Multivariat statistik

Morten Dam Jørgensen

2011-05-15

Multivariat statistik

I virkelige eksperimenter måles typisk mere end én type observabel. Måles elektriske signaler, kan man f.eks. måle spænding og strømstyrke for forskellige signaler i forsøgsopstillingen. Hvis de forskellige typer målinger afhængerer af hinanden kaldes de for korrelerede. Målinger der derimod ikke viser nogen korrelation kaldes for uafhængige, og beskriver unikke egenskaber ved forsøget.

Kovarians

I univariat statistik, har gentagende målinger af samme størrelse en sandsynlighsfordeling, der beskriver de udfald som hver enkelt måling kan have. Hvis fordelingen er normalt-fordelt (gaussisk) kan man udtale sig om dens middelværdi og varians. Variansen angiver udstrækningen af udfaldsrummet omkring middelværdien, og kan for en enkelt måling fortælle os noget om kvaliteten af målingerne (målingens usikkerhed), og for sande tilfældige observable, noget om den intrinsieke variation af værdien.

I multivariat statistik, hvor sandsynlighedsfordelingen har dimensioner større end 1, kan man indføre en ko-varians, eller variationen af en observabel givet variationen af en anden.

ko-variansen har mange interessante egenskaber, og bliver ofte brugt i forbindelse med maskinlæring og andre avancerede statistiske metoder, til at beskrive forventningen til udfaldet af en kompliceret måling.

I denne uges øvelser vil vi se nærmere på hvordan kovariansen kan bruges til at finde dybere sammenhænge mellem flere målinger af samme system. Målet er at finde frem til om K forskellige målinger på samme system, i virkeligheden kan beskrives med $J \ll K$ observable. Da flere af teknikkerne benytter Linær Algebra, er MATLAB et oplagt værktøj til at udforske metoderne.

Kovarianserne mellem flere observable beregnes for hver kovarians som vist i ligning 1, hvor $E[\cdot]$ angiver forventningsværdien, og μ er middelværdien for den enkelte observabel.

$$\Sigma_{i,j} = \mathrm{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \tag{1}$$

For at gøre den videre analyse nemmere, opstiller man en kovarians matrice som vist i ligning 2,

$$\Sigma = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}(X_n) \end{bmatrix} \tag{2}$$

Bemærk at langs diagonalen finder vi kovariansen for hver enkelt observabel med sig selv, hvilket blot er variansen.

I MATLAB, beregnes kovariansen ved enten at benytte den indbyggede metode *cov(M)*, eller manuelt ved at beregne kovariansen for hver enkelt element som beskrevet ovenfor, og indføre dem i en matrice.

- Beregn manuelt kovariansen for følgende $M \times N$ matrice, hvor M er antallet af observable, og N er antallet af målinger.

$$\begin{bmatrix} 12 & 44 & 6 \\ 6 & 23 & 2 \\ 4 & 34 & 2 \\ -2 & 54 & 0 \\ 4 & 32 & 2 \\ 5 & 14 & 2 \end{bmatrix}$$

  Beregn matricen igen med MATLABs indbyggede funktion.

- Indlæs datasættet "blabla.txt", og plot det med 'scatter()'. Bestem visuelt om de to observable x, y har er indbyrdes kovariante. Beregn kovariansen med 'cov()'.

Ved at normalisere kovariansen med produktet af standard afvigelsen for de to observable (ligning 3), kan man få et korrelationsmål. Denne type korrelation kaldes for lineær, da den ikke kan beskrive komplicerede afhænigigheder, som f.eks. symmetrier.

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3}$$

For linære sammenhænge, vil $\rho_{X,Y}$ give et tal mellem -1 og 1, hvor 1 svarer til at de to observable er 100% korrelerede, hvilket betyder at den ene variable beskriver det samme som den anden. Modsat vil $\rho_{X,Y} = -1$ betyde at en måling af en positiv værdi for X vil resultere negativ måling af Y, X og Y siges at være modsat korrelerede. Er $\rho_{X,Y} = 0$, er de to observable ikke korrelerede.

- Beregn korrelationsmatricen for datasættet i opgave 1 ved hjælp af ligning 3. Benyt derefter den indyggede funktion corr(). Er de tre observable korrelerede? Stemmer dine plots i opgave 2 med resultatet?

- Korrelation kan være tegn på et kausalt forhold mellem to eller flere variable, men det er ikke garanteret, Undersøg datasættet 'buh.txt' ved at beregne korrelationen mellem de forskellige observable, Er de korrelerede? Plot de forskellige observable mod hinanden, findes der en sammenhæng mellem de forskellige observable? Forklar hvad du ser.

Principal Component Analysis

Ved hjælp af lineær algebra, kan man finde en basis for et systems observable, der maksimerer variansen i hver dimension. En stor fordel ved denne metode er at antallet af observable kan reduceres hvis en eller flere beskriver samme fænomen. Teknikken kaldes for principal component analysis.

- Forbered datasættet med N datapunkter af $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i}, ...\mathbf{x}_{ni})$ som en matrix $\mathbf{X}$ med dimensionerne $N \times M$.

- Fratræk middelværdien af hver søjle (observabel), $\mathbf{B} = \mathbf{X}_i - \mu_i$

- Beregn kovarians matricen, $\mathbf{C} = \frac{1}{N}\mathbf{B}^{\mathrm{T}}\mathbf{B}$

- Beregn egenværdierne og egenvektorerne af $\mathbf{C}$, sådan at $\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$, hvor $\mathbf{V}$ er egenvektorerne af $\mathbf{C}$ og $\mathbf{D}$ er en $M \times M$ matrie med egenværdierne af $\mathbf{C}$ langs diagonalen.

- Sorter søjlerne af $\mathbf{D}$ med de største egenværdier først. Benyt samme sortering på egenvektorene i $\mathbf{V}$.

- Fravælg egenværdierne som falder under en grænseværdi η, der sættes sådan at de tilbageværende komponenter beskriver mest muligt (typisk 90%).

- projekter det oprindelige datasæt med egenvektorerne: $\mathbf{X}' = \mathbf{V}^{\mathrm{T}}\mathbf{B}$.