Föreläsning 7: Stokastiska vektorer

Johan Thim (johan.thim@liu.se)

2 oktober 2018

1 Repetition

Definition. Låt X och Y vara stokastiska variabler med $E(X) = \mu_X$, $V(X) = \sigma_X^2$, $E(Y) = \mu_Y$ samt $V(Y) = \sigma_Y^2$. Kovariansen $C(X,Y)$ definieras enligt

$$C(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big)$$

och korrelationen mellan X och Y enligt

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma_X\,\sigma_Y}.$$

Både kovarians och korrelation är ett mått på linjärt beroende mellan X och Y där korrelationen är normerad så det går att jämföra olika fall. Vi listar lite kända egenskaper.

(i) Om $C(X,Y) = 0$ kallas X och Y för okorrelerade.

(ii) $C(X,Y) = E(XY) - E(X)E(Y)$.

(iii) Om X och Y är oberoende så är $C(X,Y) = 0$.

(iv) $C(X,X) = V(X)$.

(v) $C\left(a_0 + \sum_{i=1}^{m} a_i X_i,\, b_0 + \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j C(X_i, Y_j)$.

(vi) $|\rho(X,Y)| \leq 1$ med likhet om och endast om det finns ett linjärt samband mellan X och Y.

Observera att $C(X,Y) = 0$ *inte* nödvändigtvis innebär oberoende. Låt till exempel X vara rektangelfördelad enligt $X \sim \mathrm{Re}(-1,1)$ och definiera $Y = X^2$. Uppenbarligen beroende variabler, men

$$C(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0\cdot E(Y) = E(X^3) = \int_{-1}^{1} x^3\cdot\frac{1}{2}\,dx = 0,$$

så X och Y är okorrelerade.

2 Vektorer av stokastiska variabler

Låt $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_n)^T$ vara en vektor vars komponenter är stokastiska variabler. Vi strävar efter att skriva vektorer som kolonnvektorer. Det faller sig naturligt att definiera väntevärdet av $\boldsymbol{X}$ genom **väntevärdesvektorn**

$$E(\boldsymbol{X}) = (E(X_1)\ E(X_2)\ \cdots\ E(X_n))^T.$$

På samma sätt definierar vi väntevärdet av en matris av stokastiska variabler. Variansen blir lite konstigare så vi introducerar **kovariansmatrisen** mellan två vektorer (av samma dimension). Låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ och definiera $C(\boldsymbol{X}, \boldsymbol{Y})$ enligt

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} = \begin{pmatrix} C(X_1, Y_1) & C(X_1, Y_2) & \cdots & C(X_1, Y_n) \\ C(X_2, Y_1) & C(X_2, Y_2) & \cdots & C(X_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_n, Y_1) & C(X_n, Y_2) & \cdots & C(X_n, Y_n) \end{pmatrix}$$

där c_{ij} är kovariansen mellan X_i och Y_j.
En stor anledning att blanda in vektorer och matriser är givetvis att få tillgång till maskineriet från linjär algebra. Kovariansen mellan två vektorer $\boldsymbol{X}$ och $\boldsymbol{Y}$ kan då lite mer kompakt skrivas

$$C(\boldsymbol{X}, \boldsymbol{Y}) = E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T) = E(\boldsymbol{X}\boldsymbol{Y}^T) - E(\boldsymbol{X})E(\boldsymbol{Y})^T,$$

där $(\,\cdot\,)^T$ innebär transponering. En produkt $A = \boldsymbol{x}\boldsymbol{y}^T$ brukar kallas för den yttre produkten och består av element $(a)_{ij} = x_i y_j$, $i, j = 1, 2, \ldots, n$. Detta är alltså *inte* skalärprodukten $(\boldsymbol{X}^T\boldsymbol{Y})$.
Låt $A, B \in \mathbf{R}^{n\times n}$ vara matriser. Då är $A\boldsymbol{X}$ en linjärkombination av $X_1, X_2, \ldots, X_n$ och $B\boldsymbol{Y}$ en linjärkombination av $Y_1, Y_2, \ldots, Y_n$. Dessutom kan *alla* linjärkombinationer skrivas på detta sätt. Vidare gäller nu tack varje linjäriteten att $E(A\boldsymbol{X}) = AE(\boldsymbol{X})$ och

$$\begin{aligned} C(A\boldsymbol{X}, B\boldsymbol{Y}) &= E(A\boldsymbol{X}(B\boldsymbol{Y})^T) - E(A\boldsymbol{X})E(B\boldsymbol{Y})^T = AE(\boldsymbol{X}\boldsymbol{Y}^T)B^T - AE(\boldsymbol{X})E(\boldsymbol{Y})^T B^T \\ &= AC(\boldsymbol{X}, \boldsymbol{Y})B^T. \end{aligned}$$

