PHÂN TÍCH THÀNH PHẦN CHÍNH

Phân tích thành phần chính tổng thể 2

Giới thiệu

- Phân tích thành phần chính (Principle Components Analysis - PCA) là bài toán về việc giảm số chiều của một tập dữ liệu có rất nhiều biến trong khi vẫn giữ lại nhiều nhất thông tin có thể từ tập dữ liệu ban đầu.
- Việc giảm số chiều được thực hiện bởi việc biến đổi tập hợp các biến gốc sang một tập hợp các biến mới, gọi là các "thành phần chính" (principal components), các thành phần chính này không tương quan và có thứ tự sao cho một số ít thành phần đầu tiên sẽ mô tả được hầu hết sự biến thiên trong dữ liệu.
- Trong phân tích thành phần chính tổng thể, các thành phần chính được tính từ ma trận hiệp phương sai Σ.

3

Ứng dụng của PCA

- Suy luận (nghiên cứu về cấu trúc).
- Tạo ra một tập các biến mới (số lượng biến nhỏ hơn ban đầu và không tương quan). Các biến này được sử dụng trong các xử lý khác (như hồi quy bội).
- Chọn ra một tập con của các biến gốc để sử dụng trong các xử lý nhiều chiều khác.
- Phát hiện các điểm outlier và chùm (cluster) trong các quan trắc.
- Kiểm tra giả định phân phối chuẩn nhiều chiều.

4

Hai cách tiếp cận

- Đại số: Các thành phần chính (PCs) là các tổ hợp tuyến tính của p biến gốc $X_1, X_2, \ldots, X_p$ sao cho
  - Thành phần chính thứ nhất có phương sai lớn nhất có thể.
  - Thành phần chính thứ hai có phương sai lớn nhất có thể (sau thành phần chính thứ nhất) và trực giao với thành phần chính thứ nhất.
  - ...
- Hình học: (ít nhất) 2 cách tiếp cận
  - Quay các trục, chuyển sang một hệ trục mới.
  - Chọn một siêu phẳng (hyper-plane) ướm tốt nhất ("best fit").

5

Phương pháp hình học cho PCA: phép quay trục tọa độ

- Các thành phần chính được biểu diễn bằng một sự lựa chọn hệ trục tọa độ mới, hệ trục tọa độ này thu được bởi việc quay các trục gốc sang 1 tập hợp các trục mới (để cung cấp 1 cấu trúc đơn giản hơn).
  - Thành phần chính thứ nhất biểu diễn hướng có sự biến thiên lớn nhất.
  - Thành phần chính thứ hai biểu diễn hướng có sự biến thiên lớn nhất và trực giao với thành phần chính thứ nhất.
  - Cứ tiếp tục như vậy, cho đến khi thành phần chính cuối cùng biểu diễn hướng có sự biến thiên nhỏ nhất và trực giao với tất cả các thành phần chính khác.

