TAMS39 omgång 2

Jonathan Vikbladh - jonvi252

October 2021

Omgång 2

8

a)

Vi testar

$$H: \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$$

mot $A \neq H$. Vi antar att datan för respektive skalbagge art är normalfördelad $\boldsymbol{X}_i \sim \mathcal{N}_{p,n_i}(\boldsymbol{\mu}_i \mathbf{1}'_{n_i}, \boldsymbol{\Sigma}_i, \boldsymbol{I})$ och tar fram within-sum of squares för varje grupp $i = 1, \ldots, r = 3$

$$\boldsymbol{V}_i = \boldsymbol{X}_i(\boldsymbol{I} - \frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}'_{n_i})\boldsymbol{X}'_i$$

och under H kan vi poola totala sum of squares

$$\boldsymbol{V} = \boldsymbol{V}_1 + \boldsymbol{V}_2 + \boldsymbol{V}_3$$

vi har olika antal observationer för varje grupp så vi tar fram en modifierad LRT teststorhet, där $f_i = n_i - 1$

$$\Lambda = \frac{\det(\boldsymbol{V}_1)^{f_1/2}\det(\boldsymbol{V}_2)^{f_2/2}\det(\boldsymbol{V}_3)^{f_3/2}(f)^{pf/2}}{\det(\boldsymbol{V})^{f/2}(f_1)^{pf_1/2}(f_2)^{pf_2/2}(f_3)^{pf_3/2}}$$

och använder Box-korrigering med någon term m, det ger teststorheten $z = -2f^{-1}m\ln\Lambda$, vi förkastar då $z > \chi^2_{df}(0.95)$ där $df = \frac{1}{2}p(p+1)(r-1)$. Vi beräknar allt och får $z = 49.27496 < 58.12404$, vi kan inte förkasta att arterna har samma kovariansmatriser. Vi fortsätter med klassificeringen

b)

Vi har varken kända $\boldsymbol{\mu}_i$ eller $\boldsymbol{\Sigma}_i$ men vi antar att $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$ och att den i:te skalbaggepopulationen är multivariatnormalfördelad $\boldsymbol{x}^{(i)} \sim \mathcal{N}_6(\boldsymbol{\mu}_i, \Sigma)$ $i = 1, 2, 3$. Vi poolar våra stickprovskovariansmatriser $\boldsymbol{S}_p = (n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + (n_3 - 1)\boldsymbol{S}_3$ och ställer upp klassificeringsregeln baserat på Mahalanobis avstånd till vardera population från observationen som

$$\boldsymbol{x}_0 \text{ ny observation, klassas som pop } i, \text{ dvs } \pi_i \text{ om } T_i = \min_{j=1,2,3} T_j$$

där $T_j = (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_j)' \boldsymbol{S}_p^{-1} (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_j)$ och $\bar{\boldsymbol{x}}_j$ som vanligt skattade väntevärdet för population $j = 1, 2, 3$. Det finns inget känt uttryck för felklassificeringssannolikheterna

c)

För 2 populationer finns ett känt asympotiskt uttryck, vi skattar Δ^2 med

$$\hat{\Delta}^2 = \frac{f - p - 1}{f} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

och skattar

$$e_1 = P(\boldsymbol{x}_0 \mapsto \pi_2 | \boldsymbol{x}_0 \in \pi_1) \approx \Phi(\frac{\Delta}{2}) + \phi(\frac{\Delta}{2}) \left(\frac{1}{16\Delta} (\frac{a_1}{n_1} + \frac{a_2}{n_2}) + \frac{a_3}{f} \right)$$

samt

$$e_2 = P(\boldsymbol{x}_0 \mapsto \pi_1 | \boldsymbol{x}_0 \in \pi_2) \approx \Phi(\frac{\Delta}{2}) + \phi(\frac{\Delta}{2}) \left(\frac{1}{16\Delta} (\frac{a_2}{n_1} + \frac{a_1}{n_2}) + \frac{a_3}{f} \right)$$

vi beräknar och får $\hat{\Delta}^2 = 36.65988$ och $e_1 = 0.002067724$ och $e_2 = 0.002015956$ så ca 0.2% sannolikhet för båda.

