Föreläsning 13

Statistisk inlärning och dataanalys (Kungliga Tekniska Högskolan)

Föreläsning 13

Nu har vi sett flera sätt att konstruera en statistisk modell och i många situationer har vi sett att vi kan konstruera flera modellera för en viss situation som ger oss ett stickprov $x_1, \ldots, x_n$. Från det frekventistiska perspektivet antar vi först att datan kommer ifrån en $f_X(x|\boldsymbol{\theta})$-fördelning där vi bestämma formen av täthetsfunktion (eller sannolikhetsfunktion) $f_X(x|\boldsymbol{\theta})$. När vi gör ett sådant val beror modellen då på parameterarna $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Det är ett val som vi gör men vi kunde har tagit en hel annan form för täthetsfunktionen $f'_X(x|\boldsymbol{\eta})$ där modellen beror på olika parameterar $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ där k och θ är inte nödvändigtvis lika med. Idag ska vi betrakta frågan om hur man kan bestämma mellan sådana olika modeller. Hur man svara denna frågan är en del av problemet av *modellval*.

Vi kan fråga samma frågan även för bayesianska modeller. När vi börjar att modellera en situation med en bayesiansk modell bestämmer vi formen för datafördelningen $f_X(x|\boldsymbol{\theta})$ och formen för aprioriförderlningen $f_\Theta(\theta)$. Då blir den statistiska modellen aposteriorfördelningen $f_\Theta(\theta|\boldsymbol{x})$ och den aposterioriprediktiva fördelningen $f_{Y|X}(y|\boldsymbol{x})$. Från det bayesianska perspektivet blir frågan hur man välja formen för datafördelningen och/eller aprioriförderlningen. Vi börjar med en diskussion kring problemet från det bayesianska perspektivet.

13.1 Aposteriorprediktiv validering

I bayesiansk statistik är vårt största intresse kvaliteten av de hela aposteriorifördelningen (inte bara en viss punktskattning; exempelvis, aposterioriväntevärdet). Det är eftersom vi skulle vilja använda aposteriorifördelningen för att göra förutsägelser. Kvaliteten av en bayesiansk modell (dvs., val för datafördelningen och/eller aprioriförderlningen) kan därmed vara utvärderat baserad på hur väl modellen göra förutsägelser om (nya) data. I idealiska situationer ska vi ha mer data (oberoende från datan som används att konstruera modellen) som vi kan använda för utvärdering. Detta är dock inte alltid möjligt. En bra start är ett verifiera att data som används för att konstruera modellen ser rimligt ut under den aposterioriprediktiva fördelningen. Detta är verkligen en självständighetskontroll: om vi observera en stor skillnad mellan data och data genereras av aposterioriprediktiva fördelningen kan det vara att vi har bestämt att använda en dålig aprioriförderlningen eller datafördelningen. En sådan kontroll kallas för ett *aposterioriprediktiv validering*.

För att konstruera ett aposterioriprediktiv validering behöver vi genererar data från aposterioriprediktiva fördelningen. Det kan vi göra ibland genom att bestämma formen av denna fördelning men inte alltid. Generellt sätt är det oftast lättare att generera stickprov med hjälp av föjlande formeln: Givet data $\boldsymbol{x}$ och $\boldsymbol{y}$ betingat oberoende givet θ är har vi att aposterioriprediktiva fördelningen $f_{Y|X}(y|x)$ uppfyller

$$\begin{aligned} f_{Y|X}(y|x) &= \int f_{Y,\Theta|\boldsymbol{X}}(y,\theta|\boldsymbol{x})d\theta, \\ &= \int \frac{f_{Y,\Theta,\boldsymbol{X}}(y,\theta,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \frac{f_{\Theta,\boldsymbol{X}}(\theta,\boldsymbol{x})}{f_{\Theta,\boldsymbol{X}}(\theta,\boldsymbol{x})} d\theta, \\ &= \int f_{Y|\Theta,\boldsymbol{X}}(y|\theta,\boldsymbol{x}) f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})d\theta, \\ &= \int f_{Y|\Theta}(y|\theta) f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})d\theta. \qquad (Y \perp\!\!\!\perp \boldsymbol{X}|\theta) \end{aligned}$$

Den sista formeln ger oss ett sätt att generera stickprov från aposterioriprediktiva fördelningen: Generera först ett stickprov $\theta_1, \ldots, \theta_n$ från $\Theta|\boldsymbol{X} = \boldsymbol{x}$ och därefter generera ett stickprov $y_1, \ldots, y_k$ från $Y|\Theta = \theta_i$ för alla $i \in [n]$. Då får vi ett stickprov från aposterioriprediktiva fördelningen med storlek nk.

