Lecture Stat 461-561
Maximum Likelihood Estimation

A.D.

January 2008

- Maximum Likelihood Estimation
- Invariance
- Consistency
- Efficiency
- Nuisance Parameters

Parametric Inference

- Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, ..., X_n)$ parametrized by $\theta \in \Theta$. Then given that $\mathbf{X} = \mathbf{x}$ is observed, the function $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ is the likelihood function.
- The most common estimate is the Maximum Likelihood Estimate (MLE) given by

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, L(\theta|\mathbf{x}).$$

- *Example*: Gaussian distribution

$$f\left(x_i|\,\theta\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right).$$

Then we have with $\theta=\left(\mu,\sigma^2\right)$

$$\begin{aligned}\log L\left(\theta|\,\mathbf{x}\right) &= \sum_{i=1}^{n}\log f\left(x_i|\,\theta\right)\\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2.\end{aligned}$$

- By taking the derivatives and setting them to zero

$$\begin{aligned}\frac{\partial\log L\left(\theta|\,\mathbf{x}\right)}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)=0,\\ \frac{\partial\log L\left(\theta|\,\mathbf{x}\right)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2=0.\end{aligned}$$

- By solving these equations, we obtain

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \widehat{\mu} \right)^2.$$

- Note that $\widehat{\mu}$ is an unbiased estimate but $\widehat{\sigma^2}$ is biased.

- *Example*: Laplace Distribution (Double Exponential)

$$f\left(x_i|\,\theta\right)=\frac{1}{2}\exp\left(-\left|x_i-\theta\right|\right).$$

Then we have

$$\log L\left(\theta|\,\mathbf{x}\right)=-n\log 2-\sum_{i=1}^{n}\left|x_i-\theta\right|.$$

- By taking the derivative, we obtain

$$\frac{d\log L\left(\theta|\,\mathbf{x}\right)}{d\theta}=\sum_{i=1}^{n}\operatorname{sgn}\left(x_i-\theta\right)$$

hence

$$\widehat{\theta}=\operatorname{med}\left\{x_1,...,x_n\right\}$$

for $n=2p+1$.

- *Example* (Uniform Distribution): Consider $X_i \sim \mathcal{U}(0, \theta)$, i.e.

$$f(x_i|\theta) = \begin{cases} 1/\theta & \text{if } 0 \leq x < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

- We have

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) = \begin{cases} (1/\theta)^n & \text{if } \theta \geq x_{(n)} \\ 0 & \text{if } \theta < x_{(n)} \end{cases}.$$

- It follows that $\widehat{\theta} = x_{(n)}$ where $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

- *Example* (Linear Regression): Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be a set of n data where $\mathbf{x}_i = \left(x_1^i, x_2^i, ..., x_p^i\right)^\mathsf{T}$ is a set of explanatory variables and $y_i \in \mathbb{R}$ is the response. We assume

$$y_i = \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

thus

$$f\left(y_i | \mathbf{x}_i, \beta\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

- We have for $\theta = \left(\boldsymbol{\beta}, \sigma^2\right)$

$$\begin{aligned}
\log L(\theta) &= \sum_{i=1}^n \log f\left(y_i | \mathbf{x}_i, \boldsymbol{\beta}\right) \\
&= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right)^2 \\
&= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\mathsf{T}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)
\end{aligned}$$

where $\mathbf{y} = \left(y_1, ..., y_n\right)^\mathsf{T}$ and $\mathbf{X} = \left(\mathbf{x}_1, ..., \mathbf{x}_n\right)^\mathsf{T}$.

- By taking the derivatives and setting them to zero

$$\begin{aligned}
\frac{\partial \log L\left(\theta|\mathbf{x}\right)}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^{\mathsf{T}}\beta + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right) = 0, \\
\frac{\partial \log L\left(\theta|\mathbf{x}\right)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) = 0.
\end{aligned}$$

- Thus we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}, \\
\widehat{\sigma^2} &= \frac{1}{n}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right).
\end{aligned}$$

- *Example* (Time Series): Consider the following autoregression

$$X_0 = x_0,\ X_n = \alpha X_{n-1} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

  where $\theta = \left(\alpha, \sigma^2\right)$.
- We have

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \alpha x_{i-1})^2}{2\sigma^2}\right)$$

- Thus we have

$$\log L(\theta|\mathbf{x}) = cst - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n} \frac{(x_i - \alpha x_{i-1})^2}{2\sigma^2}$$

- It follows that

$$\begin{aligned}\frac{\partial 2\log L\left(\theta|\,\mathbf{x}\right)}{\partial\sigma^2} &= -\frac{n}{\sigma^2}+\sum_{i=1}^{n}\frac{\left(x_i-\alpha x_{i-1}\right)^2}{\sigma^4},\\ \frac{\partial 2\log L\left(\theta|\,\mathbf{x}\right)}{\partial\alpha} &= \frac{2}{\sigma^2}\sum_{i=1}^{n}x_{i-1}\left(x_i-\alpha x_{i-1}\right)^2.\end{aligned}$$

- Thus we have

$$\widehat{\alpha}=\frac{\sum_{i=1}^{n}x_{i-1}x_i}{\sum_{i=1}^{n}x_{i-1}^2},\ \widehat{\sigma^2}=\frac{\sum_{i=1}^{n}\left(x_i-\widehat{\alpha}x_{i-1}\right)^2}{n}.$$

Invariance

Consider $\eta = g(\theta)$. We introduce the induced likelihood function L^*

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

- **Invariance property**: If $\widehat{\theta}$ is the MLE of θ then for any function $\eta = g(\theta)$ then $g\left(\widehat{\theta}\right)$ is the MLE of η.

"*Proof*": The MLE of η is defined by

$$\widehat{\eta} = \arg\sup_{\eta} \sup_{\{\theta:g(\theta)=\eta\}} L(\theta|\mathbf{x})$$

Define

$$g^{-1}(\eta) = \{\theta : g(\theta) = \eta\}.$$

Then clearly $\widehat{\theta} \in g^{-1}(\widehat{\eta})$ and cannot be in any other preimage so $\widehat{\eta} = g\left(\widehat{\theta}\right)$.

