TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA KHOA HỌC TỰ NHIÊN

BÁO CÁO TỔNG KẾT

HỌC PHẦN THỐNG KÊ NHIỀU CHIỀU

PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ ỨNG DỤNG TRONG PHÂN TÍCH DỊCH TỄ

TN441 - 2021

Cần Thơ, 2021

TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA KHOA HỌC TỰ NHIÊN

BÁO CÁO TỔNG KẾT

HỌC PHẦN THỐNG KÊ NHIỀU CHIỀU

PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ ỨNG DỤNG TRONG PHÂN TÍCH DỊCH TỄ

Giảng viên hướng dẫn:	TS. Trần Văn Lý	
Trưởng nhóm:	Trần Nam Hưng	B1906052
Các thành viên:	Lý Ngọc Thanh	B1906074
	Lê Phát Tài	B1906071
	Mai Quốc Vinh	B1906101
	Huỳnh Thị Nhật Linh	B1906058

Cần Thơ, 2021

Tóm tắt nội dung

In this research, the number of patients with Covid-19 due to this disease in some ofprovinces/cities are considered. First, the relations between the considered provinces/cities are studied using Pearson's correlation. Then, based on the spread rate of Covid-19, these provinces/cities are categorized using principal component analysis and factor analysis.

Title: *The principal component algorithm and its application to epidemiological analysis.*

Keyword:

Tóm tắt nội dung

Trong nghiên cứu này, số lượng bệnh nhân mắc Covid-19 do bệnh này ở một số tỉnh/thành phố được xem xét. Đầu tiên, mối quan hệ giữa các tỉnh/thành phố được xem xét được nghiên cứu bằng cách sử dụng tương quan Pearson. Sau đó, dựa trên tỷ lệ lây lan của Covid-19, các tỉnh/thành phố này được phân loại bằng cách sử dụng phân tích thành phần chính và phân tích nhân tố.

Mục lục

PHẦN MỞ ĐẦU **iv**

1 TỔNG QUAN CƠ SỞ LÝ THUYẾT **1**

1.1 Lý thuyết đại số tuyến tính 1

1.1.1 Ma trận và các phép tính trên ma trận 1

1.1.2 Chuẩn 2

1.1.3 Véc-tơ riêng và giá trị riêng. Thuật toán tìm véc-tơ riêng 3

1.2 Lý thuyết xác suất 3

1.3 Phương pháp chuẩn hóa dữ liệu 5

2 THUẬT TOÁN PHÂN TÍCH THÀNH PHẦN CHÍNH **7**

2.1 Dẫn nhập 7

2.2 Thuật toán phân tích thành phần chính 7

2.3 Tiêu chí giảm thiểu số chiều dữ liệu 10

3 PHƯƠNG PHÁP PHÂN TÍCH NHÂN TỐ **12**

3.1 Dẫn nhập 12

3.2 Thuật toán phân tích nhân tố 12

3.2.1 Kiểm định Barlett và kiểm định KMO 12

3.2.2 Xoay nhân tố 13

4 THỰC NGHIỆM **15**

4.1 Viêm phổi do vi-rút Corona 15

4.2 Tổng quan về việc thực hiện 16

4.2.1 Dữ liệu nghiên cứu 16

4.2.2 Các tiêu chuẩn đánh giá mô hình 16

4.2.3 Thiết kế nghiên cứu 17

4.3 Đọc và xử lý số liệu 18

4.4 Một số thống kê mô tả cho hai dữ liệu 19

4.5 Mối tương quan đối với số ca nhiễm bệnh giữa các tỉnh 23

4.6 Phân tích thành phần chính 27

4.6.1 Dữ liệu **case_data** 27

4.6.2 Dữ liệu **cul_data** 36

4.7 Kiểm định Bartlett – KMO 46

4.8 Phân tích nhân tố 48

4.9 Ma trận xoay 50

4.10 Bàn luận 53

5 KẾT LUẬN **54**
5.1 Kết luận 54
5.2 Nhận xét sơ bộ bài báo cáo 54

6 PHỤ LỤC **56**
6.1 Thông tin phần mềm 56
6.2 Nguồn mã lập trình 57

TÀI LIỆU THAM KHẢO **58**

INDEX **59**

Danh sách hình vẽ

2.1 Mô tả thuật toán phân tích thành phần chính 8
2.2 Thuật toán phân tích thành phần chính 10

4.1 Đồ thị số ca nhiễm hằng ngày 21
4.2 Đồ thị số ca nhiễm tích lũy 22
4.3 Tương quan đồ thể hiện tương quan dữ liệu hằng ngày các ca xác nhận nhiễm ở các tỉnh/thành phố. 23
4.4 Tương quan đồ thể hiện tương quan dữ liệu tích lũy các ca xác nhận nhiễm 24
4.5 Mạng tương quan pcor đối với dữ liệu ca bệnh thu nhập hằng ngày 25
4.6 Mạng tương quan pcor đối với dữ liệu ca bệnh tích lũy hằng ngày 26
4.7 Biểu đồ tương quan giữa Thành phố Hồ Chí Minh và các tỉnh lân cận 26
4.8 Sơ đồ sàng lọc với phân tích song song dữ liệu ca nhiễm hằng ngày 27
4.9 Sơ đồ sàng lọc dữ liệu ca nhiễm hằng ngày và giá trị riêng tương ứng 30
4.10 Biểu đồ biplot cho dữ liệu hằng ngày 31
4.11 Biểu đồ biplot tổng hợp mật độ cos2 giữa hai thành phần chính của dữ liệu hằng ngày 32
4.12 Đồ thị biểu diễn các thông số theo hai chiều dữ liệu đầu tiên. 33
4.13 Sơ đồ sàng lọc với phân tích song song dữ liệu ca nhiễm hằng ngày 36
4.14 Sơ đồ sàng lọc dữ liệu ca nhiễm hằng ngày và giá trị riêng tương ứng 39
4.15 Biểu đồ biplot cho dữ liệu hằng ngày 40
4.16 Biểu đồ biplot tổng hợp mật độ cos2 giữa hai thành phần chính của dữ liệu tích lũy 41
4.17 Đồ thị biểu diễn các thông số theo hai chiều đầu tiên với dữ liệu tích lũy 42
4.18 Biểu đồ giá trị cos2 đối với 5 biến đã được chọn làm thành phần chính đối với các biến khi chưa phân tích. 45
4.19 Phân cụm nhân tố và hệ số nhân tố tương ứng của hai dữ liệu. 51
4.20 Tương quan nhân tố của dữ liệu nhân tố được xác định với hệ số tải nhân tố 0.55 52

6.1 Đường dẫn cụ thể cho mã vạch QR 57

PHẦN MỞ ĐẦU

Trong chương này chúng tôi muốn giới thiệu mục tiêu nghiên cứu và bố cục của bài báo cáo. Đầu tiên mục tiêu nghiên cứu sẽ mô tả các thành phần liên quan đến tình hình của 18 tỉnh thành đang có dịch. Cuối cùng là bố cục bài báo cáo chúng tôi sẽ nêu rõ tên và trọng tâm của 5 chương.

Bài báo cáo sử dụng các phương pháp thành phần chính để phân loại biến dựa theo bài báo khoa học [4]

1. Mục tiêu nghiên cứu

Mô tả dữ liệu với các thông số về trung bình, phương sai cung cấp các thông tin dịch tễ cơ bản về 18 tỉnh/thành phố đang có dịch bệnh.

Ứng dụng thuật toán phân tích thành phần chính để giảm thiểu số chiều dữ liệu dịch tễ các trường hợp xác nhận nhiễm covid-19 đối với 18 tỉnh/thành phố miền Nam và phân tích nhân tố vào dữ liệu để phân cụm các tỉnh có các tính chất tương tự nhau.

2. Bố cục báo cáo

Đề tài này bao gồm năm chương với trọng tâm như sau.

Chương 1. Tổng quan cơ sở lý thuyết tập trung tổng kết có hệ thống một vài lý thuyết đại số tuyến tính và xác suất và chuẩn hóa dữ liệu để thiết lập các thống kê mô tả cũng như thuật toán phân tích thành phần chính và phân tích nhân tố.

Chương 2. Thuật toán phân tích thành phần chính được dành giải thích và trình bày thuật toán phân tích thành phần chính theo lý thuyết đại số tuyến tính với các định nghĩa trong thống kê. Ngoài ra, một số tiêu chí giảm thiểu số chiều dữ liệu cũng được trình bày để thiết lập thuật toán phân cụm vùng tỉnh/thành phố có bệnh dịch.

Chương 3. Phương pháp phân tích nhân tố dành trọn vẹn cho việc khảo cứu thuật toán phân tích nhân tố và cách ứng dụng vào dữ liệu dịch bệnh.

Chương 4. Thực nghiệm đầu tiên trình bày tổng quan dữ liệu và các tiêu chuẩn đánh giá tham số đối với các kiểm định. Phần chính yếu nêu các kết quả ứng dụng các thuật toán vào hai loại dữ liệu thứ cấp thể hiện số ca nhiễm hằng ngày và tích lũy đối với các tỉnh/thành phố đang bùng phát dịch.

Chương 5. Kết luận và định hướng nghiên cứu trình bày kết luận và lượng giá về bài báo cáo.

Chương 1

TỔNG QUAN CƠ SỞ LÝ THUYẾT

1.1 Lý thuyết đại số tuyến tính

Lý thuyết đại số tuyến tính cung cấp các định nghĩa về ma trận và tập trung vào các khái niệm có liên quan đến thuật toán phân tích thành phần chính và phân tích nhân tố. Ngoài ra bài báo cáo cũng đưa ra quy trình trực giao hóa và cách xác định véc-tơ riêng nhằm đi sâu giải thích thuật toán phân tích thành phần chính.

1.1.1 Ma trận và các phép tính trên ma trận

Định nghĩa 1.1 (Ma trận). Giả sử $\mathbb{F}$ là một trường tùy ý, mỗi bảng có dạng

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

trong đó $a_{ij} \in \mathbb{F}$ với $1 \leq i \leq m$ và $1 \leq j \leq n$, được gọi là một *ma trận* m hàng n cột (hay ma trận cấp $m \times n$) với các yếu tố trong trường $\mathbb{F}$. Các vô hướng $a_{ij} \in \mathbb{F}$ được gọi là *phần tử* (hay hệ tử) của hàng i cột j của ma trận $\mathbf{A}$.

Ma trận trên thường được ký hiệu gọn là $\mathbf{A} = (a_{ij})_{m \times n}$.

Định nghĩa 1.2 (Phép cộng và nhân vô hướng đối với hai ma trận). Ta định nghĩa hai phép toán cộng hai ma trận và nhân ma trận với một vô hướng trên tập hợp các ma trận $\mathbf{M}$ như sau

$$(a_{ij}) + (b_{ij}) = (a + b)_{ij}$$
$$\alpha(a_{ij}) = (\alpha a_{ij})$$

Định nghĩa 1.3 (Tích của hai ma trận). Giả sử ma trận $\mathbf{A} = (a_{ij}) \in \mathbf{M}(m \times n, \mathbb{F})$ và ma trận $\mathbf{B} = (b_{ij}) \in \mathbf{M}(n \times p, \mathbb{F})$, ta có tích của hai ma trận $\mathbf{A}$ và $\mathbf{B}$, ký hiệu $\mathbf{AB}$, là ma trận $\mathbf{C} = (c_{ij}) \in \mathbf{M}(m \times p, \mathbb{F})$ với các phần tử được xác định như sau

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}, (1 \leq i \leq m, 1 \leq k \leq p).$$

Định nghĩa 1.4 (Ma trận đơn vị). Ma trận $\mathbf{I}_n$ là phần tử trung hòa của phép nhân hai ma trận. Nếu $\mathbf{A} \in \mathbf{M}(n \times n, \mathbb{F})$ và $\mathbf{I}_n$ là ma trận đơn vị bậc n thì $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.

Định nghĩa 1.5 (Ma trận khả nghịch). Ma trận vuông $\mathbf{A} \in \mathbf{M}(n \times n\,, \mathbb{F})$ được gọi là *ma trận khả nghịch* (hoặc *ma trận không suy biến*) nếu có ma trận $\mathbf{B} \in M(n \times n\,, \mathbb{F})$ sao cho $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$. Khi đó, ta nói $\mathbf{B}$ là *ma trận nghịch đảo* của $\mathbf{A}$ và ký hiệu $\mathbf{B} = \mathbf{A}^{-1}$

Định nghĩa 1.6 (Ma trận đường chéo). Một ma trận vuông $\mathbf{A} = (a_{ij})$ với $1 \leq i\,, j \leq n$ thuộc $\mathbf{M}(n\,, n)$ được gọi là ma trận đường chéo khi và chỉ khi các phần tử khác đường chéo đều bằng 0. Ta ký hiệu ma trận đường chéo là $diag(\lambda_1 \ \ldots\ , \lambda_n)$

Định nghĩa 1.7 (Vết của ma trận vuông). Với mọi ma trận vuông $\mathbf{A} = (a_{ij}) \in \mathbf{M}(n\,, n)$, vết của ma trận vuông $\mathbf{A}$, ký hiệu $trace(\mathbf{A})$ được định nghĩa là tổng các phần tử trong đường chéo của $\mathbf{A}$, tức là $trace(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

1.1.2 Chuẩn

Phần này định nghĩa chuẩn của một véc-tơ trên tập số thực $\mathbb{R}^d$ có d-chiều và quy trình trực giao hóa.

Định nghĩa 1.8 (Tích vô hướng của hai véc-tơ). Cho hai véc-tơ $\mathbf{x}\,, \mathbf{y} \in \mathbb{R}^d$ được định nghĩa bởi

$$\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{d} x_i y_i$$

Nếu tích vô hướng của hai véc-tơ khác $\mathbf{0}$ bằng 0 (không) thì ta nói hai véc-tơ đó trực giao với nhau.

Định nghĩa 1.9 (Độ đo phân biệt giữa các phần tử rời rạc). Cho X là tập tùy ý khác rỗng. Hàm số $\mathrm{d} : X \times X \to \mathbb{R}$ là độ đo phân biệt nếu d thỏa mãn ba tiên đề

(i) $\mathrm{d}(x\,, y) \geq 0\,, \forall x\,, y \in X$;

(ii) $\mathrm{d}(x\,, y) = 0 \Leftrightarrow x = y$;

(iii) $\mathrm{d}(x\,, y) = \mathrm{d}(y\,, x)$.

Nếu ta thêm một tiên đề độ đo phân biệt thỏa mãn bất đẳng thức tam giác

$$\mathrm{d}(x\,, y) \leq \mathrm{d}(x\,, z) + \mathrm{d}(z\,, y)\,, \forall x\,, y\,, z \in X$$

thì khi đó độ đo phân biệt là một metric (khoảng cách).

Định nghĩa 1.10 (Chuẩn). Hàm số $f : \mathbb{R}^d \to \mathbb{R}$ được gọi là một chuẩn nếu nó thỏa mãn ba tiên đề sau đây

(i) $f(\mathbf{x}) \geq 0\,, \forall \mathbf{x} \in \mathbb{R}^d$;

(ii) $f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x})\,, \forall \alpha \in \mathbb{R}$;

(iii) $f(\mathbf{x}_1) + f(\mathbf{x}_2) \geq f(\mathbf{x}_1 + \mathbf{x}_2)\,, \forall \mathbf{x}_1\,, \mathbf{x}_2 \in \mathbb{R}^d$.

Định nghĩa 1.11 (Chuẩn trong không gian Euclid). Giả sử E là không gian véc-tơ Euclid với tích vô hướng $\langle \cdot, \cdot \rangle$. Khi đó, độ dài (hay chuẩn) của véc-tơ $\mathbf{v} \in E$ là số thực không âm được định nghĩa $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.

Định nghĩa 1.12 (Chuẩn của ma trận). Giả sử hàm số $\|\mathbf{x}\|_\alpha$ là một chuẩn bất kỳ của vector $\mathbf{x}$. Ứng với chuẩn này, định nghĩa chuẩn tương ứng cho ma trận $\mathbf{A}$ là

$$\|\mathbf{A}\|_\alpha = \max_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha}$$

chú ý rằng ma trận $\mathbf{A}$ có thể không vuông và số cột của nó bằng với số chiều của $\mathbf{x}$.

Chúng ta sẽ quan tâm nhiều hơn tới chuẩn bậc 2. Chuẩn bậc 2 của ma trận được định nghĩa là

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2}$$

Định nghĩa 1.13 (Hình chiếu của véc-tơ). Cho hai véc-tơ $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, ta gọi hình chiếu của véc-tơ $\mathbf{x}$ lên véc-tơ $\mathbf{y}$ là véc-tơ $Proj_{\mathbf{y}}(\mathbf{x})$ được xác định bởi công thức

1.1.3 Véc-tơ riêng và giá trị riêng. Thuật toán tìm véc-tơ riêng

Định nghĩa 1.14 (Véc-tơ riêng). Cho $\mathbf{A} \in \mathbf{M}(n \times n, \mathbb{R})$, véc-tơ $\mathbf{v} \in \mathbb{C}^d, \mathbf{v} \neq \mathbf{0}$ được gọi là véc-tơ riêng của $\mathbf{A}$ nếu tồn tại vô hướng λ sao cho $\mathbf{Av} = \lambda\mathbf{v}$. Khi đó, vô hướng λ được gọi là giá trị riêng của $\mathbf{A}$ và $\mathbf{v}$ được gọi là véc-tơ riêng ứng với giá trị riêng λ đó.

Định nghĩa 1.15 (Đa thức đặc trưng). Đa thức bậc n của một ẩn λ trong ma trận $\mathbf{A}$ với hệ số trong $\mathbb{F}$ là

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda I_n)$$

được gọi là đa thức đặc trưng của ma trận. Ta có nghiệm của đa thức đặc trưng trong ma trận chính là giá trị riêng của ma trận $\mathbf{A}$.

Từ định nghĩa trên, ta cũng có $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{x} = 0$ là phương trình đặc trưng của ma trận, tức $\mathbf{x}$ là một véc-tơ nằm trong không gian $\mathcal{N}(\mathbf{A} - \lambda\mathbf{I}_n)$. Về lý thuyết, phương trình đặc trưng có khả năng có nghiệm phức, nghĩa là $\mathbf{A}$ có giá trị riêng phức. Trên thực tế, ta không xét trường hợp này.

Định lý 1.1 (Tổng các giá trị riêng). *Tổng các giá trị riêng của một ma trận vuông bất kỳ luôn bằng vết của ma trận đó.*

1.2 Lý thuyết xác suất

Bài báo cáo trình bày các khái niệm có liên quan đến thuật toán chính của đề tài. Làm tiền đề để xây dựng lý thuyết đại số tuyến tính trên các đối tượng của lý thuyết thống kê, hình thành thuật toán phân tích thành phần chính.

Định nghĩa 1.16 (Phân phối xác suất). Một phân phối xác suất hay thường gọi hơn là một hàm phân phối xác suất là quy luật cho biết cách gán mỗi xác suất cho mỗi khoảng giá trị của tập số thực, sao cho các tiên đề xác suất được thỏa mãn.

Tiên đề thứ nhất Xác suất của một biến số là một số thực không âm. Với hai tập bất kỳ $E \in F, \mathbb{P}(E) \geq 0$

Tiên đề thứ hai Xác suất một biến cố sơ cấp nào đó trong tập mẫu sẽ xảy ra là 1. $\mathbb{P}(\Omega) = 1$.

Tiên đề thứ ba Xác suất của một tập biến cố là hợp của các tập con không giao nhau bằng tổng các xác suất của các tập con đó. Một chuỗi đếm được bất kỳ gồm các biến cố đôi một không giao nhau $E_1, E_2 \ldots$ thỏa mãn

$$\mathbb{P}(E_1 \cup E_2 \cup \cdots) = \sum \mathbb{P}(E_i)$$

Định nghĩa 1.17 (Phương sai). Phương sai của đại lượng ngẫu nhiên $\mathbf{X}$, ký hiệu là $Var(\mathbf{X})$, là trung bình bình phương độ lệch so với trung bình

$$Var(\mathbf{X}) = \mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X}))^2$$

Trong thống kê, phương sai đặc trưng cho khoảng cách giữa mỗi số liệu với nhau và đến giá trị trung bình của tập dữ liệu được thể hiện qua công thức

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

trong đó,

x_i là giá trị của quan sát thứ i trong mẫu,

$\overline{x}$ là giá trị trung bình của tập dữ liệu, được tính theo công thức $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

n là số quan sát trong tập dữ liệu.

Định nghĩa 1.18 (Ma trận hiệp phương sai). của tập hợp n biến ngẫu nhiên là một ma trận vuông hạng $n \times n$, trong đó các phần tử nằm trên đường chéo (từ trái sang phải, từ trên xuống dưới) lần lượt là phương sai tương ứng của các biến này (ta chú ý rằng $Var(\mathbf{X}) = Cov(\mathbf{X}, \mathbf{X})$), trong khi các phần tử còn lại (không nằm trên đường chéo) là các hiệp phương sai của đôi một hai biến ngẫu nhiên khác nhau trong tập hợp.

Trong trường hợp chúng ta có một tập hợp dữ liệu với hơn hai chiều, sẽ có nhiều hơn một phép đo hiệp phương sai có thể được tính toán.

Ví dụ, từ một bộ dữ liệu được đo trên ba biến $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, ta có thể tính toán $cov(\mathbf{X}, \mathbf{Y}), cov(\mathbf{X}, \mathbf{Z})$ và $cov(\mathbf{Y}, \mathbf{Z})$. Trong thực tế, đối với một bộ dữ liệu d chiều, ta có thể tính toán $\frac{n!}{(n-2)!}$ giá trị hiệp phương sai khác nhau.

Một cách hữu ích để có được tất cả các giá trị hiệp phương sai có thể có giữa tất cả các biến khác nhau là đặt tất cả các tính toán trong một ma trận. Điều này dẫn đến định nghĩa khái niệm ma trận hiệp phương sai cho một tập hợp dữ liệu $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ kích thước n

$$C = (c_{ij})m \times n, c_{ij} = cov(\mathbf{X}_i, \mathbf{X}_j)$$

Định nghĩa 1.19 (Hệ số tương quan). Hệ số tương quan Pearson đặc trưng cho mối quan hệ tương quan giữa hai biến số với công thức được xác định như sau

$$\rho_{xy} = \frac{Cov(\mathbf{X_1}, \mathbf{X_2})}{\sigma_{\mathbf{X_1}}\sigma_{\mathbf{X_2}}},$$

trong đó,

$Cov(\mathbf{X_1}, \mathbf{X_2})$ là hiệp phương sai của biến $\mathbf{X_1}$ và $\mathbf{X_2}$ được tính bằng công thức.