Exempel 13.1. Låt oss betrakta en beta-binomial modell; dvs. $X|\Theta = \theta \sim \text{Bin}(m, \theta)$ och $\Theta \sim \text{Beta}(\alpha, \beta)$. Vi låter $m = 10$, $\alpha = 2$ och $\beta = 3$. Givet data

$$x_1 = 3, x_2 = 5$$

får vi aposterioriförde lningen

$$\Theta|\boldsymbol{x} \sim \text{Beta}\left(2 + (3+5), 3 + (10)(2) - (3+5)\right) = \text{Beta}(10, 15).$$

Vi kan också bestämma en sluten form för aposterioriprediktiva fördelningen:

$$\begin{aligned} f_{Y|X}(y|x) &= \int f_{Y|\Theta}(y|\theta) f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) d\theta, \\ &= \int \binom{10}{y} \theta^y (1-\theta)^{10-y} \frac{\Gamma(10+15)}{\Gamma(10)\Gamma(15)} \theta^{10-1}(1-\theta)^{15-1} d\theta, \\ &= \binom{10}{y} \frac{\Gamma(10+15)}{\Gamma(10)\Gamma(15)} \int \theta^{10+y-1}(1-\theta)^{25-y-1} d\theta, \\ &= \binom{10}{y} \frac{\Gamma(10+15)}{\Gamma(10)\Gamma(15)} \frac{\Gamma(10+y)\Gamma(25-y)}{\Gamma(35)}. \end{aligned}$$

Denna sluten form är inte så hjälpsam om vi skulle vilja använda Python för att generera stickprov från $f_{Y|X}(y|x)$. Det går istället att använda formen i den första raden av ekvationerna. Exempelvis, om vi skulle vilja har ett stickprov med storlek 10 från $f_{Y|X}(y|x)$ kan vi börja med att generera ett stickprov med storlek 2 från $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$; dvs. vi ta ett stickprov θ_1, θ_2 från Beta$(10, 15)$:

$$\theta_1 = 0.227, \theta_2 = 0.389.$$

Då kan vi ta ett stickprov med storlek 5 från $f_{Y|\Theta}(y|\theta_1)$ respektive $f_{Y|\Theta}(y|\theta_2)$. För det första stickprovet får vi

$$2, 3, 2, 1, 4$$

och för det andra får vi

$$2, 4, 4, 1, 3.$$

Eftersom marginalisering ut Θ motsvarar att glömma vilken θ ges oss vilken y får vi ett stickprov

$$2, 3, 2, 1, 4, 2, 4, 4, 1, 3$$

från $f_{Y|X}(y|x)$.

Nu att vi har ett sätt att generera data från aposterioriprediktiva fördelningen kan vi jämföra sådana stickprov $y_1, \ldots, y_n$ med observerade datan $x_1, \ldots, x_n$ och frågor om vi tror att $x_1, \ldots, x_n$ ser ut som data som kunde har kommit från modellen. Detta är det grundliggande idé av ett aposterioriprediktiv validering.

Exempel 13.2. Anta att studenterna i en viss klass ta en muntlig tentamen som är betygsätt som P (motsvarande 1) eller F (motsvarande 0). Det finns 25 studenter is klassen och sekvensen av betyg i ordning som de tog tentamen är

$$1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0.$$

Vi antar att studenternas betyg är oberoende givet θ vilken är sannolikheten att en student får P på tentamen. Vi tar modellen där $X|\Theta = \theta \sim \text{Bin}(25, \theta)$ med aprioriförde lningen $\Theta \sim \text{U}(0, 1)$. Det följer att aposterioriförde lningen är

$$\Theta|\boldsymbol{X} = \boldsymbol{x} \sim \text{Beta}(14, 9)$$

Om vi tittar på datan kan vi se någon "autokorrelation" (dvs. det verkar att det i:te utfallet beror på det (i-1):te utfallet.) Att vara specifik ser vi att utfallet byter värde bara fem gånger i hela sekvensen. Det är en signal att vårt antagande som vi konstruerade modellen på (att studenternas betyg är oberoende) var fel.

För att testa detta antagande kan vi generera flera stickprov av storlek 25 från aposterioriprediktiv fördelningen (som anta oberoende betyg) och ser om resultatande sekvenser har detsamma autokorrelation mönster. Exempelvis, får vi ett sekvens $y_1, \ldots, y_{25}$:

$$1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0.$$

I den simulerade datan byter värdet av utfallet 13 gånger vilken är ju mycket mer än som vi såg i datan $x_1, \ldots, x_{25}$. Detta kan vara en slump. Så genererar vi 1000 sådana sekvenser $y_1, \ldots, y_{25}$ och räkna antalet gånger statistiken $T(\boldsymbol{y})$ (som är antalet gånger utfallet byter värde) är större än eller lika med 5. Vi kan då skatta sannolikheten

$$P(T(\boldsymbol{y}) \geq T(\boldsymbol{x})) \approx \frac{993}{1000} = 0.993.$$

Denna sannolikhet föreslår att modellen representera datan inte så väl.

I Exempel 13.2 förkasta vi modellen baserad på ett aposterioriprediktiv validering. I detta exempel har vi ingen annan modell att väjla att använda men det händer ofta att vi ha ett annat alternativ. Till exempel, om vi har ett modell som bildas på en aprioriförderlning som bestämds med hjälp av apriori kunskap kan vi alltid använder en model med en icke-informativa apriori fördelningen om vi upptäcker att den valda aprioriförderlningen ledar till en dålig modell.