Consistency

Definition. A sequence of estimators $\widehat{\theta}_n = \widehat{\theta}_n(X_1, ..., X_n)$ is consistent for the parameter θ if, for every $\varepsilon > 0$ and every $\theta \in \Theta$

$$\lim_{n\to\infty} P_\theta\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) = 1 \text{ (equivalently } \lim_{n\to\infty} P_\theta\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0).$$

- *Example*: Consider $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ and $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ then $\widehat{\theta}_n \sim \mathcal{N}(\theta, 1/n)$ and

$$P_\theta\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) = \int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \to 1.$$

- It is possible to avoid this calculations and use instead Chebychev's inequality

$$P_\theta\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = P_\theta\left(\left|\widehat{\theta}_n - \theta\right|^2 \geq \varepsilon^2\right) \leq \frac{\mathbb{E}_\theta\left(\left(\widehat{\theta}_n - \theta\right)^2\right)}{\varepsilon^2}$$

where $\mathbb{E}_\theta\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = var_\theta\left(\widehat{\theta}_n\right) + \left(\mathbb{E}_\theta\left(\widehat{\theta}_n - \theta\right)\right)^2$.

- Example of *inconsistent* MLE (Fisher)

$$(X_i, Y_i) \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right).$$

- The likelihood function is given by

$$L(\theta) = \frac{1}{(2\pi\sigma^2)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right]\right)$$

- We obtain

$$\begin{aligned} l(\theta) &= cste - n\log\sigma^2 \\ &\quad -\frac{1}{2\sigma^2}\left[2\sum_{i=1}^{n}\left(\frac{x_i+y_i}{2}-\mu_i\right)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2\right]. \end{aligned}$$

- We have

$$\widehat{\mu}_i = \frac{x_i+y_i}{2},\ \widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-y_i)^2}{4n} \to \frac{\sigma^2}{2}.$$

Consistency of the MLE

- Kullback-Leibler Distance: For any density f, g

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

- We have

$$D(f,g) \geq 0 \text{ and } D(f,f) = 0.$$

- Indeed

$$-D(f,g) = \int f(x) \log\left(\frac{g(x)}{f(x)}\right) dx \leq \int f(x) \left(\frac{g(x)}{f(x)} - 1\right) dx = 0$$

- $D(f,g)$ is a very useful 'distance' and appears in many different contexts.

Alternative measures of similarity

- Hellinger distance

$$D(f,g) = \int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx$$

- Generalized information

$$D(f,g) = \frac{1}{\lambda} \int \left(\left(\frac{f(x)}{g(x)} \right)^{\lambda} - 1 \right) f(x)\, dx$$

- L1-norm / Total variation

$$D(f,g) = \int |f(x) - g(x)|\, dx$$

- L2-norm / Total variation

$$D(f,g) = \int (f(x) - g(x))^2\, dx$$

Example

- Suppose we have $f(x) = \mathcal{N}\left(x; \xi, \tau^2\right)$ and $g(x) = \mathcal{N}\left(x; \mu, \sigma^2\right)$.
- We have

$$\begin{aligned}\mathbb{E}_f\left[(X-\mu)^2\right] &= \mathbb{E}_f\left[(X-\xi)^2 + 2(X-\xi)(\xi-\mu) + (\xi-\mu)^2\right] \\ &= \tau^2 + (\xi-\mu)^2\end{aligned}$$

- So it follows that

$$\begin{aligned}\mathbb{E}_f\left[\log g(X)\right] &= \mathbb{E}_f\left[-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\tau^2 + (\xi-\mu)^2}{2\sigma^2}\end{aligned}$$

and

$$\mathbb{E}_f\left[\log f(X)\right] = -\frac{1}{2}\log\left(2\pi\tau^2\right) - \frac{1}{2}.$$

- It follows that

$$\begin{aligned} D(f,g) &= \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \\ &= \mathbb{E}_f\left[\log f(X)\right] - \mathbb{E}_f\left[\log g(X)\right] \\ &= \frac{1}{2}\left\{\log\left(\frac{\sigma^2}{\tau^2}\right) + \frac{\tau^2 + (\xi-\mu)^2}{\sigma^2} - 1\right\} \end{aligned}$$

- It can be easily checked that $D(f,f) = 0$ (less easy to show $D(f,g) \geq 0$).

Example

- Assume we have $f(x) = \frac{1}{2}\exp(-|x|)$ and $g(x) = \mathcal{N}(x; \mu, \sigma^2)$.
- We obtain

$$\begin{aligned}
\mathbb{E}_f[\log f(X)] &= -\log 2 - \frac{1}{2}\int_{-\infty}^{\infty} |x| \exp(-|x|)\, dx \\
&= -\log 2 - \int_0^{\infty} x \exp(-|x|)\, dx \\
&= -\log 2 - 1,
\end{aligned}$$

$$\mathbb{E}_f[\log g(X)] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}(4 + 2\mu^2)$$

- It follows that

$$D(f, g) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(2 + \mu^2) - \log 2 - 1.$$

- Assume the pdfs $f(x|\theta)$ have common support for all θ and $f(x|\theta) \neq f(x|\theta')$ for $\theta \neq \theta'$; i.e. $S_\theta = \{x : f(x|\theta) > 0\}$ is independent of θ.
- Denote

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i|\theta)}{f(X_i|\theta_*)}$$

- As the MLE $\widehat{\theta}_n$ maximises $L(\theta|\mathbf{x})$, it also maximizes $M_n(\theta)$.
- Assume $X_i \overset{\text{i.i.d.}}{\sim} f(x|\theta_*)$. Note that by the law of large numbers $M_n(\theta)$ converges to

$$\begin{aligned}\mathbb{E}_{\theta_*}\left(\log \frac{f(X|\theta)}{f(X|\theta_*)}\right) &= \int f(x|\theta_*) \log \frac{f(x|\theta)}{f(x|\theta_*)} dx \\ &= -D(f(\cdot|\theta_*), f(\cdot|\theta)) := M(\theta).\end{aligned}$$

- Hence, $M_n(\theta) \approx -D(f(\cdot|\theta_*), f(\cdot|\theta))$ which is maximized for θ^* so we expect that its maximizer will converge towards θ_*.

