

Principal component analysis

تحلیل مؤلفه اصلی:

$$Y \sim x_1, \dots, x_n \quad n=12$$

$$X'X \sim \omega \omega' \rightarrow |X'X| = 0 \rightarrow (X'X)^{-1} \text{ وجود ندارد}$$

$$r(X) \leq \min(12, \omega) = 12 \quad r(X'X) = r(X) = 12 \quad n \leq p$$

تکریف: فرض کنید $X = (x_1, \dots, x_p)'$ یک بردار تصادفی از جابجایی n و p فاکتور

دارای n که دارای Σ باشد در این صورت تبدیل $Y = \Pi(X - \mu)$ را که در آن $\Pi = (\gamma_1, \dots, \gamma_p)$ یک فاکتور متعامد و (بردارها فاکتور Σ) به شکل زیر باشد

$$\Sigma = \Pi \Delta \Pi', \quad \Delta = \text{diag}(\lambda_1, \dots, \lambda_p)$$

$$\lambda_1 > \dots > \lambda_p$$

به تبدیل Y تبدیل مؤلفه اصلی می گویند که مؤلفه اصلی X را بر اساس λ :

$$y_i = \gamma_i' (X - \mu) \quad i = 1, \dots, p$$

که در آن γ_i ها ستون $n \times p$ فاکتور Π می باشد

فرض کنید بردار تصادفی $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ دارای متغیر $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ و فاکتور $\gamma_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

$\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ اولین دو مؤلفه اصلی را بیابید.

$$|\Sigma - \lambda I| = 0 \rightarrow \begin{vmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{vmatrix} = 0 \quad (5-\lambda)(2-\lambda) - 4 = 0$$

$$\lambda_1 = 7 \quad \lambda_2 = 1$$

$$\Sigma x_1 = \lambda_1 x_1 \rightarrow \begin{cases} 5x_1 + 2x_2 = 7x_1 \\ 2x_1 + 2x_2 = 7x_2 \end{cases}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{cases} x_1 = 2x_2 \\ x_1 = x_2 \end{cases} \xrightarrow{x_2 = C} \begin{cases} x_1 = 2C \\ x_1 = C \end{cases}$$

$$C \begin{bmatrix} r \\ 1 \end{bmatrix} \rightarrow \underline{y}_{(1)} = \begin{bmatrix} r/\sqrt{\omega} \\ 1/\sqrt{\omega} \end{bmatrix} \quad \underline{y}_{(2)} = \begin{bmatrix} 1/\sqrt{\omega} \\ -r/\sqrt{\omega} \end{bmatrix}$$

$$y_i = \begin{bmatrix} 1/\sqrt{\omega} & \frac{1}{\sqrt{\omega}} \end{bmatrix} \begin{bmatrix} x_i - r \\ x_i - r \end{bmatrix} = \frac{1}{\sqrt{\omega}} (r x_i + x_i - r)$$

$$y_i = \frac{1}{\sqrt{\omega}} (x_i - r x_i + r)$$

خواص مولفۀ اصلی:

$$E(y_i) = 0, \text{Var}(y_i) = \lambda_i; i = 1, \dots, p$$

خاصیت ۱

$$\text{Cov}(y_i, y_j) = 0; i \neq j$$

خاصیت ۲

$$\text{Var}(y_1) \geq \text{Var}(y_2) \geq \dots \geq \text{Var}(y_p)$$

خاصیت ۳

$$\text{tr}(\Sigma) = \sum_{i=1}^p \lambda_i; |\Sigma| = \prod_{i=1}^p \lambda_i$$

خاصیت ۴

$$\rightarrow E(y) = \Gamma' (E(x) \Gamma) = 0$$

اثبات خاصیت ۱:

$$\text{Var}(y) = \Gamma' \text{Var}(x) \Gamma = \Gamma' \Sigma \Gamma = \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_p \end{bmatrix}$$

$$\Rightarrow \begin{cases} \text{Var}(y_i) = \lambda_i \\ \text{Cov}(y_i, y_j) = 0 \end{cases} \quad \square$$

