Föreläsning 05

Statistisk inlärning och dataanalys (Kungliga Tekniska Högskolan)

Föreläsning 5

5.1 Maximum-likelihoodmetoden

Medan momentmetoden är snabb och enkel ger den inte alltid de bästa skattningarna. En annan metod är baserad på *likelihoodfunktionen*:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{x}) &= L(\boldsymbol{\theta}|x_1,\ldots,x_n), \\ &= f_{\boldsymbol{X}}(x_1,\ldots,x_n|\boldsymbol{\theta}), \\ &= \prod_{i=1}^{n} f_{X_i}(x_i|\boldsymbol{\theta}), \end{aligned}$$

för stickprovet $x_1,\ldots,x_n$ från en population $f_X(x|\boldsymbol{\theta})$. Likelihoodfunktionen är bara den simultan täthetsfunktionen (eller sannolikhetsfunktionen) men där värdena $x_1,\ldots,x_n$ är fixa och $\boldsymbol{\theta}$ varierar. I diskreta fallet ger likelihoodfunktionen sannolikheten av stickprovet för olika varden av parametrarna. I detta sätt svarar likelihoodfunktionen frågan om hur troligt är datan för varje givet val av $\boldsymbol{\theta}$.

Låt $\hat{\boldsymbol{\theta}}$ vara valet för $\boldsymbol{\theta}$ som maximeras sannolikheten att se datan $\boldsymbol{x}$; det vill säga

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Funktionen $\hat{\boldsymbol{\theta}}$ är en funktion $\hat{\boldsymbol{\theta}} : \mathbb{R}^n \longrightarrow \mathbb{R}^d$ som skickar möjliga observationer till parametervärden. När vi tillämpar funktionen till stickprovet $\boldsymbol{x}$ får vi den *maximum-likelihoodskattning* (ML-skattning) $\hat{\boldsymbol{\theta}}$ av $\boldsymbol{\theta}$.

Maximum-likelihoodskattningen är ofta unik; det vill säga, det finns bara ett val för $\boldsymbol{\theta}$ som maximeras $L(\boldsymbol{\theta}|\boldsymbol{x})$. Detta behöver dock inte fallet. (ML-skattningen behöver egentligen inte existera!) Det vore mer korrekt att kalla $\hat{\boldsymbol{\theta}}$ *en* ML-skattning men det är brukligt att kalla den ML-skattningen. ML-skattningar är ganska lätta att definiera men de kan vara svåra att beräkna i pratiken. Dock eftersom ML-skattningen maximerar en funktion har vi ett generellt recept för att beräkna dem:

1. Beräkna de partiella derivaten $\frac{\partial}{\partial\theta_j} L(\boldsymbol{\theta}|\boldsymbol{x})$ där $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)$. Det är vanligt att beräkna de partiella derivaten av *log-likelihoodfunktionen* $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ istället. Båda metoderna ger det samma svar eftersom log är en strikt växande funktion.

2. Beräkna kritiska punkter av $L(\boldsymbol{\theta}|\boldsymbol{x})$ (eller $\log L(\boldsymbol{\theta}|\boldsymbol{x})$); dvs, lösningarna till systemet av ekvationerna ges igenom sätter den partiella derivaten lika med 0.

3. Bestämma vilka kritiska punkter är (lokala) maxima igenom att beräkna matrisen

$$\left[\frac{\partial^2}{\partial\theta_i\theta_j} L(\boldsymbol{\theta}|\boldsymbol{x})\right]_{i,j=1}^{d}$$

   och bestämma för vilka kritiska punkter matrisen är negativt definit (alla egenvärden negativa). (Detta går också att gör med $\log L(\boldsymbol{\theta}|\boldsymbol{x})$. Ibland kan vi använda algebraisk metoder för att verifiera att en kritisk punkt är en lokal maximum istället.

4. Jämför värdet av $L(\boldsymbol{\theta}|\boldsymbol{x})$ för alla lokal maxima och gränspunkter (inklusive oändlighet) och ta punkten med det största värdet.