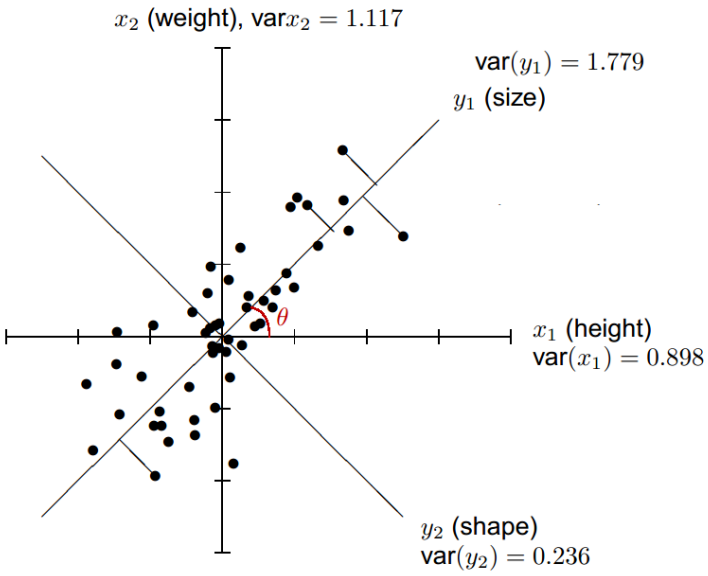
6

Phương pháp hình học cho PCA: siêu phẳng ướm tốt nhất

- Siêu phẳng "ướm" tốt nhất ("Best fit" hyper-plane) được định nghĩa bởi việc cực tiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và mặt phẳng xác định bởi các thành phần chính.
  - Thành phần chính thứ nhất xác định một đường thẳng. Tổng bình phương các khoảng cách giữa các điểm và đường thẳng này là nhỏ nhất.
  - Thành phần chính thứ hai xác định một mặt phẳng. Tổng bình phương các khoảng cách giữa các điểm đến mặt phẳng này là nhỏ nhất.
  - ...

7

Biểu diễn hình học cho PCA



8

Dạng đại số của PCA

Ta muốn biến đổi p biến sang q tổ hợp tuyến tính trực giao nhau với $q \ll p$:

$$\mathbf{X}'_{1\times p} = (X_1, X_2, \ldots, X_p)$$

sang

$$\mathbf{Y}'_{1\times q} = (Y_1, Y_2, \ldots, Y_q)$$

Tổng quát, ta có p khả năng

$$\begin{aligned} Y_1 &= \mathbf{a}'_1\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \mathbf{a}'_2\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \quad \vdots \\ Y_p &= \mathbf{a}'_p\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \\ \mathbf{Y} &= \mathbf{A}\mathbf{X} \end{aligned} \tag{1}$$

Với ma trận hiệp phương sai $\mathbf{\Sigma}_X$ của $\mathbf{X}$ đã biết, ta có

$$\mathbb{V}\mathrm{ar}\, Y_i = \mathbf{a}'_i\mathbf{\Sigma}_X\mathbf{a}_i \quad \text{và } \mathbb{C}\mathrm{ov}\,(Y_i, Y_k) = \mathbf{a}'_i\mathbf{\Sigma}_X\mathbf{a}_k$$

9

Định nghĩa PCAs

Định nghĩa 1. *Xét véc tơ p chiều $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$ và ma trận thực $\mathbf{A}$ sao cho $\mathbf{Y} = \mathbf{A}\mathbf{X}$. Các thành phần chính $Y_1, Y_2, \ldots, Y_p$ là những tổ hợp tuyến tính không tương quan, $\mathbb{C}\mathrm{ov}\,(Y_i, Y_k) = 0\ \forall\ i \neq k$, trực giao với nhau với các phương sai lớn nhất có thể.*

Cụ thể, $\mathbb{V}\mathrm{ar}\, Y_1$ là lớn nhất $\Rightarrow$ tìm $\mathbf{a}_1$ sao cho $\mathbf{a}'_1\mathbf{\Sigma}_X\mathbf{a}_1 = \max \mathbf{a}'\mathbf{\Sigma}_X\mathbf{a}$.

$\mathbb{V}\mathrm{ar}\, Y_2$ là lớn nhất và $\perp Y_1 \Rightarrow$ tìm $\mathbf{a}_2$ cho sao $\mathbf{a}'_2\mathbf{\Sigma}_X\mathbf{a}_2 = \max \mathbf{a}'\mathbf{\Sigma}_X\mathbf{a}$ và $\mathbf{a}_1\mathbf{\Sigma}_X\mathbf{a}_2 = 0$.

Tại mỗi bước, ta chọn $\mathbf{a}_i$ sao cho $\mathbf{a}_i\mathbf{X}$ có phương sai lớn nhất có thể và không tương quan với các tổ hợp tuyến tính khác.

Thông thường (nhưng không phải luôn luôn), ta chỉ sử dụng Y_1, Y_2,..., Y_q khi q nhỏ hơn nhiều so với p (mục tiêu: rút gọn dữ liệu).