9

a)

Vi ser att antalet symptom x_1 är (förhållandevis) starkt positivt korrelerad med (i fallande ordning) mängden aktivitet, aptiteten, hur mycket man äter, hur mycket man sover. Nästan okorrelerad med hudreaktionen dock. Mängden aktivitet är ganska starkt positivt korrelerad med aptit och hur mycket man äter, mindre så med hur mycket man sover och även denna nästan okorrelerad med hudreaktionen. Samma mönster med mängden sömn, att den är ganska positivt korrelerad med de andra variablerna förutom hudreaktionen. Aptiten och hur mycket man äter är starkast positivt korrelerade vilket är rimligt. Alltså verkar hudreaktionen inte ha så mycket med de andra variablerna att göra.

b)

Vi genomför en PCA på stickprovskorrelationsmatrisen R, dvs vi standardiserar variablerna först, eftersom variablerna mäts i olika skalor. Vi antar även att datan är normalfördelad. För att förklara mer än 85% av variansen behöver vi 4 principalkomponenter

- PC1 är en likaviktad kombination av variablerna $x_1, \ldots, x_5$ som alla var ganska positivt korrelerade med varandra. Om man vill ha ett endimensionellt mått på patienternas respons så skulle man använda denna.

Tolkningen är att den vanligaste typen av variation i datan är att man de flesta antingen har färre symptom, tränar mindre, äter mindre osv eller har fler symptom, tränar mer, äter mer osv.

- PC2 är tung i variabeln x_6, som vi såg innan var den ganska okorrelerad med de andra så denna principalkomponenten uppstår för att kunna förklara variationen i x_6. Tolkningen är att hudreaktionerna varierar ganska fritt bland patienterna oavsett de andra variablerna, så den behöver nästan mätas separat. Och om man sen vill ha ett tvådimensionellt mått på patienternas respons så skulle detta vara den andra komponenten.
- PC4 och PC4 är svårare att tolka, vilket kan förklaras av att de har lägre varians (syns i Skree-plotten) och uppstår ungefär för att förklara den ännu oförklarade variationen.

Vi ser i scree-plotten att PC1 är den enskilda PCn som förklara absolut mest. Vi ser att det hade kunnat vara intressant att testa om ex. PC5 och PC6 har

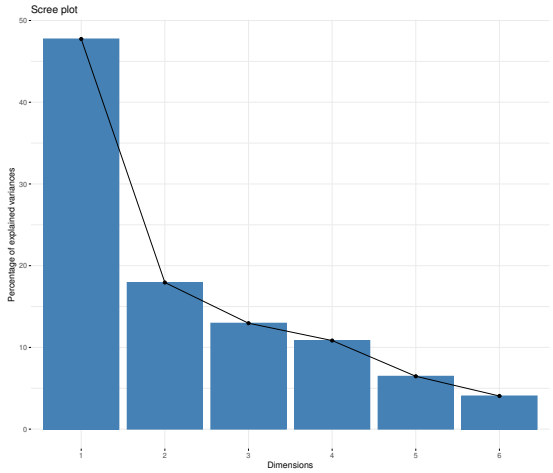

Figure 1: Scree plot

samma varians, eller om PC4, PC5 och PC6 har samma varians eller kanske PC3 och PC4. Alternativt alla principalkomponenter förutom PC1 som har mycket större variation än de andra. Vi ser i biplotten att observationerna (projicerade på underrummet som spänns upp av PC1 och PC2) är ganska spridda, vilket är bra eftersom det betyder att en stor del av variansen fångas.

c)

Vi har $n = 98$ observationer och antar att det räcker får att vi ska kunna anta att

$$\frac{\sqrt{n}(\hat{\lambda}_1 - \lambda_1)}{\sqrt{2}\lambda_1} \sim \mathcal{N}(0, 1)$$

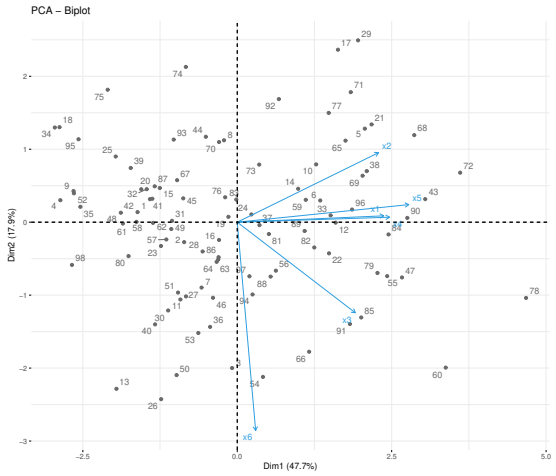

Figure 2: Biplot observationer och ursprungliga variabler projicerade på underrmmet som spänns upp av PC1 och PC2