Notationen $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$ är också vanligt förekommande, och i fallet då $\boldsymbol{Y} = \boldsymbol{X}$ skriver vi ofta

$$C_{\boldsymbol{X}} = \operatorname{cov}(\boldsymbol{X}) = C(\boldsymbol{X}, \boldsymbol{X}).$$

Exempel

Låt $\boldsymbol{X} = (X_1\ X_2)^T$ vara en stokastisk variabel med $E(\boldsymbol{X}) = (1\ 2)^T$ och $C_{\boldsymbol{X}} = \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}$.
Hitta en prediktor $\widehat{X}_2 = aX_1 + b$ så att $E(\widehat{X}_2) = E(X_2)$ och $V(X_2 - \widehat{X}_2)$ är minimal.

Lösning. Vi ser direkt att

$$E(aX_1 + b) = aE(X_1) + b = a + b \quad \text{och} \quad E(X_2) = 2,$$

så $a + b = 2$. Vidare gäller att

$$X_2 - (aX_1 + b) = \begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - b,$$

så

$$\begin{aligned} V(X_2 - aX_1 - b) &= V\left(\begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right) = \begin{pmatrix} -a & 1 \end{pmatrix} C_{\boldsymbol{X}} \begin{pmatrix} -a \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} -a \\ 1 \end{pmatrix} = a^2 + 4a + 4 = (a+2)^2. \end{aligned}$$

Minimum sker uppenbarligen när $a = -2$, vilket ger att $b = 4$.

3 Skattningar för kovarians och korrelation

Om vi har ett stickprov (x_k, y_k), $k = 1, 2, \ldots, n$, där (X_k, Y_k) är stokastiska variabler med samma fördelning, så skattar vi kovariansen C med

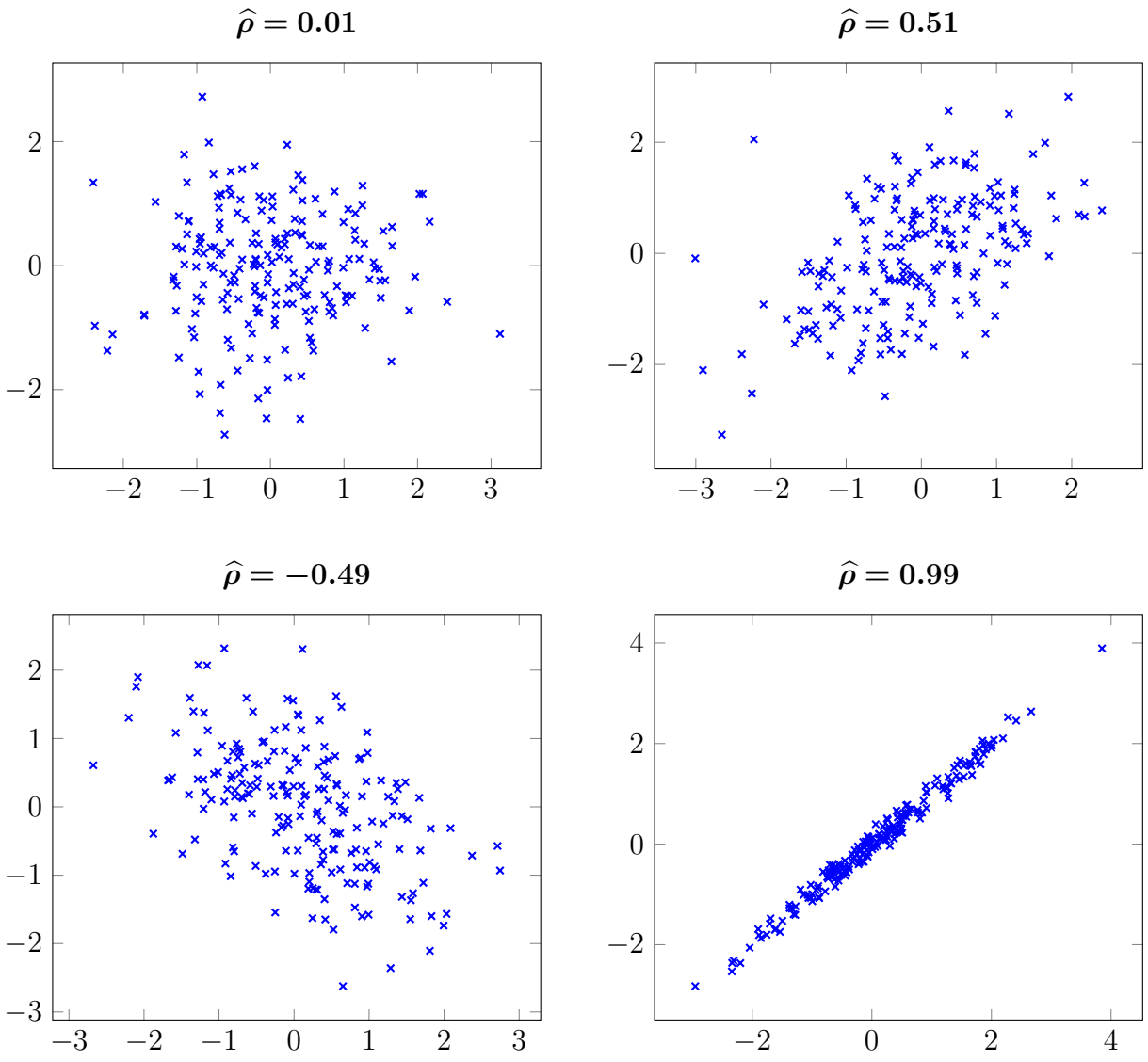
$$\widehat{c} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

