Kapitel 4

Multivariat normalfördelning

4.1 ?

En multivariat normalfördelad stokastisk vektorvariabel $X = (X_1, \ldots, X_p)^T$ med populationsväntevärde $\boldsymbol{\mu}$ och (icke-singulär) populationskovariansmatris $\boldsymbol{\Sigma}$ (notation: $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) har tätheten

$$= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(-(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2)$$

Egenskaper

- *Linjära kombinationer av komponenterna är normalfördelade* (JW, Thm 4.2, Thm 4.3, Thm 4.8)
- *Alla delvektorer har en normalfördelning* (JW, Thm 4.4).
- En populationskovarians lika med 0 implicerar oberoende mellan motsvarande komponenter (JW, Thm 4.5).
- De betingade (villkorliga) fördelningarna hos komponenterna är normalfördelade (JW, Thm 4.6).
- Den stokastiska variabeln $(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ har en chi-tvåfördelning med p frihetsgrader, $\chi^2(p)$, (JW Thm 4.7a). Härav följer att sannolikheten att X finns i området

  $$\{x; (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi^2_p(\alpha)\}$$

  är α (JW, Thm 4.7b)

Området kallas en konfidensellipsoid. Axlarna i ellipsoiden bestäms av egenvektorerna till $\mathbf{\Sigma}^{-1}$ som i sin tur bestäms av egenvektorerna till $\mathbf{\Sigma}$.

Om λ och $\boldsymbol{e}$ är egenvärde och egenvektor till $\mathbf{\Sigma}$ så är $1/\lambda$ och $\boldsymbol{e}$ egenvärde och egenvektor till $\mathbf{\Sigma}^{-1}$.

(Om $\mathbf{\Sigma}\boldsymbol{e} = \lambda\boldsymbol{e}$ så $\boldsymbol{e} = \lambda\mathbf{\Sigma}^{-1}\boldsymbol{e}$ och $\mathbf{\Sigma}^{-1}\boldsymbol{e} = \frac{1}{\lambda}\boldsymbol{e}$)

Maximum likelihoodskattningar

Om de stokastiska vektorerna $X_1, \ldots, X_n$ är ett slumpmässigt stickprov från en $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$-fördelning gäller att

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \quad \text{och} \quad \hat{\mathbf{\Sigma}} = \boldsymbol{S}_n$$

är maximum likelihoodskattningarna av $\boldsymbol{\mu}$ och $\mathbf{\Sigma}$.

Gemensam fördelning för $\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$

$\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$ är oberoende.

$\bar{\boldsymbol{X}}$ har en $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}/n)$-fördelning

$(n-1)\boldsymbol{S}$ har en Wishartfördelning med $n-1$ frihetsgrader

Allmänt

Det finns multivariata versioner av stora talens lag (JW, Thm 4.12) och av centrala gränsvärdessatsen (JW, Thm 4.13).

4.2 4.2

Ex: 2-dim normalfördelning:

Om

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

gäller att

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} = \frac{1}{|\mathbf{\Sigma}|}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} =$$

$$= \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Egenskaper för multivariat normalfördelning:

Linjärkombinationer

$\boldsymbol{X} \in \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma}) \quad \Rightarrow$ Varje linjärkombination

$$Y = \boldsymbol{a}'\boldsymbol{X} = \sum_{i=1}^{k} a_i X_i$$

är $\mathcal{N}(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\mathbf{\Sigma}\boldsymbol{a})$ (Result 4.2)

Speciellt $a_i = 1/k$: Ett medelvärde av komponenterna hos en multivariat normalfördelad variabel är (univariat) normalfördelad.

Flera linjärkombinationer: Om $\boldsymbol{A}$ är en $k \times p$ matris gäller att

$$\boldsymbol{AX} \in N_k(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}')$$

(Result 4.3)

Uppdelning

Speciellt gäller alltså att alla komponenterna i $\boldsymbol{X}$ är normalfördelade.