Example

- Assume $f(x) = g(x) = \mathcal{N}(x; 0, 1)$.
- We approximate

$$\mathbb{E}_f[\log g(X)] = -\frac{1}{2}\log(2\pi) - \frac{1}{2} = -1.4189$$

through

$$\mathbb{E}_{\widehat{f}}[\log g(X)] = -\frac{1}{2}\log(2\pi) - \frac{1}{2n}\sum_{i=1}^{n} X_i^2$$

- Numerical examples

n	10	100	1,000	10,000	$\mathbb{E}_f[\log g(X)]$
Mean	-1.4188	-1.4185	-1.4191	-1.4189	-1.4189
Variance	0.05079	0.00497	0.00050	0.00005	-
Standard deviation	0.22537	0.07056	0.02232	0.00696	-

Mean, variance and standard deviation by running 1,000 Monte Carlo trials

Theorem. Suppose

$$\sup_{\theta \in \Theta} \left| M_n(\theta) - M(\theta) \right| \xrightarrow{\mathrm{P}} 0$$

and that, for every $\varepsilon > 0$,

$$\sup_{\theta : |\theta - \theta^*| \geq \varepsilon} M(\theta) < M(\theta_*)$$

then

$$\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta_*$$

Proof. Since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n\left(\widehat{\theta}_n\right) \geq M_n(\theta_*)$. Thus,

$$\begin{aligned} M(\theta_*) - M\left(\widehat{\theta}_n\right) &= M_n(\theta_*) - M\left(\widehat{\theta}_n\right) + M(\theta_*) - M_n(\theta_*) \\ &\leq M_n\left(\widehat{\theta}_n\right) - M\left(\widehat{\theta}_n\right) + M(\theta_*) - M_n(\theta_*) \\ &\leq 2\sup_{\theta} \left|M_n(\theta) - M(\theta)\right| \\ &\xrightarrow{\mathrm{P}} 0. \end{aligned}$$

Thus it implies that for any $\delta > 0$, we have

$$\Pr\left(M\left(\widehat{\theta}_n\right) < M(\theta_*) - \delta\right) \to 0.$$

Now for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|\theta - \theta^*| \geq \varepsilon$ implies $M(\theta) < M(\theta_*) - \delta$. Hence,

$$\Pr\left(\left|\widehat{\theta}_n - \theta_*\right| > \varepsilon\right) \leq \Pr\left(M\left(\widehat{\theta}_n\right) < M(\theta_*) - \delta\right) \to 0.$$

Asymptotic Normality

- Assuming we have $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta_*$, what can we say about $\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right)$?
- **Lemma.** Let $s\left(x|\,\theta\right) := \frac{\partial \log f\left(x|\theta\right)}{\partial \theta}$ be the score function, then we have for any θ

$$\mathbb{E}_\theta\left[s\left(X|\,\theta\right)\right] = 0.$$

- *Proof.* We have

$$\begin{aligned}
& \int \frac{\partial \log f\left(x|\,\theta\right)}{\partial \theta} f\left(x|\,\theta\right) dx \\
= & \int \frac{\frac{\partial f\left(x|\theta\right)}{\partial \theta}}{f\left(x|\,\theta\right)} f\left(x|\,\theta\right) dx = \int \frac{\partial f\left(x|\,\theta\right)}{\partial \theta} dx \\
= & \frac{\partial}{\partial \theta} \underbrace{\int f\left(x|\,\theta\right) dx}_{=1} = 0.
\end{aligned}$$

- **Lemma**. We also have

$$var_\theta \left[s\left(X|\,\theta \right) \right] = \mathbb{E}_\theta \left[s\left(X|\,\theta \right)^2 \right] = -\mathbb{E}_\theta \left[\frac{\partial^2 \log f\left(X|\,\theta \right)}{\partial \theta^2} \right] := I\left(\theta \right)$$

- *Proof*. This follows from

$$\int \frac{\partial \log f\left(x|\,\theta \right)}{\partial \theta} f\left(x|\,\theta \right) dx = 0$$

thus by taking the derivative once more with respect to θ

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int \frac{\partial \log f\left(x|\,\theta \right)}{\partial \theta} f\left(x|\,\theta \right) dx \\ &= \int \frac{\partial^2 \log f\left(x|\,\theta \right)}{\partial \theta^2} f\left(x|\,\theta \right) + \int \frac{\partial \log f\left(x|\,\theta \right)}{\partial \theta} \frac{\partial f\left(x|\,\theta \right)}{\partial \theta} dx \end{aligned}$$

- *Heuristic Derivation*. We have for $l(\theta) := \log L(\theta|\mathbf{x})$

$$\begin{aligned} 0 &= l'\left(\widehat{\theta}_n\right) \approx l'(\theta_*) + \left(\widehat{\theta}_n - \theta_*\right) l''(\theta_*) \\ \Rightarrow & \left(\widehat{\theta}_n - \theta_*\right) = -\frac{l'(\theta_*)}{l''(\theta_*)} \end{aligned}$$

That is

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) = \frac{\frac{1}{\sqrt{n}} l'(\theta_*)}{-\frac{1}{n} l''(\theta_*)}.$$

- Now remember that $l'(\theta_*) = \sum_{i=1}^{n} s(X_i|\theta_*)$ where $\mathbb{E}_{\theta_*}[s(X_i|\theta_*)] = 0$ and $var_{\theta_*}[s(X_i|\theta_*)] = I(\theta_*)$ so the CLT tells us that

$$\frac{1}{\sqrt{n}} l'(\theta_*) \xrightarrow{\mathrm{D}} \mathcal{N}(0, I(\theta_*))$$

- Now the law of large number yields

$$-\frac{1}{n} l''(\theta_*) \xrightarrow{\mathrm{P}} I(\theta_*)$$

so by Slutsky's theorem

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \frac{1}{I(\theta_*)}\right) \Leftrightarrow \sqrt{n}\sqrt{I(\theta_*)}\left(\widehat{\theta}_n - \theta_*\right) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

- Note that you have already seen this expression when establishing the Cramer-Rao bound.
- It is important to remember that depending on θ_* the parameter can be more or less easy to estimate.

- Similarly, we can prove that

$$\sqrt{n}\sqrt{I\left(\widehat{\theta}_n\right)}\left(\widehat{\theta}_n-\theta_*\right)\xrightarrow{\mathrm{D}}\mathcal{N}\left(0,1\right).$$

- We can also prove that

$$\sqrt{n}\left|g'\left(\widehat{\theta}_n\right)\right|\sqrt{I\left(\widehat{\theta}_n\right)}\left(g\left(\widehat{\theta}_n\right)-g\left(\theta_*\right)\right)\xrightarrow{\mathrm{D}}\mathcal{N}\left(0,1\right).$$

- This allows us to derive some confidence intervals.