Exempel 5.1. Låt $X_1, \ldots, X_n$ vara oberoende och Ber(p)-fördelade. Så har vi likelihood-funktionen

$$L(p|\boldsymbol{x}) = L(p|x_1, \ldots, x_n) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

För att förenkla beräkningen ska vi använda log-likelihoodfunktionen

$$\log L(p|\boldsymbol{x}) = \log(p)\left(\sum_{i=1}^n x_i\right) + \log(1-p)\left(n - \sum_{i=1}^n x_i\right).$$

Vi har att

$$\frac{d}{dp}\log L(p|\boldsymbol{x}) = \frac{1}{p}\left(\sum_{i=1}^n x_i\right) - \frac{1}{1-p}\left(n - \sum_{i=1}^n x_i\right).$$

För att bestämma de kritiska punkterna sätter vi detta lika med 0 och löser för p:

$$\begin{aligned} 0 &= \frac{1}{p}\left(\sum_{i=1}^n x_i\right) - \frac{1}{1-p}\left(n - \sum_{i=1}^n x_i\right), \\ &= \frac{n}{p}\bar{x} - \frac{n}{1-p}(1-\bar{x}), \\ pn(1-\bar{x}) &= n\bar{x}(1-p), \\ p &= \bar{x}. \end{aligned}$$

Det finns bara en kritisk punkt, $p = \bar{x}$. Denna punkt är bara möjlig när $0 < n\bar{x} < n$. Annat skulle en term i ovanstående ekvationen försvinna och lämna oss med en ekvation utan en lösning. Så om vi anta att $0 < n\bar{x} < n$ ser vi att $p = \bar{x}$ är en lokal maximum eftersom

$$\frac{d^2}{dp^2}\log L(p|\boldsymbol{x}) = -\frac{1}{p^2}n\bar{x} - \frac{1}{(1-p)^2}(n - n\bar{x}) < 0$$

där vi använder att $n\bar{x} > 0$ och $n - n\bar{x} > 0$. För $n\bar{x} = 0$ respektiv $n\bar{x} = n$ är log-likelihoodfunktionerna strikt avtagande och strikt växande. Därför får vi ML-skattningarna 0 respektiv 1 vilka är lika med $\bar{x}$ i båda fall. Det följer att ML-skattningen för p är $\hat{p} = \bar{X}$.

Vi kan också använda det samma metoden med modellen som beror på mer än en parameter.

Exempel 5.2. Låt $X_1, \ldots, X_n$ vara oberoende och N(μ, σ^2)-fördelade med båda μ och σ^2 okända. Likelihoodfunktionen är

$$L(\mu, \sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right),$$

och så har vi log-likelihoodfunktionen

$$\log L(\mu, \sigma^2|\boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

Det följer att de partiella derivaten är

$$\begin{aligned} \frac{\partial}{\partial\mu}\log L(\mu, \sigma^2|\boldsymbol{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu), \\ \frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2|\boldsymbol{x}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

(Notisera att den andra derivatan är med avseende på σ^2 (inte σ).) Det följer att

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n}s^2 \end{aligned}$$

är den kritiska punkten. Vi behöver att verifiera att denna punkt är en lokal maximum. Istället för att beräkna andra derivat notiserar vi att för alla μ vi har

$$\sum_{i=1}^{n}(x_i-\mu)^2 > \sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Så har vi att

$$L(\mu,\sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \leq \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = L(\bar{x},\sigma^2|\boldsymbol{x})$$

för alla $\sigma^2 > 0$. Så är det tillräckligt att visa likelihoodfunktionen maximerar på $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ när $\mu = \bar{x}$. Den andra derivatan med avseende på σ^2 är

$$\begin{aligned}\frac{d^2}{d(\sigma^2)^2}\log L(\bar{x},\sigma^2|\boldsymbol{x}) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\bar{x})^2,\\ &= -\frac{2}{\sigma^6}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n\sigma^2}{2}\right).\end{aligned}$$

Eftersom faktoren

$$\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n\sigma^2}{2}\right)$$

är positiv när $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ vet vi att den andra derivatan är negative. Därför har vi en lokal maximum. Vi har också at $L(\bar{x},\sigma^2|\boldsymbol{x}) \longrightarrow 0$ när $\sigma^2 \longrightarrow \infty$. Så har vi en global maximum. Det har redan konstaterats att $\hat{\mu}$ ger en global maximum för μ för vilken σ^2 som helst. Därför är ML-skattningen av (μ,σ^2) lika med $(\hat{\mu},\hat{\sigma^2}) = (\bar{x},\frac{n-1}{n}s^2)$.

Resultatet i Exempel 5.2 generaliserar även till multivariatnormalfördelningar. Matrisen $\hat{\Sigma}$ i följande satsen kallas *stickprovscovariansmatrisen* och den är garanterat att vara positivt definit.