Exempel 13.3. Vår kompis ska singla slant men hen är en fuskare. Vi har apriori kunskap att hen måste har gjort någonting till slanten så att hen får krona oftare än klave. Så vi modellera utfallet som $X \sim \text{Ber}(\theta)$ med aprioriförderlningen $\Theta \sim \text{Beta}(20, 11)$. Hen singlar slant 10 gånger och vi får datan

$$0, 1, 0, 1, 1, 0, 0, 0, 0, 1$$

där 0 motsvarar klave och 1 motsvarar krona. Vi har då aposterioriförderlningen

$$\Theta|\boldsymbol{x} \sim \text{Beta}(20 + 4, 11 + 10 - 4) = \text{Beta}(24, 17).$$

Nu kan vi utföra ett aposterioriprediktiv validering för att bestämma om vi har tagit en bra aprioriförderlning. Om vi har inte kan vi istället använda den icke-informativa aprioriförderlningen $\Theta \sim \text{U}(0, 1)$. Så bestämmet vi mellan två hypoteser:

$$H_0 : \Theta \sim \text{Beta}(20, 11) \qquad \text{och} \qquad H_1 : \Theta \sim \text{U}(0, 1).$$

Precis liksom ett hypotestest ska vi acceptera H_1 om och endast om vi förkasta H_0. Vi behöver liknande en regel som berätta för oss när vi borde förkasta H_0. Sådana regeler är ofta baserad på en testvariabel $T(\boldsymbol{y})$ (som i Exempel 13.2). Idéen är att vi betrakta sannolikheten $P(T(\boldsymbol{y}) \geq T(\boldsymbol{x})|H_0)$. Om värdet för testvariablen är mer extremt för många $\boldsymbol{y}$ dras från den aposterioriprediktiva fördelningen än för $\boldsymbol{x}$ blir sannolikheten stor. Det föreslår att modellen passar inte så bra med observerade datan $\boldsymbol{x}$ och vi kan förkasta H_0. Å andra sidan om sannolikheten är tillräcklig liten kan vi acceptera H_0.

I detta exempel är en bra testvariabel

$$T(\boldsymbol{y}) = \text{antalet 1 i } \boldsymbol{y}$$

och vi kan använder reglen

$$\text{Acceptera } H_0 \text{ om } P(T(\boldsymbol{y}) \geq T(\boldsymbol{x})|H_0) \approx 0.5.$$

Till skillnad från frekventistiska tester som vi såg förut behöver vi inte beräkna sannolikheten i ett exakt form eftersom vi bara simulerar stickprov från aposterioriprediktiva fördelningen. Eftersom vi kan generera jätte många sådana stickprov ganska snabbt kan vi approximera denna sannolikhet och använder approximeringen för att bestämma oss att acceptera eller förkasta H_0.

Exempelvis, kan vi generera 100 stickprov $y_1, \ldots, y_{10}$ från $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ och se att

$$P(T(\boldsymbol{y}) \geq T(\boldsymbol{x})|H_0) \approx \frac{92}{100} = 0.92.$$

Denna sannolikhet är ganska stor och därför föreslår det är datan $\boldsymbol{x}$ är inte data som kunde har genereras från modellen. Därför förkastar vi modellen H_0 och använder $\Theta \sim \text{U}(0, 1)$. Det ser ut som vår kompis fuskar inte trots allt... denna gång.

På samma sätt som i Exempel 13.3 kan vi formulera aposterioriprediktia validering för att bestämma om ett viss val för datafördelningen är bättre än ett annat val. Vi ska se ett exempel av ett sådant validering i övningsuppgifterna.

13.2 Informationskriterium

Nu att vi ha en metod för att bestämma om en viss bayesiansk modell är ett bra val kan vi betrakta metoder för att svara frågan om modellval för frekventistiska modeller. När vi konstruerar en frekventistisk modell börjar vi med ett val av täthetsfunktion (eller sannolikhetsfunktion) och därmed ett val för möjligaparametrar. Om vi har gjort bra med valet för täthetsfunktionen är frågan om modellval då en fråga om vilken värde för parametrarna vi ska använda. Vi har sett ganska många sätt för att bestämma oss sådana värden (exempelvis, ett val för en punktskattning eller ett val för att acceptera en nollhypotes $H_0 : \theta = \theta_0$.) För denna metoder gav vi alltid metoder för att utvärdera olika val så att vi kunde ta den bäste. Till exempel, Cramér-Rao Nedre Gränsen (Sats 5.3 gav

oss en metod för att hitta bra val för väntevärdesriktiga punktskattningar. Bakom alla dessa metoder för modellval var idéen att vi baserar vårt beslut på *informationen* i stickprovet $x_1, \ldots, x_n$ om parametern θ. Vi har vanligtvis tillgång till denna information i bara ett sätt: genom den simultana fördelningnen:

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

När vi tog en punktskattning, exempelvis. ML-skattningen $\hat{\theta}$, använde vi den simultana fördelningen för att extrahera som mycket information som möjligt om parametern θ och ge en skattning för θ. Hur bra en skattning $\hat{\theta}$ var berodde på hur informativ var stickprovet. Vi såg även med styrkefunktionen att vi kan byta stickprovet (exempelvis, dess storlek) för att får ett starkare test för $H_0 : \theta = \theta_0$. På ett sätt kan vi tänka om detta som vi ökade informationen i stickprovet när vi ökade dess storlek. Det ser ut som vår förmåga att göra ett bra val för modellen beror på hur mycket information ligger i stickprovet.