$\sigma_{\mathbf{X_1}}$ và $\sigma_{\mathbf{X_2}}$ lần lượt là độ lệch chuẩn của biến $\mathbf{X_1}$ và $\mathbf{X_2}$, được tính bằng công thức.

Hệ số tương quan là đại lượng đo lường mức độ quan hệ giữa hai biến ngẫu nhiên, lấy giá trị từ -1 đến 1. Quan hệ giữa hai biến càng chặt nếu hệ số tương quan càng gần ± 1 và càng lỏng nếu hệ số tương quan càng gần 0. Quan hệ giữa hai biến là đồng biến nếu tương quan dương, ngược lại nghịch biến nếu tương quan âm.

Định nghĩa 1.20 (Ma trận tương quan). Với một tập biến $\mathbf{X_1}, \mathbf{X_2} \ldots, \mathbf{X_n}$ với hệ số tương quan đơn giữa $\mathbf{X}_i$ và $\mathbf{X}_j$ viết dưới dạng ma trận vuông ρ_{ij} được gọi là ma trận tương quan n dòng n cột mà các phần tử dòng i và cột j là ρ_{ij}.

1.3 Phương pháp chuẩn hóa dữ liệu

Chuẩn hóa cơ sở dữ liệu là một phương pháp khoa học để phân tách một bảng có cấu trúc phức tạp thành những bảng có cấu trúc đơn giản theo những quy luật đảm bảo không làm mất thông tin dữ liệu.

Trong phân tích thành phần chính, các biến thường được chia tỷ lệ (tức là được chuẩn hóa). Điều này được khuyến kích khi các biến đo lường ở các thang đo khác nhau (vd: kilogram, kilometer, centimeter,...). Nếu không được chuẩn hóa thì đầu ra của thuật toán phân tích thành phần chính sẽ bị ảnh hưởng nghiêm trọng.

Mục đích chuẩn hóa số liệu là làm cho các biến có thể so sánh được. Có rất nhiều kiểu chuẩn hóa được phát triển riêng biệt cho các loại phân tích Machine Learning khác nhau. Đối với thuật toán phân tích thành phần chính, ta sử dụng biến được chia tỷ lệ để có độ lệch chuẩn là 1 và trung bình là 0.

Định nghĩa 1.21 (Chuẩn hóa chuẩn). Cho tập dữ liệu

$$z - score = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}},$$

trong đó,

$\mathbf{x}$ là biến dữ liệu (véc-tơ),

$\boldsymbol{\mu}$ là trung bình của biến dữ liệu tương ứng,

$\boldsymbol{\sigma}$ là độ lệch chuẩn

Tổng kết chương

Tổng kết, trong chương dẫn nhập này,

Đầu tiên, bài báo cáo trình bày tổng quan lý thuyết đại số. Các định nghĩa về ma trận trên trường $\mathbb{F}$ được phát biểu lại để tập trung vào thuật toán xác định giá trị riêng và véc-tơ riêng. Thuật toán chéo hóa nhằm xoay các trục chính cho thẳng hàng với các vectơ riêng từ đó định hình nên phương pháp phân tích thành phần chính.

Thêm nữa, chúng tôi cũng phát biểu lại một số yếu điểm trong lý thuyết xác suất và chuẩn hóa dữ liệu đã được sử dụng rất nhiều trong Machine Learning. Các thực nghiệm trong bài báo cáo cũng sử dụng loại chuẩn hóa trên để xử lý dữ liệu bằng các ngôn ngữ lập trình R.

Chương tiếp theo tập trung nghiên cứu phương pháp phân tích thành phần chính cũng như đưa ra thuật toán giảm thiểu số chiều dữ liệu.

Chương 2

THUẬT TOÁN
PHÂN TÍCH THÀNH PHẦN CHÍNH

Trong thực tế một đối tượng luôn chịu tác động của nhiều yếu tố khác nhau. Để rút ra được bản chất của sự tương tác, sự ảnh hưởng, mức độ quan hệ của chúng, người ta phải xử lý số liệu nhiều chiều. Cùng với sự phát triển của thống thống kê nhiều chiều, sự phát triển của công nghệ thông tin đã giúp chúng ta xử lý khá hiệu quả số liệu nhiều chiều.

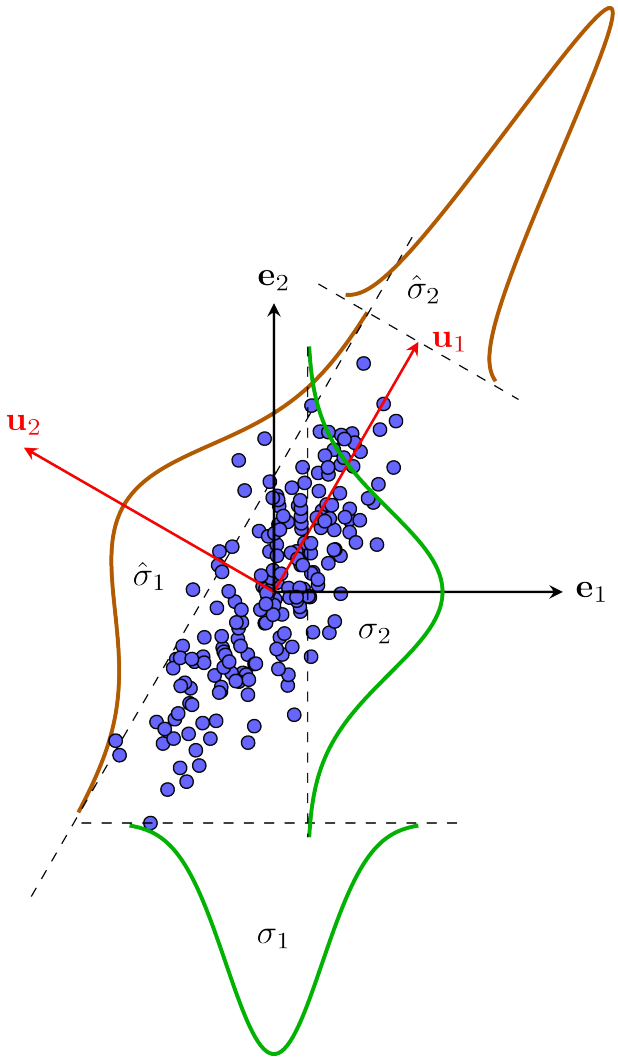
2.1 Dẫn nhập

Phân tích thành phần chính (principle component analysis, PCA) là phương pháp dựa trên việc tối đa lượng thông tin được giữ lại. Nó đi tìm một phép xoay trục toạ độ để được một hệ trục toạ độ mới sao cho trong hệ mới này, thông tin của dữ liệu chủ yếu tập trung ở một vài thành phần. Phần còn lại chứa ít thông tin hơn có thể được lược bỏ.

Hình 2.1 minh hoạ các thành phần chính với dữ liệu hai chiều. Trong không gian ban đầu với các vector cơ sở $\mathbf{e}_1, \mathbf{e}_2$, phương sai theo mỗi chiều dữ liệu (tỉ lệ với độ rộng của các hình chuông màu nâu) đều lớn. Trong hệ cơ sở mới $\mathbf{O}\mathbf{u}_1\mathbf{u}_2$, phương sai theo chiều thứ hai $\hat{\sigma}_2^2$ nhỏ so với $\hat{\sigma}_1^2$. Điều này chỉ ra rằng khi chiếu dữ liệu lên $\mathbf{u}_2$, ta được các điểm rất gần nhau và gần với giá trị trung bình theo chiều đó. Trong trường hợp này, vì giá trị trung bình theo mọi chiều bằng 0, ta có thể thay thế toạ độ theo chiều $\mathbf{u}_2$ bằng 0. Rõ ràng là nếu dữ liệu có phương sai càng nhỏ theo một chiều nào đó thì khi xấp xỉ chiều đó bằng một hằng số, sai số xấp xỉ càng nhỏ. PCA thực chất là đi tìm một phép xoay tương ứng với một ma trận trực giao sao cho trong hệ toạ độ mới, tồn tại các chiều có phương sai nhỏ có thể được bỏ qua; ta chỉ cần giữ lại các chiều/thành phần khác quan trọng hơn. Như đã khẳng định ở trên, tổng phương sai theo toàn bộ các chiều trong một hệ cơ sở bất kỳ là như nhau và bằng tổng các trị riêng của ma trận hiệp phương sai. Vì vậy, PCA còn được coi là phương pháp giảm số chiều dữ liệu sao cho tổng phương sai còn lại là lớn nhất. [6].

2.2 Thuật toán phân tích thành phần chính

Phân tích thành phần chính là kĩ thuật biểu diễn số liệu dựa theo các tiêu chuẩn về đại số và hình học mà không đòi hỏi một giả thuyết thống kê hay mô hình đặc biệt nào. Mục đích của phân tích thành phần chính là rút ra thông tin chủ yếu chứa trong bảng số liệu bằng cách xây dựng một biểu diễn đơn giản hơn sao cho đám mây số liệu được thể hiện rõ nhất. Cụ thể hơn, phân tích thành phần chính tức là đi tìm những trục hay mặt phẳng "phản ánh" tốt nhất, trung thực nhất đám mây điểm - biến, điểm - cá thể.



Hình 2.1: Phân tích thành phần chính có thể được coi là phương pháp đi tìm một hệ cơ sở trực chuẩn đóng vai trò một phép xoay, sao cho trong hệ cơ sở mới này, phương sai theo một số chiều nào đó là không đáng kể và có thể lược bỏ. Trong hệ cơ sở ban đầu $\mathbf{Oe}_1\mathbf{e}_2$, phương sai theo mỗi chiều (độ rộng của các đường hình chuông màu xanh lá) đều lớn. Trong không gian mới với hệ cơ sở $\mathbf{Ou}_1\mathbf{u}_2$, phương sai theo hai chiều (độ rộng của các đường hình chuông) chênh lệch nhau đáng kể. Chiều dữ liệu có phương sai nhỏ có thể được lược bỏ vì dữ liệu theo chiều này ít phân tán. Nguồn: Machine Learning cơ bản [6]

Với bảng số liệu có rất nhiều cột dòng, mỗi cột là một biến, mỗi dòng là một cá thể, trên đó đo đồng thời giá trị các biến, giữa các cá thể qua thể hiện rõ nhất trong một không gian con số chiều ít hơn.

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

1) Tính véc-tơ trung bình của toàn bộ dữ liệu: $\bar{\mathbf{x}} = \frac{1}{n}\sum_{n=1}^{n} \mathbf{x}_n$.

2) Trừ mỗi điểm dữ liệu đi véc-tơ trung bình của toàn bộ dữ liệu để được dữ liệu chuẩn hoá:

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}$$

3) Đặt $\widehat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_d]$ là ma trận dữ liệu chuẩn hoá, tính ma trận hiệp phương sai:

$$\mathbf{S} = \frac{1}{n}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top$$

4) Tính các trị riêng và véc-tơ riêng tương ứng có ℓ_2 chuẩn bằng 1 của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.

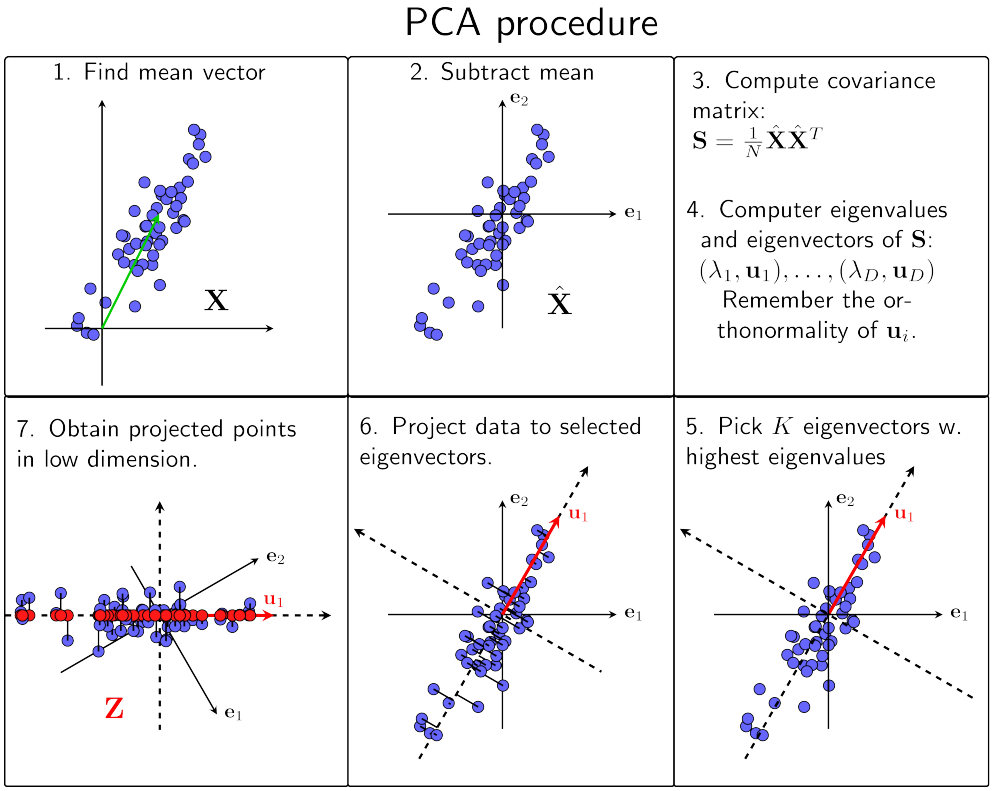
5) Chọn K véc-tơ riêng ứng với K giá trị riêng lớn nhất để xây dựng ma trận $\mathbf{U}_K$ có các cột tạo thành một hệ trực giao. K vector này được gọi là các thành phần chính, tạo thành một không gian con (gần) với phân bố của dữ liệu ban đầu đã chuẩn hoá.

6) Chiếu dữ liệu ban đầu đã chuẩn hoá $\hat{\mathbf{x}}$ xuống không gian con tìm được.

7) Dữ liệu mới là toạ độ của các điểm dữ liệu trên không gian mới:

$$\mathbf{Z} = \mathbf{U}_K^\top \widehat{\mathbf{X}}$$

Như vậy, thuật toán phân tích thành phần chính là thuật toán kết hợp của phép tịnh tiến, xoay trục toạ độ và chiếu dữ liệu lên hệ toạ độ mới.



Hình 2.2: Thuật toán phân tích thành phần chính. Bước 1. Tìm véc-tơ trung bình; Bước 2. Lấy dữ liệu trừ lần lượt cho véc-tơ trung bình; Bước 3. Tính ma trận hiệp phương sai $\mathbf{S} = \frac{1}{n}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\top}$; Bước 4. Tìm giá trị riêng và véc-tơ riêng của ma trận hiệp phương sai $\mathbf{S}$ lần lượt là $(\lambda_1, \mathbf{u}_1), \ldots, (\lambda_D, \mathbf{u}_D)$, các véc-tơ riêng được chọn phải tạo thành một hệ trực chuẩn; Bước 5. Chọn K véc-tơ riêng ứng với giá trị riêng lớn nhất; Bước 6. Chiếu dữ liệu ban đầu xuống các véc-tơ riêng đó; Bước 7. Dữ liệu giảm chiều (các điểm màu đỏ)

2.3 Tiêu chí giảm thiểu số chiều dữ liệu

Phân tích thành phần chính (viết tắt là PCA) là một cách tiếp cận đa biến nổi tiếng chuyển đổi một số biến tương quan thành một số biến không tương quan tuyến tính được đặt tên là các thành phần chính. Trong chuyển đổi này, các thành phần chính đầu tiên chứa nhiều thông tin nhất về tập dữ liệu. Trong các ứng dụng, PCA được áp dụng để chuyển đổi tập dữ liệu chiều cao thành tập dữ liệu chiều thấp hơn, bằng cách chỉ sử dụng một số thành phần chính đầu tiên để giảm kích thước của dữ liệu được biến đổi. Dựa trên Chỉ số Kaiser, số lượng các thành phần chính quan trọng bằng số lượng các giá trị riêng của ma trận tương quan với các giá trị lớn hơn 1.

1. Lựa chọn những thành phần chính để giải thích một tỷ lệ nhất định (ví dụ 95%) của $trace(\mathbf{\Lambda})$. Đây là một tiêu chí đơn giản nhưng không được khuyến cáo.

2. Hầu hết, cách tiếp cận để xác định số lượng thành phần chính bằng cách xác định giá trị riêng thông qua ma trận hệ số tương quan giữa dần đến khi số lượng thành phần chính bằng số biến). Kaiser – Harris đề xuất, thành phần chính được xác định khi giá trị riêng có giá trị lớn hơn 1.

3. Tiêu chuẩn Guttma – Kaise loại bỏ các giá trị riêng dưới mức trung bình $\frac{trace(\mathbf{\Lambda})}{d}$ (dưới 1 đối với dữ liệu chuẩn hóa), điều này có nghĩa là giảm các thành phần có phương sai được đóng góp bởi một biến nếu biến tổng được phân phối đều nhau

Tổng kết chương

Trong chương thuật toán phân tích thành phần chính này ta sẽ nói về 4 thành phần chính

Đầu tiên là dẫn nhập, ta sẽ tìm hiểu rõ phân tích thành phần chính là gì và chức năng của nó như thế nào. Phân tích thành phần chính có tên tiếng Anh là (principle component analysis, PCA) là phương pháp dựa trên việc tối đa lượng thông tin được giữ lại. Chức năng của nó là đi tìm một phép xoay trục toạ độ để được một hệ trục toạ độ mới sao cho trong hệ mới này, thông tin của dữ liệu chủ yếu tập trung ở một vài thành phần.

Thứ hai là thuật toán phân tích thành phần chính, ta sẽ đi sâu vào mục đích của phân tích thành phần chính, đặc biệt là với bảng số liệu sẽ như thế nào. Mục đích của phân tích thành phần chính là rút ra thông tin chủ yếu chứa trong bảng số liệu bằng cách xây dựng một biểu diễn đơn giản hơn sao cho đám mây số liệu được thể hiện rõ nhất. Còn với bảng số liệu có rất nhiều cột dòng, mỗi cột là một biến, mỗi dòng là một cá thể, trên đó đo đồng thời giá trị các biến, giữa các cá thể qua thể hiện rõ nhất trong một không gian con số chiều ít hơn.

Thứ ba là tiêu chí giảm thiểu số chiều dữ liệu. Ở đây ta sẽ có 2 tiêu chí là 1) Lựa chọn những thành phần chính để giải thích một tỷ lệ nhất định (ví dụ 95%) của $trace(\mathbf{\Lambda})$; 2) Tiêu chuẩn Guttma – Kaise loại bỏ các giá trị riêng dưới mức trung bình $trace(\mathbf{\Lambda})$.

Và cuối cùng là thuật toán phân loại vùng bệnh

Chương 3

PHƯƠNG PHÁP PHÂN TÍCH NHÂN TỐ

Phân tích nhân tố là các phương pháp rút gọn dữ liệu trên cơ sở tìm mối liên quan của các biến liên tục để từ đó giải thích chúng bằng vài nhân tố hoặc thành tố. Điều kiện của phân tích nhân tố là các biến phải có liên quan với nhau (nếu mối liên quan mà nhỏ - không thích hợp cho phương pháp này).

3.1 Dẫn nhập

Phân tích nhân tố nói chung là một nhóm các thuật toán được sử dụng chủ yếu để thu gọn và tóm tắt các dữ liệu. Các biến có liên quan với nhau được nhóm lại và tách ra khỏi các biến ít liên quan. Trong nghiên cứu, chúng ta có thể thu thập một lượng biến khá lớn, dẫn đến khó khăn trong xử lý, trong đánh giá bản chất. Liên hệ giữa các nhóm biến cố có tương quan được xem xét và trình bày dưới dạng tổ hợp một số các nhân tố cơ bản. Phân tích nhân tố thường được sử dụng trong các trường hợp sau

- Nhận diện một tập hợp gồm một số ít lượng biến mới, không tương quan với nhau để thay thế tập biến gốc có tương quan với nhau để thực hiện một phân tích đa biến tiếp theo.
- Nhận diện các khía cạnh hay nhân tố giải thích được các liên hệ tương quan trong một tập biến.
- Nhận diện một tập hợp gồm một số ít các biến nổi trội từ một tập hợp nhiều biến để sử dụng trong các phân tích thống kê đa biến

3.2 Thuật toán phân tích nhân tố

3.2.1 Kiểm định Barlett và kiểm định KMO

Phân tích nhân tố là một phương pháp thống kê dùng để mô tả sự biến thiên của những biến có tương quan được quan sát bằng một số nhỏ hơn các biến không quan sát được gọi là nhân tố. Ví dụ, sự biến thiên của bốn biến quan sát được có thể chỉ thể hiện sự biến thiên của hai biến không quan sát được. Những biến quan sát được mô hình hoá bằng tổ hợp tuyến tính của những nhân tố tiềm năng, cộng với số hạng lỗi. Để thực hiện được phân tích nhân tố có sự hiệu quả, bài báo cáo đề nghị các kiểm định sau.

a. Kiểm định Barlett

Kiểm định Barlett cho phép chúng ta so sánh phương sai của hai hoặc nhiều mẫu để xác định xem chúng có được rút trích từ các tập hợp có phương sai như nhau hay không.

Kiểm định Barlett phù hợp với dữ liệu phân phối chuẩn. Kiểm định có giả thiết không nếu các phương bằng nhau và kiểm định giả thiết đối nếu chúng không bằng nhau. Kiểm định có thể thực hiện trên các giá trị số (không bao gồm dữ liệu chuỗi).

Kiểm định này hữu ích để kiểm tra các giả định của một phân tích phương sai. Ta dựa vào $p - value$ để xác định kết luận với giả thiết thống kê là

H_0 : các mẫu có phương sai bằng nhau.