اثبات خصیة

$$\text{tr}(\Sigma) = \text{tr}(P \Lambda P') \Rightarrow \text{tr}(\overbrace{P P'}^I \Lambda)$$

$$\rightarrow \text{tr}(\Lambda) = \sum_{i=1}^P \lambda_i$$

$$|\Sigma| = |P \Lambda P'| = |P| |\Lambda| |P'| = |P P'| |\Lambda| = |I| = \prod_{i=1}^P \lambda_i \quad \square$$

مؤلفه‌های اصلی :

باتوجه به اینکه قبلاً فاکتورسی Var و Cov جامع یعنی Σ نداریم لذا بهترین راه حل استخراج مؤلفه‌ها اصلی بر روی فاکتورسی Σ به کمک شمولیات نمونه‌ای بر مبنای سرآورد سر S (فاکتورسی Var و Cov نمونه‌ای) لذا به راحتی می‌توان می‌توان استیانت بر مبنای فاکتورسی S را با λ ها به این گونه دیدیم ما در فاکتورسی داده‌ها یعنی X فاکتورسی Var و Cov نمونه‌ای از رابطه زیر بدست می‌آید :

$$\underline{S} = \frac{1}{n} X' X$$

و بنابراین مؤلفه اصلی λ که از λ با λ نشان می‌دهیم از رابطه زیر بدست می‌آید :

$$\underline{y}_i = (\underline{x} - \bar{x}) g_i \quad i = 1, \dots, P$$

که در آن $\underline{y}_i$ مؤلفه‌ی اصلی نمونه‌ای λ و g_i بردار P شده‌ی متناظر با معادله‌ی λ فاکتورسی

Cov و Var نمونه یعنی $\underline{S}$ است .

بنابراین در حالتی که می‌توان تمام مؤلفه‌های اصلی را با هم به شکل زیر در نظر گرفت

$$\underline{Y} = (y_1, \dots, y_P) = H X G$$

$$\underline{G} = (g_1, \dots, g_P) \quad , \quad \underline{H} = I_n - \frac{1}{n} \mathbf{1} \mathbf{1}'$$

✓ اگر تفسیر داده‌ها به شکل زیر باشد، کوفاکتورهای اصلی به صورت زیر است.

$$\underline{X} = \begin{bmatrix} 4 & 1.5 \\ 5 & 1.4 \\ 7 & 1.1 \\ 1 & 0.9 \\ 2 & 1.0 \end{bmatrix}$$

$$H = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n'$$

$$H = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} - \frac{1}{5} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{pmatrix} 0.18 & -0.12 & & & \\ -0.12 & 0.18 & & & \\ & & 0.18 & & \\ & & & 0.18 & \\ & & & & 0.18 \end{pmatrix}$$

$$S = \frac{1}{n} \underline{X}' H \underline{X} = \begin{pmatrix} 4 & 0.12 \\ 0.12 & 1.4 \end{pmatrix} \rightarrow \begin{cases} \lambda_1 = 1.75 \\ \lambda_2 = 1.4 \end{cases}$$

$$G = \begin{pmatrix} 0.141 & -0.191 \\ 0.191 & 0.141 \end{pmatrix}$$

$$\underline{Y} = H \underline{X} G = \begin{pmatrix} 1.12 & 0.181 \\ 1.05 & 0.12 \\ 1.14 & -1.12 \\ -1.12 & -0.112 \\ -1.12 & 0.191 \end{pmatrix}$$

$$y_1 = \gamma_{(1)}' (x - \mu) = (0.141 + 0.191) \begin{pmatrix} x_1 - 4 \\ x_2 - 1.0 \end{pmatrix}$$

$$= 0.141 (x_1 - 4) + 0.191 (x_2 - 1.0)$$

AKSH

$$y_2 = \gamma_{(2)}' (x - \mu) = -0.191 (x_1 - 4) + 0.141 (x_2 - 1.0)$$

عامل اندازه در مؤلفه‌ها اصلی

اگر تابع خطی به صورت $y = \alpha x$ باشد به سبب مقیاس افزایش در x سبب مقیاس افزایش در y می‌شود. در این صورت آن را عامل اندازه می‌نامیم.

عامل شکل در مؤلفه‌ها اصلی

مؤلفه اصلی به شکل $y = \alpha x$ باشد و اگر تغییر در x سبب تغییر در y می‌شود، در این صورت تغییر می‌شود که α عامل شکل است.