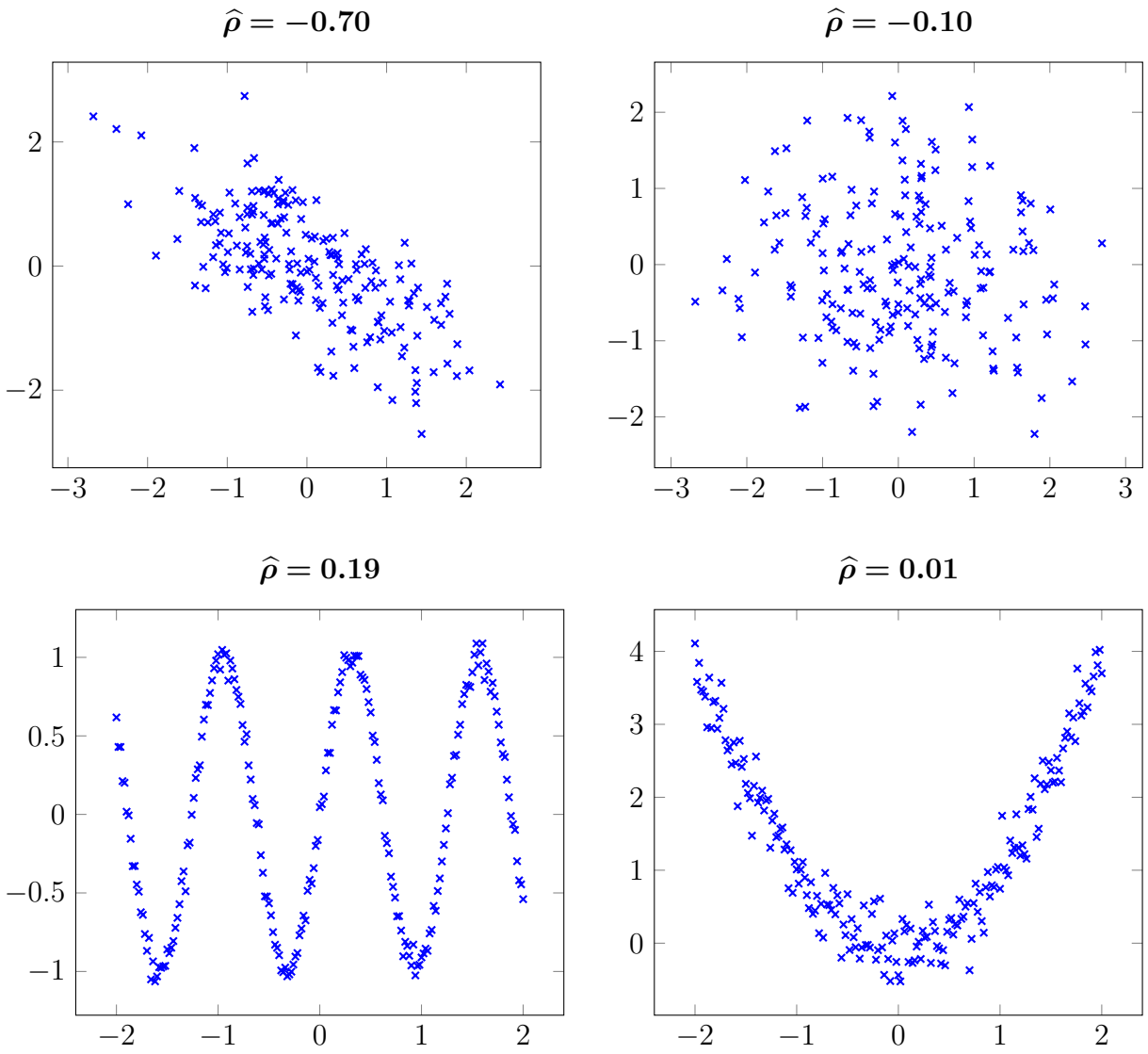
och korrelationen med

$$\widehat{\rho} = \frac{\widehat{c}}{s_x \, s_y} = \frac{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})}{\left(\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2 \right)^{1/2} \left(\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2 \right)^{1/2}}.$$

Av tradition betecknar man ofta $\widehat{\rho} = r$. En naturlig fråga i detta skede är om vi kan säga något om fördelningen för den skatta korrelationen under något lämpligt antagande om det slumpmässiga stickprovet. Vi återkommer i fallet med normalfördelning i nästa avsnitt.

