IMM, 2001-11-16
Poul Thyregod

Supplement til kapitel 4

Om sandsynlighedsmodeller for flere stokastiske variable

Todimensionale stokastiske variable

Lærebogens afsnit 4 introducerede sandsynlighedsmodeller formuleret ved stokastiske variable. I afsnittet betragtede man også fordelinger for flere end én stokastisk variabel (fx når man betragtede fordelingen af summer mv), men kun under antagelse af at de indgående stokastiske variable var *uafhængige*. I denne note vil vi udvide nogle af betragtningerne fra afsnit 4 til at omfatte fordelinger af flere stokastiske variable, der ikke nødvendigvis er uafhængige. Vi vil fortrinsvis betragte fordelinger af to stokastiske variable, X og Y.

Simultan tæthedsfunktion

I analogi med tæthedsfunktionen (probability density function, side 175) for en endimensional kontinuert stokastisk variabel, kan man indføre den *simultane tæthedsfunktion* for en flerdimensional stokastisk variabel.

Den simultane tæthedsfunktion for en todimensional stokastisk variabel (X, Y) er en funktion, $p_{X,Y}(x, y)$, defineret for alle reelle talpar (x, y) som opfylder

- $p_{X,Y}(x, y) \geq 0$
- $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x, y) \mathrm{d}y \mathrm{d}x = 1$$

- For to vilkårlige talpar, (a, b) og (c, d) med $a < b$ og $c < d$ gælder

$$\mathsf{P}[\{a < X \leq b\} \cap \{c < Y \leq d\}] = \int_a^b \left[\int_c^d p_{X,Y}(x, y) \mathrm{d}y \right] \mathrm{d}x$$

Sandsynligheden bestemmes altså som et *volumen* under den flade, der er angivet ved tæthedsfunktionen.

Specielt har man de *marginale* sandsynligheder for X,

$$\mathsf{P}[a < X \leq b] = \mathsf{P}[\{a < X \leq b\} \cap \{-\infty < Y < \infty\}] = \int_a^b \left[\int_{-\infty}^{\infty} p_{X,Y}(x, y) \mathrm{d}y \right] \mathrm{d}x$$

og den *marginale tæthedsfunktion* for X

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y) \mathrm{d}y$$

som fås ved at flytte den todimensionale sandsynlighedsmasse ud i *margenen*.

Stokastisk uafhængighed

Definitionen på stokastisk uafhængighed (Petrucelli side 180) udtrykker, at stokastiske variable er uafhængige, hvis man kan bestemme sandsynligheder i den simultane fordeling ud fra de marginale sandsynlighedsfordelinger.

Momenter

Det følger af betragtningerne om de marginale sandsynligheder, at det er underordnet om man bestemmer middelværdier og varianser for X og Y (μ_X, σ_X^2 og μ_Y, σ_Y^2) i den simultane fordeling, eller i de marginale fordelinger.

Man indfører *kovariansen* (covariance) mellem X og Y som

$$\mathsf{COV}[X, Y] = \mathsf{E}[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) p_{X,Y}(x, y) \mathrm{d}y \mathrm{d}x$$

Ofte bruges symbolet σ_{XY} for kovariansen mellem X og Y.

Vi bemærker, at kovariansen mellem X og Y er den samme som kovariansen mellem Y og X. Kovariansen måles i enheden for X gange enheden for Y. Hvis X og Y er uafhængige, er kovariansen mellem dem nul (men det omvendte gælder ikke nødvendigvis)

Undertiden opstilles varianser og kovarianser på matrixform i en såkaldt *kovariansmatrix* (eller *dispersionsmatrix*)

$$\Sigma\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \mathsf{E}\left[\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}'\right] = \begin{bmatrix} \mathsf{V}[X] & \mathsf{Cov}[X,Y] \\ \mathsf{Cov}[X,Y] & \mathsf{V}[Y] \end{bmatrix}$$

hvor $'$ angiver den *transponerede* (se side 511).
Udtrykket generaliseres let til flere end to stokastiske variable.