روش خارج کردن مؤلفه‌ها اصلی

برای خارج کردن مؤلفه‌های اصلی روش‌ها متعددی وجود دارد که یکی به طور تجربی بدست آمده و دیگری نظری ندارد. این قسمت به بررسی روش از محققین آنها می‌پردازیم:

۱. مؤلفه‌های اصلی با دارایی کمتر از متوسط را خارج می‌کنیم:

$$\text{Var}\left(\sum_{i=1}^p y_i\right) = \text{tr}(S) = \sum_{i=1}^p \lambda_i$$

$$\frac{1}{p} \sum_{i=1}^p \lambda_i = \text{دارایی متوسط} = 1 \rightarrow \text{Var}(y_i) = \lambda_i$$

if $\lambda_i < 1$ حذف

تذکره: چنانچه در تحلیل مؤلفه‌ها اصلی به جای که از فاکتورس همبستگی غویزی یعنی R استفاده کنیم در این صورت را

باید آن مؤلفه‌هایی را خارج کنیم که دارایی آن‌ها کمتر از ۱ باشد

$$\frac{1}{p} \text{tr}(R) = p/p$$

۲. آن دسته از مؤلفه‌های اصلی به حدود ۸۰ از تغییرات اطلاعات را بیان می‌کنند و در بررسی دیگر داده

و تغییرات خارج می‌کنیم

$$tr(s) = \sum_{i=1}^P \lambda_i = \sum_{i=1}^P n_i$$

$$\frac{n_1}{\sum n_i}, \frac{n_1+n_2}{\sum n_i}, \dots, \frac{n_1+\dots+n_k}{\sum n_i} \approx 1/80 \rightarrow k < P$$

در این صورت مؤلفه اصلی p ، $k+1, \dots$ را نشان‌دهنده و کار تحلیل داده‌ای را براساس k مؤلفه

اصلی اول که تغییراتی ۸۰٪ از کل اطلاعات داده‌ها هستند انجام می‌دهیم

فرض‌های تحلیل مؤلفه اصلی:

۱. فرض نریدل بودن داده‌ها در تحلیل مؤلفه اصلی لازم نیست مگر اینکه بتوانیم از روشی را بر روی داده‌ها انجام دهیم.

۲. محاسبه مؤلفه‌های اصلی به وسیله کامپیوتر بسیار ساده است و هیچ ابزاری پذیرفته نیست.

۳. مقادیر معمولی‌های مؤلفه اصلی یعنی مقادیر بردارهای مؤلفه‌های اصلی به تحلیل مرتبط می‌شوند و داده‌ها را تغییر می‌دهند.

$$x_p = x_{p1} + x_{p2} + \dots + x_{pP} = x_{p1} + x_{p2} + \dots + x_{pP}$$

تذکر مهم: اگر زمانی به تحلیل مؤلفه اصلی داران است که معیاس یا سبب یعنی اگر واحد مقیاس‌ها را تغییر دهیم آن ده

مؤلفه‌های اصلی متفاوتی بدست می‌آیند به عبارت دیگر جواب‌ها مؤلفه اصلی به واحد اندازه‌گیری داده‌ها وابسته است.

همچنین اگر مقیاس‌ها دارای واحد یکسان بوده و تغییر پذیری آنها از مقیاس اهمیت یکسان برخوردار نباشد و بار هم

تفسیر مؤلفه‌های اصلی مشکل می‌شود زیرا آنکه مشکل فوق حل شود کافی است تا آن مقیاس‌ها را استاندارد می‌کنیم

یعنی به جای استفاده از مقیاس داران از مقیاس استاندارد می‌گیریم و به این معنی R استفاده می‌کنیم

خارج کردن مقیاس‌ها و داده‌های طبعی آنها

اگر تحلیل مؤلفه اصلی بر روی مقیاس‌های استاندارد می‌شود یعنی R می‌باشد و به این معنی که k مقیاس را به دست آورده و

بقیه را خارج می‌کنیم (k عدد معلوم است) یعنی $P-k$ مؤلفه اصلی را خارج کرده ایم به این منظور بردارهای متناظر