vi stänger in $P(\left|\frac{\sqrt{n}(\hat{\lambda}_1-\lambda_1)}{\sqrt{2}\lambda_1}\right| < 1.96) = 0.95$, som blir

$$P(\frac{\hat{\lambda}_1\sqrt{n}}{\sqrt{n}-1.96\sqrt{2}} > \lambda_1 > \frac{\hat{\lambda}_1\sqrt{n}}{\sqrt{n}+1.96\sqrt{2}}) = 0.95$$

så konfidensintervallet blir

$$CI_{0.95,\lambda_1} = (\frac{\hat{\lambda}_1\sqrt{n}}{\sqrt{n}+1.96\sqrt{2}}; \frac{\hat{\lambda}_1\sqrt{n}}{\sqrt{n}-1.96\sqrt{2}})$$

vi beräknar och får, $\hat{\lambda}_1 = 2.864308$, så

$$CI_{0.95,\lambda_1} = (2.23774; 3.978205)$$

d)

Vi har att asymptotiskt för stora n är

$$\sqrt{n}(\hat{\boldsymbol{h}}_1 - \boldsymbol{h}_1) \sim \mathcal{N}_6(\boldsymbol{0}, \boldsymbol{E}_1)$$

där $\boldsymbol{E}_1 = \lambda_1 \sum_{k=2}^{6} \frac{\lambda_k}{(\lambda_k-\lambda_1)^2} \boldsymbol{h}_1\boldsymbol{h}_1'$

e)

Vi beräknar

$$r_{h_i,x_k} = \frac{\hat{h}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}$$

Vi ställer upp det som $\boldsymbol{R} = \boldsymbol{HL}$ där

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & \cdots & h_{16} \\ h_{21} & \dots & h_{26} \\ \vdots & \vdots & \vdots \\ h_{61} & \dots & h_{66} \end{bmatrix}$$

med $\boldsymbol{L} = \text{diag}(\sqrt{\frac{\hat{\lambda}_1}{s_{11}}}, \dots, \sqrt{\frac{\hat{\lambda}_6}{s_{66}}})$ Vi plottar korrelationerna mot PC1 och PC2 och ser som konstaterades tidigare att x_6 är starkt korrelerad med PC2 som just var tung i variabeln x_6, men nästan okorrelerad med PC1 vilket är rimligt eftersom PC1 och PC2 ska vara oberoende. Vi ser att x_1, x_2, x_4, x_5 är starkt korrelerade med PC1 vilket är rimligt eftersom den bestod av en likaviktning av dessa. x_3 har en viss korrelation med PC2 vilket också är rimligt eftersom den hade störst vikt i den principalkomponenten -0.38 förutom x_6.

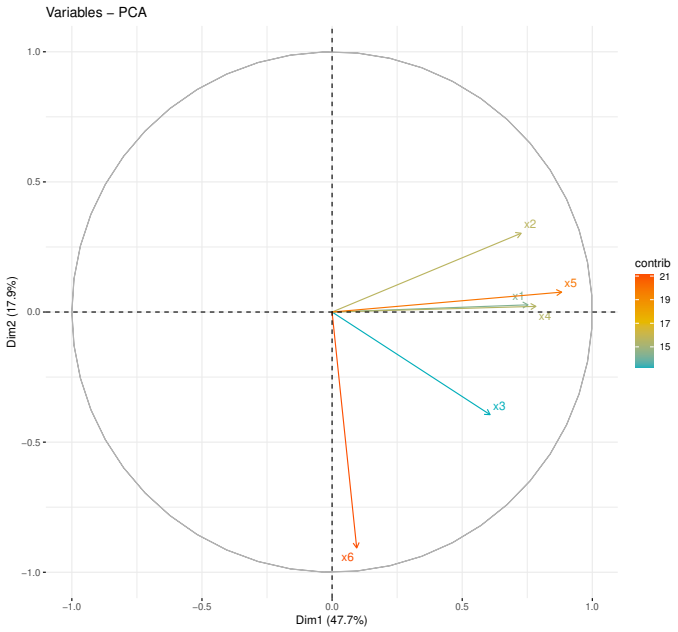

Figure 3: Korrelationerna mellan originalvariablerna och PC1 på x-axeln, och korrelationerna med PC2 på y-axeln