3.1 Vad innebär korrelationen grafiskt?





4 Multivariat normalfördelning

"Pain has a face. Allow me to show it to you."
–Pinhead

Vi har stött på den flerdimensionella normalfördelningen tidigare, men vi kan formulera det hela lite mer kompakt på följande sätt.

Multivariat normalfördelning

Definition. Vi säger att $\boldsymbol{Y}$ har en **multivariat normalfördelning** om det finns en konstant vektor $\boldsymbol{\mu} \in \mathbf{R}^n$ och en konstant matris $A \in \mathbf{R}^{n \times m}$ så att $\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{X}$, där $\boldsymbol{X}$ är en vektor med stokastiska variabler, $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_m)^T$, och $X_i \sim N(0,1)$ är oberoende.

Är definitionen vettig? Ja, den reducerar åtminstone till det förväntade resultatet om $n = 1$: $Y = \mu + \sigma^2 X$ där $X \sim N(0,1)$. Vidare gäller så klart att

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} + AE(\boldsymbol{X}) = \boldsymbol{\mu}$$

och

$$C_{\boldsymbol{Y}} = AC_{\boldsymbol{X}}A^T = AA^T$$

eftersom $C_{\boldsymbol{X}}$ är identitetsmatrisen (variablerna är oberoende om har varians 1).

Exempel

Låt $\boldsymbol{X} \sim \mathrm{N}\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$. Bestäm fördelningen för $Y = X_1 + X_2$.

Lösning. Vi skriver $Y = (1\ 1)(X_1\ X_2)^T = A\boldsymbol{X}$. Då blir

$$E(Y) = AE(X) = (1\ 1)(1\ \ -1)^T = 0$$

och

$$C_Y = AC_XA^T = (1\ 1)\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}(1\ 1)^T = 5.$$

Sats. Om $\boldsymbol{Y}$ har väntevärdesvektorn $\boldsymbol{\mu}$ och en kovariansmatris C som uppfyller att $\det C \neq 0$ så gäller att $\boldsymbol{Y}$ har multivariat normalfördelning om och endast om $\boldsymbol{Y}$ har den simultana täthetsfunktionen

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad \boldsymbol{y} \in \mathbf{R}^n.$$

Bevis. Eftersom kovariansmatrisen C alltid är positivt semidefinit (varför?) och vi antar att determinanten $|C| := \det C \neq 0$, så är C positivt semidefinit och då finns alltid en inverterbar matris $A \in \mathbf{R}^{n\times n}$ sådan att $C = AA^T$. Definiera $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{\mu}$, där $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_n)^T$ och $X_k \sim N(0,1)$ är oberoende. Täthetsfunktionen för $\boldsymbol{X}$ ges då av

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right), \quad \boldsymbol{x} \in \mathbf{R}^n.$$

Enligt transformationssatsen för flerdimensionella stokastiska variabler så kommer

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)\left|\frac{d(x_1, x_2, \ldots, x_n)}{d(y_1, y_2, \ldots, y_n)}\right|\right.$$

eftersom $\boldsymbol{X} = A^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})$. Vi ser att jacobianen ges av

$$\frac{d(x_1, x_2, \ldots, x_n)}{d(y_1, y_2, \ldots, y_n)} = |A^{-1}| = |A|^{-1},$$

så

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}}|A|^{-1}\exp\left(-\frac{1}{2}\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^T\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)\right) \\ &= \frac{1}{(2\pi)^{n/2}}|AA^T|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T(A^{-1})^TA^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{n/2}}|AA^T|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^TC^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \end{aligned}$$

där vi utnyttjat att $|A|^{-1} = |A|^{-1/2}|A^T|^{-1/2} = |AA^T|^{-1/2}$.
Omvänt, om $\boldsymbol{Y}$ är normalfördelad så säger definitionen att det finns en matris $A \in \mathbf{R}^{n\times m}$ och en vektor $\boldsymbol{\mu} \in \mathbf{R}^n$ så att $Y = A\boldsymbol{X} + \boldsymbol{\mu}$ för $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_m)^T$ där $X_k \sim N(0,1)$ är oberoende. Faktum är att $m = n$ är nödvändigt då $C = AA^T$ antas vara inverterbar, eftersom $n = \operatorname{rank}(AA^T) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(A^T)\}$ (ty vid produkter av matriser vinner alltid den med lägst rank) och $\operatorname{rank}(A^T) = \operatorname{rank}(A)$, så $\operatorname{rank}(A) = n$ eftersom vi har n kolonner. Samma argument som ovan visar nu att täthetsfunktionen ges av uttrycket i satsen. □

Exempel

Låt $X_1,\ X_2 \sim \mathrm{N}(0,1)$ vara oberoende och definiera $\boldsymbol{Y} = (X_1 - X_2,\ 2X_1 + X_2)$. Bestäm täthetsfunktionen för $\boldsymbol{Y}$.