Sats 5.1. *Låt $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ vara oberoende och $\mathrm{N}(\boldsymbol{\mu},\Sigma)$-fördelade. ML-skattningen för $(\boldsymbol{\mu},\Sigma)$ är*

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i = \bar{\boldsymbol{X}}, \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^T.$$

Nu kanske börjar vi märka att det verkar finnas en koppling mellan ML-skattningar och tillräckliga statistikor. Exempelvis, såg vi att $\bar{X}$ respectiv S^2 är tillräckliga statistikor för μ och σ^2 när $X_1,\ldots,X_n$ är $\mathrm{N}(\mu,\sigma^2)$-fördelade och vi såg även att ML-skattningarna för μ respektiv σ^2 är $\bar{X}$ och $\frac{n-1}{n}s^2$. Denna koppling är inte en slump. Med hjälp av Sats 4.4 vet vi att

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x})g(T(\boldsymbol{x})|\boldsymbol{\theta})$$

för några funktioner h och g när $T(\boldsymbol{x})$ är en tillräcklig statistika för $\boldsymbol{\theta}$ där $\boldsymbol{x}$ är ett stickprov från en $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$-population. Det följer att $L(\boldsymbol{\theta}|\boldsymbol{x})$ maximera på $\hat{\boldsymbol{\theta}}$ om och endast om $g(T(\boldsymbol{x})|\boldsymbol{\theta})$ maximera på $\hat{\boldsymbol{\theta}}$. Det vill säga att ML-skattningen kan alltid vara skriven som en funktion av tillräckliga statistikor.

5.1.1 Invarians egenskap hos ML-skattningar

En användbar egenskap av ML-skattningar är att de är invariant under transformationer. In synnerhet är ML-skattningen av $\tau(\boldsymbol{\theta})$ lika med $\tau(\hat{\boldsymbol{\theta}})$ när $\tau:\mathbb{R}^d \longrightarrow \mathbb{R}^d$ är injektiv och $\hat{\boldsymbol{\theta}}$ är ML-skattningen för $\boldsymbol{\theta}$. Om τ är inte injektiv (och därmed inte inverterbar) är likelihoodfunktionen $L(\tau(\boldsymbol{\theta})|\mathbf{x})$ inte väldefinierad. Istället definierar vi *inducerad likelihoodfunktionen*

$$L^*(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\boldsymbol{\eta}} L(\boldsymbol{\theta}|\boldsymbol{x}).$$

På ett sätt säger detta att från alla $\boldsymbol{\theta}$ som kunde ge oss η tar vi den som ger oss den största likelihooden. Valet $\hat{\eta}$ som maximerar inducerad likelihoodfunktionen definieras som ML-skattningen för η. Medan det kan låter svårt att bestämma sådana ML-skattningar går det lättare med följande:

Sats 5.2. *Om $\hat{\boldsymbol{\theta}}$ är ML-skattningen av $\boldsymbol{\theta}$ så har vi att $\tau(\hat{\boldsymbol{\theta}})$ är ML-skattningen för $\tau(\boldsymbol{\theta})$ för vilken funktion $\tau : \mathbb{R}^d \longrightarrow \mathbb{R}^\ell$ som helst.*

Bevis. Låt $\hat{\boldsymbol{\eta}}$ maximera $L^*(\boldsymbol{\eta}|\boldsymbol{x})$. Vi skulle vilja visa att

$$L^*(\hat{\boldsymbol{\eta}}|\boldsymbol{x}) = L^*(\tau(\hat{\boldsymbol{\theta}})|\boldsymbol{x}).$$

Då gäller det att $\tau(\hat{\boldsymbol{\theta}})$ maximera också $L^*(\boldsymbol{\eta}|\boldsymbol{x})$ och därför är en ML-skattning.

Eftersom $\hat{\boldsymbol{\eta}}$ maximerar $L^*(\boldsymbol{\eta}|\boldsymbol{x})$ har vi

$$L^*(\hat{\boldsymbol{\eta}}|\boldsymbol{x}) = \sup_{\boldsymbol{\eta}} L^*(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\eta}} \sup_{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\boldsymbol{\eta}} L(\boldsymbol{\theta}|\boldsymbol{x})$$

som följer direkt från definitionen av $L^*(\boldsymbol{\eta}|\boldsymbol{x})$. Att maximera $L(\boldsymbol{\theta}|\boldsymbol{x})$ först över en delmängd $\{\boldsymbol{\theta} : \tau(\boldsymbol{\theta}) = \boldsymbol{\eta}\}$ och då över alla $\boldsymbol{\eta}$ är ekvivalent till att maximera över alla $\boldsymbol{\theta}$. Så har vi

$$L^*(\hat{\boldsymbol{\eta}}|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x}) = L(\hat{\boldsymbol{\theta}}|\boldsymbol{x})$$

där $\hat{\boldsymbol{\theta}}$ är ML-skattningen för $\boldsymbol{\theta}$. Vi har även att

$$L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\tau(\hat{\boldsymbol{\theta}})} L(\boldsymbol{\theta}|\boldsymbol{x}) = L^*(\tau(\hat{\boldsymbol{\theta}})|\boldsymbol{x}).$$