10

a)

Vi ställer upp en faktormodell med k faktorer som

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{Lf} + \boldsymbol{\epsilon}$$

Vi antar normalfördelning för de gemensamma faktorerna $\boldsymbol{f} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$ och de unika faktorerna, det som inte kan förklaras med de gemensamma latenta faktorerna, $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi})$ där $p = 15$, detta gör vi för att kunna maximumlikelihood-estimera $\boldsymbol{L}$ och $\boldsymbol{\Psi}$. Vi standardiserar variablerna så att kovariansmatrisen blir korrelationsmatrisen.

b)

Vi ser att vi får 3 stycken egenvärden strikt större än 1 för stickprovskorrelationsmatrisen så vi använder $k = 3$ faktorer.

c)

Vi testar om $\boldsymbol{\Sigma}$ kan beskrivas med $k = 3$ gemensamma faktorer dvs om 3 faktorer räcker

$$H : \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$$

mot $A \neq H$. Med $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\boldsymbol{\Psi}}$ och stickprovskovariansmatrisen (stickprovskorrelationsmatrisen pga standardiserad data) $\boldsymbol{R}$ har vi teststorheten

$$z = -\left(n - \frac{2p + 4k + 11}{6}\right) \ln\left(\frac{\det(\boldsymbol{R})}{\det(\hat{\boldsymbol{\Sigma}})}\right)$$

vi förkastar då $z > \chi^2_g(0.95)$ med g frihetsgrader och får $z = 71.42444 < 82.52873$ dvs vi kan inte förkasta. Alltså verkar antalet faktorer vara tillräckligt.

11

a)

Vi löser egenvärdesproblemen

$$\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{\alpha} = \rho^2\boldsymbol{\alpha}$$

och

$$\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{\beta} = \rho^2\boldsymbol{\beta}$$

vi erhåller de kanoniska korrelationerna $\hat{\rho}_1 = 0.3266219, \hat{\rho}_2 = 0.1710696$

b)

Vi löser för egenvektorerna i egenvärdesproblemet ovan och får de kanoniska korrelationskoefficienterna (stickprovs) för $\boldsymbol{x}^{(1)}$

$$\hat{\boldsymbol{\alpha}}_1 = \begin{bmatrix} 0.999996661 \\ -0.002584248 \end{bmatrix}$$

$$\hat{\boldsymbol{\alpha}}_2 = \begin{bmatrix} -0.5228640 \\ 0.8524161 \end{bmatrix}$$

och för $\boldsymbol{x}^{(2)}$

$$\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} -0.5243952 \\ -0.8514750 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} -0.9233797 \\ 0.3838878 \end{bmatrix}$$

detta ger $\hat{u}_1 = \hat{\boldsymbol{\alpha}}_1' \boldsymbol{x}^{(1)}$ och $\hat{v}_1 = \hat{\boldsymbol{\beta}}_1' \boldsymbol{x}^{(2)}$. Vi får att $\hat{u}_1 \approx x_1^{(1)}$ och att $\hat{v}_1 \approx -0.52x_1^{(2)} - 0.85x_2^{(2)}$ så om första gruppen är "mord 1973" och andra är "straff 1970" så ser vi att den maximala korrelationen blir ca 0.3, och att vi bara behöver statistiken över "nonprimary" mord, (mord mot andra än familjemedlemmer osv) och statistiken över straffen 3 år tidigare (antal avtjänade månader, osv). Vi ser att $\hat{u}_1$ och $\hat{v}_2$ väl egentligen är negativt korrelerade, vilket innebär att om straffen 1970 blir längre så sker färre "nonprimary" mord 1973. Däremot "primary" mord dvs mord mellan folk som känner varandra, verkar inte bidra eller vara korrelerade med straffen.