10

Xác định phương sai lớn nhất

Xác định $\max(\mathbf{a}'\mathbf{\Sigma}_X\mathbf{a})$:
Ta có thể nhân $Y_1 = \mathbf{a}'\mathbf{Y}$ với một hằng số $|c| > 1$, mà sẽ làm tăng phương sai,
$\mathbb{V}\mathrm{ar}(cY_1) = \mathbb{V}\mathrm{ar}(c\mathbf{a}'\mathbf{X}) = c^2\,\mathbb{V}\mathrm{ar}(\mathbf{a}'\mathbf{X})$.
Do đó, ta chuẩn hóa véc-tơ tổ hợp

$$\mathbf{a}'\mathbf{a} = 1 = L_{\mathbf{a}}^2 = L_{\mathbf{a}}$$

Vấn đề là tìm $\mathbf{a}_1$ sao cho phương sai lớn nhất với một ràng buộc

$$\max_{\mathbf{a}}\left(\frac{\mathbf{a}'\mathbf{\Sigma}_X\mathbf{a}}{\mathbf{a}'\mathbf{a}}\right) = \mathbb{V}\mathrm{ar}(Y_1)$$

Ta chứng minh được rằng

$$\max_{\mathbf{a}}\left(\frac{\mathbf{a}'\mathbf{\Sigma}_X\mathbf{a}}{\mathbf{a}'\mathbf{a}}\right) = \lambda_1 \tag{2}$$

đặt được khi $=\mathbf{e}_1$ với $(\lambda_1, \mathbf{e}_1)$ là trị riêng lớn nhất và véc-tơ riêng tương ứng của ma trận $\mathbf{\Sigma}_X$.

11

Thành phần chính tổng thể: kết quả 1

Định lý 1. *Xét $\mathbf{\Sigma}$ là ma trận hiệp phương sai tương ứng với véc-tơ $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$. Gọi $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$, …, $(\lambda_p, \mathbf{e}_p)$ lần lượt là các cặp (trị riêng, véc-tơ riêng) tương ứng của ma trận $\mathbf{\Sigma}$ với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Thì thành phần chí thứ i xác định bởi*

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \tag{3}$$

với $i = 1, 2, \ldots, p$. Và,

$$\mathbb{V}\mathrm{ar}(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \tag{4}$$
$$\mathbb{C}\mathrm{ov}(Y_i, Y_k) = \mathbf{e}_i\mathbf{\Sigma}e_k = 0 \quad \textit{với } i \neq k \tag{5}$$

Nếu có một vài λ_i bằng nhau, thì sự lựa chọn véc-tơ riêng tương ứng $\mathbf{e}_i$ (và cả Y_i) sẽ không duy nhất.

12

Ví dụ

Cho véc-tơ $\mathbf{X} = (X_1, X_2)$ có ma trận hiệp phương sai:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}$$

Xác định các thành phần chính của $\mathbf{X}$ và tính các phương sai của các thành phần chính.

13

Thành phần chính tổng thể: kết quả 2

Định lý 2. *Xét $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai tương ứng với véc-tơ $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$. Gọi $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$, $\ldots$, $(\lambda_p, \mathbf{e}_p)$ lần lượt là các cặp (trị riêng, véc-tơ riêng) tương ứng của ma trận $\boldsymbol{\Sigma}$ với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Gọi $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}$, $\ldots$, $Y_p = \mathbf{e}_p'\mathbf{X}$ là các thành phần chính, thì*

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \sigma_{ii} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \lambda_i \tag{6}$$

Sự biến thiên tổng thể toàn phần (Total Population Variance) $= \sum_{i=1}^{p} \lambda_i$.

Tỷ lệ biến thiên toàn phần (Proportion of total variance) gây ra bởi thành phần chính thứ k là

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad k = 1, 2, \ldots, p \tag{7}$$

14

Tỷ lệ sự biến thiên ...

Ta thường chọn q thành phần chính đầu tiên mà có tỷ lệ biến thiên toàn phần có tổng gần bằng 1.

Tỷ lệ sự biến thiên (Proportion of Variance) được kiểm soát bởi q thành phần chính đầu tiên bằng

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{8}$$

Ta cần cân bằng giữa phần trăm sự biến thiên (lượng thông tin) được giữ lại và số thành phần chính (đơn giản nhất) và thay thế $\mathbf{X}$ bằng $\mathbf{Y}$.