□

Exempel 5.3. Vi såg i Exempel 5.2 att

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

är ML-skattningen för σ^2 när $X_1, \ldots, X_n$ oberoende och $\mathrm{N}(\mu, \sigma^2)$-fördelade. Eftersom $\sigma = \tau(\sigma^2)$ där $\tau(\theta) = \sqrt{\theta}$ har vi att $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ är ML-skattningen för standardavvikelsen σ.

Exempel 5.4. Låt $\tau(\theta) = \theta(1-\theta)$ och låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{Ber}(p)$-fördelade. Variansen av $X \sim \mathrm{Ber}(p)$ är $p(1-p)$ och vi såg i Exempel 5.1 att $\hat{p} = \bar{X}$ är ML-skattningen för p. Så följer det från Sats 5.2 att $\tau(\bar{X}) = \bar{X}(1-\bar{X})$ är ML-skattningen för variansen av modellen.

Vi avslutar diskussionen om ML-skattningar med en observation om deras relation till momentmetodens punktskattningar. För ett $\mathrm{N}(\mu, \sigma^2)$-fördelad stickprov såg vi att

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{n-1}{n}s^2\right) = (\tilde{\mu}, \tilde{\sigma}^2).$$

Det vill säga att båda punktskattningar är det samma. Det är också snabbt att verifiera att ML-skattningen för p är det samma som momentmetodens punktskattningen när $x_1, \ldots, x_n$ är ett stickprov från en $\mathrm{Ber}(p)$-population. Medan det kan ofta hända dessa är dessa två punktskattningar inte alltid det samma.

Exempel 5.5. Låt $X_1, \ldots, X_n$ vara obereonde och $\mathrm{U}(0, \theta)$-fördelade. Momentmetodens punktskattningen för θ beräknas som

$$\frac{\theta}{2} = \mathrm{E}[X|p] = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$

och därför är $\tilde{\theta} = 2\bar{X}$. Å andra sidan har vi att likelihoodfunktionen för θ är

$$L(\theta|\boldsymbol{X}) = \frac{1}{\theta^n}$$

vilken är en strikt avtagande funktion av θ. Det följer att ML-skattningen för θ är $\hat{\theta} = \max_{i\in[n]}(X_i)$. Generellt sett är $\tilde{\theta} \neq \hat{\theta}$. Exemplevis, om vi har ett stickprov

$$x_1 = 0.28, x_2 = 0.43, x_3 = 0.88, x_4 = 0.48, x_5 = 0.2$$

sa får vi att $\tilde{\theta} = 0.908$ och $\hat{\theta} = 0.88$.

Nu har vi olika metoder för att bestämma en punktskattning för parametrarna som kan ge oss olika skattningar. Frågan blir ”vilken punktskattning borde vi använda?”

5.2 Utvärdering av skattningar

För att bestämma oss vilken skattning är bättre behöver vi metoder för att utvärdera skattningar. Utvärdering av skattningar använder vanligtvis funktioner som kallas förlustfunktioner. Studiet av utförande och optimalitet av skattningar som utvärderas igenom förlustfunktioner kallas *beslutsteori*. Given data $x_1, \ldots, x_n$ från en $f_X(x|\boldsymbol{\theta})$-population (där $\boldsymbol{\theta}$ är okänt) kan vi fatta ett val (eller *beslut*) om $\boldsymbol{\theta}$. Exempelvis, vi fattar beslutet att $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. *Handlingsrummet* $\mathcal{A}$ är mängden av alla tillåtna beslut. Handlingsrummet är vanligtvis lika med parameterrummet; dvs, $\mathcal{A} = \Omega$ men det behöver inte vara så.

En *förlustfunktion* är en funktion som kvantifiera vårt beslut $\boldsymbol{a} \in \mathcal{A}$ relativt till $\boldsymbol{\theta}$. Om $\boldsymbol{a}$ är långt borta från $\boldsymbol{\theta}$ så ska förlusten (dvs, värdet av förlustfunktionen) vara stor. Samtidigt om $\boldsymbol{a}$ är nära till $\boldsymbol{\theta}$ så ska förlusten vara liten. I detta sätt ger förlustfunktioner oss ett mått på hur mycket vi förlorar när vi skatta $\boldsymbol{\theta}$ med $\boldsymbol{a}$.