H_1 : ít nhất một mẫu có phương sai khác nhau có ý nghĩa.

$p - value \leq 0.05$ bác bỏ giả thuyết và $p - value > 0.05$ không bác bỏ giả thiết.

b. Kiểm định KMO

Kiểm định Bartlett (Bartlett's test of sphericity) dùng để xem xét các biến quan sát trong nhân tố có tương quan với nhau hay không. Chúng ta cần lưu ý, điều kiện cần để áp dụng phân tích nhân tố là các biến quan sát phản ánh những khía cạnh khác nhau của cùng một nhân tố phải có mối tương quan với nhau. Điểm này liên quan đến giá trị hội tụ trong phân tích EFA được nhắc ở trên.

Do đó, nếu kiểm định cho thấy không có ý nghĩa thống kê thì không nên áp dụng phân tích nhân tố cho các biến đang xem xét. Kiểm định Bartlett có ý nghĩa thống kê (sig Bartlett's Test < 0.05), chứng tỏ các biến quan sát có tương quan với nhau trong nhân tố.

3.2.2 Xoay nhân tố

Trong phần này ta xét ma trận nhân tố (Component Matrix). Ma trận này chứa hệ số biểu diễn các biến chuẩn hóa bằng các nhân tố (mỗi biến là một đa thức của các nhân tố). Những hệ số tải này (factor loading) biểu diễn tương quan giữa các nhân tố và các biến. Hệ số này lớn cho biết nhân tố và biến có liên hệ chặt chẽ với nhau. Các hệ số này được dùng để giải thích các nhân tố. Hệ số tải nhân tố $FactorLoading > 0.5$. Nếu biến quan sát nào có hệ số tải nhân tố thấp hơn 0.5 sẽ bị loại nhằm đảm bảo tập dữ liệu đưa vào là có ý nghĩa cho phân tích nhân tố.

Trong ma trận nhân tố, nếu có nhiều biến có hệ số tải 0.5 ta tiến hành xoay nhân tố để các hệ số lớn hơn hơn 0.5. Có nhiều phương pháp xoay nhưng phương pháp xoay varimax là phổ biến nhất và thường được sử dụng để xoay các phương pháp thành phần chính.

Tải trọng dương cho biết một biến và một thành phần chính có tương quan thuận: sự gia tăng của một trong những kết quả là sự gia tăng của thành phần kia. Tải trọng âm cho thấy mối tương quan âm. Tải trọng lớn (có thể là tích cực hoặc tiêu cực) cho thấy rằng một biến có ảnh hưởng mạnh mẽ đến thành phần chính đó.

Varimax là một vòng quay trực giao của các trục nhân tố để tối đa hóa sự thay đổi của các tải trọng bình phương của một nhân tố (cột) trên tất cả các biến (hàng) trong một ma trận nhân tố.

Trong bài báo cáo này, chúng tôi sử dụng hệ số tải ứng với 90 quan sát là 0.6 và sử dụng phương pháp xoay nhân tố varimax

Tổng kết chương

Trong chương này ta sẽ nói về phương pháp phân tích nhân tố với hai ý chính là dẫn nhập và thuật toán phân tích nhân tố. Đầu tiên ta biết được phân tích nhân tố là các phương pháp rút gọn dữ liệu trên cơ sở tìm mối liên quan của các biến liên tục để từ đó giải thích chúng bằng vài nhân tố hoặc thành tố. Thứ hai là dẫn nhập ta sẽ phát biểu phân tích nhân tố nói chung là gì và phân tích nhân tố được sự dụng trong các

trường hợp nào. Phân tích nhân tố nói chung là một nhóm các thuật toán được sử dụng chủ yếu để thu gọn và tóm tắt các dữ liệu. Phân tích nhân tố được sử dụng trong ba trường hợp

- Nhận diện một tập hợp gồm một số ít lượng biến mới, không tương quan với nhau để thay thế tập biến gốc có tương quan với nhau để thực hiện một phân tích đa biến tiếp theo.
- Nhận diện các khía cạnh hay nhân tố giải thích được các liên hệ tương quan trong một tập biến.
- Nhận diện một tập hợp gồm một số ít các biến nổi trội từ một tập hợp nhiều biến để sử dụng trong các phân tích thống kê đa biến.

Thứ ba là thuật toán phân tích nhân tố trong đó bao gồm kiểm định Bartlett, KMO và xoay nhân tố.

Chương 4

THỰC NGHIỆM

Trong những tháng cuối năm 2019, các nhà khoa học đã báo cáo một chủng mới của vi-rút corona, được lấy tên là 2019-nCov (hoặc Covid-19). COVID-19 là bệnh do một loại coronavirus mới có tên là SARS-CoV-2 gây ra. Sars-CoV-2 là loại vi-rút dòng Corona thứ 7 lây nhiễm sang người. Trong đó SARS, MERS và Sars-CoV-2 là loại vi-rút nguy hiểm, gây tổn thương nghiêm trọng đến đường hô hấp của cơ thể. Còn HKU1, NL63, OC43 và 229E hầu như để lại rất ý triệu chứng. WHO lần đầu tiên biết đến loại vi-rút mới này vào ngày 31 tháng 12 năm 2019, sau một báo cáo về một nhóm các trường hợp "viêm phổi do vi rút" ở tỉnh Vũ Hán thuộc Cộng hòa Nhân dân Trung Hoa.

Từ tháng một đến tháng tư năm 2020, bệnh dịch đã trở thành đại dịch lan rộng ra toàn thế giới và số người bệnh cũng như tử vong đối với bệnh này tăng rất nhanh qua từng ngày ở hầu khắp các quốc gia.

Trải qua bốn đợt dịch, Việt Nam hiện nay đang phải hứng chịu tác động tiêu cực về kinh tế, xã hội. Đến giờ phút này nguy cơ lan nhanh của dịch bệnh rất lớn với biến chủng mới delta phát hiện gần đây. Hậu quả của đại dịch COVID 19 là chưa từng có trong lịch sử loài người.

4.1 Viêm phổi do vi-rút Corona

Đại dịch coronavirus 2019 (COVID-19) do coronavirus 2 (SARS-CoV-2) gây ra hội chứng hô hấp cấp tính nghiêm trọng đã gây ra nhiều tác hại cho sức khỏe và nền kinh tế toàn cầu. Sự hiểu biết về quá trình phát sinh bệnh SARS-CoV-2 đã tiến bộ với tốc độ chưa từng có, nhưng những lỗ hổng quan trọng vẫn còn và những phát hiện sơ bộ cần được xác nhận, nhất là đối với phía Việt Nam

Viêm phổi do vi-rút Covid-19 tác động đến mỗi người theo những cách khác nhau. Hầu hết những người nhiễm vi-rút sẽ có triệu chứng bệnh từ nhẹ đến trung bình và có thể hồi phục mà không cần nhập viện. Những triệu chứng thường gặp nhất khi nhiễm vi-rút này là sốt, ho khan và mệt mỏi; Ít gặp hơn là đau nhức, đau họng, tiêu chảy, viêm kết mạc, đau đầu, mất vị giác hoặc khứu giác hoặc da nổi mẩn hay ngón tay hoặc ngón chân bị tấy đỏ hoặc tím tái.

Trong quá trình tầm soát, một chủng mới delta có khả năng lây lan rất cao được phát hiện đã bắt đầu đợi dịch thứ IV kéo dài cho đến hiện tại (tháng Tám, 2021). Các báo cáo trường hợp xác nhận có bệnh được nhiều trang web cập nhật hằng ngày.

Chúng tôi nghiên cứu tình hình dịch tễ đối với 18 tỉnh/thành phố thuộc phía Nam (Nam bộ) bao gồm các tỉnh xếp theo mức độ nguy hiểm hiện nay gồm TP. Hồ Chí Minh, Tiền Giang, Long An, An Giang, Bến Tre, Cần Thơ, Vĩnh Long, Trà Vinh, Cà Mau, Hậu Giang, Kiên Giang, Sóc Trăng, Bạc Liêu, Đồng Tháp, Bình Dương, Bà Rịa - Vũng Tàu (viết tắt Vũng Tàu) và Bình Phước.

4.2 Tổng quan về việc thực hiện

4.2.1 Dữ liệu nghiên cứu

Dữ liệu nghiên cứu bao gồm toàn bộ các trường hợp ghi nhận nhiễm bệnh cộng dồn cũng như theo dõi theo ngày trên 18 tỉnh/thành phố phía nam kể từ ngày bắt đầu đợt dịch thứ IV ngày 27/4/2021 đến ngày 31/7/2021 (tức 96 ngày). Dữ liệu được thu thập từ trang web infographics (cập nhật mỗi 6h và 18h) và tham khảo thêm các nguồn từ trang An toàn Covid từ Bộ Y tế (cập nhật mỗi 11h) và trang báo Vnexpress.net liên tục cập nhật dữ liệu tích lũy trong suốt đợt dịch thứ IV. Còn dữ liệu theo dõi theo ngày được suy ra từ bộ dữ liệu cộng dồn bằng cách tính số ca xác nhận nhiễm bệnh hôm sau trừ cho số ca nhiễm hôm trước.

Phần tiếp theo, chúng tôi nêu lên một số tiêu chuẩn đánh giá khác nhau phục vụ cho tác vụ phân tích thành phần chính và phân tích nhân tố.

4.2.2 Các tiêu chuẩn đánh giá mô hình

1. Tiêu chuẩn đánh giá dựa trên giá trị p-value

- Khi $p - value > 0.05$: Sự khác biệt không có ý nghĩa thống kê;
- Khi $p - value < 0.05$: Sự khác biệt có ý nghĩa thống kê;
- Khi $p - value < 0.01$: Sự khác biệt rất có ý nghĩa thống kê;
- Khi $p - value < 0.001$: Sự khác biệt rất có ý nghĩa thống kê rất lớn.

2. Tiêu chuẩn đánh giá hệ số tương quan dựa trên giá trị ρ

- Khi $-1 < \rho < -0.5$: Tương quan nghịch khá cao;
- Khi $-0.5 < \rho < 0.5$: Không có tương quan;
- Khi $0.5 < \rho < 0.8$: Tương quan thuận khá cao;
- Khi $0.8 < \rho < 1$: Tương quan thuận rất cao.

3. Tiêu chuẩn đánh giá thích hợp của phân tích nhân tố trong kiểm định KMO

- Khi Overall MSA ≥ 0.6: Phù hợp để phân tích nhân tố;
- Khi Overall MSA ≥ 0.7: Rất phù hợp để phân tích nhân tố;
- Khi Overall MSA ≥ 0.8: Sự phù hợp để phân tích nhân tố là rất lớn.

4. Tiêu chuẩn chọn hệ số tải

- Khi $FactorLoading = 0.60$ khi kích thước mẫu tối thiểu 85;
- Khi $FactorLoading = 0.55$ khi kích thước mẫu tối thiểu 100;
- Khi $FactorLoading = 0.5$ khi kích thước mẫu tối thiểu 120;

4.2.3 Thiết kế nghiên cứu

1. Đầu tiên, tổng hợp mô tả các biến trong dữ liệu để có cái nhìn tổng quát đối với dữ liệu.
2. Tiếp theo, mối quan hệ của các ca xác nhận nhiễm bệnh viêm phổi do vi-rút Corona gây ra giữa các tỉnh/thành phố được thiết lập sử dụng hệ số tương quan Pearson.
3. Sau đó, dựa trên tỷ lệ lây lan, các tỉnh/thành phố được phân loại sử dụng phân tích thành phần chính.
4. Tiếp theo, tiến hành kiểm định Kaiser-Meyer-Olkin (KMO) xem xét sự thích hợp của phân tích nhân tố đến dữ liệu.
5. Cuối cùng, phân tích nhân tố được sử dụng để thiết lập các yếu tố quan trọng

4.3 Đọc và xử lý số liệu

Số liệu được lưu trữ trên trang Github nên khi tải và lưu giải nén trong ổ đĩa cá nhân (khuyến khích sử dụng dữ liệu được lưu ở ổ đĩa **D**), ta thực hiện đọc dữ liệu vào ngôn ngữ lập trình thống kê R như sau

Dữ liệu được lưu ở ổ đĩa **D** với tên file là PCA_for_Covid. Ta sử dụng lệnh **setwd** để truy cập vào dữ liệu dựa trên đường dẫn như sau

```
setwd("D:/PCA_for_Covid/PCA/Data")
```

Dữ liệu được lưu dưới dạng tệp covid_case.csv (dữ liệu hằng ngày) và covid_cul.csv (dữ liệu tích lũy), ta sử dụng lệnh **read.csv()** để đọc dữ liệu vào R.

```
covid_cul <- read.csv("covid_case.csv",
                 header = TRUE,
                 sep = ",",
                 stringsAsFactors = FALSE)
covid_case <- read.csv("covid_cul.csv",
                 header = TRUE,
                 stringsAsFactors = FALSE)
```

Tiếp theo ta sử dụng hàm **as.Data()** để chuyển định dạng của dữ liệu về đúng dạng với dữ liệu thời gian.

```
covid_case$Day <- as.Date(covid_case$Day, format = "%d/%m/%Y")
covid_cul$Day <- as.Date(covid_cul$Day, format = "%d/%m/%Y")
```

Ta tiếp tục chọn tất cả các biến dữ liệu mà không cần sử dụng đến biến *Day* để dễ dàng trong các phân tích tiếp theo hơn.

```
case_data <- covid_case %>% select(., -Day)
cul_data <- covid_cul %>% select(., -Day)
```

Dữ liệu bao gồm 19 biến với cỡ mẫu là 90. Ta có tổng quan dữ liệu tích lũy các ca xác nhận nhiễm Covid được thể hiện qua lệnh **dim()**

```
covid_case %>% dim()
```

```
## [1] 96 19
```

```
covid_cul %>% dim()
```

```
## [1] 96 19
```

Tên các biến được trình bày bằng dòng lệnh dưới đây

```
case_data %>% names()
```

```
##  [1] "Day"            "TP.Ho.Chi.Minh" "Tien.Giang"     "Long.An"
##  [5] "An.Giang"       "Ben.Tre"        "TP.Can.Tho"     "Vinh.Long"
##  [9] "Tra.Vinh"       "Ca.Mau"         "Hau.Giang"      "Kien.Giang"
## [13] "Soc.Trang"      "Bac.Lieu"       "Dong.Thap"      "Binh.Duong"
## [17] "Vung.Tau"       "Tay.Ninh"       "Binh.Phuoc"
```

4.4 Một số thống kê mô tả cho hai dữ liệu

Trong phần này, mô tả về tập dữ liệu của nghiên cứu và giới thiệu phân tích thành phần chính được trình bày.

Trong khuôn khổ bài báo cáo ngắn, chúng tôi chỉ trực quan dữ liệu với 6 mẫu ngẫu nhiên được chọn từ dữ liệu. Ta có cột đầu tiên trong dữ liệu là ngày bắt đầu đợt dịch thứ IV từ 27/4/2021 đến ngày 31/7/2021. Các cột còn lại lần lượt là các 18 tỉnh/thành phố phía Nam được chọn để phân tích gồm TP. Hồ Chí Minh, Tiền Giang, Long An, An Giang, Bến Tre, Cần Thơ, Vĩnh Long, Trà Vinh, Cà Mau, Hậu Giang, Kiên Giang, Sóc Trăng, Bạc Liêu, Đồng Tháp, Bình Dương, Bà Rịa - Vũng Tàu (viết tắt Vũng Tàu) và Bình Phước.

Ta xem xét 6 dòng dữ liệu được lấy ngẫu nhiên từ dữ liệu như sau

```
covid_case %>% sample_n(., 6)
```

```
##          Day TP.Ho.Chi.Minh Tien.Giang Long.An An.Giang Ben.Tre Can.Tho
## 1 2021-07-23           4913         94      602        2      23      34
## 2 2021-07-01            464         38       28        5       0       0
## 3 2021-07-04            599         29       72        6       1       0
## 4 2021-07-09           1229         34       77        5       0       6
## 5 2021-06-24            162          9        2        0       0       0
## 6 2021-07-15           2691          0       41        8      30      11
##   Vinh.Long Tra.Vinh Ca.Mau Hau.Giang Kien.Giang Soc.Trang Bac.Lieu Dong.Thap
## 1        12       15       2         4         13         0         0       129
## 2         1        0       0         0          0         0         0         1
## 3         2        1       0         0          0         0         0         6
## 4         0        8       0         4          0         2         2        32
## 5         0        0       0         0          0         0         0         0
## 6        17        3       1         0          0         4         0        99
##   Binh.Duong Vung.Tau Tay.Ninh Binh.Phuoc
## 1        608       58      212          4
## 2         90        0        0          1
## 3         87        2        2          0
## 4         73        4        0          1
## 5         27        0        2          0
## 6        122       17        0         13
```

Đối với dữ liệu tích lũy, đầu tiên, ta xem xét 6 dòng cuối cùng của dữ liệu như sau

```
covid_cul %>% tail()
```

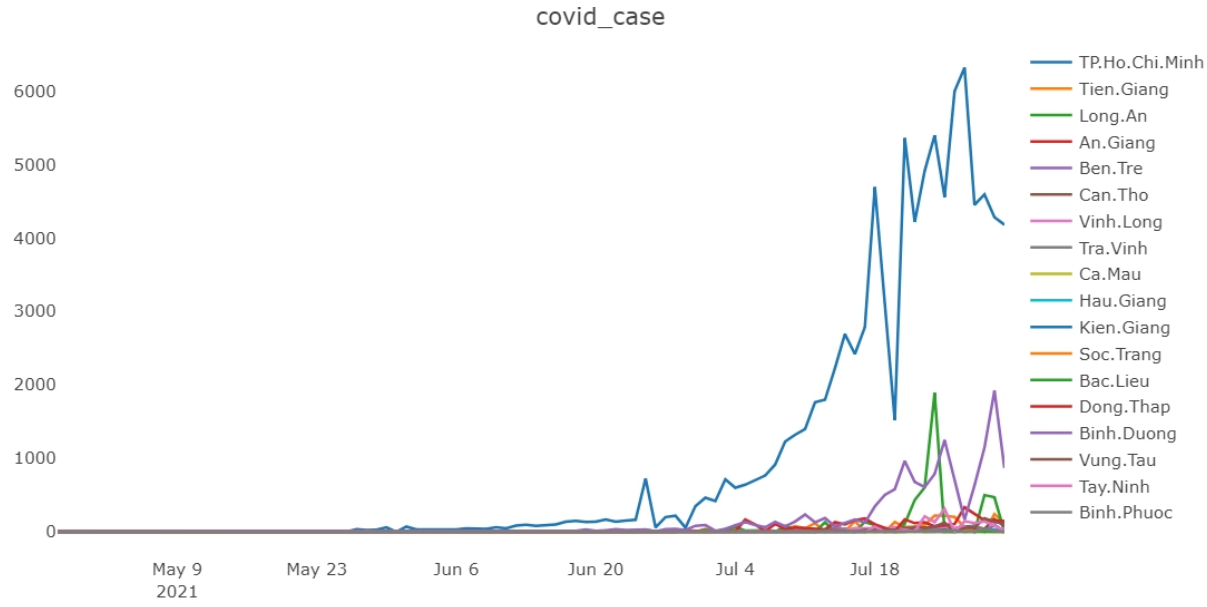
```
##           Day TP.Ho.Chi.Minh Tien.Giang Long.An An.Giang Ben.Tre TP.Can.Tho
## 91 2021-07-26          66422       1762     3856      183     491         376
## 92 2021-07-27          72740       1825     3931      226     551         447
## 93 2021-07-28          77189       1855     3931      250     635         518
## 94 2021-07-29          81781       1855     4430      260     635         557
## 95 2021-07-30          86063       2097     4899      276     732         731
## 96 2021-07-31          90243       2220     4899      278     732         803
##    Vinh.Long Tra.Vinh Ca.Mau Hau.Giang Kien.Giang Soc.Trang Bac.Lieu Dong.Thap
## 91       635      142     24        92        148        87       23      2064
## 92       708      145     25       102        161       109       24      2397
## 93       708      237     27       108        161       121       24      2641
## 94       739      255     31       121        182       121       28      2798
## 95       754      291     31       149        199       121       28      2955
## 96       802      291     31       168        215       121       28      3101
##    Binh.Duong Vung.Tau Tay.Ninh Binh.Phuoc
## 91       8743      555      794        133
## 92       8909      607      938        133
## 93       9540      663     1058        136
## 94      10684      848     1197        171
## 95      12604      981     1285        183
## 96      13472     1096     1285        183
```

Biểu đồ sau đây hình 4.1 thể hiện số ca nhiễm hằng ngày và số ca nhiễm tích lũy.

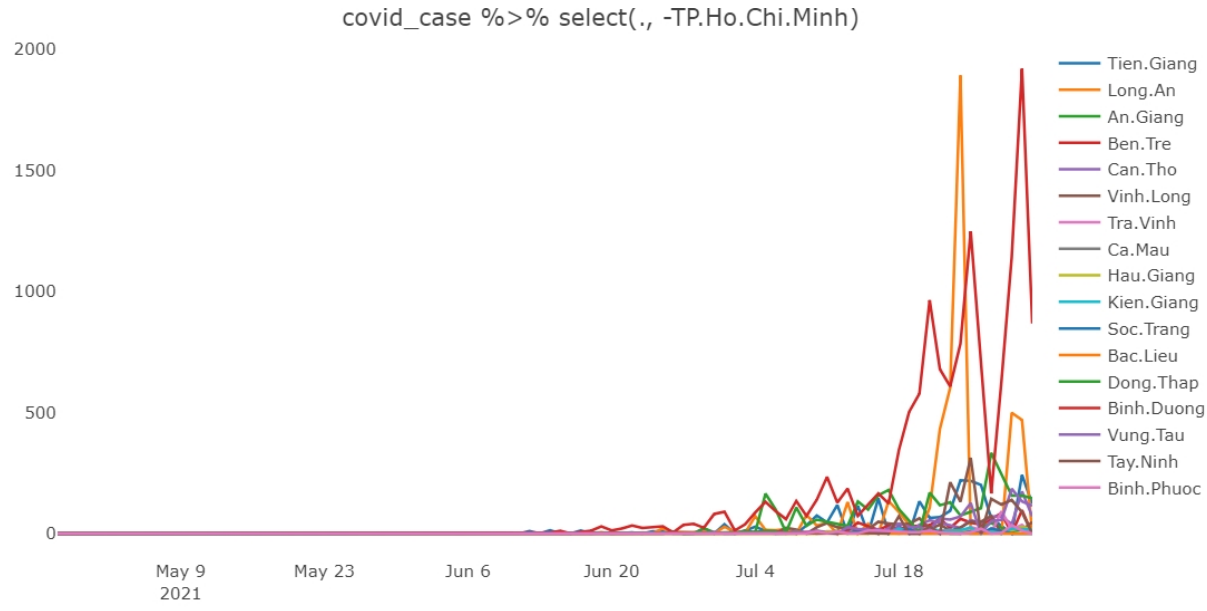
Khi xem xét toàn bộ dữ liệu, ta thấy các trường hợp xác nhận nhiễm bệnh ở các tỉnh phía Nam bắt đầu có dấu hiệu bùng phát từ khoảng cuối tháng năm, tức sau khi bắt đầu đợt dịch lớn ở Bắc Giang khoảng một tháng. Bắt đầu nhiễm mạnh ở thành phố Hồ Chí Minh sau đó, dịch bệnh lan rộng ra các tỉnh Đông Nam Bộ và bùng phát toàn phía Nam.