با توجه به مقدار λ_P در نظر بگیریم و معیاری به این ترتیب ضریب از نظر قدر مطلق این ضرایب می‌گیریم به عبارت دیگر اگر λ_P اصل که متناظر با کوچکترین مقدار λ_P است را به نظر زیر در نظر بگیریم.

$$\lambda_P : \lambda_P = \lambda_P x = \lambda_{P1} x_1 + \dots + \lambda_{Pi} x_i + \dots + \lambda_{PP} x_P$$

حذف (λ_i, P)

در ادامه مؤلفه اصلی $P-1$ در نظر گرفته و به همین طریق عمل می‌کنیم تا $P-K$ کوچکترین مقدار درجه مورد بررسی قرار گیرد. قلاً در مرحله بعد که λ_P را از طریق دامنه ارزش متغیرها $\{x_1, \dots, x_P\}$ یک معیار را ضایع می‌کنیم و همین طور ادامه می‌دهیم.

نکته: به طور کلی با این روش معیاری را ضایع می‌کنیم که اولاً دارای بیشترین ضریب از نظر قدر مطلق در آن مؤلفه باشد و ثانیاً قبلاً هم ضایع نشده باشد.

برای امتحان از ضایع کردن یک متغیر به دو ضریب همسایه را نیز بدست آوریم به عنوان مثال اگر λ_P را به x_P حذف شد آنگاه برای امتحان مقدار $P(\lambda_P, x_P) = \frac{cov(x_P, y_P)}{\sqrt{var(x_P) var(y_P)}}$ را نیز حساب می‌کنیم و اگر مقدار آن نزدیک به یک شد آنگاه x_P را حذف می‌کنیم.

$$P(\lambda_P, x_P) = \frac{\lambda_P \cdot \lambda_P}{\sqrt{\lambda_P \cdot \lambda_P}} = \frac{\lambda_P \cdot \lambda_P}{\lambda_P} = \lambda_P$$

ازمونه هردوی: در این از مونه فرض H_0 به صورت زیر است:

$$H_0: \lambda_{k+1} = \dots = \lambda_{P-1} = \lambda_P$$

آماره آزمون فوق به صورت زیر است:

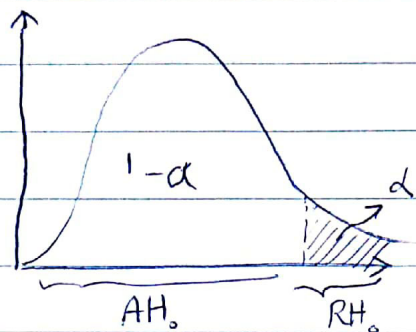
$$\chi^2_{\frac{(P-K+1)(P-K-1)}{2}} \sim \left(n - \frac{P+1}{2} \right) (P-K) \ln \left(\frac{a_0}{g_0} \right)$$

$\leftarrow$ ضرایب جدیدی $\leftarrow$ ضرایب هردوی

$$a_0 = \frac{\lambda_{k+1} + \dots + \lambda_P}{P-K}$$

$$g_0 = \sqrt[P-K]{\lambda_{k+1} + \dots + \lambda_P}$$

فرض H_0 (قبول می‌نماید) مقدار داده از میان صفر قبل از حاصل در جدول χ^2 (دست‌یافت از توزیع χ^2) در سطح $1-\alpha$ کمتر باشد.



✓ تحلیل نمونه اصلی داده‌ها مربوط به ۲۷۶ دفعه نتایج زیر داشته است.

با استفاده از دوروش خوانده شده‌شان دهد که چند نمونه اصلی راجی توان بر تحلیل مثل بد است؟

اگر توزیع حاصله مثل باشد (فرض شود) و $\alpha = 0.05$ باشد با استفاده از جدول جدولی مشخص کنید چه نمونه خارج دقت نه داشته می‌شود؟ تا با استفاده از مقدار فرضی $\lambda = 0.7$ شود مقیاس‌های راجی با در سطح شود مشخص کنید.