Lösning. Vi skriver

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \text{där } A = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}.$$

Då blir

$$\boldsymbol{\mu_Y} = E\left(A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) = A \cdot \mathbf{0} = \mathbf{0}$$

och

$$C_{\boldsymbol{Y}} = AC_{\boldsymbol{X}}A^T = A \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} A^T = AA^T = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}.$$

Således blir $\det C_{\boldsymbol{Y}} = 9$ och

$$C_{\boldsymbol{Y}}^{-1} = \frac{1}{9} \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}.$$

Alltså blir

$$f_{\boldsymbol{Y}}(y_1, y_2) = \frac{1}{6\pi} \exp\left(-\frac{1}{18}\left(5y_1^2 - 2y_1y_2 + 2y_2^2\right)\right)$$

ty

$$(y_1\ y_2)\, \frac{1}{9} \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{9}\left(5y_1^2 - 2y_1y_2 + 2y_2^2\right).$$

Sats. Låt $\boldsymbol{Z} = \boldsymbol{d} + B\boldsymbol{Y}$, där $\boldsymbol{Y}$ är multivariat normalfördelad. Då är även $\boldsymbol{Z}$ multivariat normalfördelad.

Bevis. Följer direkt från definitionen.

Sats. För $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, C)$ gäller att komponenterna i $\boldsymbol{Y}$ är oberoende om och endast om C är en diagonalmatris (under förutsättning att A är inverterbar).

Bevis. Kravet på att A ska vara inverterbar följer av att om så icke är fallet så är fördelningen degenererad eftersom $A\boldsymbol{x} = 0$ har oändligt många lösningar. Det är alltså självklart i detta läge att komponenterna i $\boldsymbol{Y}$ inte kan vara oberoende. Så antag nu att det $A \neq 0$.
Den ena riktningen är mer eller mindre självklar eftersom om komponenterna i $\boldsymbol{Y}$ är oberoende kommer $C(Y_i, Y_j) = 0$ för $i \neq j$ och $C(Y_i, Y_i) = \sigma_i^2$, så $C_{\boldsymbol{Y}}$ blir en diagonalmatris.
Antag nu att $C_{\boldsymbol{Y}}$ är en diagonalmatris, säg

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

Eftersom $C_{\boldsymbol{Y}} = AA^T$ kommer $C_{\boldsymbol{Y}}$ att vara inverterbar, vilket innebär att samtliga $\sigma_i^2 \neq 0$. Inversen $C_{\boldsymbol{Y}}^{-1}$ är även den en diagonalmatrisen med diagonalelementen σ_i^{-2}. Således blir den simultana täthetsfunktionen

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det C_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma_1\sigma_2\cdots\sigma_n} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu_j)\sigma_j^{-2}(y_j-\mu_j)\right) \\ &= \prod_{j=1}^{n} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(y_j-\mu_j)^2}{2\sigma_j^2}\right) = \prod_{j=1}^{n} f_{Y_j}(y_j). \end{aligned}$$

Eftersom den simultana täthetsfunktionen ges av produkten av täthetsfunktionerna för Y_j följer det att variablerna är oberoende. □

4.1 Bivariat normalfördelning

Specialfallet när $n = 2$ förtjänar lite kommentarer eftersom den situationen frekvent dyker upp. Låt (X, Y) vara normalfördelad med väntevärdesvektor och kovariansmatris enligt

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \text{och} \quad \begin{pmatrix} \sigma_X^2 & C(X,Y) \\ C(Y,X) & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Täthetsfunktionen ges enligt ovan av

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\,\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right),$$

för $(x, y) \in \mathbf{R}^2$.
Vi ser direkt att om $\rho = 0$ blir det produkten av täthetsfunktionerna för två oberoende variabler, precis som satsen i föregående avsnitt påstod. Men vad händer om variablerna inte är oberoende, dvs om $\rho \neq 0$ (oberoende och okorrelerade är ekvivalent i *normalfördelningsfallet*)? Låt oss beräkna den marginella tätheten $f_X(x)$ bara för kul (fast vi har nytta av den snart..). För att underlätta notationen låter vi

$$u = \frac{x-\mu_X}{\sigma_X} \quad \text{och} \quad v = \frac{y-\mu_Y}{\sigma_Y}.$$