Making the proof more rigourous

- We have

$$l'\left(\widehat{\theta}_n\right) = l'\left(\theta_*\right) + \left(\widehat{\theta}_n - \theta_*\right) l''\left(\theta_*\right) + \frac{1}{2}\left(\widehat{\theta}_n - \theta_*\right)^2 l'''\left(\theta_{*,n}\right)$$

  where $\theta_{*,n}$ lies between $\widehat{\theta}_n$ and θ_* so that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) = \frac{\frac{1}{\sqrt{n}} l'\left(\theta_*\right)}{-\frac{1}{n} l''\left(\theta_*\right) - \frac{1}{2n}\left(\widehat{\theta}_n - \theta_*\right)^2 l'''\left(\theta_{*,n}\right)}.$$

- To proof the result, we need to check that $\frac{1}{2n}\left(\widehat{\theta}_n - \theta_*\right)^2 l'''\left(\theta_{*,n}\right) \xrightarrow{\mathrm{P}} 0$. As $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta_*$, we just need to prove that $\frac{1}{n} l'''\left(\theta_{*,n}\right)$ is bounded (in probability). So we need an additional condition of the form say for any θ

$$\left|\frac{\partial^3 \log f\left(x|\,\theta\right)}{\partial \theta^3}\right| \leq C\left(x\right)$$

  with $\mathbb{E}_{\theta_*}\left[C\left(X\right)\right] < \infty$.

Multiparameter Case

- The extension to the multiparameter case $\theta = (\theta_1, ..., \theta_d)$ is straightforward

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, J\left(\theta_*\right)\right)$$

where $J\left(\theta_*\right) = I\left(\theta_*\right)^{-1}$ where

$$\left[I\left(\theta_*\right)\right]_{k,l} = -\mathbb{E}_{\theta_*}\left[\frac{\partial^2 \log f\left(x|\,\theta\right)}{\partial\theta_k \partial\theta_l}\right].$$

- We define $\nabla g := \left(\frac{\partial g}{\partial\theta_1}, \ldots, \frac{\partial g}{\partial\theta_d}\right)^{\mathsf{T}}$ then if $\nabla g\left(\theta_*\right) \neq 0$

$$\sqrt{n}\left(g\left(\widehat{\theta}_n\right) - g\left(\theta_*\right)\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \nabla g\left(\theta_*\right) J\left(\theta_*\right) \nabla g\left(\theta_*\right)^{\mathsf{T}}\right)$$

- Example: If $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\theta = (\mu, \sigma)$ then

$$\log f\left(x|\,\theta\right) = cst - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

$$s\left(x|\,\theta\right) = \begin{pmatrix} \frac{(x-\mu)}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} \end{pmatrix},$$

$$I\left(\theta\right) = \begin{pmatrix} \frac{1}{\sigma^2} & \mathbb{E}_\theta\left[\frac{2(x-\mu)}{\sigma^3}\right] \\ \mathbb{E}_\theta\left[\frac{2(x-\mu)}{\sigma^3}\right] & \mathbb{E}_\theta\left(-\frac{1}{\sigma^2} + \frac{3(x-\mu)^2}{\sigma^4}\right) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

- The MLE of μ is given by

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow var\left[\widehat{\mu}\right] = \frac{\sigma^2}{n}$$

and the MLE is indeed efficient (it reaches Cramer-Rao lower bound).

- Assume we observe a vector $X = (X_1, ..., X_k)$ where $X_j \in \{0, 1\}$, $\sum_{j=1}^{k} X_j = 1$ with

$$f\left(x \middle| p_1, ..., p_{k-1}\right) = \left(\prod_{j=1}^{k-1} p_j^{x_j}\right)\left(1 - \sum_{j=1}^{k-1} p_j\right)^{x_k}$$

  where $p_j > 0$ and $p_k := 1 - \sum_{j=1}^{k-1} p_j < 1$. We have $\theta = (p_1, ..., p_{k-1})$.

- We have

$$\begin{aligned}
\frac{\partial \log f\left(x \middle| \theta\right)}{\partial p_j} &= \frac{x_j}{p_j} - \frac{x_k}{p_k}, \\
\frac{\partial^2 \log f\left(x \middle| \theta\right)}{\partial p_j^2} &= -\frac{x_j}{p_j^2} - \frac{x_k}{p_k^2} \\
\frac{\partial^2 \log f\left(x \middle| \theta\right)}{\partial p_j \partial p_l} &= -\frac{x_k}{p_k^2}, j \neq l < k.
\end{aligned}$$

- Recall that X_j has a Bernoulli distribution with mean p_j so

$$I(\theta) = \begin{bmatrix} p_1^{-1} + p_k^{-1} & p_k^{-1} & \cdots & p_k^{-1} \\ p_k^{-1} & p_2^{-1} + p_k^{-1} & & \vdots \\ \vdots & p_k^{-1} & & \vdots \\ \vdots & \vdots & & p_{k-1}^{-1} + p_k^{-1} \end{bmatrix}$$

- One can check that

$$I(\theta)^{-1} = \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_{k-1} \\ -p_1 p_2 & p_2(1-p_2) & & -p_2 p_{k-1} \\ \vdots & \vdots & & \vdots \\ -p_1 p_{k-1} & -p_2 p_{k-1} & \cdots & p_{k-1}(1-p_{k-1}) \end{bmatrix}$$

- Now assume we observe $X^1, X^2, ..., X^n$ then

$$\log L(\theta|\mathbf{x}) = \sum_{j=1}^{k} t_j \log p_j = \sum_{j=1}^{k-1} t_j \log p_j + t_k \log\left(1 - \sum_{j=1}^{k-1} p_j\right)$$

  where $t_j = \sum_{i=1}^{n} x_j^i$.
- So we have

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial p_j} = \frac{t_j}{p_j} - \frac{t_k}{p_k} \text{ for } j = 1, ..., k-1 \Rightarrow \widehat{p}_j = \frac{t_j}{n}.$$

- Clearly t_j is Binomial(n, p_j) with variance $np_j(1-p_j)$ so $\widehat{p}_j$ is efficient.

Nuisance Parameters

- Assume $\theta = (\theta_1, ..., \theta_d)$ is the parameter vector but only the scalar θ_1 is of interest whereas $(\theta_2, ..., \theta_d)$ are nuisance parameters.
- We want to assess how the asymptotic precision with which we estimate θ_1 is influenced by the presence of nuisance parameters; i.e. if $\widehat{\theta}$ is an efficient estimate for θ, then how does $\widehat{\theta}_1$ as an estimator of θ_1 compare to an efficient estimation of θ_1, say $\widetilde{\theta}_1$, which would assume that all the nuisance parameters are known.
- Intuitively, we should have $var\left[\widetilde{\theta}_1\right] \leq var\left[\widehat{\theta}_1\right]$; i.e. ignorance cannot bring you any advantage.