Ett sätt för att mäta informationen om en viss parameter θ i ett stickprov $x_1, \ldots, x_n$ från $f_X(x|\theta)$ är att beräkna *Fisher informationen* $I(\theta)$. Vi såg i Föreläsningen 5 att Fisher informationen definieras som

$$I(\theta) = n\,\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X(x|\theta)\right)^2\right] = -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(x|\theta)\right].$$

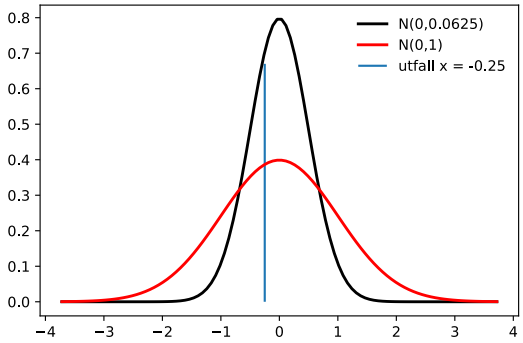
Vi noterar att det andra jämlikheten stämmer bara när vissa villkor uppfylls (som händer nästa alltid i denna kurs):

1. $\frac{\partial}{\partial\theta} f_X(x|\theta)$ existerar nästan överallt,

2. $\frac{\partial}{\partial\theta}\int f_X(x|\theta)dx = \int \frac{\partial}{\partial\theta} f_X(x|\theta)dx$, och

3. stöd av $f_X(x|\theta)$ beror inte å θ.

När dessa villkor uppfylls ser vi från formeln för $I(\theta)$ att Fisher informationen mäter (den genomsnittliga) krökningen för log-likelihoodfunktionen $\log f_X(x|\theta)$ för ett givet θ. Om vi betrakta $I(\theta)$ för den "riktiga" värdet θ får vi information om θ i formen av krökningen av log-likelihoodfunktionen kring θ. Om maximalt för log-likelihoodfunktionen är trubbigt kommer vi behöver mer data för att få en bra skattning för θ; dvs. om ett enda utfall innehåller ganska liten information om θ. Å andra sidan, om maximalt är skärpt ett enda utfall kan innehäller ganska mycket information om θ:



Vi ser i bilden att medan stickprovet är detsamma distans från den riktiga parameter värdet $\theta = 0$ stickprovet $x = -0.25$ innehåller mer "information" om θ i den meningen att det kommer att ta ganska mindre stickprov från $\mathrm{N}(0, 0.0625)$ för att uppskatta $\theta = 0$ med $\hat{\theta} = \bar{x}$ än för uppskatta detsamma parametervärdet med stickprov från $\mathrm{N}(0, 1)$. Detta är en konsekvens av faktumet att $\mathrm{N}(0, 0.0625)$ är mycket skarpare toppad på $\theta = 0$ än $\mathrm{N}(0, 1)$. Vi ser också från bilden hur Fisher informationen ger oss ett mått av variansen i modellen.

Exempel 13.4. Vi skulle vilja beräkna Fisher information $I(\theta)$ i ett stickprov från en Ber(θ)-population. Det är

$$\begin{aligned} I(\theta) &= -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(x|\theta)\right], \\ &= -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\right], \\ &= -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}(X\log(\theta)+(1-X)\log(1-\theta))\right], \\ &= -n\,\mathrm{E}\left[-\frac{X}{\theta^2}-\frac{1-X}{(1-\theta)^2}\right], \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

Vi ser från denna formel att det finns två sätt att öka informationen i stickprovet $x_1,\ldots,x_n$: (1) öka stickprovs storlek eller (2) minska variansen $\mathrm{Var}[X]=\theta(1-\theta)$. Vi kan ofta inte byta det riktiga värdet för θ och därför tar till att öka stickprovs storlek när det är möjligt.

Det är bra att notisera att vi kan beräkna Fisher informationen också för modeller som beror på flera parameter $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_d)$. Den Fisher informationsmatrisen definieras som $I(\boldsymbol{\theta})=[I(\boldsymbol{\theta})_{i,j}]$ där

$$I(\boldsymbol{\theta})_{i,j} = n\,\mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f_X(x|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f_X(x|\boldsymbol{\theta})\right)\right] = -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_X(x|\boldsymbol{\theta})\right].$$

13.2.1 Ömsesidig Information och Kullback-Leibler Divergens

Vi har redan sett att när vi bestämma mellan två modeller $f_X(x|\theta)$ och $f_X(x|\eta)$ där täthetsfunktionerna har samma form och skillnaden mellan modeller är bara ett val för parametervärdet kan Fisher information vara ganska hjälpsam. Exempelvis, en väntevärdesriktiga punktskattning med den minsta variansen har variansen begränsade med en funktion av Fisher informationen!