Några vanliga förlustfunktioner inkluderar

- *absolutfel*: $L(\theta, a) = |a - \theta|$,
- *kvadratfel*: $L(\theta, a) = (a - \theta)^2$.

Det går att generalisera dessa till flera parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ och beslut $\boldsymbol{a} = (a_1, \ldots, a_d)$ som

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_{i=1}^{n} |a_i - \theta_i| \qquad \text{respektiv} \qquad L(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_{i=1}^{n} (a_i - \theta_i)^2.$$

Eftersom $L(\boldsymbol{\theta}, \boldsymbol{a}) = 0$ om och endast om $\boldsymbol{a} = \boldsymbol{\theta}$ är vår förlust 0 om och endast om vårt beslut är korrekt. Olika val av förlustfunktioner kan ge oss naturligtvis olika förlust för det samma beslut. Exempelvis, absolutfel ger en större förlust för liten fel relativt till kvadratfel. Å andra siden ger kvadratfel oss en större förlust för större fel relativt till absolutfel. Vi kan även definiera förlustfunktioner som ger oss mindre fel när vi överskattar θ än när vi underskattar θ:

$$L(\theta, a) = \begin{cases} (a - \theta)^2 & a < \theta, \\ 10(a - \theta)^2 & a \geq \theta. \end{cases}$$

Anta att vi beräknar en skattning $W(\boldsymbol{X})$ för parametrarna $\boldsymbol{\theta}$ från ett stickprov $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ dras från en $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$-population. Given en förlustfunktion kvantifieras kvaliteten av en skattning $W(\boldsymbol{X})$ med hjälp av en *riskfunktion*:

$$R(\boldsymbol{\theta}, W(\boldsymbol{X})) = \mathrm{E}[L(\boldsymbol{\theta}, \boldsymbol{a})].$$

Noterar att väntevärdet tas mot avseende på $\boldsymbol{X}$; dvs,

$$R(\boldsymbol{\theta}, W(\boldsymbol{X})) = \int_{\boldsymbol{X}} L(\boldsymbol{\theta}, \boldsymbol{a}) f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}.$$

Värdet $R(\boldsymbol{\theta}, W(\boldsymbol{X}))$ för en given $\boldsymbol{\theta}$ är den genomsnittliga förlusten som vi ska ha när vi använder $W(\boldsymbol{X})$ som en punktskattning för $\boldsymbol{\theta}$. Eftersom $\boldsymbol{\theta}$ är okänt stämmer det att använda en punktskattning $W(\boldsymbol{X})$ för vilken $R(\boldsymbol{\theta}, W(\boldsymbol{X}))$ är liten för alla $\boldsymbol{\theta}$. Ett sätt för att jämföra olika punktskattningar är att jämföra deras risk. Exempelvis, om $W_1(\boldsymbol{X})$ och $W_2(\boldsymbol{X})$ är olika punktskattningar för $\boldsymbol{\theta}$ där

$$R(\boldsymbol{\theta}, W_1(\boldsymbol{X})) < R(\boldsymbol{\theta}, W_2(\boldsymbol{X})), \qquad \text{för alla } \boldsymbol{\theta} \in \Omega$$

så är $W_1(\boldsymbol{X})$ den föredragna punktskattningen.

Generellt sett kan det vara svårt att beräkna risken för en given punktskattning och förlustfunktion men för kvadratfel har vi en formel som görs det lättare:

$$\mathrm{E}[(W(\boldsymbol{X}) - \theta)^2] = \mathrm{Var}[W(\boldsymbol{X})] + (\mathrm{E}[W(\boldsymbol{X})] - \theta)^2.$$