- Asymptotic variance of $\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right)$ is $I^{-1}(\theta_*)$ whose (i,j) parameter is denoted $\alpha_{i,j}$.
- Asymptotic variance of $\sqrt{n}\left(\widetilde{\theta}_1 - \theta_{*,1}\right)$ is $I^{-1}(\theta_{*,1}) = 1/\gamma_{1,1}$.
- **Theorem**. We have $\alpha_{1,1} \geq 1/\gamma_{1,1}$, with equality if and only if $\alpha_{1,2} = \cdots = \alpha_{1,d} = 0$.

- Partition $I(\theta)$ as follows

$$I(\theta) = \begin{pmatrix} \gamma_{1,1} & \rho^{\mathsf{T}} \\ \rho & \Sigma \end{pmatrix}.$$

- Now we use the fact that

$$I^{-1}(\theta) = \frac{1}{\tau} \begin{pmatrix} 1 & -\rho^{\mathsf{T}}\Sigma^{-1} \\ -\Sigma^{-1}\rho & \Sigma^{-1}\rho\rho^{\mathsf{T}}\Sigma^{-1} + \tau\Sigma^{-1} \end{pmatrix}$$

  where $\tau = \gamma_{1,1} - \rho^{\mathsf{T}}\Sigma^{-1}\rho$.
- As $I(\theta)$ is definite positive then Σ^{-1} is definite positive and

$$\alpha_{1,1} = \frac{1}{\tau} \geq 1/\gamma_{1,1}$$

  with equality iff $\rho = 0$.
- To show that $\tau > 0$ we use the fact that $I(\theta)$ is p.d. and that

$$\tau = v^{\mathsf{T}} I(\theta) v \text{ where } v = \begin{pmatrix} 1 \\ -\rho^{\mathsf{T}}\Sigma^{-1} \end{pmatrix}.$$

Beyond Maximum Likelihood: Method of Moments

- MLE estimates can be difficult to compute, the method of moments is a simple alternative. The obtained estimators are typically not optimal but can be used as starting values for more sophisticated methods.
- For $1 \leq j \leq d$, define the j^{th} moment of $f\left(x|\,\theta\right)$ where $\theta = (\theta_1, ..., \theta_d)$

$$\alpha_j\left(\theta\right) = \mathbb{E}_\theta\left[X^j\right] = \int x^j f\left(x|\,\theta\right) dx$$

and, given $\mathbf{X} = (X_1, ..., X_n)$, the j^{th} sample as

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

- The idea of the method of moments method is to match the theoretical moments to the sample moments; that is we defined $\widehat{\theta}_n$ as the value of θ such that

$$\alpha_j\left(\widehat{\theta}_n\right) = \widehat{\alpha}_j \text{ for } j = 1, ..., d.$$

- Example: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\theta = \left(\mu, \sigma^2\right)$ then

$$\begin{aligned} \alpha_1(\theta) &= \mu,\ \alpha_2(\theta) = \sigma^2 + \mu^2, \\ \widehat{\alpha}_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1,\ \widehat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

- Thus we obtain

$$\widehat{\mu} = \widehat{\alpha}_1 \text{ and } \widehat{\sigma^2} = \widehat{\alpha}_2 - \left(\widehat{\alpha}_1\right)^2.$$

- Note that $\widehat{\sigma^2}$ is not unbiased.

- Assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\theta_1, \theta_2)$ where $-\infty < \theta_1 < \theta_2 < +\infty$ then

$$\alpha_1(\theta) = \frac{\theta_1 + \theta_2}{2}, \ \alpha_2(\theta) = \frac{\theta_1^2 + \theta_2^2 + \theta_1\theta_2}{3}.$$

- Now we solve and obtain

$$\begin{aligned} \theta_1 &= 2\widehat{\alpha}_1 - \theta_2, \\ 3\widehat{\alpha}_2 &= (2\widehat{\alpha}_1 - \theta_2)^2 + \theta_2^2 + (2\widehat{\alpha}_1 - \theta_2)\theta_2 \Leftrightarrow (\theta_2 - \widehat{\alpha}_1)^2 = 3\left(\widehat{\alpha}_2 - \widehat{\alpha}_1^2\right) \end{aligned}$$

Since $\theta_2 > \mathbb{E}(X)$ then

$$\widehat{\theta}_2 = \widehat{\alpha}_1 + \sqrt{3\left(\widehat{\alpha}_2 - \widehat{\alpha}_1^2\right)}, \ \widehat{\theta}_1 = \widehat{\alpha}_1 - \sqrt{3\left(\widehat{\alpha}_2 - \widehat{\alpha}_1^2\right)}.$$

- Note that $\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$ is NOT a function of the sufficient statistics $X_{(1)}, X_{(n)}$.

- Assume $X_i \overset{\text{i.i.d.}}{\sim} Bi(p, k)$ with parameters $k \in \mathbb{N}$ and $p \in (0, 1)$.
- We have
$$\alpha_1(\theta) = kp, \ \alpha_2(\theta) = kp(1-p) + k^2p^2.$$
- Thus we obtain
$$\begin{aligned} \widehat{p} &= \left(\widehat{\alpha}_1 + \widehat{\alpha}_1^2 - \widehat{\alpha}_2\right) / \widehat{\alpha}_1, \\ \widehat{k} &= \widehat{\alpha}_1^2 / \left(\widehat{\alpha}_1 + \widehat{\alpha}_1^2 - \widehat{\alpha}_2\right). \end{aligned}$$
- The estimator $\widehat{p} \in (0, 1)$ but $\widehat{k}$ is generally not an integer.

Statistical Properties of the Estimate

- Let $\widehat{\alpha} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_d)$, we have

$$\widehat{\alpha} = h\left(\widehat{\theta}\right)$$

and if the inverse function $g = h^{-1}$ exists, then

$$\widehat{\theta} = g\left(\widehat{\alpha}\right).$$

- If g is continuous at $\alpha = (\alpha_1, ..., \alpha_d)$ then $\widehat{\theta}$ is a consistent estimate of θ as $\widehat{\alpha}_j \longrightarrow \alpha_j$.
- Moreover if g is differentiable at α and $\mathbb{E}_\theta\left[X^{2d}\right] < \infty$ then

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathsf{D}} \mathcal{N}\left(0, \nabla g\left(\alpha\right)^{\mathsf{T}} V_\alpha \nabla g\left(\alpha\right)\right)$$

where

$$V_\alpha\left[i, j\right] = \alpha_{i+j} - \alpha_i \alpha_j.$$

- The result follows from

$$\sqrt{n}\left(\widehat{\alpha}_n - \alpha\right) \xrightarrow{\mathsf{D}} \mathcal{N}\left(0, V_\alpha\right)$$

as

$$\mathbb{E}_\theta\left[\widehat{\alpha}_{j,n}\right] = \alpha_j, \;\; cov\left[\widehat{\alpha}_{i,n}\widehat{\alpha}_{j,n}\right] = \frac{\alpha_{i+j} - \alpha_i\alpha_j}{n}$$