Vi kan också använder andra *informationkriterium* relaterade till Fisher informationen för att väjla mellan modeller vems täthetsfunktioner har olika former (och olika antal parametrar). Exempelvis, om vi skulle vilja bestämma om två variabler (X,Y) kan modelleras som oberoende; dvs, $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ istället för den mer komplicerad formen $f_{X,Y}(x,y)$, kan vi betrakta *ömsesidg informationen* för X och Y:

$$I(X,Y)=\int_Y\int_X \log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)f_{X,Y}(x,y)dxdy = \mathrm{E}\left[\log\left(\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right)\right].$$

Notisera att $I(X,Y)=0$ om $X \perp\!\!\!\perp Y$. Vi kan tänka om $I(X,Y)$ som ett mått om hur mycket "överraskning" vi får om vi använder en oberoende modell $f_X(x)f_Y(y)$ för att modellera $f_{X,Y}(x,y)$ när $X \not\perp\!\!\!\perp Y$ på riktigt. Det finns ingen överraskning om den riktiga modellen är $f_X(x)f_Y(y)$.

Exempel 13.5. Låt $(X_1,X_2)\sim \mathrm{N}(\mathbf{0},\Sigma)$ där

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$$

där $a\in[-1,1]$. Det föjler att

$$f_{X_1,X_2}(x_1,x_2)=\frac{1}{2\pi\sqrt{\det(\Sigma)}}e^{-\frac{1}{2(1-a^2)(x_1^2-2ax_1x_2+x_2^2)}}.$$

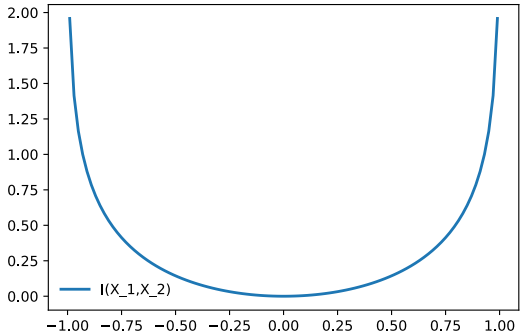
Eftersom X_1 och X_2 är N(0,1)-fördelade har vi också att

$$f_{X_1}(x_1)=\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2} \qquad \text{och} \qquad f_{X_2}(x_2)=\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}.$$

Det följer att

$$\begin{aligned}
I(X,Y) &= \mathrm{E}\left[\log\left(\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right)\right], \\
&= \mathrm{E}\left[\log\left(\frac{1}{\det(\Sigma)}\right) - \frac{1}{2(1-a^2)}\left(X_1^2 - 2aX_1X_2 + X_2^2\right) + \frac{1}{2}(X_1^2 + X_2^2)\right], \\
&= \log\left(\frac{1}{\det(\Sigma)}\right) - \frac{1}{2(1-a^2)}\left(\mathrm{E}[X_1^2] - 2a\,\mathrm{E}[X_1X_2] + \mathrm{E}[X_2^2]\right) + \frac{1}{2}(\mathrm{E}[X_1^2] + \mathrm{E}[X_2^2]), \\
&= \log\left(\frac{1}{\det(\Sigma)}\right) - \frac{1}{2(1-a^2)}\left(\mathrm{Var}[X_1] - 2a\,\mathrm{Cov}[X_1,X_2] + \mathrm{Var}[X_2]\right) + \frac{1}{2}(\mathrm{Var}[X_1] + \mathrm{Var}[X_2]), \\
&= \log\left(\frac{1}{\det(\Sigma)}\right) - \frac{1}{2(1-a^2)}\left(1 - 2a^2 + 1\right) + \frac{1}{2}(2), \\
&= \log\left(\frac{1}{\det(\Sigma)}\right), \\
&= \log\left(\frac{1}{1-a^2}\right).
\end{aligned}$$

Så ömsesidig informationen är en funktion av $a = \mathrm{Cov}[X_1, X_2]$. När a är nära 0 har vi liten överraskning om vi modellera simultan fördelningen $f_{X_1,X_2}(x_1,x_2)$ som $f_{X_1}(x_1)f_{X_2}(x_2)$. Å andra sidan, om $|a| \approx 1$ har vi ganska mycket överraskning. Det kan vi se i följande bild där vi plottar $I(X_1, X_2)$ som en funktion av a:



Med ömsesidig information kan vi kvantifiera hur dålig det skulle vara om vi välja att använda en modell där vi anta oberoende. Vi kan använda samma idéen för att hjälpa oss välja andra typer av modeller förutom oberoende modeller. *Kullback-Leibler Divergens* (eller *relativ entropin* mellan två fördelningar med täthetsfunktioner $f_{\boldsymbol{X}}(\boldsymbol{x})$ respektive $g_{\boldsymbol{X}}(\boldsymbol{x})$ är

$$D(f,g) = \int_{\boldsymbol{X}} \log\left(\frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x})}\right) f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}.$$

Ömsesidig informationen är $D(f,g)$ där $f_{X,Y}(x,y)$ och $g_{X,Y}(x,y) = f_X(x)f_Y(y)$. På samma sätt som ömsesidig information är $D(f,g)$ hur mycket "överraskning" vi får när vi använder $g_X(x)$ när den riktiga modellen är $f_X(x)$.