15

Tương quan giữa Y_i và X_k

Định lý 3. *Nếu $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}$, $\ldots$, $Y_p = \mathbf{e}_p'\mathbf{X}$ là các thành phần chính thu được từ ma trận hiệp phương sai $\boldsymbol{\Sigma}$, thì*

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \tag{9}$$

Ta có thể sử dụng ρ_{Y_i, X_k} để suy luận sự đóng góp lượng thông tin của X_k trong Y_i.

16

Ví dụ

Dữ liệu về phần trăm người làm việc trong các ngành công nghiệp ở các nước Châu Âu trong năm 1979. Số lượng nước khảo sát $N = 26$. Dữ liệu khảo sát 9 ngành công nghiệp, tuy nhiên ta chỉ xét 3 biến $(p = 3)$:

$X_1 =$ phần trăm người làm việc trong ngành sản xuất.
$X_2 =$ phần trăm người làm việc trong ngành dịch vụ.
$X_3 =$ phần trăm người làm việc trong ngành dịch vụ xã hội.

$$\boldsymbol{\mu} = \begin{pmatrix} 27.008 \\ 12.958 \\ 20.023 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 49.109 & 6.535 & 7.379 \\ 6.535 & 20.933 & 17.879 \\ 7.379 & 17.879 & 46.643 \end{pmatrix}$$

Phương sai tổng cộng = $\operatorname{tr}(\boldsymbol{\Sigma}) = 116.68432$.

17

Ví dụ: trị riêng và véc-tơ riêng

i	$\mathbb{V}\mathrm{ar}(Y_i)/\lambda_i$	Phương sai tích lũy	%	% tích lũy
1	62.62	62.62	53.66	53.66
2	42.47	105.09	36.39	90.06
3	11.60	116.68	9.94	100.00

Các véc-tơ riêng, cho biết các trọng số tương tứng cho các thành phần chính

$$\mathbf{e}_1' = (0.580, 0.396, 0.712)$$
$$\mathbf{e}_2' = (0.811, -0.207, -0.546)$$
$$\mathbf{e}_3' = (-0.069, 0.894, -0.442)$$

Các thành phần chính

$$Y_1 = 0.580X_1 + 0.396X_2 + 0.712X_3$$
$$Y_2 = 0.811X_1 - 0.207X_2 - 0.546X_3$$
$$Y_3 = -0.069X_2 + 0.894X_2 - 0.442X_3$$

18

Ví dụ: suy luận từ các thành phần chính

Ta xét tương quan giữa Y_1 và Y_2 và với mỗi X_i:

	Các thành phần chính	
Biến gốc	Y_1	Y_2
X_1	$\sqrt{\frac{62.62}{49.109}}(0.580) = 0.66$	$\sqrt{\frac{42.47}{49.109}}(0.811) = 0.75$
X_2	$\sqrt{\frac{62.62}{20.933}}(0.396) = 0.69$	$\sqrt{\frac{42.47}{20.933}}(-0.207) = -0.30$
X_3	$\sqrt{\frac{62.62}{46.643}}(0.712) = 0.82$	$\sqrt{\frac{42.47}{46.643}}(-0.546) = -0.52$

Y_1: tất cả các biến đều đóng góp vào thành phần thứ nhất; là phần trăm tổng cộng cho tỷ lệ người làm việc trong tất cả các ngành công nghiệp.

Y_2: biểu diễn tương phản giữa ngành sản xuất (X_1) và ngành dịch vụ (X_2) và ngành dịch vụ xã hội (X_3).