- We have

$$\widehat{\theta}_n - \theta = g\left(\widehat{\alpha}_n\right) - g\left(\alpha_n\right)$$

so using the delta method

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathsf{D}} \mathcal{N}\left(0, \nabla g\left(\alpha\right)^{\mathsf{T}} V_\alpha \nabla g\left(\alpha\right)\right)$$

- We can also establish the n^{-1} order asymptotic bias of the estimate as

$$\begin{aligned} g\left(\widehat{\alpha}_n\right) &= g\left(\alpha\right) + \nabla g\left(\alpha\right)^{\mathsf{T}}\left(\widehat{\alpha}_n - \alpha\right) \\ &\quad + \frac{1}{2}\left(\widehat{\alpha}_n - \alpha\right)^{\mathsf{T}} \nabla^2 g\left(\alpha\right)\left(\widehat{\alpha}_n - \alpha\right) + o\left(\frac{1}{n}\right) \end{aligned}$$

  where $\sqrt{n}\left(\widehat{\alpha}_n - \alpha\right) \xrightarrow{\mathsf{D}} Z_\Sigma$ with $Z_\Sigma \sim \mathcal{N}\left(0, \Sigma\right)$ so

$$n\left(\widehat{\alpha} - \alpha\right)^{\mathsf{T}} \nabla^2 g\left(\alpha\right)\left(\widehat{\alpha} - \alpha\right) \xrightarrow{\mathsf{D}} Z_\Sigma^{\mathsf{T}} \nabla^2 g\left(\alpha\right) Z_\Sigma$$

  as recall that $X_n \xrightarrow{\mathsf{D}} X$ implies $\varphi\left(X_n\right) \xrightarrow{\mathsf{D}} \varphi\left(X\right)$.
- Thus we have

$$\begin{aligned} \mathbb{E}\left[g\left(\widehat{\alpha}_n\right)\right] &= g\left(\alpha\right) + \frac{1}{2n}\mathbb{E}\left[Z_\Sigma^{\mathsf{T}} \nabla^2 g\left(\alpha\right) Z_\Sigma\right] + o\left(\frac{1}{n}\right) \\ &= g\left(\alpha\right) + \frac{tr\left(\nabla^2 g\left(\alpha\right)\Sigma\right)}{2n} + o\left(\frac{1}{n}\right). \end{aligned}$$

Beyond Maximum Likelihood: Pseudo-Likelihood

- Assume $X = (X_1, ..., X_q) \sim f(x|\theta)$. Given n observations $X^i \sim f(x|\theta)$, the MLE requires maximizing $L(\theta|\mathbf{x})$.
- However in some problems, it might be difficult to specify $f(x|\theta)$ and we may be only able to specify say

$$f(x_k, x_l|\theta) \text{ for } 1 \leq k < l \leq q$$

- Based on this information and n observations, we could define the pseudo-log-likelihood functions

$$\begin{aligned} l_1(\theta|\mathbf{x}) &= \sum_{i=1}^{n} l_1(\theta|x^i) = \sum_{i=1}^{n}\sum_{s=1}^{q} \log f(x_s|\theta), \\ l_2(\theta|\mathbf{x}) &= \sum_{i=1}^{n} l_2(\theta|x^i) = \sum_{i=1}^{n}\sum_{s=1}^{q}\sum_{t=s+1}^{q} \log f(x_s, x_t|\theta) + \alpha l_1(\theta|\mathbf{x}). \end{aligned}$$

- These pseudo-likelihood functions are simpler that the full likelihood.

- Under regularity conditions very similar to the ones for the MLE, solving

$$l'_k(\theta|\mathbf{x}) = 0 \text{ for } k = 1, 2$$

will provide unbiased estimates.

- To derive the asymptotic variance, we use

$$\begin{aligned} 0 &= l'_k\left(\widehat{\theta}_n\right) \approx l'_k(\theta_*) + \left(\widehat{\theta}_n - \theta_*\right) l''_k(\theta_*) \\ \Rightarrow \quad \sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) &= -\frac{\frac{1}{\sqrt{n}} l'_k(\theta_*)}{\frac{1}{n} l''_k(\theta_*)} \end{aligned}$$

where $\frac{1}{n} l''_k(\theta_*) \xrightarrow{\mathrm{P}} \mathbb{E}_{\theta_*}\left[l''_k\right]$ and $\frac{1}{\sqrt{n}} l'(\theta_*) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \mathbb{E}_{\theta_*}\left[l'^2_k\right]\right)$, thus

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_*\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \mathbb{E}_{\theta_*}\left[-l''_k\right]^{-2} \mathbb{E}_{\theta_*}\left[l'^2_k\right]\right).$$

- We have the estimates

$$\mathbb{E}_{\theta_*}\left[l''_k\right] \approx \frac{1}{n}\sum_{i=1}^{n} l''_k(\theta|x_i), \quad \mathbb{E}_{\theta_*}\left[l'^2_k\right] \approx \frac{1}{n}\sum_{i=1}^{n} l'^2_k(\theta|x_i).$$

- Example. Assume that $X = (X_1, ..., X_q) \sim \mathcal{N}(0, \Sigma)$ where $[\Sigma](i,j) = 1$ if $i = j$ and ρ otherwise. We are interested in estimating $\theta = \rho$.
- There is no information about ρ in $l_1(\theta|\mathbf{x})$ so we use $l_2(\theta|\mathbf{x})$ for $\alpha = 0$. For n observations $\left(X^1, ..., X^n\right)$, we have

$$\begin{aligned} l_2(\theta|\mathbf{x}) &= -\frac{nq(q-1)}{4}\log\left(1-\rho^2\right) - \frac{q-1+\rho}{2(1-\rho^2)}SS_W \\ &\quad -\frac{(q-1)(1-\varrho)}{2(1-\rho^2)}\frac{SS_B}{q} \end{aligned}$$

where

$$SS_W = \sum_{i=1}^{n}\sum_{s=1}^{q}\left(X_s^i - \left(\sum_{t=1}^{q}X_t^i\right)\right)^2, \; SS_B = \sum_{i=1}^{n}X_t^{i^2}.$$

- After simple but tiedous calculations, we obtain for the asymptotic variance

$$\frac{2}{nq\left(q-1\right)}\frac{\left(1-\rho^{2}\right)c\left(q,\rho\right)}{\left(1+\rho^{2}\right)^{2}}$$

where

$$\begin{aligned} c\left(q,\rho\right) &= \left(1-\rho^{2}\right)\left(1+3\rho^{2}\right)+q\rho\left(-3\rho^{3}+8\rho^{2}-3\rho+2\right) \\ &\quad +q^{2}\rho^{2}\left(1-\rho^{2}\right) \end{aligned}$$

whereas for MLE we have

$$\frac{2}{nq\left(q-1\right)}\frac{\left(1+\left(q-1\right)\rho\right)^{2}\left(1-\rho\right)^{2}}{1+\left(q-1\right)\rho^{2}}.$$

- The ratio is 1 for $q=2$ as expected and also 1 if $\rho=0$ or 1. For any other values, there is a loss of efficiency for $I_{2}\left(\theta\middle|\mathbf{x}\right)$ which increases as $q\rightarrow\infty$.