Vi har nu

$$\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 = u^2 - 2\rho uv + v^2 = (v-\rho u)^2 + (1-\rho^2)u^2,$$

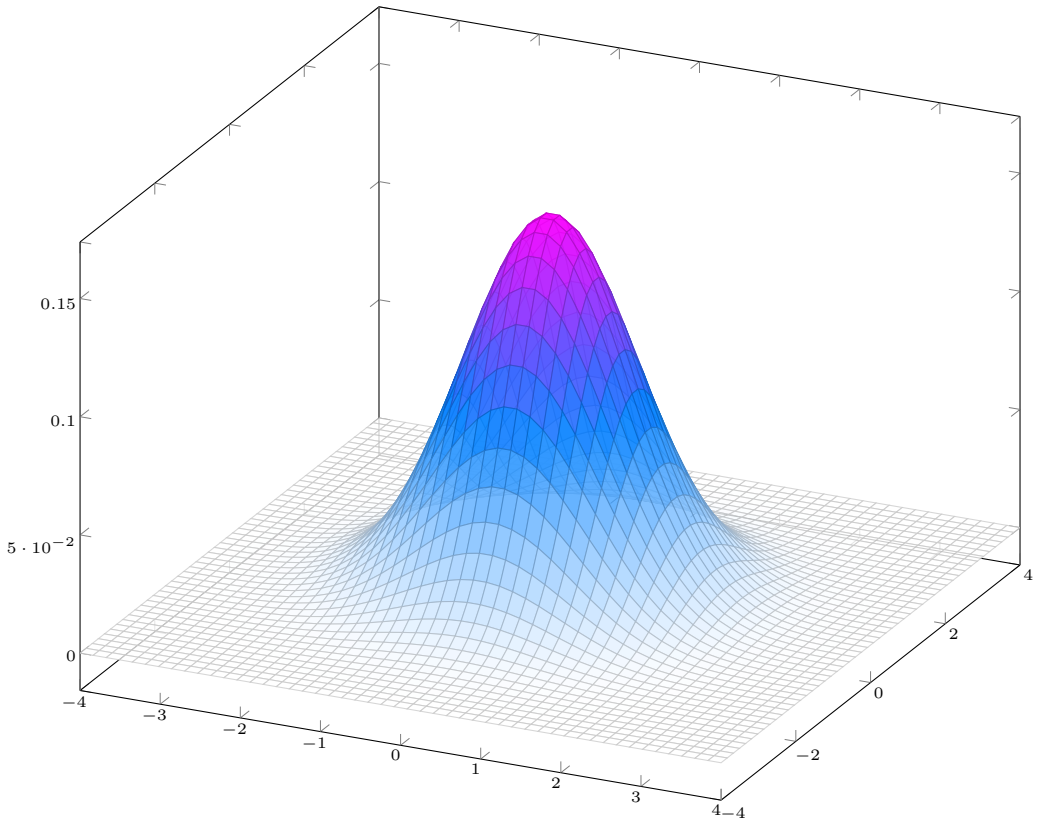
så

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}u^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}u^2\right) \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dv \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}u^2\right), \end{aligned}$$