Nhìn tổng quan đồ thị hình 4.1, ta nhận thấy sự khác biệt rõ ràng những ca có bệnh giữa các tỉnh. Trong đó, thành phố Hồ Chí Minh có số lượng người bệnh được xác nhận là cao nhất mỗi ngày đỉnh điểm lên đến gần 6000 ca/ngày. Các tỉnh Bình Dương và Long An cũng có xu hướng tăng mạnh vào những tuần gần đây nhất và cao nhất gần 2000 ca mắc một ngày. Các tỉnh còn lại có số ca mắc không quá cao (dưới 100 ca/ngày) nhưng vẫn có xu hướng tăng và tăng dài kỳ.

Đồ thị tích lũy các ca xác nhận bệnh có Covid-19 thể hiện xu hướng tăng chưa có dấu hiệu đỉnh dịch. Bình Dương có số ca nhiễm bệnh cao sau thành phố Hồ Chí Minh Xu hướng tăng mạnh ở tỉnh Long An khi trong khoảng thời gian ngắn (từ ngày 21/7 đến ngày 24/7, tức 3 ngày) nhưng số ca mắc tăng đột ngột 2926 ca nhiễm.

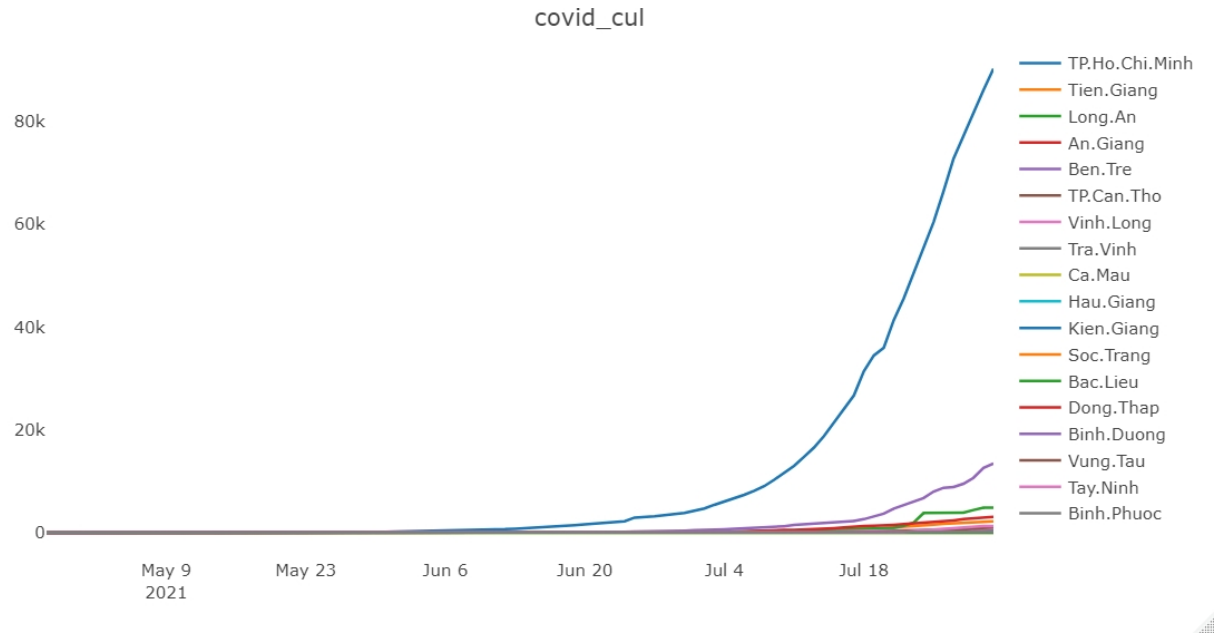


(a) Đồ thị thể hiện số lượng ca nhiễm hằng ngày tính từ ngày 27/4

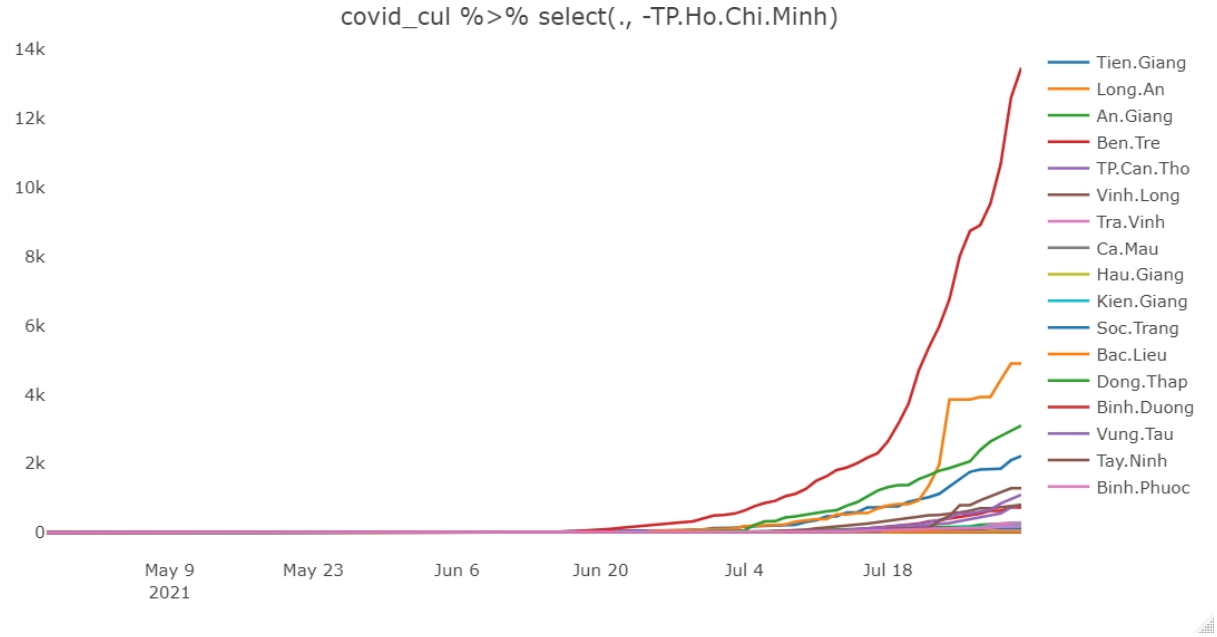


(b) Đồ thị thể hiện số lượng ca nhiễm hằng ngày tính từ ngày 27/4 trừ thành phố Hồ Chí Minh

Hình 4.1: Đồ thị số ca nhiễm hằng ngày. Để được tiện trong tra cứu các số liệu, chúng tôi trực quan đồ thị thể hiện số lương ca nhiễm trừ thành phố Hồ Chí Minh. Ta nhận ra có ba tỉnh/thành phố có số ca nhiễm trong ngày khá cao và khác biệt với các tỉnh/thành phố khác là Thành phố Hồ Chí Minh, Bình Dương và Long An đều nằm tập trung ở Đông Nam Bộ và đều có ranh giới với nhau.



(a) Đồ thị thể hiện số lượng ca nhiễm tích lũy tính từ ngày 27/4



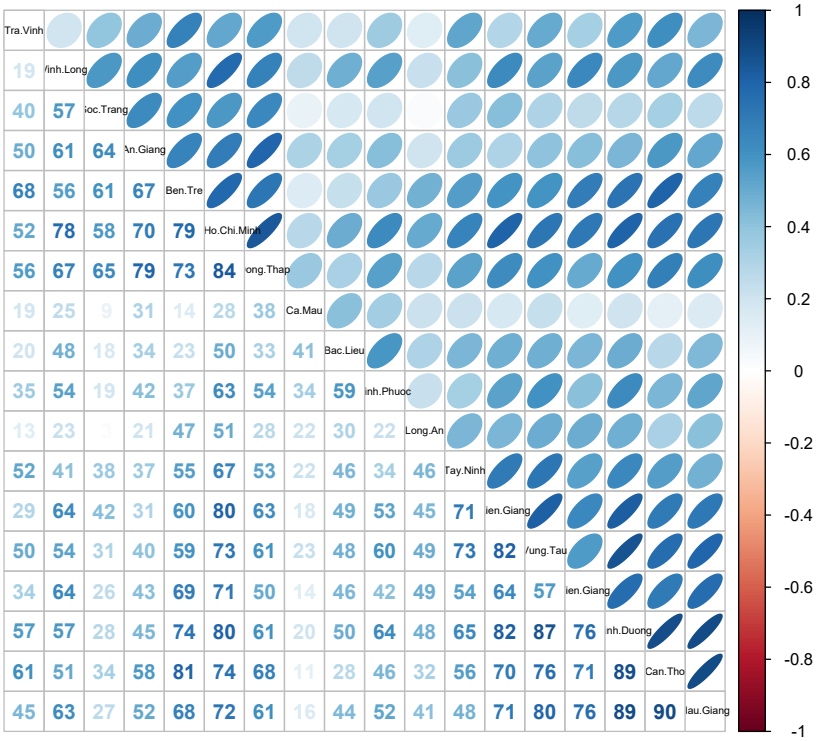
(b) Đồ thị thể hiện số lượng ca nhiễm tích lũy tính từ ngày 27/4 trừ thành phố Hồ Chí Minh

Hình 4.2: Đồ thị số ca nhiễm tích lũy nhằm thể hiện tốc độ tăng nhanh các ca nhiễm. Đồ thị (b) cho ta thấy tỉnh Bình Dương có xu hướng tăng sau thành phố Hồ Chí Minh. Cũng trong đồ thị (b) ta thấy rõ hơn mức độ tăng bất thường của tỉnh Long An.

4.5 Mối tương quan đối với số ca nhiễm bệnh giữa các tỉnh

Đối với dữ liệu các ca nhiễm thu thập hằng ngày, tương quan đồ dưới đây thể hiện quan hệ giữa các biến ứng với số ca nhiễm mỗi ngày. Tương quan thuận cao cho biết các biến có xu hướng đồng biến giữa hai cặp biến và nhằm dự báo xu hướng tăng trong tương lai.

```
corrplot::corrplot.mixed(cor_data <- case_data %>% cor(),
                tl.cex = 0.5,
                tl.col = "black",
                order ="hclust",
                addCoefasPercent = TRUE,
                upper="ellipse")
```

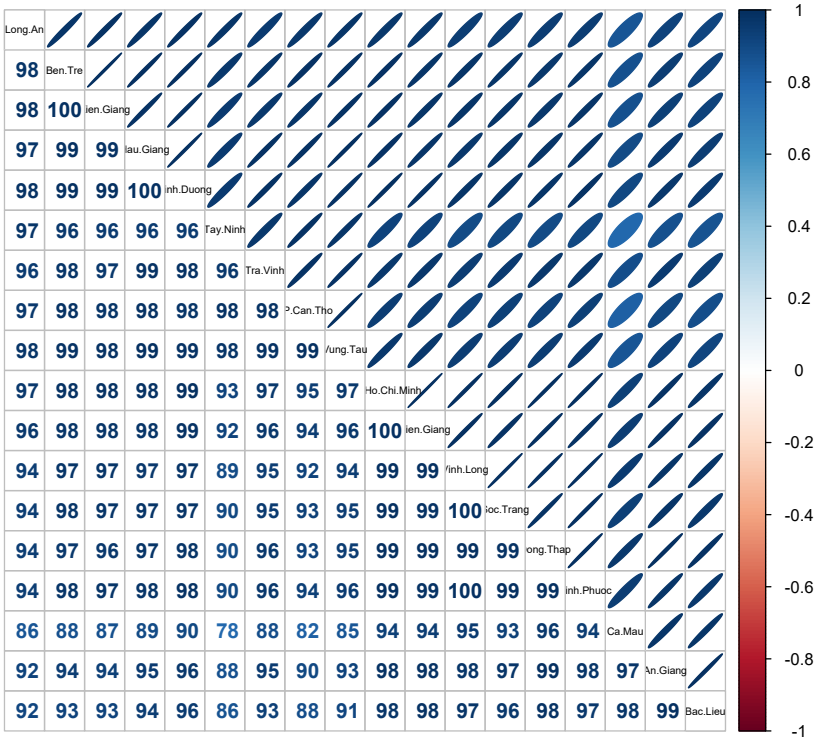


Hình 4.3: Tương quan đồ thể hiện tương quan dữ liệu hằng ngày các ca xác nhận nhiễm ở các tỉnh/thành phố. Trong đồ thị này, hệ số tương quan càng lớn thể hiện xu hướng tăng các ca nhiễm theo ngày càng cao giữa hai biến bất kỳ. Mức độ (độ lớn) của tương quan thể hiện bởi mức độ đậm nhạt của màu sắc được chú thích trong phổ màu bên phải.

Qua tương quan đồ, không có một cặp biến nào có tương quan âm. Tức là mỗi ngày, dự báo về số ca nhiễm sẽ có thể tiếp tục tăng. Một số cặp tỉnh/thành phố có hệ số tương quan khá cao như Hậu Giang – Cần Thơ (90%), Bình Dương – Tp. Hồ Chí Minh (80%) hoặc Bình Dương – Hậu Giang (89%) là những tỉnh/thành phố có vị trí địa lý giáp nhau hoặc chịu ảnh hưởng bởi thành phố có nhiều ca bệnh.

Từ dữ liệu tích lũy, ta có ma trận tương quan Pearson cho 18 biến phụ thuộc tạo thành tương quan đa điểm. Các hệ số tương quan trong từng cặp biến được thể hiện qua tương quan đồ.

```
corrplot::corrplot.mixed(cor_data <- cul_data %>% cor(),
                 tl.cex = 0.5,
                 tl.col = "black",
                 order ="hclust",
                 addCoefasPercent = TRUE,
                 upper="ellipse")
```

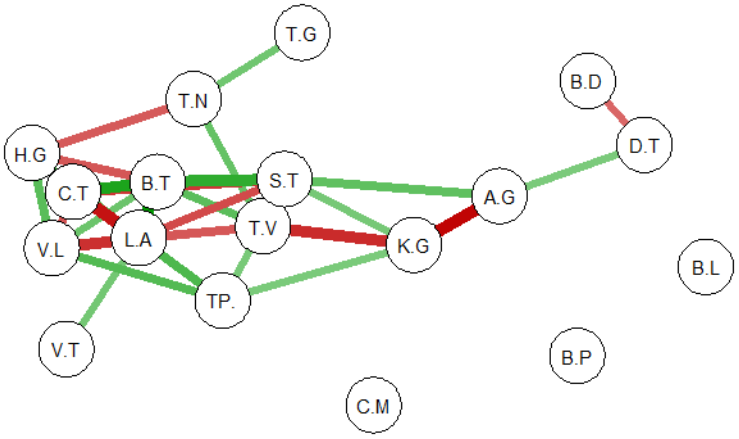


Hình 4.4: Tương quan đồ thể hiện tương quan dữ liệu tích lũy các ca xác nhận nhiễm ở các tỉnh/thành phố. Trong biểu đồ này, các số liệu thể hiện phần trăm tương quan ở mỗi biến tuân theo phổ màu; Các hình ellipse mô tả tương quan thuận ứng với số phần trăm tương quan khi diện tích càng nhỏ; Tương quan đồ đã được phân cụm và tuân theo phổ màu phía bên phải.

Ta dễ dàng nhận thấy và cũng không quá ngạc nhiên rằng từ hình 4.4 hầu hết các cặp biến tương quan thuận rất cao. Đặc biệt ở một số tỉnh tương quan chặt chẽ có hệ số tương quan gần như 100%. Một số cặp biến tương quan không quá cao như Cà Mau – Tây Ninh (78%) hoặc Cần Thơ – Cà Mau (82%) có thể được suy diễn nguyên nhân khác biệt khoảng cách địa lý và nhiều lệnh giãn cách xã hội được đặt ra nối tiếp nhau khi dịch bệnh bắt đầu bùng phát nhanh ở các tỉnh miền Nam.

Trong trường hợp phân tích tương quan, mối liên kết giữa 2 biến được xác lập nhờ vào giá trị của r và p-value. Ta sẽ mượn các công cụ và lý thuyết Network analysis cho mục tiêu phân tích tương quan

```
Graph_pcor <- case_data %>% cor() %>%
  qgraph::qgraph(.,
                 graph = "pcor",
                 layout = "spring",
                 threshold = "bonferroni",
                 sampleSize = nrow(case_data),
                 alpha = 0.05)
```

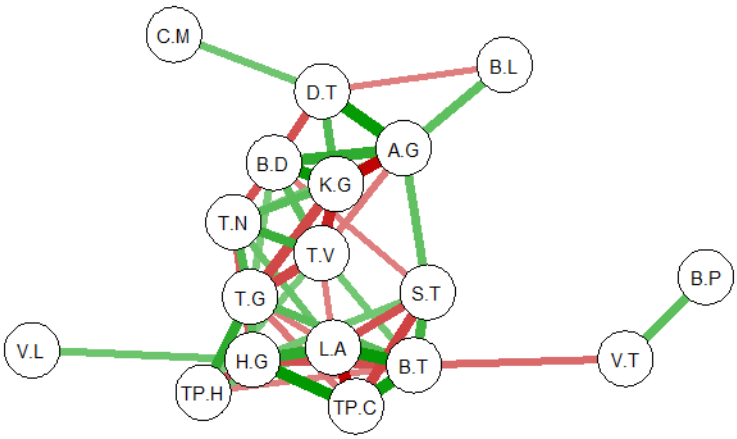


Hình 4.5: Mạng tương quan pcor đối với dữ liệu ca bệnh thu nhập hằng ngày. Đồ thị nhằm đánh giá sự tương tác giữa các biến dữ liệu qua khoảng cách các biến, độ đậm nhạt của các dây cung và màu sắc

Từ đồ thị mạng tương quan, ta thấy rõ hơn sự tương tác giữa các tỉnh/thành phố từ dữ liệu thu thập hằng ngày. Các tỉnh tập trung thành nhóm biểu thị sự tương tác cao. Các tỉnh không có tương tác nghĩa là hệ số tương quan không có ý nghĩa thống kê với mức ý nghĩa 0.05. Độ dài và độ đậm của dây cung biểu thị mức độ tương tác cụ thể giữa các cặp biến. Tỷ lệ của các cạnh theo chiều rộng và độ bão hòa màu. Các cạnh có trọng lượng tuyệt đối lớn hơn giá trị này sẽ có cường độ màu mạnh nhất và càng rộng càng mạnh và các cạnh có trọng lượng tuyệt đối dưới giá trị này sẽ có chiều rộng nhỏ nhất và càng mờ thì trọng lượng càng yếu. Màu xanh thể hiện tương tác thuận, màu đỏ thể hiện tương tác nghịch.

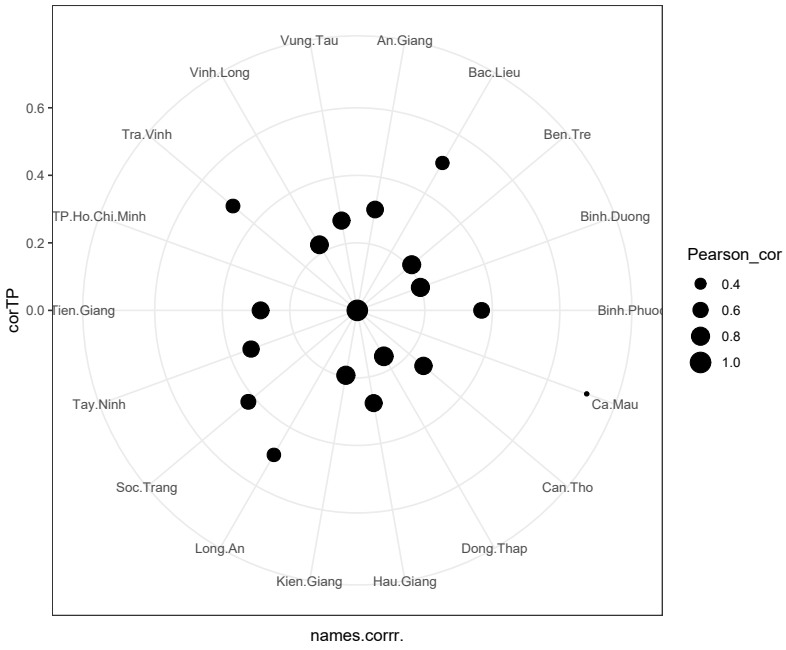
Trong đồ thị mạng tương quan, cụm các biến Can.Tho, Ben.Tre, Vinh.Long, Hau.Giang tương tác cao với nhiều tỉnh thành khác và tương tác lẫn nhau. TP. Hồ Chí Minh tương tác thuận cao với các tỉnh Vĩnh Long, Long An, Trà Vinh và Kiên Giang. Ba biến Ca.Mau, Binh.Phuoc và Bạc Liêu tương quan không có ý nghĩa thống kê.

```
Graph_pcor <- cul_data %>% cor() %>%
  qgraph::qgraph(.,
                 graph = "pcor",
                 layout = "spring",
                 threshold = "bonferroni",
                 sampleSize = nrow(cul_data),
                 alpha = 0.05)
```



Hình 4.6: Mạng tương quan pcor đối với dữ liệu ca bệnh tích lũy hằng ngày.