Formellen stämmer eftersom

$$\begin{aligned}
\mathrm{E}[(W(\boldsymbol{X}) - \theta)^2] &= \mathrm{E}[(W(\boldsymbol{X}) - \mathrm{E}[W(\boldsymbol{X})] + \mathrm{E}[W(\boldsymbol{X})] - \theta)^2], \\
&= \mathrm{E}[(W(\boldsymbol{X}) - \mathrm{E}[W(\boldsymbol{X})])^2] + \mathrm{E}[(W(\boldsymbol{X}) - \mathrm{E}[W(\boldsymbol{X})])(\mathrm{E}[W(\boldsymbol{X})] - \theta)] + \mathrm{E}[(\mathrm{E}[W(\boldsymbol{X})] - \boldsymbol{\theta})^2], \\
&= \mathrm{Var}[(W(\boldsymbol{X})] + (\mathrm{E}[W(\boldsymbol{X})] - \mathrm{E}[W(\boldsymbol{X})])(\mathrm{E}[W(\boldsymbol{X})] - \theta) + (\mathrm{E}[W(\boldsymbol{X})] - \theta)^2, \\
&= \mathrm{Var}[(W(\boldsymbol{X})] + (0)(\mathrm{E}[W(\boldsymbol{X})] - \theta) + (\mathrm{E}[W(\boldsymbol{X})] - \theta)^2, \\
&= \mathrm{E}[(W(\boldsymbol{X}) - \theta)^2] = \mathrm{Var}[W(\boldsymbol{X})] + (\mathrm{E}[W(\boldsymbol{X})] - \theta)^2.
\end{aligned}$$

Risken $\mathrm{E}[(W(\boldsymbol{X}) - \theta)^2]$ kallas vanligtvis *medelkvadratfel*. Med hjälp av ovanstående formel kan vi jämföra våra två punktskattningar för parametern θ i Exempel 5.5:

Exempel 5.6. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{U}(0, \theta)$-fördelade. Vi bestämde att $\tilde{\theta} = 2\bar{X}$ respektiv $\hat{\theta} = \max_{i \in [n]}(X_i)$ var den momentmetodens punktskattningen och ML-skattningen i Exempel 5.5. Medelkvadratfelet för $\tilde{\theta}$ beräknas som

$$\begin{aligned}
\mathrm{E}[(2\bar{X} - \theta)^2] &= \mathrm{Var}[2\bar{X}] + (\mathrm{E}[2\bar{X}] - \theta)^2, \\
&= \frac{4}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i] + \left(\frac{2}{n} \sum_{i=1}^{n} \mathrm{E}[X_i] - \theta \right)^2, \qquad (X_1, \ldots, X_n \text{ oberoende}) \\
&= \frac{4}{n^2} \frac{n\theta^2}{12} + \left(\frac{2}{n} \frac{n\theta}{2} - \theta \right)^2, \\
&= \frac{\theta^2}{3n}.
\end{aligned}$$

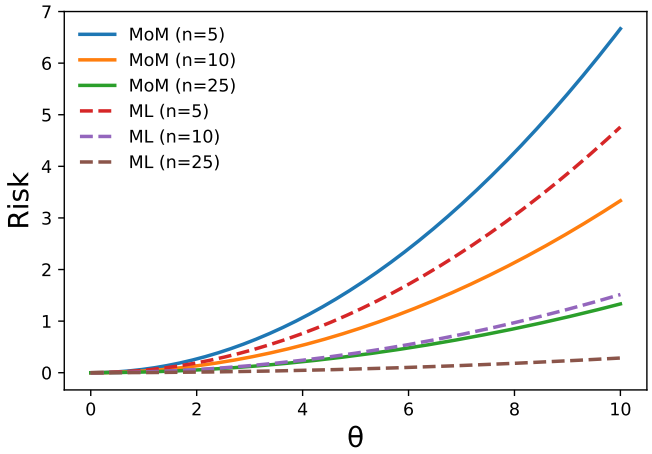
För att beräkna medekvadratfelet för $\hat{\theta}$ behöver vi $\mathrm{Var}[\hat{\theta}]$ och $\mathrm{E}[\hat{\theta}]$. Det kan visas att samplingfördelningen för $\hat{\theta}$ är

$$f_{\hat{\theta}}(w) = \frac{1}{\theta^n} n w^{n-1}.$$

Med hjälp av denna formel kan vi beräkna $\mathrm{E}[\hat{\theta}] = \frac{n}{n+1}\theta$ och $\mathrm{E}[\hat{\theta}] = \frac{n}{n+2}\theta^2$. Det följer att

$$\begin{aligned}
\mathrm{E}[(\hat{\theta} - \theta)^2] &= \mathrm{Var}[\hat{\theta}] + (\mathrm{E}[\hat{\theta}] - \theta)^2, \\
&= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1} \right)^2 + \left(\frac{\theta}{n+1} \right)^2, \\
&= \frac{2}{(n+2)(n+1)} \theta^2.
\end{aligned}$$

Eftersom båda $\frac{1}{3n}$ och $\frac{2}{(n+2)(n+1)}$ går till 0 som $n \longrightarrow \infty$ blir risken för båda punktskattningar mindre för större stickprovor. Risken för ML-skattningen blir dock mindre snabbare än momentmetodens punktskattningen. Det visas i följande



Så ser ML-skattningen ut som en bättre punktskattning för θ.