12

a)

Vi standardiserar variablerna för varje variabel i och observation j

$$x_{i,j} \mapsto \frac{x_{i,j} - \bar{x}_i}{s_i}$$

där $\bar{x}_i$ och s_i^2 är de vanliga konsistenta skattningarna. Vi beräknar stickprovskorrelationsmatrisen och får

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.87 & -0.37 & -0.39 & -0.49 & -0.23 \\ & 1 & -0.35 & -0.55 & -0.65 & -0.19 \\ & & 1 & 0.15 & 0.23 & 0.03 \\ & & & 1 & 0.70 & 0.50 \\ & & & & 1 & 0.67 \\ & & & & & 1 \end{bmatrix}$$

Vi ser att vikt och midja är starkt positivt korrelerade vilket är rimligt, och att både vikten och midjan är relativt negativt korrelerade med pulsen. (Antar att det är pulsen som mäts under träningen och inte vilopulsen). Vi ser även att vikt och midja är mer eller mindre negativt korrelerade med träningsvariablerna vilket också är rimligt. Pulsen är lite förvånande inte särskilt korrelerad med träningsvariablerna. Och var för sig är träningsvariablerna ganska starkt positivt korrelerade med varandra.

b)

Vi genomför en kanonisk korrelationsanalys på gruppen med fysiologiska variabler mot gruppen träningsvariabler. Först antar vi att datan är normalfördelad

så att den fysiologiska datan $\boldsymbol{X}^{(1)} = \left[\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_n^{(1)}\right] \sim \mathcal{N}_{3,n}(\boldsymbol{\mu}_1 \mathbf{1}_n', \boldsymbol{\Sigma}_{11}, \boldsymbol{I}_n)$ och träningsdatan $\boldsymbol{X}^{(2)} = \left[\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_n^{(2)}\right] \sim \mathcal{N}_{3,n}(\boldsymbol{\mu}_2 \mathbf{1}_n', \boldsymbol{\Sigma}_{22}, \boldsymbol{I}_n)$
Vi erhåller de kanoniska korrelationerna

$$\hat{\rho}_1 = 0.7956, \hat{\rho}_2 = 0.2005, \hat{\rho}_3 = 0.072$$

vi ser att det första $\hat{\rho}_1$ är relativt stor men de andra, särskilt $\hat{\rho}_3$ är små. Låt $\boldsymbol{\alpha}_1$ vara för koefficienterna för linjärkombinationen för grupp 1 x_1, x_2, x_3 och $\boldsymbol{\beta}_1$ vara för grupp 2 x_4, x_5, x_6 som ger $\text{corr}(\hat{\boldsymbol{\alpha}}_1' \boldsymbol{x}^{(1)}, \hat{\boldsymbol{\beta}}_1' \boldsymbol{x}^{(2)}) = \hat{\rho}_1 = 0.7956$

$$\hat{\boldsymbol{\alpha}}_1 = \begin{bmatrix} -0.1779 \\ 0.3623 \\ -0.0136 \end{bmatrix}$$

och

$$\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} -0.0802 \\ -0.2418 \\ 0.1644 \end{bmatrix}$$

Koefficienterna är svåra att tolka, men vi ser att $\boldsymbol{\alpha}_1$ har mest vikt i midjevariablen, och sen viktvariablen, men nästan ingen vikt i pulsvariablen. Och för $\boldsymbol{\beta}_1$ är den mesta vikten i situps och jumps, men nästan ingen vikt i chins. Det är också intressant att koefficienterna som fås då de beräknas explicit som lösningar till egenvärdesproblem som i uppgift 11 b), inte är samma som dessa.

c)

Vi har från b) antagit normalfördelning, och nu testar vi

$$H_0^k : \rho_{k+1} = \cdots = \rho_p = 0$$

mot $H_1^k \neq H_0^k$, för $k = 0, 1, 2$. Vi förkastar nollhypotesen för test k om

$$m \ln \Lambda > \chi^2_{(p-k)(q-k)}(0.95)$$

Vi har $p = q = 3$ antal variabler i grupp 1 resp. grupp 2, och

$$\Lambda = \prod_{i=k+1}^{p} (1 - \hat{\rho}_i^2)$$

och m är en Box-korrektion. Vi får

- k=0 gav $z = 15.59113 < 16.91898$ kan ej förkasta
- k=1 gav $z = 0.6406557 < 9.487729$ kan ej förkasta
- k=2 gav $z = 0.06770227 < 3.841459$ kan ej förkasta alltså kan vi inte förkasta att samtliga kanoniska korrelationer är noll (hypotestestet för $k = 0$), och inte heller för de två mindre och inte heller för den minsta korrelationskoefficienten. Alltså verkar grupperna med variabler vara okorrelerade