- Consider the following time series

$$X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right), \; X_n = \alpha X_{n-1} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

  where $\theta = \sigma^2$.
- We can show that we have for any $i = 0, 1, ..., n$

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(1-\alpha^2\right) x_i^2}{2\sigma^2}\right)$$

  and we consider

$$2l_1(\theta|\mathbf{x}) = 2\sum_{i=1}^{n} \log f(x_i|\theta) = cste - n\log\sigma^2 - \frac{\left(1-\alpha^2\right)}{2\sigma^2}\sum_{i=1}^{n} x_i^2$$

- This pseudo-likelihood can easily be maximized

$$\widehat{\sigma^2} = \frac{\left(1-\alpha^2\right)\sum_{i=1}^{n} x_i^2}{n}.$$

- If one is interested in estimating α, it would be necessary to introduce $f(x_i, x_{i+1}|\theta)$.

- Pseudo-likelihood is widely used for Markov random fields since its introduction by Besag (1975). In the Gaussian context, we have $X = (X_1, ..., X_n)$ where d is extremely large and Gaussian and the model is specified by

$$\mathbb{E}_\theta \left[X_i \middle| x_{-i} \right] = \lambda \sum_{i=1}^{n} H_{ij} x_j, \;\; var_\theta \left[X_i \middle| x_{-i} \right] = \kappa.$$

- Computing the likelihood for $\theta = (\lambda, \kappa)$ can be too computatonally intensive so the pseudo-likelihood is defined through

$$\widetilde{l}\left(\theta \middle| \mathbf{x} \right) = \sum_{i=1}^{n} \log f\left(x_i \middle| \theta, x_{-i} \right)$$

thus

$$\widehat{\lambda} = \frac{\mathbf{x}^\mathsf{T} \mathbf{H} \mathbf{x}}{\mathbf{x}^\mathsf{T} \mathbf{H}^2 \mathbf{x}}, \;\; \kappa = d^{-1} \left(\mathbf{x}^\mathsf{T} \mathbf{x} - \frac{\left(\mathbf{x}^\mathsf{T} \mathbf{H} \mathbf{x} \right)^2}{\mathbf{x}^\mathsf{T} \mathbf{H}^2 \mathbf{x}} \right)$$

- In this context, it can be show that the estimate is consistent and has a reasonable asymptotic variance.

Summary of Pseudo Likelihood

- In many applications, the log-likelihood $l(\theta; y_{1:n})$ is very complex to compute.
- Instead we maximize a surrogate function $l_S(\theta; y_{1:n})$.
- If possible, we pick this function such that if θ^* is the 'true' parameter then

$$\mathbb{E}_{\theta^*}\left(l_S(\theta; Y_{1:n})\right)$$

is maximized for $\theta = \theta^*$ and solving

$$\nabla l_S\left(\widehat{\theta}_n; Y_{1:n}\right) = 0$$

is 'easy'.

- Under regularity assumptions, we have

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \Rightarrow \mathcal{N}\left(0, G_n^{-1}\left(\theta^*\right)\right)$$

where

$$G_n^{-1}\left(\theta\right) = H_n^{-1}\left(\theta\right) J_n\left(\theta\right) H_n^{-\mathsf{T}}\left(\theta\right)$$

with

$$\begin{aligned} J_n\left(\theta\right) &= \mathbb{V}\left\{\nabla l_S\left(\theta; Y_{1:n}\right)\right\}, \\ H_n\left(\theta\right) &= \mathbb{E}\left\{\nabla^2 l_S\left(\theta; Y_{1:n}\right)\right\}. \end{aligned}$$

- When $l_S\left(\theta; Y_{1:n}\right) = l\left(\theta; Y_{1:n}\right)$ and the model is correctly specified then $G_n\left(\theta\right)$ is the Fisher information matrix.
- When $l_S\left(\theta; Y_{1:n}\right) \neq l\left(\theta; Y_{1:n}\right)$, we typically lose in terms of efficiency.

Application to General State-Space Models

- Consider the following general state-space model. Let $\{X_k\}_{k\geq 1}$ be an Markov process defined by

$$X_1 \sim \mu_\theta \text{ and } X_k | (X_{k-1} = x_{k-1}) \sim f_\theta(\cdot | x_{k-1}).$$

- Then we have that for any $n > 0$

$$\begin{aligned} p_\theta(x_{1:n}) &= p_\theta(x_1) \prod_{k=2}^{n} p_\theta(x_k | x_{1:k-1}) \\ &= \mu_\theta(x_1) \prod_{k=2}^{n} f_\theta(x_k | x_{k-1}). \end{aligned}$$

- We are interested in estimating θ from the data but we do not have access to $\{X_k\}_{k\geq 1}$. We only have access to a process $\{Y_k\}_{k\geq 1}$ such that, conditional upon $\{X_k\}_{k\geq 1}$, the observations are statistically independent and

$$Y_k | (X_k = x_k) \sim g_\theta (\cdot | x_k) .$$

- That is we have for any $n > 0$

$$p_\theta (y_{1:n} | x_{1:n}) = \prod_{k=1}^{n} p_\theta (y_k | x_k) = \prod_{k=1}^{n} g_\theta (y_k | x_k) .$$

Examples

- *Linear Gaussian model*. Consider say for $|\alpha| < 1$

$$\begin{aligned} X_1 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right), \; X_k = \alpha X_{k-1} + \sigma V_k \text{ where } V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ Y_k &= \beta + X_k + \tau W_k \text{ where } W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