När vi betraktar $f_X(x|\theta_0)$ och $f_X(x|\theta_1)$ med samma formen men olika värden för parametern ser vi något bekant:

$$D(f_X(x|\theta_0), f_X(x|\theta_1)) = \int_X \log\left(\frac{f_X(x|\theta_0)}{f_X(x|\theta_1)}\right) f_X(x|\theta_0)dx.$$

Vi såg i övningen om hypotestest att ett likelihood-kvottest för

$$H_0 : \theta = \theta_0 \qquad \text{mot} \qquad H_1 : \theta = \theta_1$$

ges av

$$\text{Förkasta } H_0 : \theta = \theta_0 \text{ om } \log\left(\tfrac{f_X(x|\theta_0)}{f_X(x|\theta_1)}\right) \leq c.$$

Om vi felaktigt förkastar H_0 kommer vi använda θ_1 för att modellera fördelningen istället för θ_0. Kullback-Leibler Divergens berätter därför för oss den förväntade överraskningen vi kommer att få när vi använda θ_1 istället för

den "riktiga" parametern θ_0. Man ofta använder Kullback-Leibler Divergens i maskininlärning för modellval där vi försöker att minimera KL-divergens mellan en modell och den empiriska fördelningen som ges av datan. Vi ska dock inte säga mer om det i denna kursen.

Kullback-Leibler divergens (och därmed ömsesidig information) är naturligt kopplade till Fisher information. I synnerhet Fisher informationen ger oss krökningen av Kullback-Leibler divergens:

$$\begin{aligned}\frac{d^2}{d\theta_1^2} D(f_X(x|\theta_0), f_X(x|\theta_1)) &= \frac{d^2}{d\theta_1^2} \int_X \log\left(\frac{f_X(x|\theta_0)}{f_X(x|\theta_1)}\right) f_X(x|\theta_0)dx, \\ &= \frac{d^2}{d\theta_1^2} \int_X \log(f_X(x|\theta_0)) f_X(x|\theta_0)dx - \frac{d^2}{d\theta_1^2} \int_X \log(f_X(x|\theta_1)) f_X(x|\theta_0)dx, \\ &= -\int_X \left(\frac{d^2}{d\theta_1^2} \log(f_X(x|\theta_1))\right) f_X(x|\theta_0)dx.\end{aligned}$$

Det sista ekvationen är ett väntevärdet i formen av Fisher informationen. Så följer det att

$$\left(\frac{d^2}{d\theta_1^2} D(f_X(x|\theta_0), f_X(x|\theta_1))\right)\Big|_{\theta_1=\theta_0} = I(\theta_0).$$

Då ser vi att Fisher information kan vara användbar i modellval, både när vi skulle vilja välja den bäste parametervärdet och när vi skulle vilja verfiera om vi kan använder enklare formen av en simultan fördelningen (exempelvis, oberoende variabler). På liknande sätt kan vi definiera andra informationkriterium som kan hjälpa oss välja mellan olika modeller – särkilt modeller med olika antal parametrar.

13.2.2 Akaike Informationkriterium

En viktig fråga med avseende på modellval är hur många parametrar borde man använda. Vi motiverar frågan med ett exempel

Exempel 13.6. Anta att vi har två simultan fördelade stokastisk variabler X_1 och X_2 som har möjliga utfall $\{0, 1\}$. Mängden av möljiga simultana utfall (x_1, x_2) är därför $\{(0,0), (0,1), (1,0), (1,1)\}$. Anta att vi observerar bara sju utfall i ett stickprov med frekvenserna:

Utfall	Frekvens
(0,0)	1
(0,1)	3
(1,0)	2
(1,1)	1

Vi kan beräkna ML-skattningen för sannolikheter $P(x_1, x_2)$ som

$$M_1 : \qquad P(0,0) = \frac{1}{7}, \qquad P(0,1) = \frac{3}{7}, \qquad P(1,0) = \frac{2}{7}, \qquad \text{och} \qquad P(1,1) = \frac{1}{7}.$$

