao cho mỗi cột có giá trị trung bình là 0 và phương sai là 1 . PCA của X với r thành phần là tìm ma trận trực giao $A \in \mathbb{R}^{m\times r}$ và ma trận $B \in \mathbb{R}^{n\times r}$ sao cho X có thể biểu diễn được dưới dạng ma trận

$$X = AB^T + E$$

trong đó $\|E\|^2$ đạt giá trị nhỏ nhất. Ma trận A được gọi là ma trận thành phần, và các cột của nó là các thành phần chính. Ma trận B được gọi là ma trận tải, các tải trọng là các trọng số cho phép tái cấu trúc các biến ban đầu dưới dạng tổ hợp tuyến tính của các thành phần chính. Cặp (A, B) được gọi là nghiệm PCA.

Ngoài ra, PCA được trình bày theo một cách khác dưới dạng véctơ như sau:

$$X = \sum_{i=1}^{r} a_i b_i^T + E$$

Điều này cho thấy PCA là xấp xỉ X với tổng của r ma trận có hạng 1 . Mục tiêu của PCA là làm giảm tối thiểu $\|E\|^2 = \left\|X - AB^T\right\|^2$. Vì rank $\left(AB^T\right) \leq r$ nên AB^T là SVD "chặt cụt" của X, nghĩa là nếu $U_r S_r \left(V_r\right)^T$ là SVD "chặt cụt" của X thì $A = m^{\frac{1}{2}} U_r$ và $B^T = m^{-\frac{1}{2}} S_r \left(V_r\right)^T$. Số lượng thành phần tối thiểu phù hợp là số giá trị kì dị khác 0 của X, hay là hạng của X. Vì vậy, không cần thiết phải lấy số thành phần r lớn hơn số lượng biến n. Trên thực tế, r thường được lấy nhỏ hơn nhiều so với n.

Do thứ tự giảm dần của các giá trị kì dị của X trong SVD, thành phần chính đầu tiên giải thích phương sai nhiều nhất có thể và từng thành phần chính tiếp theo giải thích phương sai với ràng buộc là nó trực giao (không tương quan với các thành phần trước đó). Khi đó tổng phương sai được giải thích là tr $\left(BB^T\right)$ = tr $\left(m^{-\frac{1}{2}} S_r^2\right)$ với tr $\left(S_r^2\right)$ là tổng bình phương của r giá trị kì dị lớn nhất của X.

3.4 Tìm các thành phần chính của bài toán PCA thông qua SVD

Xét một vectơ x bất kì. Thành phần chính là tổ hợp tuyến tính $s = \sum_{i=1}^{m} w_i x_i$ có chứa càng nhiều phương sai của dữ liệu đầu vào càng tốt. Như vậy, thành phần chính đầu tiên được định nghĩa bằng trực giác là tổ hợp tuyến tính của các biến quan sát, trong đó có phương sai lớn nhất.

Chúng ta cần đưa ra ràng buộc cho chuẩn của vectơ $w = (w_1, w_2, \ldots, w_m)$. Để đơn giản, chúng ta ràng buộc w có chuẩn bằng 1 , tức là

$$\|w\| = \sqrt{\sum_{i=1}^{m} w_i^2} = 1$$

Các ràng buộc khác về giá trị chuẩn của w chúng ta có thể đưa về ràng buộc trên.

Chú ý rằng phương sai của một tổ hợp tuyến tính bất kì đều có thể được tính thông qua ma trận hiệp phương sai của dữ liệu. Xét một tổ hợp tuyến tính $w^T x = \sum_{i=1}^{m} w_i x_i$. Giả sử giá trị trung bình bằng 0 , tức là $E\{x\} = 0$. Khi đó

$$\begin{aligned} E\left\{\left(w^T x\right)^2\right\} = E\left\{\left(w^T x\right)\left(w^T x\right)\right\} = E\left\{w^T\left(xx^T\right)w\right\} &= w^T E\left\{xx^T\right\} \\ &= w^T C w \end{aligned}$$

trong đó $C = E\left\{xx^T\right\}$ là ma trận hiệp phương sai. Vì vậy, bài toán cơ bản PCA được xác định như sau:

$$\max_{w:\|w\|=1} w^T C w$$

Vì C là ma trận đối xứng nên theo Định lí phổ của ma trận đối xứng, tồn tại ma trận trực giao $U \in \mathbb{R}^{m\times n}$ và ma trận đường chéo $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ sao cho $C = UDU^T$, trong đó $\lambda_1, \ldots, \lambda_n$ là các giá trị riêng của C, và các véctơ cột của U là các véctơ riêng của C ứng với giá trị riêng đó. Thực hiện đổi biến $v = U^T w$. Khi đó ta nhận được

$$w^T C w = w^T U D U^T w = v^T D v = \sum_{i=1}^{n} v_i^2 \lambda_i$$

Vì U trực giao nên $\|v\| = \|w\|$, do đó, v có ràng buộc $\|v\| = 1$. Tiếp tục thực hiện phép đổi biến $m_i = v_i^2, i = 1, \ldots, n$. Khi đó ràng buộc $\|v\| = 1$ tương đương với ràng buộc $m_i \geq 0$ và $\sum_{i=1}^{n} m_i = 1$. Bài toán được chuyển sang dạng

$$\max \sum_{i=1}^{n} m_i \lambda_i, \text{ với } m_i \geq 0, \sum_{i=1}^{n} m_i = 1.$$

Rõ ràng, bài toán cho thấy giá trị lớn nhất tìm được khi m_i tương ứng với λ_i lớn nhất bằng 1 và các m_i còn lại bằng 0 . Kí hiệu i^* là chỉ số của giá trị riêng lớn nhất. Trở lại biến w, điều này tương đương với w bắt đầu từ vectơ riêng thứ i^*, tức là cột thứ i^* của U. Như vậy, thành phần chính đầu tiên được tìm một cách dễ dàng thông qua phân tích giá trị riêng.