Trong đồ thị mạng tương quan trên, tất cả các biến đều có sự tương quan tập trung.



Hình 4.7: Biểu đồ tương quan giữa Thành phố Hồ Chí Minh và các tỉnh lân cận. Biểu đồ thể hiện mối tương quan Pearson giữa biến Tp.Ho.Chi.Minh và các biến khác trong hệ tọa độ cực. Tâm điểm của biểu đồ tương ứng với giá trị $r = 1$ (tương quan mạnh nhất) còn ngoại vi biểu đồ tương ứng với $r = 0$ (không có tương quan). Từ đó nhận thấy, khi càng gần về tâm biểu đồ, ta có tương quan mạnh giữa Tp. Hồ Chí Minh và ngược lại.

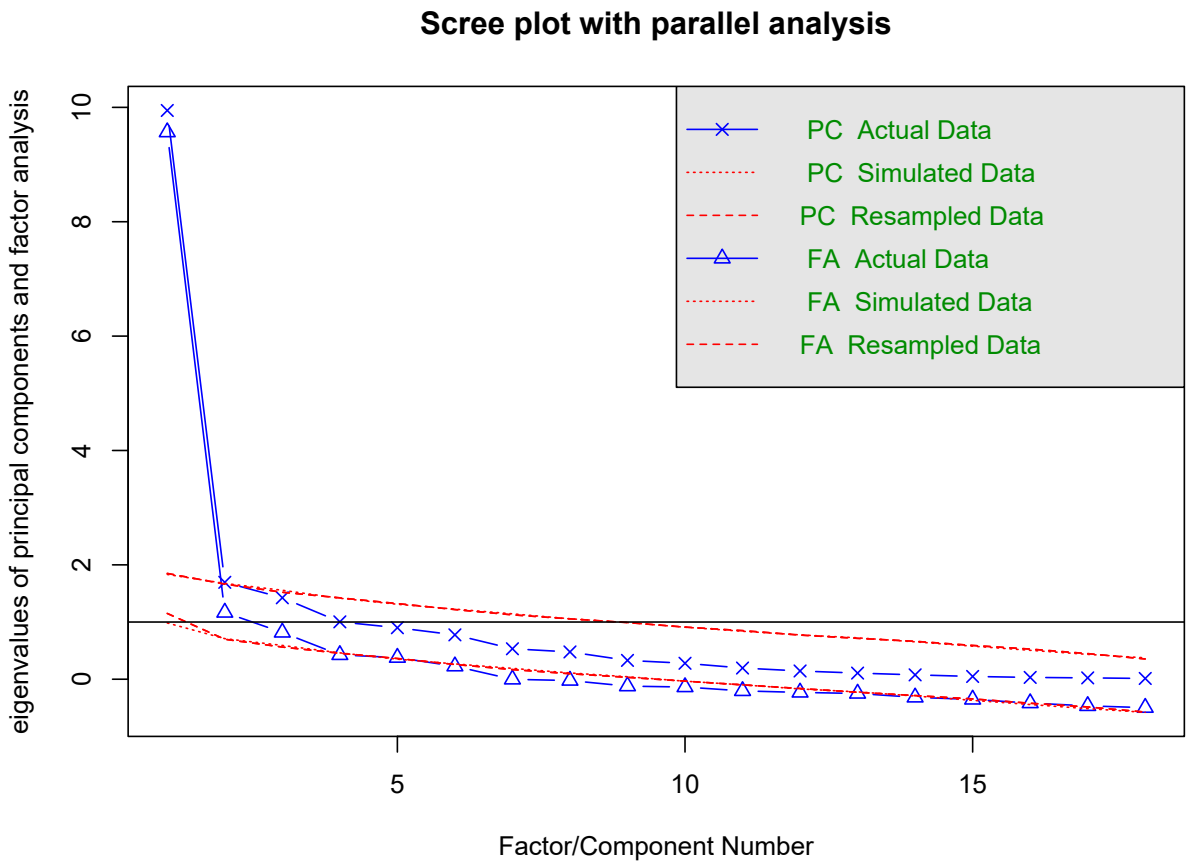
4.6 Phân tích thành phần chính

Ta tiến hành thuật toán phân tích thành phần chính để chọn ra số thành phần với hy vọng phản ánh cao nhất phần trăm phương sai dữ liệu. Để dễ dàng trong phân tích, phần này chia thành hai mục phân tích riêng từng loại dữ liệu.

4.6.1 Dữ liệu case_data

```
case_data %>% psych::fa.parallel(.,
                main = "Scree plot with parallel analysis")
```

```
## Parallel analysis suggests that the number of factors =  3  and the number of
## components =  1
```



Hình 4.8: Sơ đồ sàng lọc với phân tích song song dữ liệu ca nhiễm hằng ngày. Sơ đồ cũng đề nghị số thành phần là 1 và số nhân tố là 3. Đường vạch đen thể hiện ngưỡng giá trị riêng 1.0 theo tiêu chuẩn chọn các thành phần chính.

Qua sơ đồ sàng lọc giữa số thành phần chính và các giá trị riêng khi phân tích thành phần, ta thấy số thành phần có giá trị riêng lớn hơn 1 là 2 thành phần. Để có thể tính toán chính xác giá trị của các giá trị riêng, bài báo cáo sẽ để ở phần phân tích tiếp theo. Ngoài ra phân tích trong lệnh này cũng cho ta gợi ý lựa chọn số các nhân tố sẽ hữu dụng khi phân tích nhân tố cũng như số thành phần chính.

Ta chọn số thành phần là nhiều hơn hai (thành phần) với mục tiêu tìm được trên 60% tổng phương sai. Khi

đó các thành phần chính sẽ giải thích được tương đối đầy đủ dữ liệu.

Tiếp theo ta xem xét các giá trị độ lệch chuẩn, phần trăm phương sai cũng như phần trăm phương sai tích luỹ sau khi dùng lệnh **prcomp** có sẵn trong R để phân tích thành phần chính.

```
prcomp(case_data, center = TRUE, scale = TRUE) %>%
        summary()
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     3.1537 1.30172 1.19334 1.00157 0.94887 0.8808 0.72960
## Proportion of Variance 0.5525 0.09414 0.07911 0.05573 0.05002 0.0431 0.02957
## Cumulative Proportion  0.5525 0.64667 0.72578 0.78151 0.83153 0.8746 0.90420
##                           PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     0.69075 0.57371 0.52808 0.44200 0.3796 0.32740 0.27600
## Proportion of Variance 0.02651 0.01829 0.01549 0.01085 0.0080 0.00595 0.00423
## Cumulative Proportion  0.93071 0.94899 0.96449 0.97534 0.9833 0.98930 0.99353
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.21968 0.17769 0.15258 0.11543
## Proportion of Variance 0.00268 0.00175 0.00129 0.00074
## Cumulative Proportion  0.99621 0.99797 0.99926 1.00000
```

Thành phần chính đầu tiên có phần trăm phương sau giải thích 55.3% dữ liệu. Khi thành phần chính là 2 ta thấy phần trăm tích lũy ở PC2 đạt 64.7%. Việc giải thích được trên 60% dữ liệu là kết quả khá tốt khi phân tích thành phần chính. Nếu ta chọn số thành phần chính là 3 thì tổng phần trăm tích lũy chỉ tăng nhẹ (khoảng 8%) và nâng việc giải thích dữ liệu khi tăng thêm một chiều thành 72.6%.

Để tổng phần trăm phương sai trên 80%, ta chọn số thành phần chính là 5 thành phần từ 18 biến dữ liệu. Chúng tôi cũng khảo sát tỷ lệ đóng góp được trích gọn kết quả như sau

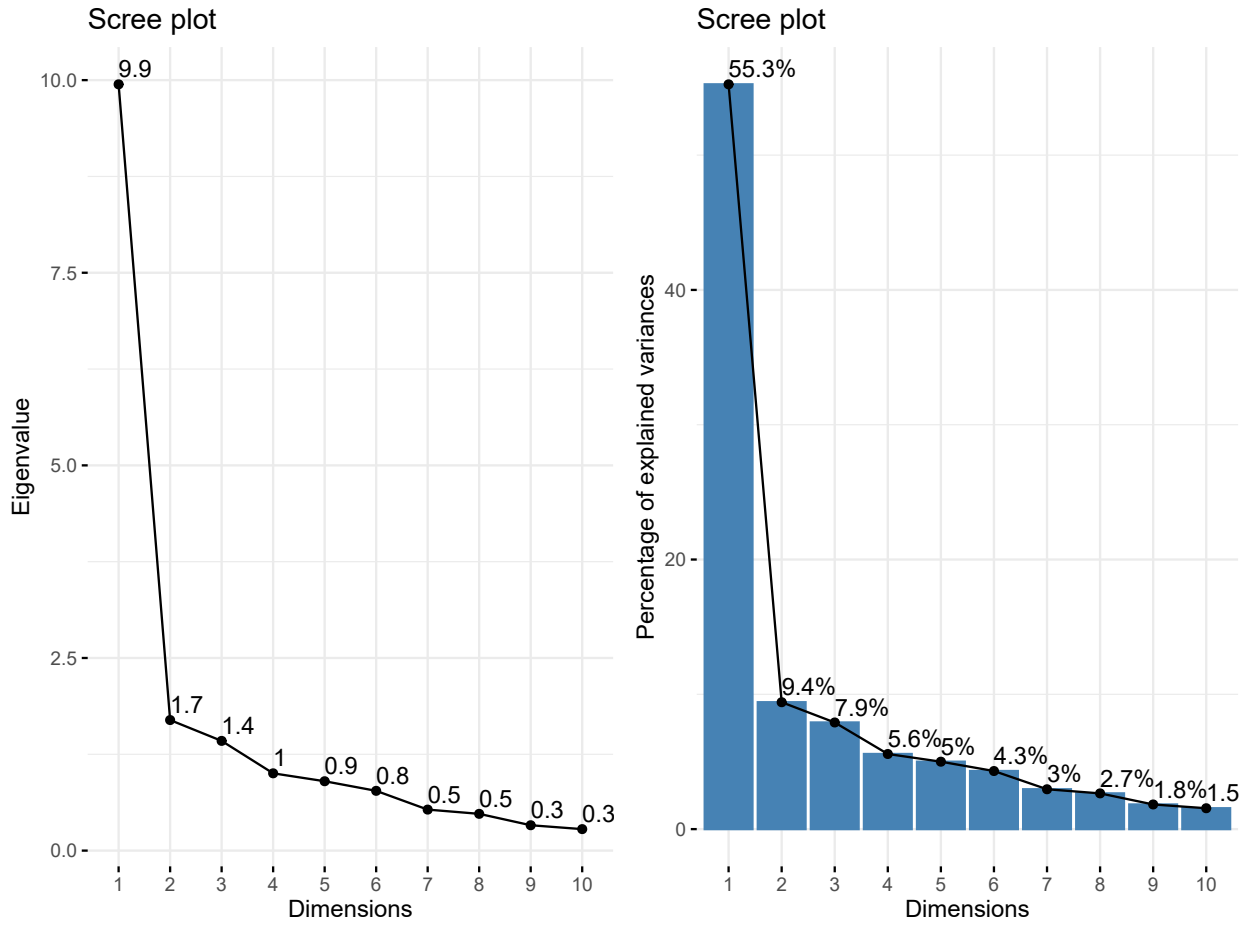
```
princomp(case_data, scores = TRUE) %>%
        loadings()
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## TP.Ho.Chi.Minh  0.985  0.160
## Tien.Giang                            0.563  0.436 -0.653 -0.126         0.129
## Long.An               -0.633  0.767
## An.Giang
## Ben.Tre                                             -0.166  0.481  0.419 -0.516
## Can.Tho                                             -0.166  0.255  0.586  0.351
## Vinh.Long                                                   -0.408        -0.726
## Tra.Vinh                                                     0.337
## Ca.Mau
## Hau.Giang
## Kien.Giang
## Soc.Trang                                                                 -0.168
## Bac.Lieu
## Dong.Thap                           -0.516 -0.406 -0.706 -0.104 -0.223
```

```
## Binh.Duong      0.154 -0.744 -0.634
## Vung.Tau                           -0.264                -0.605  0.608  0.136
## Tay.Ninh                           -0.571  0.791          0.102 -0.110
## Binh.Phuoc
##                Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
## TP.Ho.Chi.Minh
## Tien.Giang      0.152
## Long.An
## An.Giang               -0.414  -0.755   0.277          -0.284          -0.264
## Ben.Tre                 0.440  -0.141  -0.155   0.163           0.156
## Can.Tho        -0.509  -0.291   0.136  -0.148  -0.119          -0.167
## Vinh.Long      -0.277  -0.356   0.266
## Tra.Vinh        0.660  -0.464   0.385          -0.185  -0.105
## Ca.Mau                                                 -0.597           0.783
## Hau.Giang              -0.134           0.106  -0.291   0.225   0.889
## Kien.Giang     -0.139   0.342   0.216          -0.510  -0.597          -0.422
## Soc.Trang               0.183  -0.154   0.356  -0.669   0.328  -0.330   0.342
## Bac.Lieu                                               -0.153   0.170
## Dong.Thap
## Binh.Duong
## Vung.Tau        0.362   0.142
## Tay.Ninh       -0.130
## Binh.Phuoc      0.150  -0.104  -0.299  -0.854  -0.338   0.105
##                Comp.18
## TP.Ho.Chi.Minh
## Tien.Giang
## Long.An
## An.Giang
## Ben.Tre
## Can.Tho
## Vinh.Long
## Tra.Vinh
## Ca.Mau         -0.131
## Hau.Giang      -0.149
## Kien.Giang     -0.115
## Soc.Trang
## Bac.Lieu        0.967
## Dong.Thap
## Binh.Duong
## Vung.Tau
## Tay.Ninh
## Binh.Phuoc
```

Khi vẽ sơ đồ sàng lọc cụ thể cho các giá trị riêng và phần trăm phương sai trích, ta có số chiều có sự phân hóa cụ thể

```
pca_case <- FactoMineR::PCA(case_data, graph = FALSE)
left <- pca_case %>% factoextra::fviz_eig(.,
                choice = 'eigenvalue',
                geom = 'line',
                addlabels = TRUE,
                repel = TRUE)
right <- pca_case %>% factoextra::fviz_screeplot(.,
                addlabels = TRUE,
                repel = TRUE)
gridExtra::grid.arrange(left, right, ncol = 2)
```

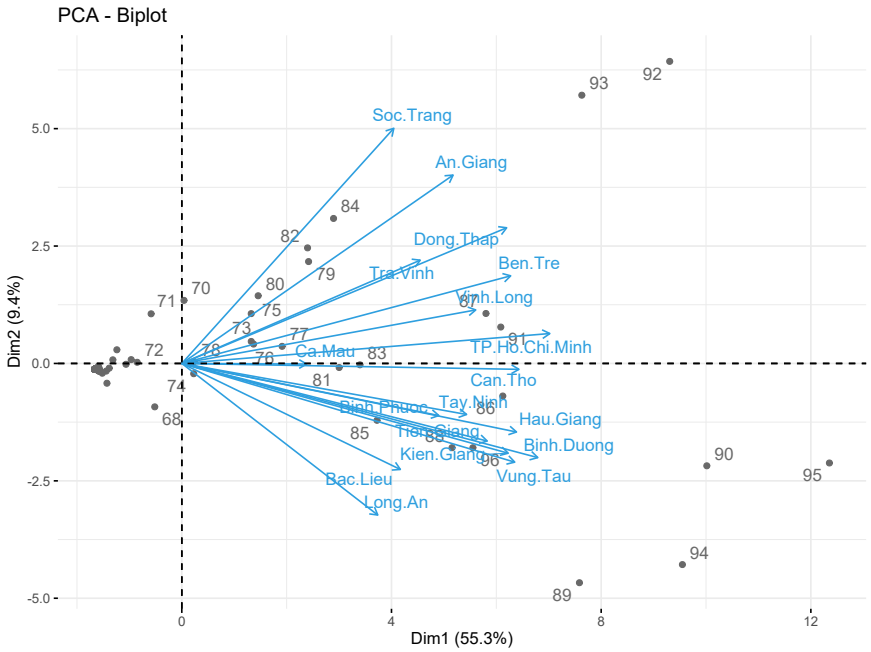


Hình 4.9: Sơ đồ sàng lọc cho kết quả đồ thị bên trái thể hiện giá trị riêng của từng thành phần chính. Trong đồ thị này, thành phần thứ nhất có giá trị riêng lớn nhất, thành phần thứ năm trở đi thì có giá trị riêng bé hơn 1. Sơ đồ bên phải thể hiện phần trăm phương sai được giải thích đối với từng thành phần.

Qua sơ đồ trên, hơn 80% các phương sai có thể được giải thích bởi chỉ 5 chiều (thành phần) đầu tiên, với thành phần thứ nhất giải thích 53.3% như đã biết.

Bước cuối cùng là hình dung sự phân bố của các mẫu trong không gian sắp xếp mới này, nó chỉ là một phép quay của không gian biến ban đầu của chúng ta. Điểm số mô tả vị trí của các mẫu và chúng ta vẽ biểu đồ này dưới dạng biểu đồ phân tán, bắt đầu với cặp thành phần chính quan trọng nhất (1 và 2). Lưu ý hai chiều đại diện này chỉ chiếm hơn 60% sự giải thích dữ liệu.

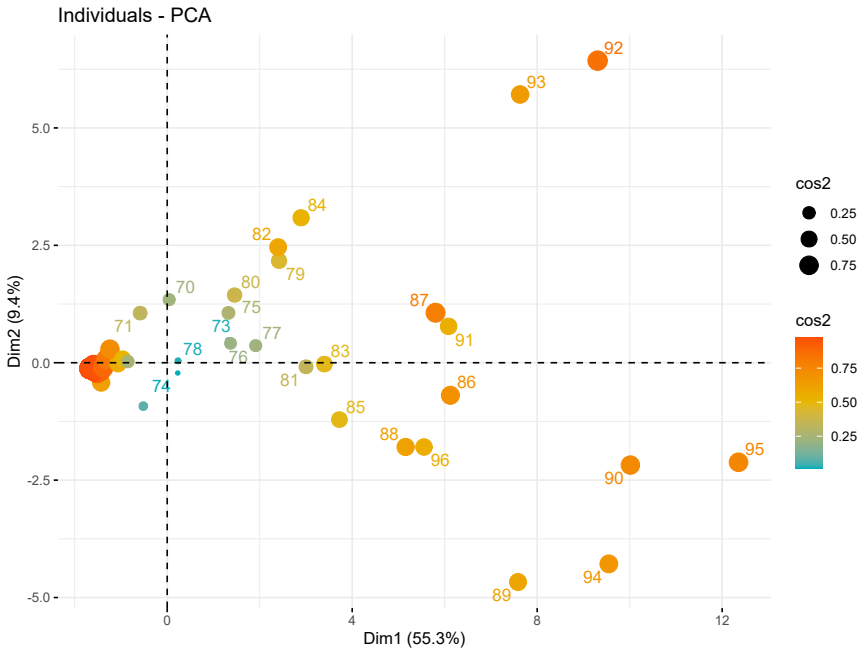
```
pca_case %>% factoextra::fviz_pca_biplot(.,
                repel = TRUE,
                col.var = "#2E9FDF",
                col.ind = "#696969")
```



Hình 4.10: Biểu đồ biplot cho biết mối quan hệ giữa các biến ban đầu và các thành phần chính. Đây được gọi là biểu đồ khoảng cách và nó hiển thị các quan sát riêng lẻ cũng như các véc-tơ tương ứng với tải. Độ dài của vector cho biết độ mạnh của mối tương quan của biến ban đầu với thành phần chính.

Biểu đồ cho biết vị trí các ngày được đánh số thứ tự tăng dần đến ngày thứ 96 khi chỉ sử dụng hai thành phần chính làm hệ trục tọa độ. Ngoài ra, biểu đồ biplot cũng chiếu các biến lên hệ trục tọa độ.

```
pca_case %>% fviz_pca_ind(.,
                col.ind = "cos2",
                pointsize = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)
```

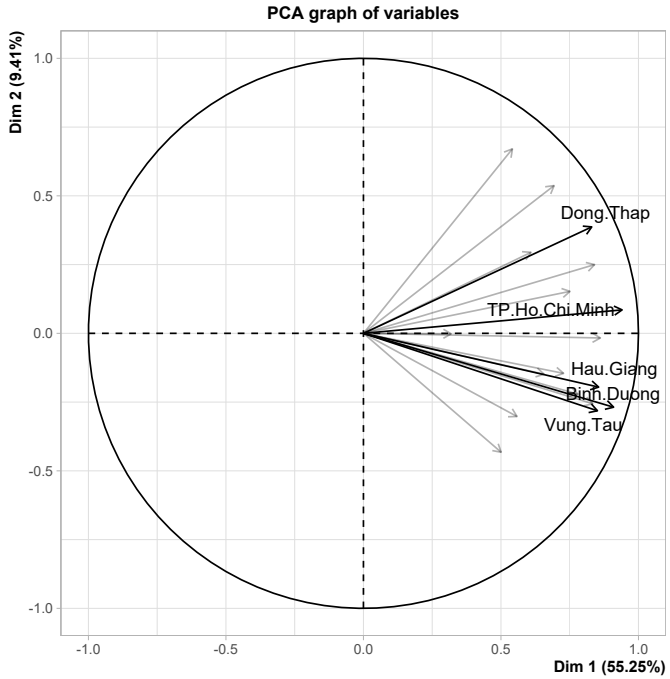


Hình 4.11: Biểu đồ biplot tổng hợp mật độ cos2 giữa hai thành phần chính của dữ liệu hằng ngày. Biểu đồ giải thích chính xác hơn độ lớn của giá trị cos2 qua kích thước và phổ màu.