I denna modell har vi tre parameterar; exempelvis, $P(0,0)$, $P(0,1)$ och $P(1,0)$. Notisera att vi behöver inte alla fyra sannolikheter för att parametriserar modellen eftersom

$$P(0,0) + P(0,1) + P(1,0) + P(1,1) = 1.$$

Givet att vi har så liten data är det möjligt att ML-skattningen är inte så bra. Till exempel är det möjligt att sannolikheter $P(0,0)$ och $P(1,1)$ är inte så liten på riktigt. Det är möjligt att ML-skattningen har "överanpassat" till datan och därmed ges orealistiska skattningarna för $P(0,0)$ och $P(1,1)$. För att undvika problemet av överanpassning kan vi minska antalet parametrar som vi använda i modellen. Exempelvis, om vi modellera fördelningen som $X_1 \perp\!\!\!\perp X_2$ har vi bara två parameter $P(X_1 = 0)$ och $P(X_2 = 0)$. ML-skattningen för denna modell ges av

$$P(X_1 = 0) = \frac{4}{7}, \qquad P(X_1 = 1) = \frac{3}{7}, \qquad P(X_2 = 0) = \frac{3}{7}, \qquad \text{och} \qquad P(X_2 = 1) = \frac{4}{7}.$$

Det följer att den skattande simultan fördelningen under modellen $X_1 \perp\!\!\!\perp X_2$ är

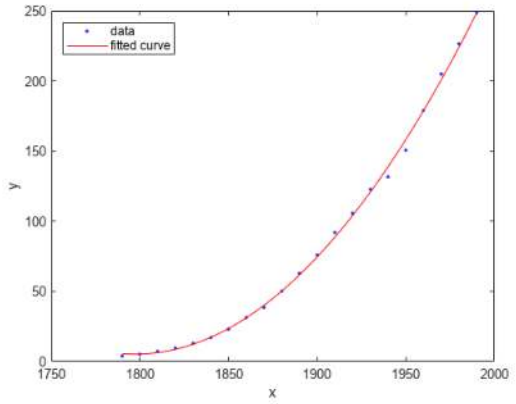
$$M_2: \qquad P(0,0) = \frac{12}{49}, \qquad P(0,1) = \frac{16}{49}, \qquad P(1,0) = \frac{9}{49}, \qquad \text{och} \qquad P(1,1) = \frac{12}{49}.$$

Vi ser att sannolikheter $P(0,0)$ och $P(1,1)$ har ökat under denna modell och därför kanske har undvikit problemet av överanpassning."

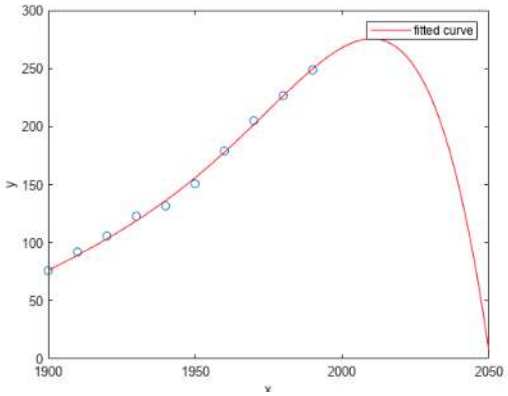
Generellt sett är det bra om modellen maximerer likelihooden men problemet blir att det kan ofta ökar likelihooden att använder modeller med ytterligande parametrar och där har vi risken att överanpassa modellen till datan. Om vi skulle vilja ha en balans mellan en hög likelihood och låg risk av överanpassning kan vi välja en modell som optimerar en funktion som straffar likelihooden för att undvika överanpassning.

När vi försöker att hitta en modell som representerar data-generering processen bra ska modellen vi använder nästa alltid inte vara den riktiga modell. (Det är därför att vi kallas det för en modell!) Om vi använder mer parameterar kan vi få kanske en modell som passar jättebra med datan. Det är även möjligt att det passar så bra att modellen är inte rimligt!

Exempel 13.7. En regressionanalys gjord av Mathworks anpassade en regressionskurva med grad 3 till US census data:



Modellen anpassar bra med datan men vi kan få en bättre anpassning om vi använda en polynom med grad 6:



Problemet är att denna model (som har ytterligare tre parametrar) förutsäger att populationen is USA blir 0 i 2050. Detta är kanske orealistiskt.

Exemplet visar att mer parametrar kan ger dålig modeller för förutsägelser. För att begränsa antalet parametrar modellen använder medan få en modell med en hög likelihood kan vi använda modellen som minimerar *Akaike informationkriterium* (AIC):

$$AIC(\boldsymbol{x}, f_X(x|\theta)) = 2k - 2\log(L(\hat{\theta}|\boldsymbol{x})).$$

Här $f_X(x|\theta)$ kan vara vilken modell som helst (någon val för formen av täthetsfunktionen och parametrar). För att förenkla notationen kan vi också skriver $AIC(\boldsymbol{x}, M)$ där M betecknar en modell med en viss täthetfunktion (eller sannolikhetsfunktion) och en viss antal parametrar. Variablen k är antalet parametrar modellen använder och $\hat{\theta}$ är

ML-skattningen för modellens parametrar. Ett liten värdet för AIC är bäst: En hög likelihood är bra men inte om det kräver för mycket parametrar.