19

Nếu tổng thể có phân phối chuẩn nhiều chiều

Nếu $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Nhắc lại contour xác suất là (các ellipsoid):

$$(\mathbf{X} - \boldsymbol{\mu})\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{X} - \boldsymbol{\mu})$$

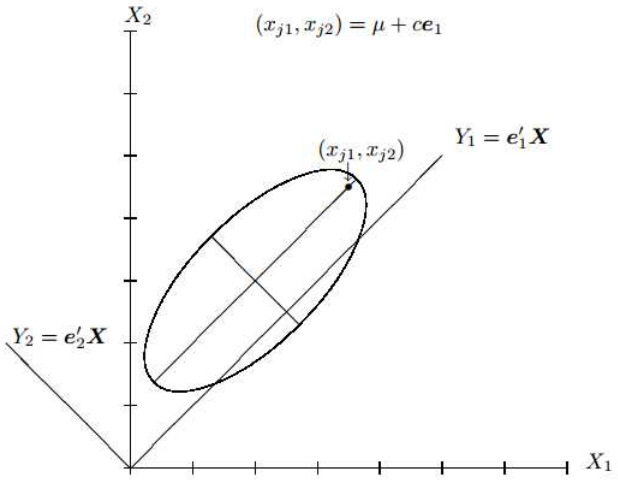
Tâm của ellipsoid là $\boldsymbol{\mu}$ và các trục là $\boldsymbol{\mu} \pm c\sqrt{\lambda_i}\mathbf{e}_i$, trong đó $(\lambda_i, \mathbf{e}_i)$ là trị riêng và véc-tơ riêng thứ i của $\boldsymbol{\Sigma}$. Các thành phần chính là

$$Y_1 = \mathbf{e}_1'\mathbf{X}$$
$$Y_2 = \mathbf{e}_2'\mathbf{X}$$
$$\vdots \quad \vdots$$
$$Y_p = \mathbf{e}_p'\mathbf{X}$$

Các thành phần chính có cùng hướng với các trục của contour xác suất (ellipsoid).

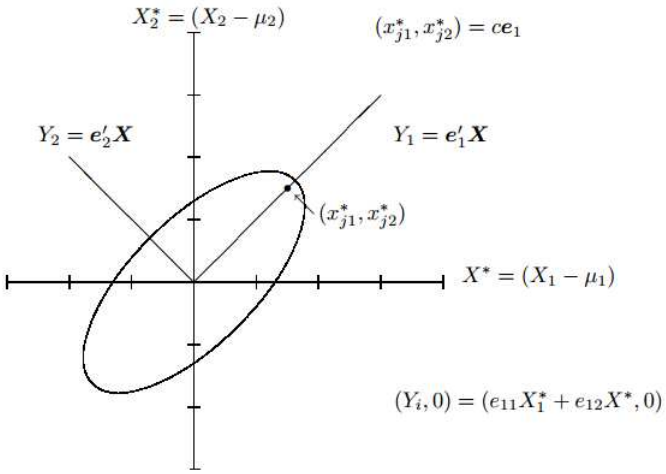
20

Contour xác suất



21

Contour xác suất



22

Chuẩn hóa các thành phần chính

Ta chuẩn hóa các thành phần chính khi phương sai của các biến khác biệt nhiều.
Sử dụng các biến chuẩn hóa ("z-scores"):

$$\begin{aligned} Z_1 &= \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} \\ Z_2 &= \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \\ &\vdots \\ Z_p &= \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}} \end{aligned} \tag{10}$$

Biểu diễn dạng ma trận:

$$\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \tag{11}$$

trong đó $\mathbf{V}^{-1/2} = \operatorname{diag}\left(1/\sqrt{\sigma_{ii}}\right)$. Do đó, $\mathbf{Z}$ là một tổ hợp tuyến tính của $\mathbf{X}$.

23

Chuẩn hóa các thành phần chính

Ta có:

$$\mathbb{E}(\mathbf{Z}) = \mathbf{0} \quad \text{và } \boldsymbol{\Sigma}_Z = \boldsymbol{\rho}$$

với $\boldsymbol{\rho}$ là ma trận hiệp phương sai tương ứng của $\mathbf{X}$.

Định lý 4. *Thành phần chính thứ i của biến chuẩn hóa $\mathbf{Z}' = (Z_1, Z_2, \ldots, Z_p)$ với $\mathbb{V}\mathrm{ar}(\mathbf{Z}) = \boldsymbol{\rho}$ cho bởi*

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{e}}_i'\mathbf{Z} = \tilde{\mathbf{e}}_i'\mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \quad \textit{với } i = 1, 2, \ldots, p \tag{12}$$

trong đó $\tilde{\mathrm{e}}_i'$ là trị véc tơ riêng thứ i ứng với trị riêng $\tilde{\lambda}_i$ của ma trận $\boldsymbol{\rho}$.