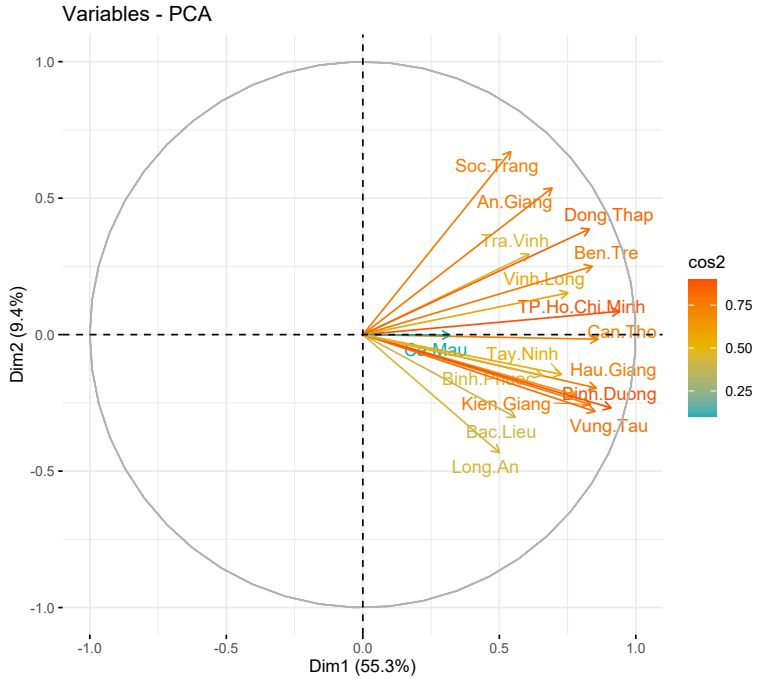
Biểu đồ biplot cá thể trên hai trục chính đầu tiên được thể hiện thông qua hai trục thành phần chính đầu tiên. Các giá trị cos2 (Cô-sin bình phương) càng cao thể hiện khả năng đóng góp vào thành phần chính càng cao. Ta thấy kể từ ngày thứ 71 trở đi, dữ liệu tán xạ ra hai phía của trục chính và có hệ số cos2 tăng đáng kể.

Để nói thêm về hệ số cos2 và chọn ra 5 contrib chứa đóng góp (tính theo phần trăm) của các biến cho các thành phần chính, ta tiến hành trực quan dữ liệu qua hai chiều đầu tiên trong hệ trục cực tọa độ.

```
pca_case %>% FactoMineR::plot.PCA(.,
                choix='var',
                select='contrib 5')
pca_case %>% factoextra::fviz_pca_var(.,
                col.var = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)
```



(a) Đồ thị biểu diễn tương quan theo hai chiều dữ liệu đầu tiên của dữ liệu theo ngày



(b) Đồ thị biểu diễn cos2 của giá trị riêng theo hai chiều dữ liệu đầu tiên của dữ liệu theo ngày

Hình 4.12: Đồ thị biểu diễn các thông số theo hai chiều dữ liệu đầu tiên.

Dựa vào biểu đồ (a) ta thấy sự khác biệt hệ số tải giữa các thành phần chính. Các biến có tương quan thuận được nhóm lại với nhau. Tất cả các biến đều có tương quan thuận khi xét thành phần thứ nhất và khi xét thành phần thứ hai, các biến được chia thành hai nhóm có tương quan âm và dương. Khi chọn số thành phần chính là 5 thành phần, ta có các thành phần chính được chia thành hai nhóm. Nhóm 1 gồm TP. Hồ Chí Minh và tỉnh Đồng Tháp có hệ số tương quan dương đối với hai chiều dữ liệu; Nhóm 2 gồm ba tỉnh Hậu Giang, Bình Dương, Vũng Tàu có hệ số tương quan ở chiều thứ nhất là hệ số dương và chiều thứ hai âm.

Đồ thị của dữ liệu (b) cho thấy giá trị cos2 cao và mật độ dày đặc nằm ở bên phải trục chính thứ nhất và chia thành hai phần khi chiếu đến thành phần thứ hai. Tỉnh Sóc Trăng và Long An có xu hướng trực giao. Tp. Cần Thơ có vị trí tương quan rất cao đối với thành phần thứ nhất mà gần như không tương quan khi xét thành phần thứ hai. Qua đồ thị (a) và (b) ta có thể chọn ra 5 biến có cos2 cao nhất làm các thành phần chính và giảm chiều dữ liệu dựa trên 5 biến này.

Ta có các giá trị riêng và phần trăm phương sai và phần trăm phương sai tích lũy được viết gọn theo thứ tự giảm dần các giá trị riêng cũng như tăng dần phần trăm phương sai tích lũy như sau

```
factoextra::get_eig(pca_case)
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  9.94554139      55.25300770                     55.25301
## Dim.2  1.69446623       9.41370128                     64.66671
## Dim.3  1.42405017       7.91138981                     72.57810
## Dim.4  1.00313508       5.57297266                     78.15107
## Dim.5  0.90036068       5.00200375                     83.15308
## Dim.6  0.77572770       4.30959832                     87.46267
## Dim.7  0.53232157       2.95734207                     90.42002
## Dim.8  0.47713645       2.65075808                     93.07077
## Dim.9  0.32914826       1.82860147                     94.89938
## Dim.10 0.27886842       1.54926898                     96.44864
## Dim.11 0.19536717       1.08537314                     97.53402
## Dim.12 0.14407396       0.80041086                     98.33443
## Dim.13 0.10718923       0.59549572                     98.92992
## Dim.14 0.07617433       0.42319071                     99.35311
## Dim.15 0.04825925       0.26810696                     99.62122
## Dim.16 0.03157511       0.17541728                     99.79664
## Dim.17 0.02328171       0.12934282                     99.92598
## Dim.18 0.01332331       0.07401838                    100.00000
```

Ta quan tâm đến 10 giá trị cá thể đầu tiên được tính toán các hệ số ctr và cos2 cho từng cá thể.

```
summary(pca_case)
```

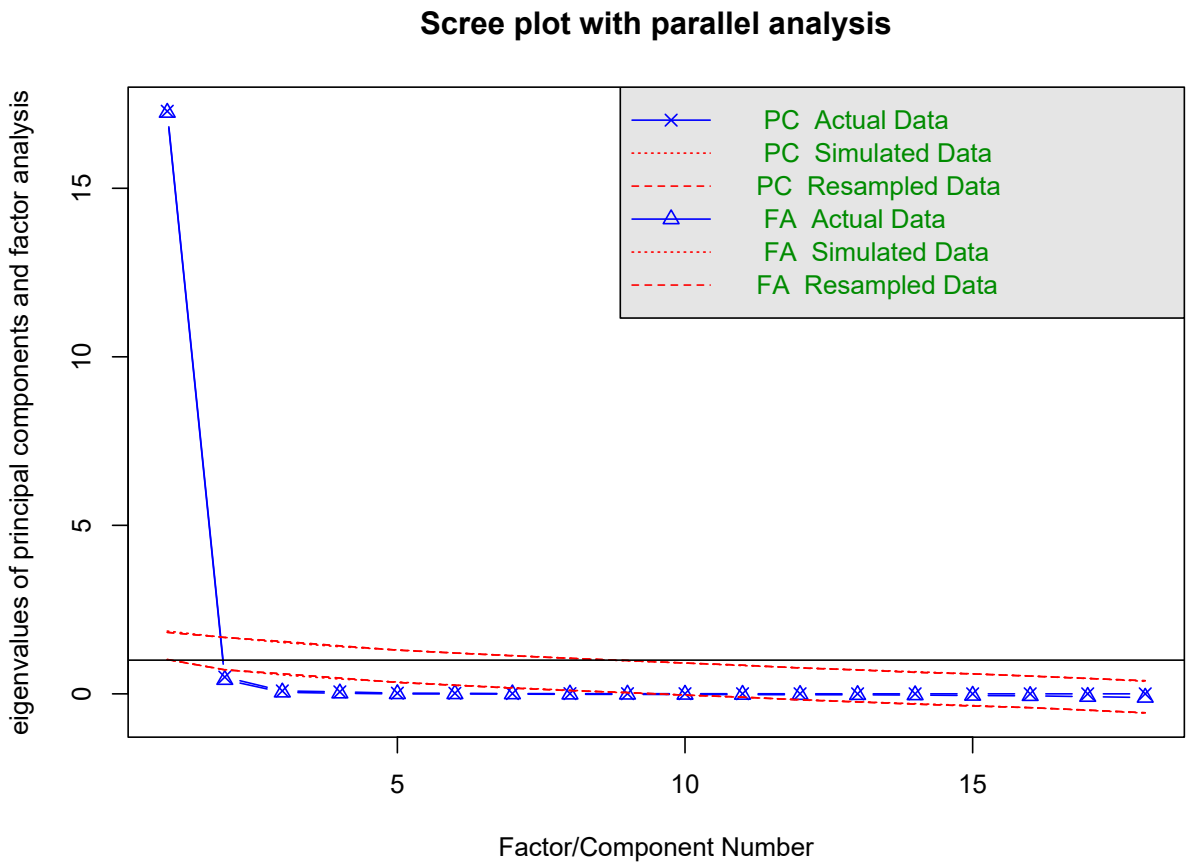
```
##
## Call:
## FactoMineR::PCA(X = case_data)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               9.946   1.694   1.424   1.003   0.900   0.776   0.532
## % of var.             55.253   9.414   7.911   5.573   5.002   4.310   2.957
## Cumulative % of var.  55.253  64.667  72.578  78.151  83.153  87.463  90.420
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.477   0.329   0.279   0.195   0.144   0.107   0.076
## % of var.              2.651   1.829   1.549   1.085   0.800   0.595   0.423
## Cumulative % of var.  93.071  94.899  96.449  97.534  98.334  98.930  99.353
##                       Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.048   0.032   0.023   0.013
## % of var.              0.268   0.175   0.129   0.074
## Cumulative % of var.  99.621  99.797  99.926 100.000
##
## Individuals (the 10 first)
```

```
##                    Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 2              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 3              |  1.689 | -1.662  0.289  0.969 | -0.122  0.009  0.005 | -0.221
## 4              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 5              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 6              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 7              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 8              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 9              |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
## 10             |  1.689 | -1.663  0.290  0.969 | -0.122  0.009  0.005 | -0.221
##                   ctr   cos2
## 1               0.036  0.017 |
## 2               0.036  0.017 |
## 3               0.036  0.017 |
## 4               0.036  0.017 |
## 5               0.036  0.017 |
## 6               0.036  0.017 |
## 7               0.036  0.017 |
## 8               0.036  0.017 |
## 9               0.036  0.017 |
## 10              0.036  0.017 |
##
## Variables (the 10 first)
##                    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## TP.Ho.Chi.Minh |  0.940  8.893  0.885 |  0.085  0.429  0.007 |  0.057  0.227
## Tien.Giang     |  0.781  6.125  0.609 | -0.221  2.892  0.049 | -0.140  1.382
## Long.An        |  0.501  2.522  0.251 | -0.432 11.036  0.187 | -0.053  0.199
## An.Giang       |  0.693  4.832  0.481 |  0.537 17.038  0.289 |  0.161  1.822
## Ben.Tre        |  0.840  7.103  0.706 |  0.250  3.686  0.062 | -0.298  6.239
## Can.Tho        |  0.862  7.464  0.742 | -0.017  0.017  0.000 | -0.350  8.585
## Vinh.Long      |  0.751  5.674  0.564 |  0.153  1.374  0.023 |  0.262  4.811
## Tra.Vinh       |  0.609  3.732  0.371 |  0.295  5.122  0.087 | -0.237  3.938
## Ca.Mau         |  0.319  1.024  0.102 | -0.002  0.000  0.000 |  0.689 33.352
## Hau.Giang      |  0.855  7.356  0.732 | -0.196  2.256  0.038 | -0.177  2.207
##                   cos2
## TP.Ho.Chi.Minh  0.003 |
## Tien.Giang      0.020 |
## Long.An         0.003 |
## An.Giang        0.026 |
## Ben.Tre         0.089 |
## Can.Tho         0.122 |
## Vinh.Long       0.069 |
## Tra.Vinh        0.056 |
## Ca.Mau          0.475 |
## Hau.Giang       0.031 |
```

4.6.2 Dữ liệu cul_data

Tương tự đối với dữ liệu ca bệnh tích lũy, ta cũng có sơ đồ sàng lọc và phân tích song song ca nhiễm tích lũy.

```
cul_data %>% psych::fa.parallel(.,
                main = "Scree plot with parallel analysis")
```



Hình 4.13: Sơ đồ sàng lọc với phân tích song song dữ liệu ca nhiễm tích lũy. Sơ đồ cũng đề nghị số thành phần là 1 và số nhân tố là 3.

Ở đây, ta thấy kể từ thành phần chính thứ hai đã có giá trị riêng không vượt qua ngưỡng 1.0. Điều này cho thấy sự nhất quán trong dữ liệu. Để có cái nhìn đối sánh hơn, ta tiến hành phân tích thành phần chính và tìm số phần trăm giải thích phương sai thích hợp đê giải thích dữ liệu.

```
prcomp(cul_data, center = TRUE, scale = TRUE) %>%
                summary()
```

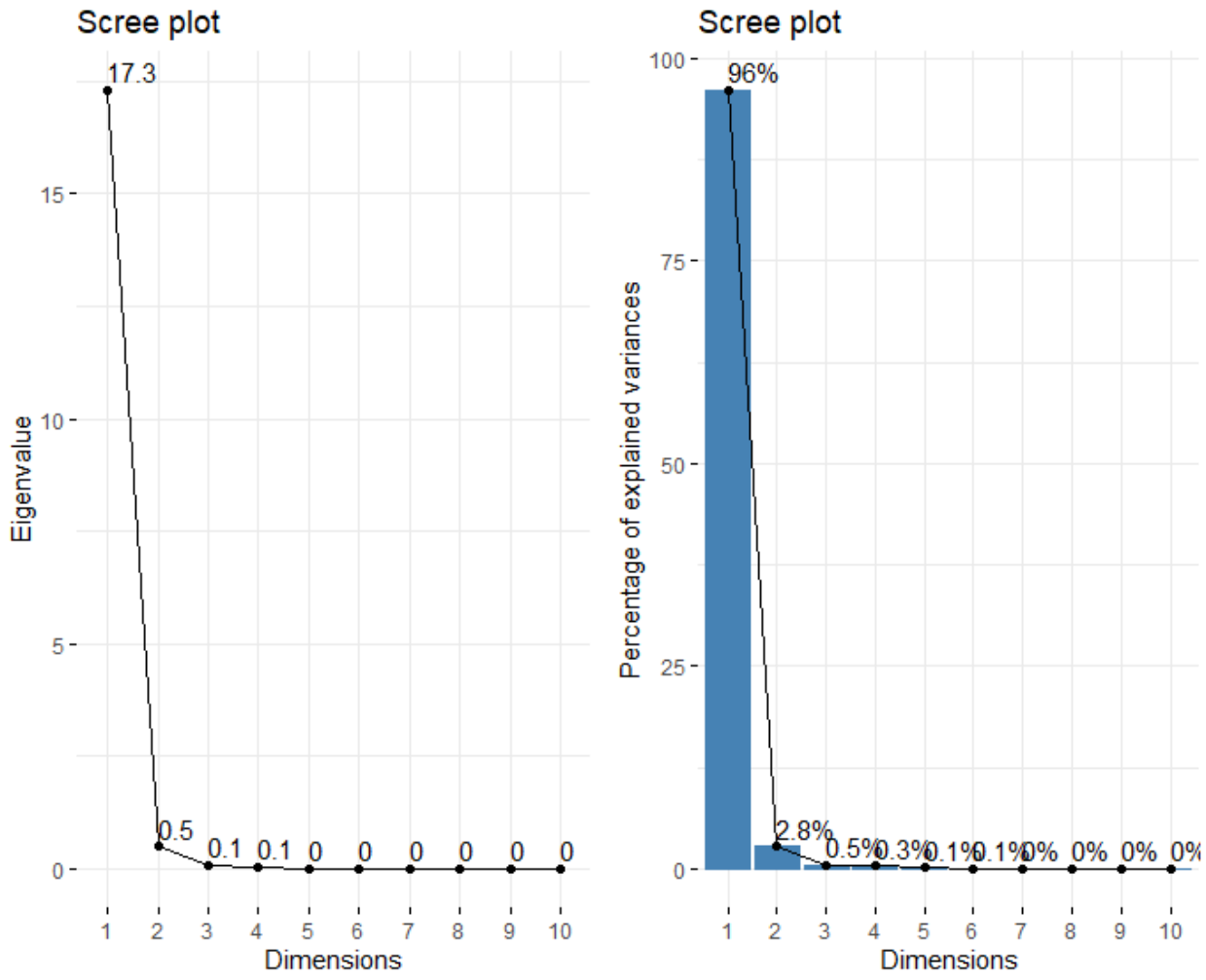
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     4.1577 0.70565 0.29500 0.24320 0.14515 0.13135 0.08728
## Proportion of Variance 0.9604 0.02766 0.00483 0.00329 0.00117 0.00096 0.00042
## Cumulative Proportion  0.9604 0.98801 0.99284 0.99613 0.99730 0.99826 0.99868
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.07794 0.06992 0.06559 0.05729 0.04386 0.03259 0.02960
## Proportion of Variance 0.00034 0.00027 0.00024 0.00018 0.00011 0.00006 0.00005
## Cumulative Proportion  0.99902 0.99929 0.99953 0.99971 0.99982 0.99988 0.99993
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.02321 0.02225 0.01241 0.01149
## Proportion of Variance 0.00003 0.00003 0.00001 0.00001
## Cumulative Proportion  0.99996 0.99998 0.99999 1.00000
```

```
princomp(cul_data, scores = TRUE) %>%
                loadings()
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## TP.Ho.Chi.Minh  0.989  0.129
## Tien.Giang                          -0.250  0.217 -0.844 -0.376  0.153
## Long.An                -0.557  0.794 -0.211
## An.Giang                            -0.112        -0.116  0.198
## Ben.Tre                              0.151 -0.190  0.187 -0.512        -0.485
## TP.Can.Tho                           0.136 -0.246        -0.262  0.400 -0.419
## Vinh.Long                                                0.119 -0.589 -0.554  0.342
## Tra.Vinh                                      -0.144                 0.181
## Ca.Mau
## Hau.Giang
## Kien.Giang                                                    -0.132
## Soc.Trang
## Bac.Lieu
## Dong.Thap              0.116 -0.121 -0.796 -0.540
## Binh.Duong      0.129 -0.772 -0.582 -0.109  0.157
## Vung.Tau               -0.103               -0.295        -0.236  0.571  0.659
## Tay.Ninh               -0.200         0.421 -0.654 -0.433  0.174 -0.346
## Binh.Phuoc                                                   -0.104         0.126
##                Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
## TP.Ho.Chi.Minh
## Tien.Giang
## Long.An
## An.Giang         0.497   0.288  -0.673   0.231                   -0.232   0.142
## Ben.Tre                 -0.494  -0.320                   0.112           0.153
```

```
## TP.Can.Tho              0.646   0.182                          -0.102  -0.167
## Vinh.Long       0.242   0.309   0.103           0.197
## Tra.Vinh        0.726  -0.277   0.481  -0.177          -0.120  -0.220
## Ca.Mau                         0.101          -0.156   0.954
## Hau.Giang       0.130   0.174   0.111  -0.100  -0.499           0.526   0.613
## Kien.Giang     -0.278                  -0.162  -0.417  -0.102  -0.740   0.273
## Soc.Trang       0.149          -0.209  -0.271  -0.623  -0.109   0.128  -0.642
## Bac.Lieu                                       -0.140   0.166  -0.183   0.220
## Dong.Thap      -0.139
## Binh.Duong
## Vung.Tau               -0.104  -0.226
## Tay.Ninh
## Binh.Phuoc             -0.160   0.221   0.888  -0.297                  -0.116
##                Comp.18
## TP.Ho.Chi.Minh
## Tien.Giang
## Long.An
## An.Giang        0.137
## Ben.Tre
## TP.Can.Tho
## Vinh.Long
## Tra.Vinh
## Ca.Mau          0.196
## Hau.Giang       0.109
## Kien.Giang      0.227
## Soc.Trang      -0.102
## Bac.Lieu       -0.929
## Dong.Thap
## Binh.Duong
## Vung.Tau
## Tay.Ninh
## Binh.Phuoc
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.056  0.056  0.056  0.056  0.056  0.056  0.056  0.056  0.056
## Cumulative Var  0.056  0.111  0.167  0.222  0.278  0.333  0.389  0.444  0.500
##                Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var   0.056   0.056   0.056   0.056   0.056   0.056   0.056   0.056
## Cumulative Var   0.556   0.611   0.667   0.722   0.778   0.833   0.889   0.944
##                Comp.18
## SS loadings      1.000
## Proportion Var   0.056
## Cumulative Var   1.000
```