For vilkårlige stokastiske variable, X og Y gælder

$$\mathsf{V}[X \pm Y] = \mathsf{V}[X] + \mathsf{V}[Y] \pm 2\mathsf{Cov}[X,Y]$$

Såfremt X og Y er uafhængige fås netop resultaterne på side 162.

Korrelationskoefficient

Undertiden betragtes et dimensionsløst mål for samvariationen mellem to variable, den såkaldte korrelationskoefficient.

Korrelationskoefficienten, ρ (ro) mellem to stokastiske variable, X og Y defineres som

$$\rho = \frac{\mathsf{Cov}[X,Y]}{\sqrt{\mathsf{V}[X]\mathsf{V}[Y]}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Det kan vises, at $-1 \leq \rho \leq 1$.

Den todimensionale normalfordeling

Den endimensionale normalfordeling kan generaliseres til flerdimensionale variable. Lige som den endimensionale normalfordeling er parametriseret ved middelværdi og varians er den flerdimensionale normalfordeling parametriseret ved middelværdier, varianser og kovarianser for de betragtede variable.

Den todimensionale normalfordeling har tæthedsfunktionen

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

hvor

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Middelværdien for X og Y er hhv μ_X og μ_Y og kovariansmatricen er

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

Det gælder, at de *betingede fordelinger* (af Y givet en værdi $X = x$, og analogt af X givet en værdi $Y = y$) også er normalfordelinger.

For den betingede fordeling af Y givet en værdi $X = x$ gælder

$$\begin{aligned} \mathsf{E}[Y|X=x] &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \\ \mathsf{V}[Y|X=x] &= \sigma_Y^2(1-\rho^2) \end{aligned}$$

Udtrykket for den betingede middelværdi knytter forbindelsen mellem korrelationskoefficienten og *regressionen af Y på X*. Udtrykket for den betingede varians viser, at i den todimensionale normalfordelingsmodel er variansen omkring regressionslinien konstant (den afhænger således ikke af værdien af x), og endvidere genfinder vi "forklaringsgraden" $(1-\rho^2)$, der netop udtrykker den del af den marginale varians, σ_Y^2 for Y, som er forklaret af samvariationen mellem X og Y.

Den k-dimensionale normalfordeling

Betragt nu en k-dimensional stokastisk variabel (skrevet som søjlevektor),

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}$$

Y siges at følge en k-dimensional normalfordeling med (den k-dimensionale) middelværdivektor $\boldsymbol{\mu}$ og (den $k \times k$-dimensionale) kovariansmatrix $\boldsymbol{\Sigma}$, hvis tætheden for Y kan skrives på formen

$$p_Y(y) = \frac{1}{(\sqrt{2\pi})^k |\boldsymbol{\Sigma}|} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right]$$

hvor $|\boldsymbol{\Sigma}|$ angiver determinanten af $\boldsymbol{\Sigma}$.

Lineære transformationer af flerdimensionale variable

Lad $\mathbf{Y}$ være en k-dimensional stokastisk variabel med middelværdivektor $\boldsymbol{\mu}$ og kovariansmatrix $\boldsymbol{\Sigma}$ og lad $\mathbf{A}$ være en $p \times k$-dimensional matrix af konstanter, da gælder

$$\mathsf{E}[\mathbf{AY}] = \mathbf{A}\mathsf{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\mu}$$

og endvidere gælder, at kovariansmatricen for $\mathbf{AY}$ er

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

Disse resultater benyttes i Appendix 8.1

Såfremt specielt $\mathbf{Y}$ følger en k-dimensional normalfordeling gælder at $\mathbf{AY}$ vil følge en p-dimensional normalfordeling med middelværdivektor og kovariansmatrix bestemt som ovenfor. Det gælder derfor specielt for den multiple regressionsmodel, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, at estimatoren $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (8.52) følger en normalfordeling med kovariansmatricen $\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ som anført nederst på side 516.