Å andra sidan såg vi i Exempel 5.6 att $\mathrm{E}[\tilde{\theta}] = \theta$ medan $\mathrm{E}[\hat{\theta}] < \theta$ för alla n. Så i genomsnitt är momentmetodens punktskattningen mer exakt än ML-skattningen. För att formaliserar denna idé definiera vi *systematiskt felet* av en punktskattning $W(\boldsymbol{X})$ för $\boldsymbol{\theta}$ som

$$\mathrm{Bias}[W(\boldsymbol{X})] = \mathrm{E}[W(\boldsymbol{X})] - \theta.$$

Så har vi att

$$\mathrm{E}[(W(\boldsymbol{X}) - \theta)^2] = \mathrm{Var}[W(\boldsymbol{X})] + \mathrm{Bias}[W(\boldsymbol{X})]^2.$$

Generellt sätt mäter systematiskt fel *riktighet* (hur långt borta är punktskattningen från målet) medan variansen mäter *precision* (punktskattningens konsekvens). Vi vill vänligtvis ha båda små riktighet och precision eftersom det

menar att risken är låg. Ibland accepterar vi en liten systematiskt fel om det låter oss reducera variansen (eller vice versa). Det gjorde vi, exempelvis, i Exempel 5.6.

Vi har också sett i Exempel 5.6 att det kan ta mycket arbete för att beräkna båda variansen och systematiskt felet även för enkla modeller. Ett sätt för att bestämma bra punktskattningar i allmänhet är att kräver att systematiskt felet är 0; dvs, $\mathrm{E}[W(\boldsymbol{X})] = \theta$. I sådant fall kallas $W(\boldsymbol{X})$ en *väntevärdesriktig skattning*.

Exempel 5.7. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{N}(\mu, \sigma^2)$-fördelade. Vi såg i Sats 4.1 att $\bar{X}$ respektiv S^2 är väntevärdesriktiga skattningar för μ och σ^2 eftersom

$$\mathrm{E}[\bar{X}] = \mu \qquad \text{och} \qquad \mathrm{E}[S^2] = \sigma^2.$$

Vi kan också beräkna deras varianser som

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n} \qquad \text{och} \qquad \mathrm{Var}[S^2] = \frac{2\sigma^4}{n-1}.$$

Eftersom dessa skattningar är väntevärdesriktiga följer det att deras varianser är like med deras medelkvadratfel.

Om vi tar istället ML-skattningen $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ för σ^2 har vi systematisktfelet

$$\mathrm{Bias}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{\sigma^2}{n}$$

och varians

$$\mathrm{Var}[\hat{\sigma}^2] = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}[S^2] = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)}{n^2}\sigma^4.$$

Så är medelkvadratfelet

$$\mathrm{E}[(\hat{\sigma}^2 - \sigma^2)^2] = \frac{2(n-1)}{n^2}\sigma^4 + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4 = \mathrm{E}[(S^2 - \sigma^2)^2].$$

Så ML-skattningen har en mindre medelkvadratfel än S^2 trots att ML-skattningen är inte väntevärdesriktig och S^2 är väntevärdesriktig.

I våra exemplar hittills har ML-skattningen varit en bättre skattning än alternativen i den meningen att den har haft mindre medelkvadratfel. Det betyder inte nödvändigtvis att man ska alltid använda ML-skattningen. Exemplvis, för stora stickprovor finns det inte så mycket skillnad mellan skattningarna. Så kan det vara lättare att använda, exempelvis, momentmetodens skattningen om det är lättare att beräkna. Vi också vet att ML-skattningen (i dessa exemplar) är inte väntevärdesriktig och underskattar parametern. Detta kan vara ett problem i vissa tillämpningar. Samtidigt vi har bara jämfört skattningarna igenom medelkvadratfelet som straffar båda underskattning och överskattning det samma. Detta stämmer nödvändigtvis inte för en skalaparameter som σ^2 som är alltid positivt och kan potentiellt vara mer känsligt till underskattning. Sist kan vi se att medelkvadratfelet är en funktion θ som kan variera över Ω. Det är möjligt att det finns en skattning som fungerar bättre för vissa värdena av θ än ML-skattningen men inte för andra.