Do các vectơ riêng của ma trận đối xứng là trực giao nên việc tìm thành phần chính thứ hai đồng nghĩa với việc tối đa hóa phương sai sao cho v_{i^*} vẫn bằng 0 . Điều này thực sự tương đương với việc tạo w trực giao mới cho vectơ riêng đầu tiên. Như vậy, về mặt m_i, chúng ta có bài toán tối ưu hóa tương tự nhưng với ràng buộc $m_{i^*} = 0$. Rõ ràng, nghiệm của bài toán tối ưu đó thu được khi w bằng với vectơ riêng tương ứng với giá trị riêng lớn nhất thứ hai. Lôgic này áp dụng cho thành phần chính thứ k.

Do đó, tất cả các thành phần chính có thể được tìm thấy bằng cách đặt các vectơ riêng $u_i, i = 1, \ldots, n$ trong U sao cho các giá trị riêng giảm dần. Chúng ta hãy giả sử U được định nghĩa như vậy. Khi đó thành phần chính thứ i có dạng

$$s_i = u_i^T x$$

Lưu ý rằng tất cả các λ_i đều không âm đối với ma trận hiệp phương sai.

3.5 Tính duy nhất nghiệm của PCA

Định lý 2.3.1. Nếu (A, B) là một nghiệm của mô hình PCA thì (AQ, BQ) cũng là một nghiệm của mô hình PCA, với Q là ma trận trực giao cấp r.

Lúc này, Q được gọi là phép quay trực giao.

Chưng minh. Giả sử (A, B) là một nghiệm của mô hình PCA.

Với Q là một ma trận trực giao cấp r, tức là $QQ^T = Q^TQ = I_r$, ta có

$$(AQ)(AQ)^T = AQQ^TA^T = AI_rA^T = AA^T = I_m$$

và

$$(AQ)^T(AQ) = Q^TA^TAQ = Q^TI_rQ = Q^TQ = I_r$$

Suy ra (AQ) là ma trận trực giao cỡ $m \times r$.

Mặt khác, ta có

$$(AQ)(BQ)^T = AQQ^TB^T = AI_rB^T = AB^T$$

và

$$\left\|X - AQQ^TB^T\right\|^2 = \left\|X - AB^T\right\|^2$$

Vậy (AQ, BQ) là một nghiệm của mô hình PCA.

Từ Định lí 2.3.1, ta có nhận xét sau:

Nhận xét 2.3.1. (i) Nghiệm (A, B) của PCA không duy nhất.

(ii) Phép quay trực giao Q sẽ cho ta ma trận tải có cấu trúc đơn giản hơn, do đó các nhân tố sẽ được diễn giải dễ dàng hơn.

3.6 Thuật toán tìm PCA của một ma trận

Giả sử X là ma trận cỡ $m \times n$, với $m \geq n$. Để tìm PCA của ma trận X, chúng ta thực hiện các bước sau:

- Bước 1: Tìm SVD "chặt cụt" của ma trận X, ta được $X = U_rS_r\left(V_r\right)^T$ với $r \leq n$.

- Bước 2: Tính ma trận A và B theo công thức sau:

$$A = m^{\frac{1}{2}}U_r, \quad B^T = m^{-\frac{1}{2}}S_r\left(V_r\right)^T$$

- Bước 3: Nếu nghiệm (A, B) chưa tốt thì chọn phép quay Q, với Q là ma trận trực giao cấp r, ta tìm được nghiệm của mô hình PCA là (AQ, BQ).

3.7 Ưu và nhược điểm của PCA

PCA có nhiều đặc tính tốt

- Giúp giảm số chiều của dữ liệu. - Thay vì giữ lại các trục tọa độ của không gian cũ, PCA xây dựng một không gian mới ít chiều hơn, nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương không gian cũ, nghĩa là đảm bảo độ biến thiên của dữ liệu trên mỗi chiều dữ liệu mới.

- Các trục tọa độ trong không gian mới là tổ hợp tuyến tính của không gian cũ, do đó về mặt ngữ nghĩa, PCA xây dựng feature mới dựa trên các feature đã quan sát được. Điểm hay là những feature này vẫn biểu diễn tốt dữ liệu ban đầu.

- Trong không gian mới, các liên kết tiềm ẩn của dữ liệu có thể được khám phá, mà nếu đặt trong không gian cũ thì khó phát hiện hơn, hoặc những liên kết như thế không thể hiện rõ.

Bên cạnh đó, PCA cũng có một vài hạn chế sau

- Chỉ làm việc với dữ liệu số (numeric).

- Nhạy cảm với các điểm nằm bên ngoài/cực trị (outlier/extreme).

- Không phù hợp với môi trường phi tuyến, do PCA hoàn toàn dựa trên các biến đổi tuyến tính.