مقیاس / رتبه	y_1	y_2	y_3	y_4	y_5	y_6
x_1	0.145	0.154	0.176	-0.105	0.104	0.100
x_2	0.133	0.17	0.174	0.100	0.100	-0.104
x_3	0.144	0.119	-0.105	0.154	0.119	0.159
x_4	0.144	-0.145	0.102	0.148	-0.145	-0.163
$\sqrt{x_5}$	0.143	-0.148	-0.106	0.151	0.177	-0.148
x_6	0.144	-0.142	-0.105	0.148	-0.17	0.155
x_7	0.157	0.171	0.141	0.17	0.108	0.151

< 0.7

$x_5 > x_2 > x_1 > x_3 > x_7 > x_4$

λ_1 : درازی جعبه ، λ_2 : پهنای جعبه ، λ_3 : ارتفاع بازوین کف و ارتفاع

λ_4 : فاصله بین طول استخوان بازو ، λ_5 : طول استخوان ران ، λ_6 : طول استخوان ران پا

کافی

$$\bar{\lambda} = \frac{1}{7} \sum \lambda_i = \frac{1}{7} = 1 \rightarrow \lambda_1$$

با استفاده از روش متوسط نسبی می شود فقط اولین مؤلفه اصلی باقی می ماند و بقیه حذف می شوند
با استفاده از روش تقریب اول طول استخوان مؤلفه اول (مؤلفه اول) حذف شود در این صورت داریم:

$$\frac{\lambda_1}{\sum \lambda_i} = \frac{F_{1st}}{7} \approx 76\%$$

دو چنانچه دو مؤلفه اول در نظر گرفته شوند تقریباً ۸۸ درصد اطلاعات موجود مشاهدات را حفظ خواهد کرد

کرد

کافی

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} \approx 88\% \rightarrow \lambda_1, \lambda_2$$

λ_2 و λ_5

آزمون همبستگی: ابتدا برای $K=0$ داریم

$$H_0: \lambda_1 = \lambda_2 = \dots = \lambda_7$$

$K=0$

$$\alpha_0 = \frac{\lambda_1 + \dots + \lambda_7}{7} = 1 \quad , \quad g_0 = \sqrt[7]{\lambda_1 \dots \lambda_7} = 0.32$$

$$\left(\frac{n - p + 1}{7} \right) (p - K) \ln \frac{\alpha_0}{g_0} = 1870.17$$

$$\chi^2_{\frac{1-\alpha}{2}} \left(\frac{1}{2} (p - K - 1)(p - K + 2) \right) = \chi^2_{0.05} (4) = 9.49$$

$$1870.17 > 9.49 \Rightarrow R H_0$$

$$H_0: \lambda_7 = \dots = \lambda_r$$

$|K|=1$

$$a_0 = \frac{\lambda_7 + \dots + \lambda_r}{\omega} = 0,19$$

$$g_0 = \sqrt{\frac{\omega}{\lambda_r \dots \lambda_7}} = 0,19$$

$$\text{جواب: } \omega \vee 7,197 > \chi^2_{0,1,\omega}(114) = 215,5 \Rightarrow RH_0$$

$$H_0: \lambda_7 = \dots = \lambda_r$$

$|K|=2$

$$a_0 = \frac{\lambda_7 + \dots + \lambda_r}{r} = 0,11$$

$$g_0 = \sqrt{\frac{r}{\lambda_r \dots \lambda_7}} = 0,11$$

$$\text{جواب: } 170,121 > \chi^2_{0,1,\omega}(17,9) = 17,9 \Rightarrow RH_0$$

$|K|=r$

$$H_0: \lambda_7 = \lambda_\omega = \lambda_r$$

$$a_0 = \frac{\lambda_7 + \lambda_\omega + \lambda_r}{r} = 0,1$$

$$g_0 = \sqrt{\frac{r}{\lambda_r \lambda_\omega \lambda_7}} = 0,1$$

$$\text{جواب: } 171,2 > \chi^2_{0,1,\omega}(\omega) = 111,14 \Rightarrow RH_0$$

$$H_0: \lambda_7 = \lambda_\omega$$

$|K|=r$

$$a_0 = \frac{\lambda_7 + \lambda_\omega}{r} = 0,105$$

$$g_0 = \sqrt{\frac{r}{\lambda_\omega \lambda_7}} = 0,1$$

$$\text{جواب: } \omega, 71 < \chi^2_{0,1,\omega}(r) = \omega, 19 \Rightarrow AH_0$$

$$\lambda_\omega + \lambda_r + \lambda_1 + \lambda_r$$