```
pca_cul <- FactoMineR::PCA(cul_data, graph = FALSE)
left <- pca_cul %>% factoextra::fviz_eig(.,
                choice = 'eigenvalue',
                geom = 'line',
                addlabels = TRUE,
                repel = TRUE)
right <- pca_cul %>% factoextra::fviz_screeplot(.,
                addlabels = TRUE,
                repel = TRUE)
gridExtra::grid.arrange(left, right, ncol = 2)
```

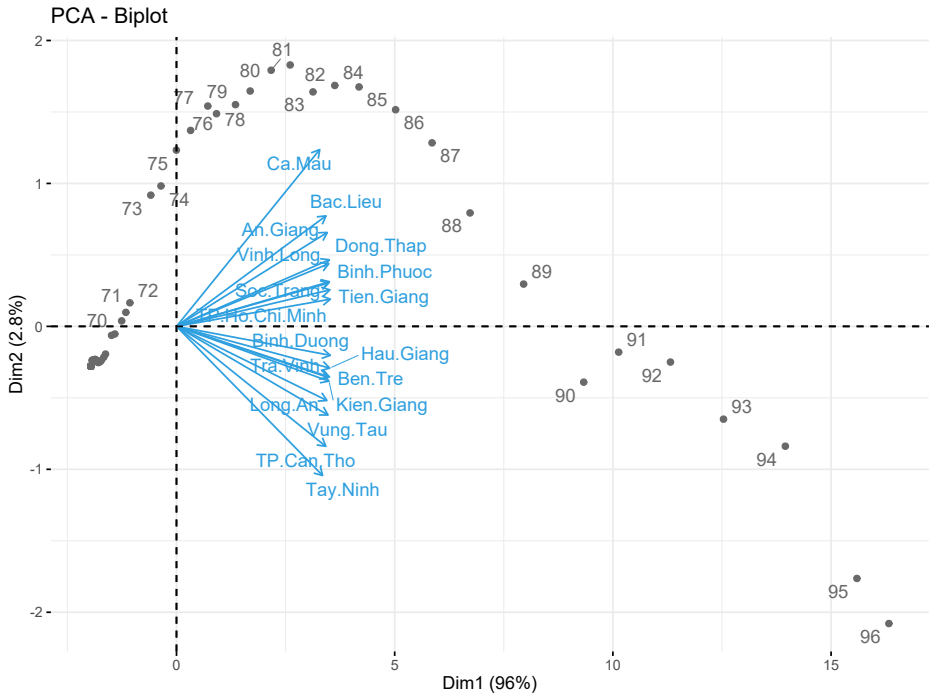


Hình 4.14: Sơ đồ sàng lọc cho kết quả đồ thị bên trái thể hiện giá trị riêng của từng thành phần chính. Trong đồ thị này, thành phần thứ nhất có giá trị riêng lớn nhất, thành phần thứ năm trở đi thì có giá trị riêng bé hơn 1. Sơ đồ bên phải thể hiện phần trăm phương sai được giải thích đối với từng thành phần.

Ở đây, thành phần chính đầu tiên đã giải thích một tỷ lệ rất lớn của phương sai, và các thành phần chính tiếp theo giải thích ít hơn nhiều. Điều này cho thấy chúng ta nên tập trung chủ yếu vào thành phần chính thứ nhất.

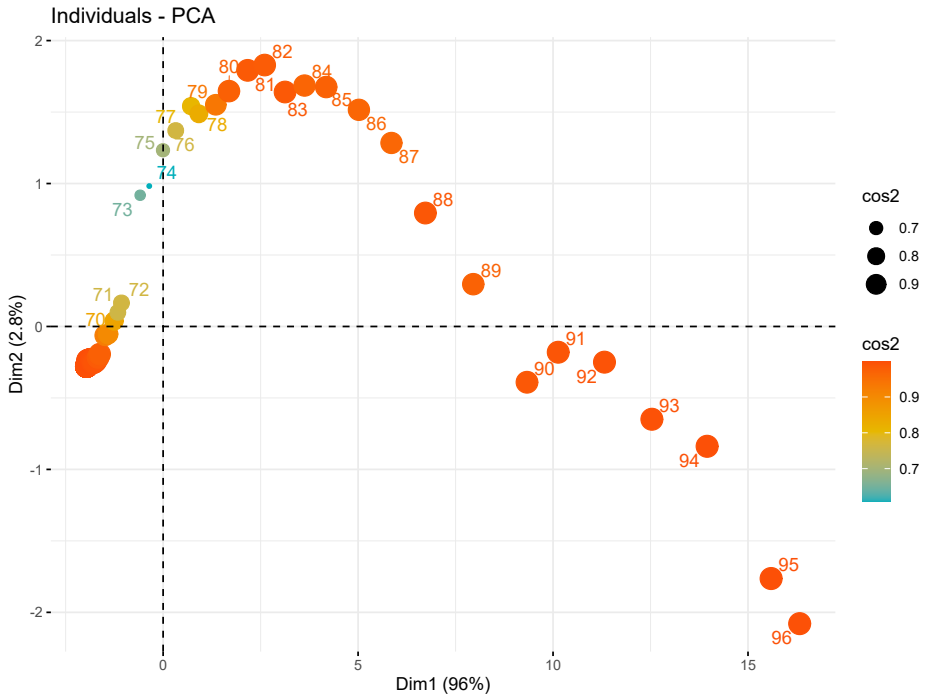
```
pca_cul %>% factoextra::fviz_pca_biplot(.,
                repel = TRUE,
                col.var = "#2E9FDF",
                col.ind = "#696969")
```

Tương tự đối với dữ liệu hằng ngày, ta nhìn rõ ràng hơn khi trực quan hóa vị trí của từng ngày được đánh số từ 1 đến 96 và biểu thị chỉ số cos2 bằng kích thước và phổ màu.



Hình 4.15: Biểu đồ biplot cho biết mối quan hệ giữa các biến ban đầu và các thành phần chính. Độ dài của vector cho biết độ mạnh của mối tương quan của biến ban đầu với thành phần chính.

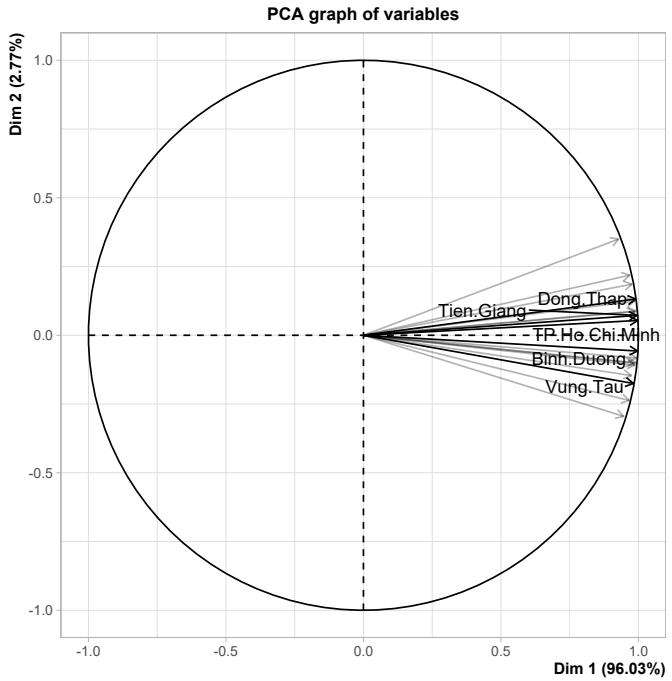
```
pca_cul %>% fviz_pca_ind(.,
                col.ind = "cos2",
                pointsize = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)
```



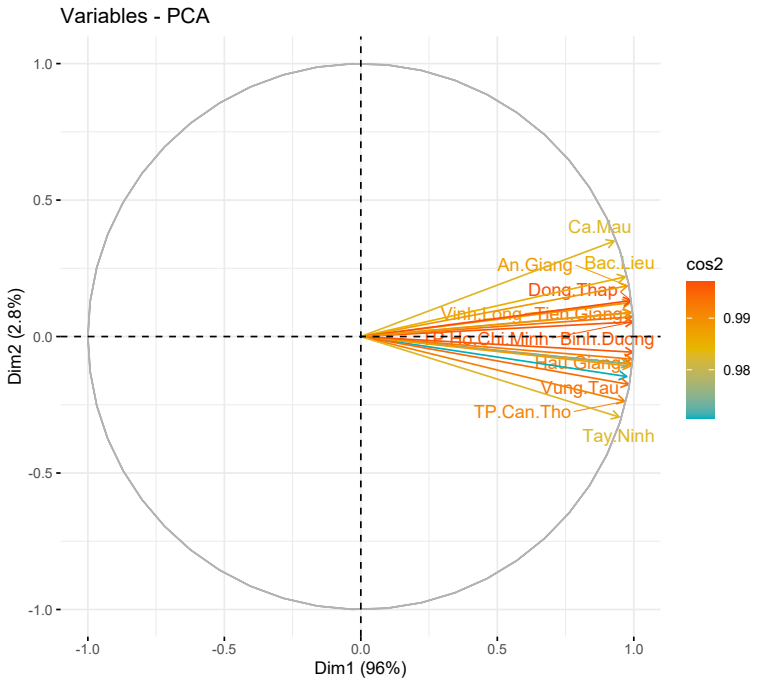
Hình 4.16: Biểu đồ biplot tổng hợp mật độ cos2 giữa hai thành phần chính của dữ liệu tích lũy. Biểu đồ giải thích chính xác hơn độ lớn của giá trị cos2 qua kích thước và phổ màu.

Không quá bất thường khi các véc-tơ biến dữ liệu tích lũy gần như có độ dài như sau do mối tương quan thuận đã phân tích phía trên.

```
pca_cul %>% FactoMineR::plot.PCA(.,
                choix='var',
                select='contrib 5')
pca_cul %>% factoextra::fviz_pca_var(.,
                col.var = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)
```



(a) Đồ thị biểu diễn tương quan theo hai chiều dữ liệu đầu tiên của dữ liệu tích lũy



(b) Đồ thị biểu diễn cos2 của giá trị riêng theo hai chiều dữ liệu đầu tiên của dữ liệu tích lũy

Hình 4.17: Đồ thị biểu diễn các thông số theo hai chiều đầu tiên với dữ liệu tích lũy.

Ta chia 5 biến được chọn thành hai nhóm. Nhóm 1 gồm hai tỉnh Đồng Tháp và Tiền Giang và Tp. Hồ Chí Minh; Nhóm 2 gồm hai tỉnh Bình Dương và Vũng Tàu.

```
factoextra::get_eig(pca_cul)
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  1.728623e+01     9.603460e+01                     96.03460
## Dim.2  4.979355e-01     2.766308e+00                     98.80091
## Dim.3  8.702444e-02     4.834691e-01                     99.28438
## Dim.4  5.914503e-02     3.285835e-01                     99.61296
## Dim.5  2.106883e-02     1.170491e-01                     99.73001
## Dim.6  1.725158e-02     9.584213e-02                     99.82585
## Dim.7  7.618441e-03     4.232467e-02                     99.86818
## Dim.8  6.073989e-03     3.374438e-02                     99.90192
## Dim.9  4.888367e-03     2.715759e-02                     99.92908
## Dim.10 4.301642e-03     2.389801e-02                     99.95298
## Dim.11 3.282633e-03     1.823685e-02                     99.97121
## Dim.12 1.923590e-03     1.068661e-02                     99.98190
## Dim.13 1.061794e-03     5.898857e-03                     99.98780
## Dim.14 8.760986e-04     4.867214e-03                     99.99267
## Dim.15 5.385863e-04     2.992146e-03                     99.99566
## Dim.16 4.952817e-04     2.751565e-03                     99.99841
## Dim.17 1.540549e-04     8.558604e-04                     99.99927
## Dim.18 1.320986e-04     7.338813e-04                    100.00000
```

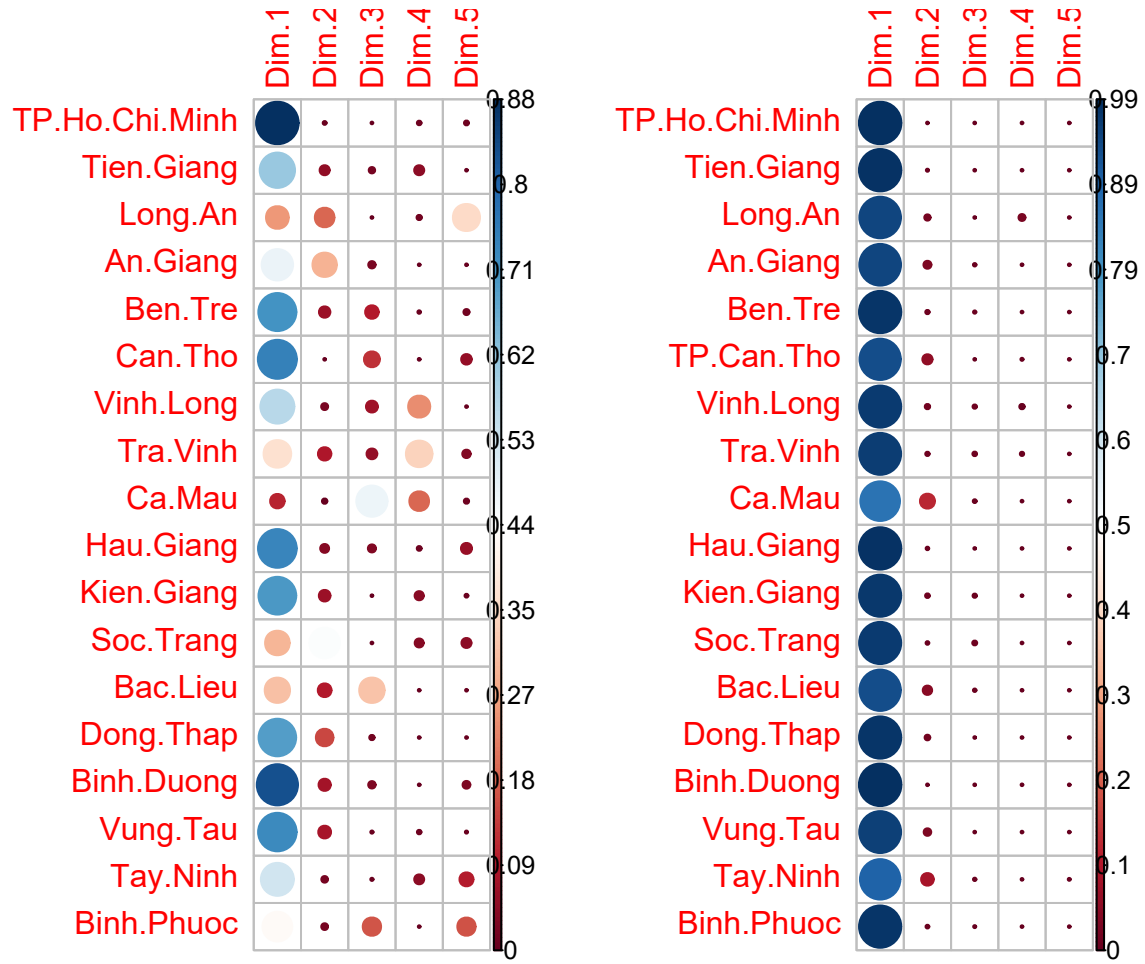
```
summary(pca_cul)
```

```
##
## Call:
## FactoMineR::PCA(X = cul_data, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance              17.286   0.498   0.087   0.059   0.021   0.017   0.008
## % of var.             96.035   2.766   0.483   0.329   0.117   0.096   0.042
## Cumulative % of var.  96.035  98.801  99.284  99.613  99.730  99.826  99.868
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.006   0.005   0.004   0.003   0.002   0.001   0.001
## % of var.              0.034   0.027   0.024   0.018   0.011   0.006   0.005
## Cumulative % of var.  99.902  99.929  99.953  99.971  99.982  99.988  99.993
##                       Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.001   0.000   0.000   0.000
## % of var.              0.003   0.003   0.001   0.001
## Cumulative % of var.  99.996  99.998  99.999 100.000
##
## Individuals (the 10 first)
##                    Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 2              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
```

```
## 3              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 4              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 5              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 6              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 7              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 8              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 9              |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
## 10             |  1.987 | -1.964  0.232  0.977 | -0.279  0.163  0.020 | -0.064
##                   ctr   cos2
## 1               0.049  0.001 |
## 2               0.049  0.001 |
## 3               0.049  0.001 |
## 4               0.049  0.001 |
## 5               0.049  0.001 |
## 6               0.049  0.001 |
## 7               0.049  0.001 |
## 8               0.049  0.001 |
## 9               0.049  0.001 |
## 10              0.049  0.001 |
##
## Variables (the 10 first)
##                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## TP.Ho.Chi.Minh |  0.997  5.748  0.994 |  0.054  0.581  0.003 | -0.033  1.247
## Tien.Giang     |  0.994  5.717  0.988 |  0.072  1.048  0.005 | -0.011  0.149
## Long.An        |  0.974  5.491  0.949 | -0.147  4.313  0.021 |  0.003  0.011
## An.Giang       |  0.977  5.518  0.954 |  0.186  6.922  0.034 |  0.081  7.603
## Ben.Tre        |  0.991  5.681  0.982 | -0.100  2.011  0.010 | -0.077  6.780
## TP.Can.Tho     |  0.967  5.410  0.935 | -0.237 11.300  0.056 |  0.031  1.105
## Vinh.Long      |  0.987  5.632  0.973 |  0.124  3.087  0.015 | -0.101 11.667
## Tra.Vinh       |  0.982  5.580  0.965 | -0.101  2.046  0.010 |  0.107 13.254
## Ca.Mau         |  0.928  4.979  0.861 |  0.350 24.550  0.122 |  0.090  9.246
## Hau.Giang      |  0.993  5.702  0.986 | -0.083  1.369  0.007 |  0.011  0.142
##                   cos2
## TP.Ho.Chi.Minh  0.001 |
## Tien.Giang      0.000 |
## Long.An         0.000 |
## An.Giang        0.007 |
## Ben.Tre         0.006 |
## TP.Can.Tho      0.001 |
## Vinh.Long       0.010 |
## Tra.Vinh        0.012 |
## Ca.Mau          0.008 |
## Hau.Giang       0.000 |
```

Để thấy độ lớn của các cos2 trong các biến chính đã được chọn, ta có

```
var <- pca_case %>% factoextra::get_pca_var()
                corrplot(var$cos2, is.corr = FALSE)
var <- pca_cul %>% factoextra::get_pca_var()
                corrplot(var$cos2, is.corr = FALSE)
```



(a) Tương quan giữa các biến trong cos2 của dữ liệu hằng (b) Tương quan giữa các biến trong cos2 của dữ liệu tích lũy ngày

Hình 4.18: Biểu đồ giá trị cos2 đối với 5 biến đã được chọn làm thành phần chính đối với các biến khi chưa phân tích.

4.7 Kiểm định Bartlett – KMO

Tiến hành kiểm định Bartlett với dữ liệu covid_case trong gói lệnh **psych**. Mục đích của kiểm định này là kiểm tra giả thiết các mẫu có phương sai bằng nhau

```
case_data %>% psych::cortest.bartlett()
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 2105.383
##
## $p.value
## [1] 0
##
## $df
## [1] 153
```

Ở đây, giá trị $p - value = 0$nhau, ta bác bỏ giả thiết phương sai bằng nhau từng đôi. Tức là dữ liệu thích hợp để phân tích nhân tố. Mặt khác ta cũng kiểm định KMO để xem

```
case_data %>% psych::KMO()
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = .)
## Overall MSA =  0.77
## MSA for each item =
## TP.Ho.Chi.Minh     Tien.Giang        Long.An       An.Giang        Ben.Tre
##           0.84           0.87           0.44           0.74           0.72
##        Can.Tho      Vinh.Long       Tra.Vinh         Ca.Mau      Hau.Giang
##           0.72           0.74           0.63           0.73           0.77
##     Kien.Giang      Soc.Trang       Bac.Lieu      Dong.Thap     Binh.Duong
##           0.77           0.63           0.79           0.87           0.90
##       Vung.Tau       Tay.Ninh     Binh.Phuoc
##           0.86           0.78           0.84
```

Giá trị KMO trung bình nằm ở mức 0.77. Theo tiêu chuẩn đánh giá phù hợp để phân tích nhân tố, giá trị $OverallMSA \geq 0.7$, ta xác định dữ liệu rất thích hợp để phân tích nhân tố. Hệ số MSA của biến Long An dưới mức phù hợp (0.44) nên khi phân tích nhân tố ta loại bỏ biến Long An. Sau khi tiến hành bỏ biến Long An trong dữ liệu thì $OverallMSA$ có sự thay đổi

```
case_data %>% select(.,-Long.An) %>%
KMO() %>%  .$MSA
```

```
## [1] 0.8227004
```

Chỉ số $MSA = 0.82$ đã được cải thiện sau khi loại trừ biến Long An có $MSA = 0.44$ khá thấp và không phù hợp để phân tích nhân tố.

Ta thực hiện kiểm định Barlett đối với dữ liệu tích lũy, với giả thiết thống kê

```
cul_data %>% psych::cortest.bartlett()
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 7978.616
##
## $p.value
## [1] 0
##
## $df
## [1] 153
```

Ở đây, giá trị $p - value = 0$ tức là bác bỏ giả thiết . Tức là dữ liệu cũng thích hợp để phân tích nhân tố. Mặt khác ta cũng kiểm định KMO để xem

```
cul_data %>% psych::KMO()
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = .)
## Overall MSA =  0.87
## MSA for each item =
## TP.Ho.Chi.Minh     Tien.Giang        Long.An       An.Giang        Ben.Tre
##           0.89           0.85           0.83           0.84           0.84
##     TP.Can.Tho      Vinh.Long       Tra.Vinh         Ca.Mau      Hau.Giang
##           0.83           0.94           0.85           0.95           0.85
##     Kien.Giang      Soc.Trang       Bac.Lieu      Dong.Thap     Binh.Duong
##           0.84           0.87           0.94           0.87           0.87
##       Vung.Tau       Tay.Ninh     Binh.Phuoc
##           0.93           0.84           0.94
```

Giá trị KMO trung bình nằm ở mức 0.87 sự thích hợp để phân tích nhân tố là rất cao. Như vậy bảng dữ liệu đủ điều kiện để phân tích nhân tố. Tất cả các biến đều có giá trị MSA trên 0.6 nên ta không cần loại biến nào. Cả hai dữ liệu trên đều sẵn sàng để phân tích nhân tố.

Đối với dữ liệu hằng ngày, ta chọn số nhân tố là 3 và đối với dữ liệu tích lũy ta chọn 2 nhân tố. Vì dữ liệu có 96 quan sát nên ta chọn hệ số tải (Factor Loading) là 0.55

Như đã phân tích, biến Long An trong dữ liệu hằng ngày có KMO thấp nên được loại khỏi dữ liệu

4.8 Phân tích nhân tố