Så länge vi kan hitta ML-skattningen är AIC lätt att beräkna:

Exempel 13.8. Vi beräknar AIC för modellerna M_1 och M_2 från Exempel 13.6. För M_1 har vi att

$$\begin{aligned} AIC(\boldsymbol{x}, M_1) &= 2(3) - \log(L(\hat{\theta}|\boldsymbol{x})), \\ &= 6 - \log\left(\left(\frac{1}{7}\right)^1 \left(\frac{3}{7}\right)^3 \left(\frac{2}{7}\right)^2 \left(\frac{1}{7}\right)^1\right), \\ &= 6 - \log\left(\left(\frac{1}{7}\right)^2 \left(\frac{3}{7}\right)^3 \left(\frac{2}{7}\right)^2\right), \\ &\approx 9.882. \end{aligned}$$

För modellen M_2 har vi att

$$\begin{aligned} AIC(\boldsymbol{x}, M_2) &= 2(2) - \log\left(\left(\frac{4}{7}\right)^4 \left(\frac{3}{7}\right)^3 \left(\frac{3}{7}\right)^3 \left(\frac{4}{7}\right)^4\right), \\ &= 4 - \log\left(\left(\frac{4}{7}\right)^7 \left(\frac{3}{7}\right)^7\right), \\ &\approx 8.277. \end{aligned}$$

Så föredrar AIC modellen där vi antar att $X_1 \perp\!\!\!\perp X_2$.

Akaike Informationkriterium (AIC) används ofta när man tror att modellerna de välja från innehåller inte den riktiga modell. Detta är eftersom straffet för antalet parametrar in modellen är relativt mild. Så är AIC ofta används i situationer där man antar att det finns oobserverade variabler som påverkar de observerade variablerna och därmed skapar korrelationer (kovariationer) in det observerade systemet. Därmed skulle vilja man lägga till några parametrar för att modellera dessa kovariationer men inte så mycket för att skapa överanpassning. Exempelvis, använder man ofta AIC i tidserier och regressions analys.

13.2.3 Bayesiansk Informationkriterium

Å andra sidan, anta att man tror att den riktiga modellen finns i mängden av möjliga modeller som vi betraktar. I detta fall kan vi vara mindre bekymrade över spurious korrelationer i systemet och därför kan vi straffar ytterligande parametrar mer. I sådana fall minimerar vi *Bayesiansk informationkriterium* (BIC):

$$BIC(\boldsymbol{x}, f_X(x|\theta)) = k\log(n) - 2\log(L(\hat{\theta}|\boldsymbol{x})).$$

I princip $L(\hat{\theta}|\boldsymbol{x})$ borde vara högst för den riktiga modellen men vi har redan sett att ytterligande parameterar kan felakigt föreslå en orealistisk modell. I fallet att vi ha bara en liten datamängd för vilken likelihooden är underväldigande är det möjligt att

$$AIC(\boldsymbol{x}, M_{true}) \approx AIC(\boldsymbol{x}, M_{False}).$$

Exempelvis, vi kan ha

$$\log(L(\hat{\theta}_{False}|\boldsymbol{x})) \approx 1 - \log(L(\hat{\theta}_{true}|\boldsymbol{x})).$$

För att undvika detta kan vi använda datamängdens storlek för att skala skillnaden mellan skattningar. Om vi skala straffet med en funktion av n, exempelvis $\log(n)$ kan vi separerar skattningar för olika modeller mer när datamängden är liten.

BIC ger också ett större straff för mer parametrar när datamängden är stor. Medan detta avskräcker överpassning kan det ledar till "underpassning." Så är det ofta en bra idé att beräkna både AIC och BIC i praktiken.

Exempel 13.9. I Exempel 10.2 vi hade simulerade data från en grafiska modell. Liam använde en viss graf att generera datan. Kan BIC berätter för oss vilken graf han använde? Med nivå $\alpha = 0.05$ i vårt t-test för betingat oberoende bestämde vi att det finns fyra olika modeller att betrakta:

Vi kallas modeller M_1, M_2, M_3, respektive M_4 från vänster till höger. Vi såg också i inlämningsuppgiften att vi kan hitta ML-skattningen för sådana modeller genom att beräkna residualkvadratsummarna. Om vi göra så får vi BIC:er

$$\begin{aligned} BIC(\boldsymbol{x}, M_1) &= 304.642, \\ BIC(\boldsymbol{x}, M_2) &= 304.642, \\ BIC(\boldsymbol{x}, M_3) &= 304.642, \\ BIC(\boldsymbol{x}, M_4) &= 311.847. \end{aligned}$$

Så BIC föreslår att den riktiga modellen är modellen M_1, M_2 eller M_3. Notisera att alla tre modeller har det samma BIC. Det är eftersom alla tre modeller är Markovekvivalent! BIC kan inte skilja mellan dem eftersom de kodar detsamma betingade oberoende modell. Liam kan också verifiera att BIC har fått den rätta modellen!