4.3 Stickprov från multivariat normalfördelning och maximum-likelihoodskattning

Om $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ är n oberoende observationer av $\boldsymbol{X} \in \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ gäller att

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}))$$

Baserad på dessa observationer är likelihooden för $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}$$

Maximum-likelihood skattningarna av $\boldsymbol{\mu}$ och $\boldsymbol{\Sigma}$ är

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \qquad \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{S} = \boldsymbol{S}_n$$

4.4 Fördelning för $\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$

Wishartfördelning

$W_m(\cdot, \Sigma)$=Wishartfördelning med m frihetsgrader = fördelningen för

$$\sum_{k=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j'$$

där $\boldsymbol{x}_j \in \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Om $\boldsymbol{A}_1 \in W_{m_1}(\cdot, \boldsymbol{\Sigma})$ och $\boldsymbol{A}_2 \in W_{m_2}(\cdot, \boldsymbol{\Sigma})$ så gäller att $\boldsymbol{A}_1 + \boldsymbol{A}_2 \in W_{m_1+m_2}(\cdot, \boldsymbol{\Sigma})$.

Om $\boldsymbol{A} \in W_m(\cdot, \boldsymbol{\Sigma})$ så gäller att $\boldsymbol{CAC}' \in W_m(\cdot, \boldsymbol{C\Sigma C}')$.

Det gäller att

1. $\bar{\boldsymbol{X}} \in \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$
2. $(n-1)\boldsymbol{S} \in W_p(n-1, \boldsymbol{\Sigma})$
3. $\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$ är oberoende.

4.5 Asymptotiska resultat för $\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$

$\bar{\boldsymbol{X}}$ och $\boldsymbol{S}$ konvergerar i sannolikhet mot $\boldsymbol{\mu}$ och $\boldsymbol{\Sigma}$.

Centrala gränsvärdessatsen:

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I})$$

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(p)$$

4.6 Fastställande av normalfördelningsantagandet

Vi vet att om $\boldsymbol{X}$ är multivariat normalfördelad så är alla marginalfördelningar (multivariat) normalfördelade. Spec:

1. 1-dimensionella fördelningar är en-dimensionell normalfördelade,
2. 2-dimensionella fördelningar är 2-dimensionellt normalfördelade.

1. Histogram
   ungefär 95 % av observationerna mellan $\mu \pm 1.96\sigma$
   Q-Q-plot: Plot av stickprovskvantiler mot förväntade kvantiler om variabeln normalfördelad: Ordna observationerna i växande ordning

   $$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

   Plotta $x_{(j)}$ mot q_j där

   $$P(Z \leq q_j) = \int_{-\infty}^{q_j} \varphi(z)\, dz = \Phi(q_j) = \frac{j - \frac{1}{2}}{n}$$

   Om variabeln normalfördelad bör $(x_{(j)}, q_j)$ ligga på ungefär rät linje.

2. Bivariat normalitet?

   Sannolikheten för $(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_{0.5}(2)$ är 0.5. Skatta $\boldsymbol{\mu}$ med $\bar{\boldsymbol{x}}$ och $\boldsymbol{\Sigma}$ med $\boldsymbol{S}$.

   Enklare att bestämma "generaliserade kvadratiska avstånd

   $$d_j^2 = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$$

   för $j = 1, 2, \ldots, n$ samt konstruera ett chi-två plot:

   (a) Ordna avstånden i växande ordning

   $$d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$$

   (b) Plotta $d_{(j)}^2$ mot $\chi^2_{\frac{j-\frac{1}{2}}{n}}(p)$, dvs $100(j-\frac{1}{2})/n$-percentilen i χ^2-fördelningen med p frihetsgrader.

   $(q_{c,p}(\alpha) = \chi^2_\alpha(p).)$

4.7 Transformation till normalitet

Några vanliga transformationer

1. Antal y: $\sqrt{y}$ (rot-transformation)
2. Proportion p: $logit(\hat{p}) = \frac{1}{2}\ln(\frac{\hat{p}}{1-\hat{p}})$
3. Korrelation r: Fishers $z = arctan(r) = \frac{1}{2}\ln(\frac{1+r}{1-r})$

Box-Cox:

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

Kan generaliseras till multivariata observationer.