```
principal(case_data %>% select(-Long.An),
                nfactors = 3
                rotate = "varimax") %>%
        print.psych(.,
                cut = 0.55,
                sort = TRUE)
```

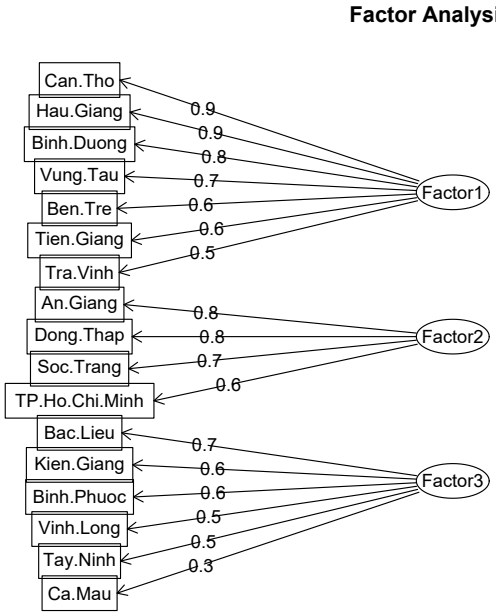
```
## Principal Components Analysis
## Call: principal(r = case_data %>% select(-Long.An), nfactors = Nfacs,
##     rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 item   RC1   RC2   RC3   h2    u2 com
## Binh.Duong        14  0.92             0.94 0.055 1.2
## Hau.Giang          9  0.86             0.82 0.176 1.3
## Can.Tho            5  0.85             0.88 0.117 1.4
## Vung.Tau          15  0.84             0.82 0.185 1.3
## Kien.Giang        10  0.80             0.77 0.229 1.4
## Tien.Giang         2  0.76             0.67 0.334 1.3
## TP.Ho.Chi.Minh     1  0.66  0.57       0.89 0.108 2.5
## Tay.Ninh          16  0.65             0.54 0.460 1.6
## Soc.Trang         11        0.85       0.74 0.257 1.0
## An.Giang           3        0.83       0.81 0.192 1.4
## Dong.Thap         13        0.76       0.85 0.146 2.0
## Ben.Tre            4  0.63  0.69       0.87 0.135 2.0
## Tra.Vinh           7        0.55       0.51 0.492 1.9
## Vinh.Long          6                   0.65 0.347 3.0
## Bac.Lieu          12              0.77 0.73 0.272 1.4
## Ca.Mau             8              0.74 0.59 0.415 1.1
## Binh.Phuoc        17              0.63 0.64 0.359 2.0
##
##                        RC1  RC2  RC3
## SS loadings           6.43 3.87 2.42
## Proportion Var        0.38 0.23 0.14
## Cumulative Var        0.38 0.61 0.75
## Proportion Explained  0.51 0.30 0.19
## Cumulative Proportion 0.51 0.81 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  117.62  with prob <  0.019
##
## Fit based upon off diagonal values = 0.99
```

```r
principal(cul_data,
                nfactors = 1,
                rotate = "varimax") %>%
        print.psych(.,
                cut = 0.55,
                sort = TRUE)
```

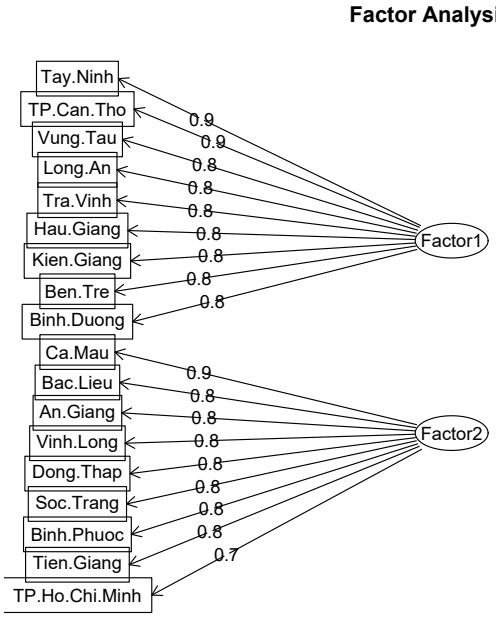
```
## Principal Components Analysis
## Call: principal(r = cul_data, nfactors = 1, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                  V  PC1   h2     u2 com
## TP.Ho.Chi.Minh   1 1.00 0.99 0.0064   1
## Binh.Duong      15 1.00 0.99 0.0070   1
## Tien.Giang       2 0.99 0.99 0.0118   1
## Hau.Giang       10 0.99 0.99 0.0144   1
## Ben.Tre          5 0.99 0.98 0.0180   1
## Binh.Phuoc      18 0.99 0.98 0.0189   1
## Dong.Thap       14 0.99 0.98 0.0202   1
## Kien.Giang      11 0.99 0.98 0.0241   1
## Vinh.Long        7 0.99 0.97 0.0265   1
## Soc.Trang       12 0.99 0.97 0.0265   1
## Tra.Vinh         8 0.98 0.96 0.0355   1
## Vung.Tau        16 0.98 0.96 0.0366   1
## An.Giang         4 0.98 0.95 0.0462   1
## Long.An          3 0.97 0.95 0.0509   1
## Bac.Lieu        13 0.97 0.94 0.0625   1
## TP.Can.Tho       6 0.97 0.94 0.0649   1
## Tay.Ninh        17 0.95 0.90 0.1041   1
## Ca.Mau           9 0.93 0.86 0.1394   1
##
##                  PC1
## SS loadings    17.29
## Proportion Var  0.96
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.03
##  with the empirical chi square  20.23  with prob <  1
##
## Fit based upon off diagonal values = 1
```

4.9 Ma trận xoay

```
fa_case <- case_data %>% select(-Long.An) %>%
        factanal(., 3,
                rotation = "varimax")
fa_cul <- cul_data %>% factanal(., 2,
                rotation = "varimax")
fa.diagram(fa_case$loadings)
fa.diagram(fa_cul$loadings)
```



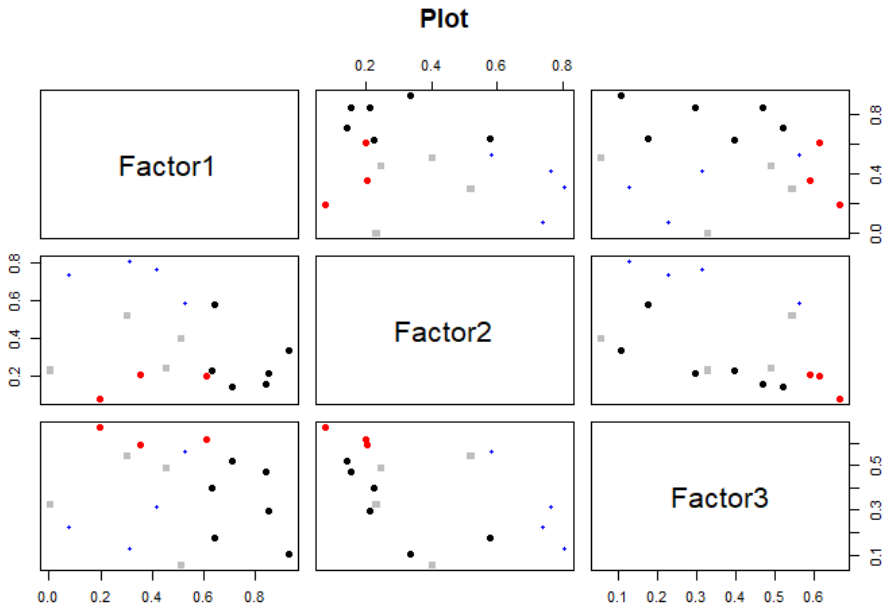
(a) Nhân tố dữ liệu hằng ngày



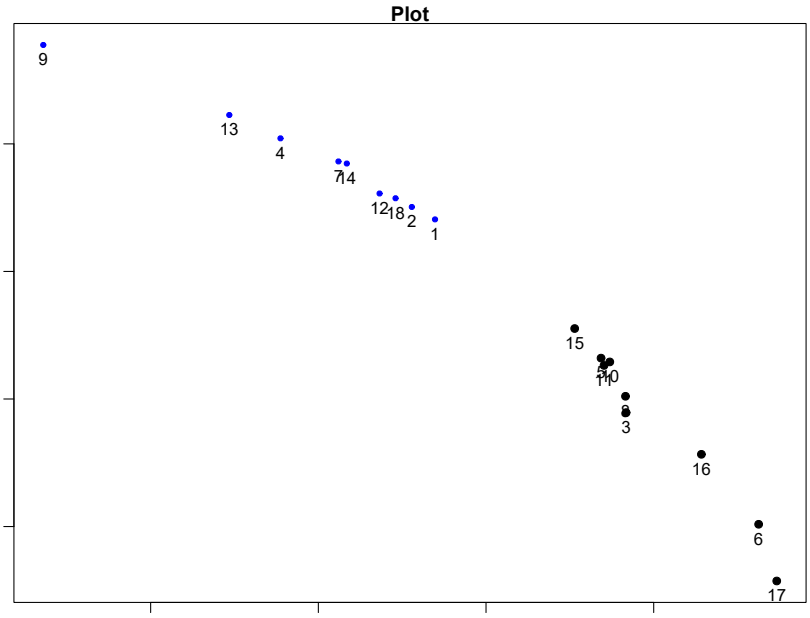
(b) Nhân tố dữ liệu tích lũy

Hình 4.19: Phân cum nhân tố và hệ số nhân tố tương ứng của hai dữ liệu.

```
fa_case %>% factor.plot(., cut = 0.55)
fa_cul %>% factor.plot(., cut = 0.55)
```



(a) Tương quan nhân tố của dữ liệu hằng ngày. Với ba nhân tố được xác định với hệ số tải nhân tố 0.55



(b) Tương quan nhân tố của dữ liệu tích lũy. Với hai nhân tố được xác định với hệ số tải nhân tố 0.55

Hình 4.20: Tương quan nhân tố của dữ liệu nhân tố được xác định với hệ số tải nhân tố 0.55

4.10 Bàn luận

Do các tác động xã hội, kinh tế và môi trường khác nhau của Covid-19, điều quan trọng là phải nghiên cứu và so sánh tốc độ lây lan của bệnh này ở các tỉnh/thành phố khác nhau. Trong nghiên cứu này, số lượng bệnh nhân có Covid-19 ở 18 tỉnh/thành phố đã được xem xét. Đầu tiên, mối quan hệ giữa các tỉnh/thành phố được xem xét được nghiên cứu bằng cách sử dụng mối tương quan của Pearson. Kết quả chỉ ra rằng có mối quan hệ thuận chiều cực kỳ cao giữa các tỉnh/thành phố được xem xét, dựa trên số lượng bệnh nhân mắc bệnh Covid-19 và tích lũy của số ca nhiễm bệnh theo ngày. Sau đó, dựa trên tốc độ lây lan của Covid-19, các tỉnh/thành phố này được phân loại bằng cách sử dụng phân tích thành phần chính. Kết quả chỉ ra rằng, đối với số lượng bệnh nhân, sự phân bố lây lan ở Sóc Trăng, An Giang, Trà Vinh, Đồng Tháp, Bến Tre, Vĩnh Long và Tp. Hồ Chí Minh là tương tự nhau và khác với các tỉnh/thành phố khác. Ngoài ra, đối với số lượng bệnh nhân tích lũy theo ngày, sự phân bố lây lan ở Bình Dương và Vũng Tàu là tương tự nhau và khác với các tỉnh Đồng Tháp, Tiền Giang và Tp. Hồ Chí Minh khi ta phân tích và chọn ra 5 thành phần chính tiêu biểu. Các tác giả đề nghị các nhà nghiên cứu xem xét nhiều tỉnh/thành phố hơn và phân loại chúng dựa trên phân tích thành phần chính hoặc các phương pháp khác như phân tích nhân tố có sự cải tiến.

Chương 5

KẾT LUẬN

5.1 Kết luận

Những kết quả chính thu được sau khi phân loại 18 biến dữ liệu các tỉnh/thành phố có ca nhiễm hằng ngày và tích lũy bao gồm

1. Cung cấp thông tin về mô tả số lượng các ca nhiễm theo ngày và các ca nhiễm tích lũy được xác nhận do vi-rút corona từ lúc bắt đầu đợt dịch thứ IV đối với 18 tỉnh/thành phố phía nam.
2. Thông tin hệ số tương quan giữa các biến được trình bày bằng tương quan đồ
3. Phân tích thành công thuật toán phân tích thành phần chính và phân loại được các biến thành hai nhóm.
4. Phân tích nhân tố với các tiêu chí tìm được.

Đặc biệt, phương pháp phân tích thành phần chính được sử dụng như một công cụ phân loại đối với các dữ liệu rời rạc cho kết quả khá ổn định.

Bài nghiên cứu cũng trực quan hóa thành công các mạng tương quan ứng dụng xét các mối quan hệ tương quan giữa các biến. Đặt biệt, chúng tôi nghiên cứu và trực quan thành công biểu đồ tương quan Pearson 4.7 giữa một biến so với các biến còn lại lấy ý tưởng từ tấm bia đạn với mục tiêu chính nằm ở biến được chọn để xét hệ số tương quan với các biến khác.

Nghiên cứu này sử dụng phần mềm lập trình thống kê R (phiên bản 4.1.0) để phân tích thống kê. R là một ngôn ngữ lập trình với nhiều lợi thế như cú pháp đơn giản, hệ thống thư viện có cấu trúc chặt chẽ, tương thích cao, đặc biệt tối ưu cho các mô hình Machine Learning,... Các chương trình lệnh và thông tin mã nguồn được lưu trữ và cập nhật trên trang web Github.

5.2 Nhận xét sơ bộ bài báo cáo

Bài báo cáo đã cơ bản hoàn thành với việc giảm thiểu từ 18 biến dữ liệu thành 5 biến với phần trăm phân tích phương sai đạt 83% đối với dữ liệu hằng ngày và 1 thành phần chính đối với dữ liệu tích lũy giải thích 96% phương sai.

Phân tích nhân tố phân chia dữ liệu hằng ngày thành ba nhân tố chính và chia dữ liệu tích lũy thành hai nhân tố chính. Với hệ số tải cho trước, các hệ số nhân tố đã được thiết lập và phân tích.

Song, vì phải vừa thu thập dữ liệu thứ cấp hằng ngày và kiểm tra trong nhiều nguồn khác nhau, chúng tôi không tránh khỏi khó khăn về thời gian hoàn thiện việc viết bài. Mặc khác, các nghiên cứu trước đây về việc ứng dụng phân tích thành phần chính và phân tích nhân tố cho dữ liệu thời gian khá ít. Điều đó càng làm tăng sự khó khăn cho chúng tôi khi nghiên cứu về vấn đề mới mẻ này. Ngoài ra, bản thân các báo cáo viên cũng phải hứng chịu những tác động tiêu cực bởi đại dịch Covid-19 nên việc trao đổi và nghiên cứu trở nên khó khăn, nhất là đối với sinh viên mới tiếp xúc với chương trình học.

Hướng nghiên cứu tiếp theo chúng tôi tập trung vào phân tích thành phần chính ứng dụng trong phân loại các hình ảnh có số chiều lớn. Sử dụng thêm các bài toán phân loại để phân loại mô hình dưới nền tảng của phân tích thành phần chính.

Chương 6

PHỤ LỤC

6.1 Thông tin phần mềm

Phần này cung cấp một số thông tin về thiết bị mà nhóm tác giả sử dụng để hoàn thành phần thực nghiệm. Trong quá trình thực hiện phân tích thành phần chính cũng như phân tích nhân tố và các tác vụ khác, chúng tôi chỉ sử dụng một máy chủ với các thông tin kỹ thuật được cho bởi bảng dưới đây.

Processor	Intel(R) Core(TM) i3-9100 CPU @ 3.60GHz 3.60 GHz
Installed RAM	8,00 GB
System type	64-bit operating system, x64-based processor (AMD64)
Edition	Windows 10

Bài báo sử dụng ngôn ngữ lập trình thống kê R phiên bản 4.1.0 (cập nhật vào ngày 18/7/2021) để thực hiện các tác vụ trong bài báo cáo. Thông tin chi tiết được cho bởi bảng dưới dây. Để hoàn thành các mục tiêu nghiên cứu, chúng tôi sử dụng một số gói chương trình lệnh trong ngôn ngữ lập trình thống kê R được liệt kê như sau.

```
pks <- c('psych', 'tidyverse', 'factoextra', 'ggplot2',
        'gridExtra', 'FactoMineR', 'igraph', 'corrplot')
install.packages(pks, dependencies = TRUE)
library(psych)
library(tidyverse)
library(factoextra, FactoMineR)
library(ggplot2, gridExtra)
library(igraph, corrplot)
```

psych gồm các chức năng chủ yếu dành cho phân tích đa biến và xây dựng thang đo bằng cách sử dụng phân tích nhân tố, phân tích thành phần chính, phân tích cụm và phân tích độ tin cậy,... ngoài ra gói "psych" còn dùng để tính toán các chỉ tiêu thống kê cơ bản theo một hoặc nhiều nhóm (hay đối tượng) thông qua một câu lệch ngắn gọn.

tidyverse có trang web chính là https://www.tidyverse.org/packages/. Nó là một tập hợp các gói lệnh R mã nguồn mở được Hadley Wickham và nhóm của ông giới thiệu

factoextra là một gói R giúp dễ dàng trích xuất và trực quan hóa đầu ra của các phân tích dữ liệu đa biến khám phá bao gồm phân tích thành phần chính; phân tích nhân tố và nhiều nhân tố; phân tích nhân tố của dữ liệu hỗn hợp

FactoMineR dùng để phân tích dữ liệu khám phá đa biến và khai thác dữ liệu. Các phương pháp phân tích dữ liệu thăm dò để tóm tắt, trực quan hóa và mô tả các bộ dữ liệu. Các phương thức thành phần chính có sẵn, những phương thức có tiềm năng lớn nhất về các ứng dụng: phân tích thành phần chính (PCA) khi các biến là định lượng, phân tích tương ứng (CA) và phân tích nhiều tương ứng (MCA) khi các biến được phân loại, Phân tích nhiều yếu tố khi các biến được cấu trúc theo nhóm, v.v. và phân tích cụm phân cấp. F. Husson, S. Le và J. Pages (2017).

ggplot2 là gói chương trình lệnh dùng chủ yếu trong các tác vụ đồ thị trực quan hóa dữ liệu mã nguồn mở cho ngôn ngữ lập trình thống kê R.

igraph là gói chương trình cho phép trực quan các mạng phức tạp, mạng tương quan pcor...

6.2 Nguồn mã lập trình

Chúng tôi lưu trữ và cập nhật các mã nguồn khi sử dụng phần mềm ngôn ngữ lập trình thống kê R trên trang web mã nguồn mở Github (truy cập không cần đăng nhập tài khoản) bao gồm

Hình 6.1: Đường dẫn cụ thể cho mã vạch QR (truy cập không cần đăng nhập tài khoản): https://github.com/hungtrannam/PCA_for_Covid19

Giải thích các tệp có trong Github

README Giới thiệu các thành viên và phần tóm tắt của bài báo

Data chứa hai bộ tập dữ liệu bao gồm 1) Dữ liệu các ca xác nhận dương tính với SAR-Cov-2 từ 18 tỉnh/thành phố được thu thập theo ngày (đặt tên là covid_case) 2) Dữ liệu các ca xác nhận dương tính với SAR-Cov-2 được tích luỹ theo ngày (đặt tên là covid_cul);

Modules Tệp R chứa trích gọn các lệnh phân tích dữ liệu. Khi sử dụng, ta lưu ý việc hạ tải nên được lưu ở ổ đĩa **D** để dễ dàng phân tích phía sau.

thesis Tệp các chương được trình bày trong cáo báo viết bằng hệ thống soạn thảo văn bản LaTeXcũng như các hình ảnh minh họa đính kèm.

LISENCE Ghi chú các chương trình lệnh trên ngôn ngữ lập trình R dùng để phân tích và phân loại dữ liệu và hướng dẫn sử dụng gói lệnh phân tích trên R.

Tài liệu tham khảo

Tài liệu Tiếng Việt

[1] Lâm Hoàng Chương, *Giáo trình Xác suất Thống kê – Toán Thống kê*, NXB Đại học Cần Thơ, năm 2019, ISBN: 978-604-965-139-7.

[2] Nguyễn Hữu Việt Hưng, *Đại số tuyến tính*, NXB Đại học Quốc gia Hà Nội, năm 2019, ISBN 978-604-9854-92-7

[3] Nguyễn Hữu Khánh, *Giáo trình Đại số Tuyến tính và Hình học*, tập II, NXB Đại học Cần Thơ, năm 2013

[4] Mohammad Reza Mahmoudi, Mohammad Hossein Heydari, Sultan Noman Qasem, Amirhosein Mosavi, Shahab S. Band, *Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries*, Alexandria Engineering Journal, Volume 60, Issue 1, 2021, Pages 457-464, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2020.09.013. (https://www.sciencedirect.com/science/article/pii/S1110016820304543)

[5] Trần Văn Lý, *Bài giảng môn học Thống kê nhiều chiều*, năm 2020.

[6] Vũ Hữu Tiệp, *Machine Learning cơ bản*, e-book: https://github.com/tiepvupsu/ebookMLCB, năm 2020.

[7] Yadolah Dodgle, *Từ điển các thuật ngữ thống kê Oxford*, NXB Đại học Quốc gia Hà Nội, năm 2018, ISBN: 978-604-961-921-2

Trang web

[8] **infographics.vn**, tct https://infographics.vn/interactive-du-lieu-dot-dich-covid-19-thu-4-tai-viet-nam-lien-tuc-cap-nhat/20981.vna, ntc: 30/7/2021

[9] **vnexpress.net**, tct https://vnexpress.net/covid-19/covid-19-viet-nam?, ntc: 30/7/2021

[10] **covid.cantho.gov.vn**, tct https://covid.cantho.gov.vn/, ntc: 29/7/2021

INDEX

biến chủng delta, 15

chuẩn, 2
chỉ số Kaiser, 10

giá trị cos2, 33
giá trị riêng, 3

hệ số tương quan Pearson, 4
hệ số tải, 13

kiểm định Kaiser-Meyer-Olkin, 17

ma trận, 1
ma trận hiệp phương sai, 4
ma trận nhân tố, 13
Ma trận vuông, 2
Machine Learning, 5
metric, 2

ngôn ngữ lập trình R, 6

phương pháp xoay varimax, 13
phần trăm phương sai, 28
phần trăm phương sai tích luỹ, 28

SARS-CoV-2, 15

tiên đề xác suất, 3
tiêu chuẩn Guttma – Kaise, 10
trang web mã nguồn mở Github, 57
trục toạ độ, 7

vi-rút corona, 15
véc-tơ riêng, 3

ý nghĩa thống kê, 16

đó độ đo phân biệt, 2
độ lệch chuẩn, 28

Trang này dành để ghi chú.