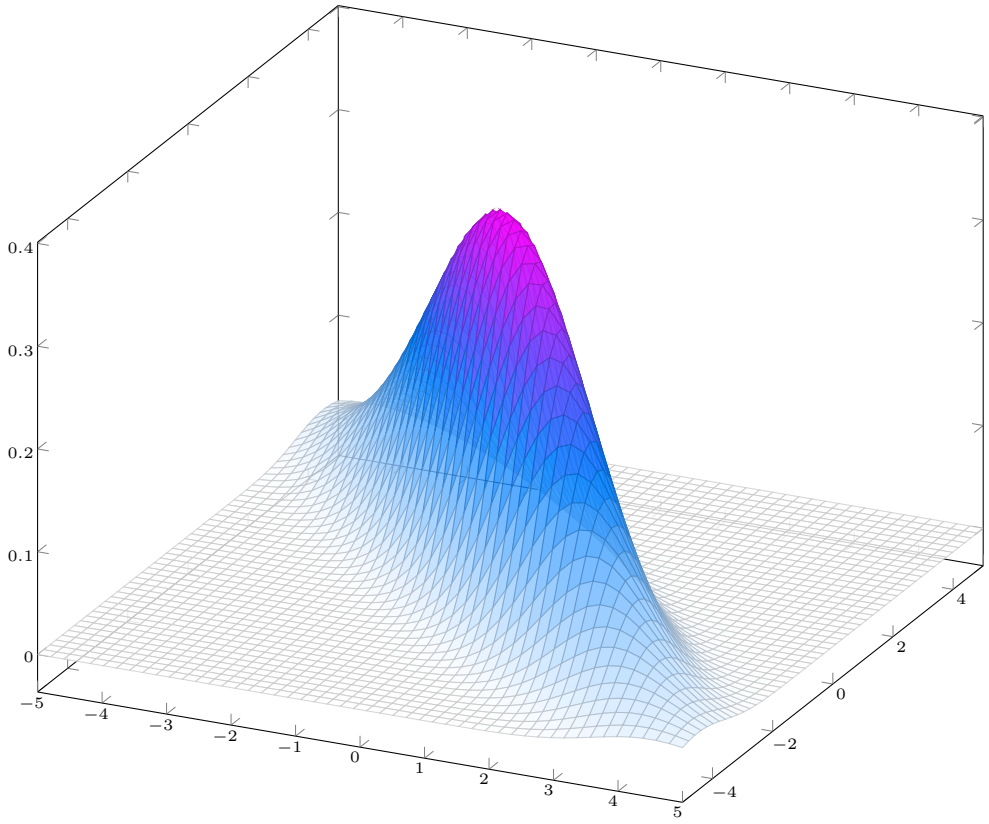
ty

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dv = 1.$$

Hur ser bivariata normalfördelningar ut? Om $\sigma_X = \sigma_Y = 1$ och $\rho = 0$ får vi följande figur:



och med $\sigma_X = \sigma_Y = 1$ och $\rho = 0.9$ erhåller vi



4.2 Test för $\rho = 0$

Att direkt ge sig på uttrycket för $\widehat{\rho}$ är komplicerat, så vi börjar lite annorlunda. Låt (X, Y) vara bivariat normalfördelad. Då har (X, Y) en simultan täthetsfunktion $f(x, y)$ och den betingade (på $X = x$) täthetsfunktionen blir

$$f_{Y \mid X=x}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(v-\rho u\right)^2\right),$$

vilket är tätheten för en normalfördelad variabel $Y \mid X = x$ med

$$E(Y \mid X = x) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X + \rho\frac{\sigma_Y}{\sigma_X}x = \beta_0 + \beta_1 x$$

och

$$V(Y \mid X = x) = \sigma_Y^2(1-\rho^2).$$

Det betingade (för givet X) väntevärdet är alltså en rät linje $y = \beta_0 + \beta_1 x$. Intressant! Åter igen något som är halvmagiskt för normalfördelningen (det finns ingen fördelning ni kan misshandla lika mycket). Den observante läsaren funderar nog även om detta har med regressionsanalysen att göra, vilket vi kommer till nästa föreläsning. För nuvarande situation, notera specifikt att

$$\beta_1 = \rho\,\frac{\sigma_Y}{\sigma_X}.$$

Anledningen till denna manöver är att vi hellre betraktar tester för β_1 än direkt för ρ. Varför? Det har med ovanstående att göra (linjär regression). Tänk tillbaka till anda föreläsningen. Där visade vi att MK-skattningen $\widehat{\beta_1}$ av β_1 ges av

$$\widehat{\beta_1} = \frac{\sum_{j=1}^n (x_j - \overline{x})(Y_j - \overline{Y})}{\sum_{j=1}^n (x_j - \overline{x})^2}$$

och att $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\,\overline{x}$. Anledning till att blanda in detta är att vi på nästa föreläsning kommer att visa att

$$\widehat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{j=1}^n (x_j-\overline{x})^2}\right).$$

Vi introducerar lite förenklande beteckningar. Låt

$$S_x^2 = \frac{1}{n-1}\sum_{j=1}^n (x_j-\overline{x})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{j=1}^n (Y_j-\overline{Y})^2 \quad \text{samt} \quad S_{xy} = \frac{1}{n-1}\sum_{j=1}^n (x_j-\overline{x})(Y_j-\overline{Y}).$$

Notera nu att vi kan skriva R (den stokastiska motsvarigheten till $\widehat{\rho}$) som

$$R = \frac{S_{xy}}{S_x\,S_y}$$

och därmed blir

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_x^2} = R\frac{S_y}{S_x}.$$

Vi introducerar det totala kvadratfelet, dvs summan av kvadraterna på skillnaden mellan mätvärden Y_j och de skattade värdena $\widehat{\beta}_0 + \widehat{\beta}_1 x_j$:

$$\mathrm{SS_E} = \sum_{j=1}^n (Y_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)^2.$$