- In this case we have say $\theta = \left(\beta, \sigma^2, \alpha, \tau^2\right)$ and

$$\begin{aligned} f_\theta\left(x_k \middle| x_{k-1}\right) &= \mathcal{N}\left(x_k; \alpha x_{k-1}, \sigma^2\right), \\ g_\theta\left(y_k \middle| x_k\right) &= \mathcal{N}\left(y_k; \beta + x_k, \tau^2\right). \end{aligned}$$

- *Stochastic Volatility model*. Consider say for $|\alpha| < 1$

$$\begin{aligned} X_1 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right), \; X_k = \alpha X_{k-1} + \sigma V_k \text{ where } V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ Y_k &= \beta \exp(X_k/2) W_k \text{ where } W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

- In this case we have say $\theta = \left(\beta, \sigma^2, \alpha\right)$ and

$$\begin{aligned} f_\theta(x_k | x_{k-1}) &= \mathcal{N}\left(x_k; \alpha x_{k-1}, \sigma^2\right), \\ g_\theta(y_k | x_k) &= \mathcal{N}\left(y_k; 0, \beta^2 \exp(x_k)\right). \end{aligned}$$

- In this case, the likelihood of the observations $y_{1:n}$ is given by

$$\begin{aligned} p_\theta\left(y_{1:n}\right) &= \int p_\theta\left(x_{1:n}, y_{1:n}\right) dx_{1:n} \\ &= \int p_\theta\left(y_{1:n} \middle| x_{1:n}\right) p_\theta\left(x_{1:n}\right) dx_{1:n} \\ &= \int \left(\prod_{k=1}^{n} g_\theta\left(y_k \middle| x_k\right)\right) \left(\mu_\theta\left(x_1\right) \prod_{k=2}^{n} f_\theta\left(x_k \middle| x_{k-1}\right)\right) dx_{1:n}. \end{aligned}$$

- If the model is linear Gaussian or finite state-space, we can compute the likelihood in closed-form but the maximization is not trivial. Otherwise, we cannot.

Pairwise likelihood for state-space models

- We consider the following pseudo-likelihood for $m \geq 1$

$$L_S(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{\min\{i+m,n\}} p_\theta(y_i, y_j)$$

where

$$p_\theta(y_i, y_j) = \int g_\theta(y_i|x_i)\, g_\theta(y_j|x_j)\, p_\theta(x_i, x_j)\, dx_i dx_j.$$

- As an alternative, if $n = pm$, we could maximize

$$L_S(\theta; y_{1:n}) = \prod_{i=1}^{p} p_\theta\left(y_{(i-1)m+1:im}\right).$$

- For the two models discussed earlier, it is possible to compute exactly $p_\theta(x_i, x_j)$ as

$$p_\theta(x_i, x_j) = p_\theta(x_i)\, p_\theta(x_j | x_i)$$

where

$$\begin{aligned} p_\theta(x_i) &= \mathcal{N}\left(x_i; 0, \frac{\sigma^2}{1-\alpha^2}\right) \\ p_\theta(x_j | x_i) &= \mathcal{N}\left(x_j; \alpha^{j-i} x_i, \sigma^2 \sum_{k=0}^{j-i-1} \alpha^{2k}\right). \end{aligned}$$

- In a general case, we could approximate $p_\theta(y_i, y_j)$ through Monte Carlo

$$\widehat{p}_\theta(y_i, y_j) = \frac{1}{N} \sum_{l=1}^{N} g_\theta\left(y_i | X_i^l\right) g_\theta\left(y_j | X_j^l\right)$$

where $\left(X_i^l, X_j^l\right) \sim p_\theta(x_i, x_j)$.

- Under regularity assumptions, we have

$$\int l_S(\theta; y_{1:n})\, p_{\theta^*}(y_{1:n})\, dy_{1:n}$$

which is maximum in θ^* so maximizing this pseudo-likelihood method makes sense.

- To prove it, note that

$$l_S(\theta; y_{1:n}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{\min\{i+m,n\}} \log p_\theta(y_i, y_j)$$

and

$$\begin{aligned} & \int \log p_\theta(y_i, y_j)\, .p_{\theta^*}(y_{1:n})\, dy_{1:n} \\ = & \int \log p_\theta(y_i, y_j)\, .p_{\theta^*}(y_i, y_j)\, dy_i dy_j \end{aligned}$$

which is maximum in $\theta = \theta^*$.

Application

- Consider

$$X_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right), \; X_k = \alpha X_{k-1} + \sigma V_k \text{ where } V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

$$Y_k = \beta + X_k + \tau W_k \text{ where } W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

where $\theta = \left(\beta, \sigma^2, \alpha, \tau^2\right)$.

- In this case, we can directly establish not only $p_\theta(x_i, x_j)$ but $p_\theta(y_i, y_j)$

$$\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \beta \end{pmatrix}, \begin{pmatrix} \tau^2 + \frac{\sigma^2}{1-\alpha^2} & \alpha^{j-i}\frac{\sigma^2}{1-\alpha^2} \\ \alpha^{j-i}\frac{\sigma^2}{1-\alpha^2} & \tau^2 + \frac{\sigma^2}{1-\alpha^2} \end{pmatrix}\right)$$

- For $m = 2, ..., 20$ we compare the performance of $\widehat{\theta}_{MLE}$ and $\widehat{\theta}_{MPL}$ where the likelihood and pseudo-likelihood are maximized using a simple gradient algorithm (EM could be used).
- 1,000 time series of length $n = 500$ with $\beta^* = 0.1$, $\tau^* = 1.0$, $\alpha^* = 0.95$ and $\sigma^* = 0.55$ are simulated.

true	$\widehat{\theta}^{(2)}_{MPL}$	$\widehat{\theta}^{(6)}_{MPL}$	$\widehat{\theta}^{(12)}_{MPL}$	$\widehat{\theta}^{(20)}_{MPL}$	$\widehat{\theta}_{ML}$
β 0.1	0.108	0.108	0.109	0.109	0.102
	(0.488)	(0.489)	(0.4908)	(0.492)	(0.481)
τ 1.0	0.994	0.997	0.990	0.981	0.995
	(0.066)	(0.048)	(0.054)	(0.068)	(0.046)
α 0.95	0.941	0.941	0.939	0.937	0.941
	(0.033)	(0.020)	(0.022)	(0.024)	(0.020)
σ 0.55	0.535	0.551	0.560	0.571	0.554
	(0.160)	(0.064)	(0.072)	(0.087)	(0.061)

Now you should...

- be able to compute MLE estimate for rather complex models,
- be able to compute the asymptotic variance of the MLE estimate,
- be able to derive the expression of the asymptotic variance of simple estimates different from MLE.