5.3 Cramér-Rao olikheten

Nu fokuserar vi på väntevärdesriktiga skattningar $W(\boldsymbol{X})$. För sådana skattningar minimerar vi medelkvadratfel precis när vi minimerar variansen $\mathrm{Var}[W(\boldsymbol{X})]$. Det finns en kraftfull sats som hjälper oss att hitta väntevärdesriktiga skattningar med den minsta möjliga variansen. Satsen ger oss en nedre gräns om variansen för sådana skattningar. Om vi når denna gräns vet vi att vi har en skattning med det minsta möjliga medelkvadratfelet.

Sats 5.3. *Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade med sannolikhetsfunktion (eller täthetsfunktion) $f_X(x|\theta)$. Låt $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ vara en väntevärdesriktig skattning av $g(\theta)$ och anta att $f_X(x|\theta)$ tillhör en exponentialfamilj. Så gäller det att*

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{g'(\theta)^2}{I(\theta)},$$

där

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

Antalet $I(\theta)$ kallas *informationstalet* eller *Fisher informationen* och värdet $\frac{1}{I(\theta)}$ kallas den *Cramér-Rao under gränsen*.

Exempel 5.8. För $X_1, \ldots, X_n$ oberoende och Po(λ)-fördelade vet vi att

$$\mathrm{E}[\bar{X}] = \lambda$$

eftersom en Po(λ)-fördelad stokastisk variabel har väntevärdet lika med λ. (Här använder vi Sats 4.1.) Sannolikhetsfunktionen för en Po(λ)-fördelad stokastisk variabel är

$$f_X(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Så har vi att Fisher informationen för λ är

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

Med hjälp av Sats 4.1 har vi att

$$\mathrm{Var}[\bar{X}] = \frac{\mathrm{Var}[X]}{n} = \frac{\lambda}{n} = \frac{1}{I(\lambda)}.$$

Det följer att $\bar{X}$ har den minsta möjliga variansen (och därmed kvadratfelet) över alla väntevärdesriktiga skattningar för λ.

Exempel 5.9. Låt $X_1, \ldots, X_n$ vara oberoende och N(μ, σ^2)-fördelade. Fisher informationen för σ^2 är

$$\begin{aligned}
I(\sigma^2) &= -n\,\mathrm{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2} f_X(X|\mu,\sigma^2)\right], \\
&= -n\,\mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right], \\
&= -n\left(\frac{1}{s\sigma^4} - \frac{\mathrm{E}[(X-\mu)^2]}{\sigma^6}\right), \\
&= -n\left(\frac{1}{s\sigma^4} - \frac{\mathrm{Var}[X]}{\sigma^6}\right), \\
&= -n\left(\frac{1}{s\sigma^4} - \frac{1}{\sigma^4}\right), \\
&= \frac{n}{2\sigma^4}.
\end{aligned}$$

Enligt Sats 5.3 har en väntevärdesriktig skattning $W(\boldsymbol{X})$ för σ^2

$$\mathrm{Var}[W(\boldsymbol{X})] \geq \frac{2\sigma^4}{n}.$$

Sats 4.1 visar att S^2 är en väntevärdesriktig skattning för σ^2 men vi såg i Exempel 5.7 att

$$\mathrm{Var}[S^2] = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}.$$

Därför uppnås inte S^2 den Cramér-Rao under gränsen.

I situationer när vi har inte hittas en skattning som uppnår den Cramér-Rao under gränsen är det oklart om det finns en skattning som gör det. Följande satsen ger oss ett sätt för att svara denna fråga.

Sats 5.4. *Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade med sannolikhetsfunktion (eller täthetsfunktion) $f_X(x|\theta)$ som tillhör en exponentialfamilj. En väntevärdesriktig skattning $W(\boldsymbol{X})$ för $g(\theta)$ uppnår den Cramér-Rao under gränsen om och endast om*

$$\frac{\partial}{\partial\theta}\log f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = a(\theta)(W(\boldsymbol{X}) - g(\theta))$$

för någon funktion $a(\theta)$ där $f_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ är simultan sannolikhetsfunktionen (eller täthetsfunktionen) för $X_1, \ldots, X_n$.

Exempel 5.10. För normalfamilijen har vi simultan täthetsfunktionen

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Det följer att

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f_{\boldsymbol{X}}(\boldsymbol{x}|\mu,\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,\\ &= \frac{n}{2\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \sigma^2\right).\end{aligned}$$

När vi tar $a(\sigma^2) = \frac{n}{2\sigma^4}$ får vi

$$W(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Det följer att en skattning som uppnår den Cramér-Rao under gränsen kräver att vi vet μ. Utan denna information kan vi inte uppnår gränsen.

I sådana fall som Exempel 5.10 säger vi att gränsen är *ouppnåeligt*.