Hơn nữa,

$$\sum_{i=1}^{p} \mathbb{V}\mathrm{ar}(Z_i) = \sum_{i=1}^{p} \tilde{\lambda}_i = \sum_{i=1}^{p} \mathbb{V}\mathrm{ar}(\tilde{Y}_i) = \operatorname{tr}(\boldsymbol{\rho}) = p$$

Chú ý rằng: $\lambda_i \neq \tilde{\lambda}_i$ và $\mathbf{e}_i \neq \tilde{\mathbf{e}}_i$.

24

Ví dụ 2: Men-track data

(Johnson & Wichern): dữ liệu về thời gian của 8 thể loại đua điền kinh ($p = 8$) ở Olympics 1984 được ghi lại đối với vận động viên của 55 quốc gia ($n = 55$). 8 thể loại này như sau:

1. 100m: thời gian cho chặng đua 100m (Đv: s).
2. 200m: thời gian cho chặng đua 200m (Đv: s).
3. 400m: thời gian cho chặng đua 400m (Đv: s).
4. 800m: thời gian cho chặng đua 800m (Đv: phút).
5. 1500m: thời gian cho chặng đua 1500m (Đv: phút).
6. 5K: thời gian cho chặng đua 5000m (Đv: phút).
7. 10K: thời gian cho chặng đua 10000m (Đv: phút).
8. Marathon: thời gian cho chặng đua Marathon (khoảng 26 dặm) (Đv: phút)

Xem code trong R.

25

Thành phần chính mẫu

Các thành phần chính mẫu (Sample Principal Components) mô tả sự biến thiên trong mẫu.
Các biến đổi đại số tương tự như thành phần chính tổng thể.
Gọi $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ là n quan trắc độc lập từ một tổng thể có kỳ vọng $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$.
Từ mẫu khảo sát, ta có:

$$\bar{\mathbf{x}}_{p\times 1} = \text{ véc-tơ trung bình mẫu}$$
$$\mathbf{S}_{p\times p} = \{s_{ij}\} \text{ ma trận hiệp phương sai mẫu}$$

$\mathbf{S}$ có các cặp (trị riêng, véc-tơ riêng) là $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$ trong đó $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$.
Thành phần chính mẫu thứ i xác định bởi:

$$\hat{y}_i = \hat{\mathbf{e}}_i \mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p \tag{13}$$

26

Thành phần chính mẫu

Phương sai của thành phần chính mẫu thứ i:

$$\mathbb{V}\text{ar}(\hat{y}_i) = \hat{\lambda}_i \quad \text{với } i = 1, 2, \ldots, p$$

Hiệp phương sai của thành phần chính thứ i và j:

$$\mathbb{C}\text{ov}(\hat{y}_i, \hat{y}_j) = 0 \; \forall i \neq j$$

Phương sai mẫu toàn phần

$$\text{tr}(\mathbf{S}) = \sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$$

Tỷ lệ sự biến thiên do ảnh hưởng của thành phần chính thứ i:

$$\frac{\hat{\lambda}_i}{\sum_{k=1}^{p} \hat{\lambda}_k}$$

27

Thành phần chính mẫu

Hệ số tương quan giữa $\hat{y}_i$ và x_k:

$$r_{\hat{y}_i, x_k} = \frac{\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \hat{e}_{ik} \tag{14}$$

Nếu dùng thành phần chính chuẩn hóa các x_k thì

$$\tilde{r}_{\hat{y}_i, x_k} = \sqrt{\hat{\tilde{\lambda}}_i} \hat{\tilde{e}}_{ik}$$

trong đó $(\hat{\tilde{\lambda}}_i, \hat{\tilde{\mathbf{e}}}_i)$ là (trị riêng, véc-tơ riêng) của ma trận hiệp phương sai mẫu $\mathbf{R}$.

28