Vi kommer (även detta nästa föreläsning) att visa att $\mathrm{SS_E}$ är oberoende av $\widehat{\beta}_1$ och att $\frac{1}{\sigma^2}\mathrm{SS_E} \sim \chi^2(n-2)$. Vidare har denna storhet egenskapen att

$$\mathrm{SS_E} = (n-1)S_Y^2(1-R^2)$$

vilket kan ses genom att expandera kvadraten i summan som definierar $\mathrm{SS_E}$:

$$\begin{aligned}\mathrm{SS_E} &= (n-1)\left(S_y^2 - 2\widehat{\beta}_1 S_{xy} + \widehat{\beta}_1{}^2 S_x^2\right)\\ &= (n-1)\left(S_y^2 - 2R\frac{S_y}{S_x}S_{xy} + R^2\frac{S_y^2}{S_x^2}S_x^2\right) = (n-1)S_y^2(1-R^2).\end{aligned}$$

Det följer då (Gossets sats) att

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2}\,\mathrm{SS_E}/(n-1)S_x^2}} \sim t(n-2).$$

Om nu $\rho = 0$ (vilket innebär att $\beta_1 = 0$ enligt ovan) så gäller att identiteten

$$\frac{\widehat{\beta}_1}{\sqrt{\frac{1}{n-2}\mathrm{SS_E}/((n-1)S_X^2)}} = \frac{rS_Y/S_X}{\sqrt{\frac{(n-1)S_Y^2(1-r^2)}{(n-2)(n-1)S_X^2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

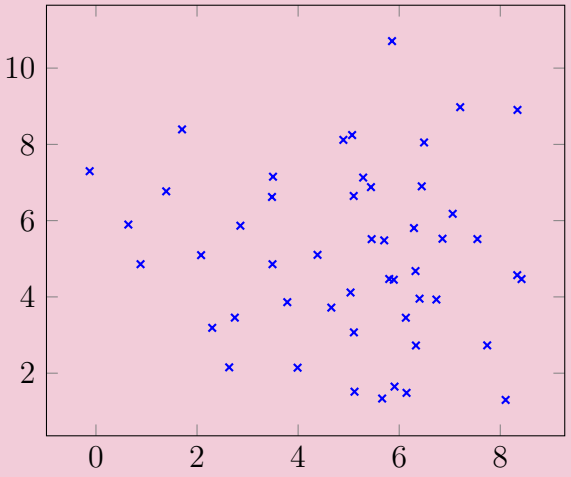
medför att

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2).$$

Vi kan alltså använda denna storhet för att testa hypotesen $H_0 : \rho = 0$.

Exempel

Astrid och Åsa grälar om två variabler är okorrelerade eller inte. Vid 50 mätningar av två variabler X och Y erhöll de diagrammet till höger som spridningsplott. Den empiriska korrelationen beräknades till $\widehat{\rho} = -0.0838$. Astrid hävdar att det tyder på att $\rho = 0$ medan Åsa anser att det absolut är signifikant (om än lågt pga slumpen). Om vi antar att X och Y är normalfördelade, testa hypotesen att $H_0 : \rho = 0$ mot $H_1 : \rho \neq 0$ med signifikansnivån 5%.



Lösning. Med 50 mätningar av (X, Y) blir

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(48)$$

om H_0 är sann. Kritiskt område erhålls därmed som

$$C = \{t \in \mathbf{R} : |t| > 2.0106\}$$

ty $P(T \leq 2.0106) = 0.975$. Med det uppmätta $r = -0.0838$ blir

$$t = \frac{-0.0838\sqrt{48}}{\sqrt{1-(-0.0838)^2}} = -0.5826.$$

Eftersom $t \notin C$ så kan vi inte förkasta H_0. Variablerna kan mycket väl vara okorrelerade (men vi vet inte det!).

5 Bonus: fördelningen för R

Sats. Om $\rho = 0$ ges fördelningen för R av täthetsfunktionen

$$f_R(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi}}(1-r^2)^{(n-4)/2}, \quad -1 < r < 1.$$

Bevis. Eftersom $g(s) = \dfrac{s}{\sqrt{1-s^2}}$ är en strängt växande funktion för $-1 < s < 1$ så gäller att

$$F_R(r) = P(R \leq r) = P\left(\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \leq \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right) = F_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right),$$

där F_T är fördelningsfunktionen för en $t(n-2)$-fördelad variabel. Denna funktion är en integral av en kontinuerlig täthet, så vi kan derivera fram

$$\begin{aligned} f_R(r) =& \frac{d}{dr}\left(F_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)\right) = f_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)\frac{\sqrt{n-2}}{(1-r^2)^{3/2}} \\ =& \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1+\frac{r^2}{1-r^2}\right)^{-(n-1)/4}(1-r^2)^{-3/2} \\ =& \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{(n-1)/4}(1-r^2)^{-3/2} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{(n-4)/4}, \end{aligned}$$

vilket är täthetsfunktionen given i satsen. □

Vad händer om $\rho \neq 0$? En fullt rimlig fråga, men fördelningen har inget trevligt utseende då (inkluderar hypergeometriska funktioner).