ST519/807: Mathematical Statistics

Bent Jørgensen
University of Southern Denmark

November 22, 2014

Abstract

These are additional notes on the course material for ST519/807: Mathematical Statistics. HMC refers to the textbook Hogg et al. (2013).

Key words: Asymptotic theory, consistency, Cramér-Rao inequality, efficiency, exponential family, estimation, Fisher's scoring method, Fisher information, identifiability, likelihood, maximum likelihood, observed information, orthogonality, parameter, score function, statistical model, statistical test, sufficiency.

Fisher (1922), under the heading "The Neglect of Theoretical Statistics", wrote: *Several reasons have contributed to the prolonged neglect into which the study of statistics, in its theoretical aspects, has fallen. In spite of the immense amount of fruitful labour which has been expended in its practical application, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved.* Fisher then went on to introduce the main ingredients of likelihood theory, which shaped much of mathematical statistics of the 20th Century, including concepts such as *statistical model, parameter, identifiability, estimation, consistency, likelihood, score function, maximum likelihood, Fisher information, efficiency, and sufficiency.* Here we review the basic elements of likelihood theory in a contemporary setting.

Prerequisites: Sample space; probability distribution; discrete and continuous random variables; PMF and PDF; transformations; independent random variables; mean, variance, covariance and correlation.

Special distributions: Uniform; Bernoulli; binomial; Poisson; geometric; negative binomial; gamma; chi-square; beta; normal; t-distribution; F-distribution.

Contents

1 Stochastic convergence and the Central Limit Theorem

- **Setup:** Let X denote a random variable (r.v.) and let $\{X_n\}_{n=1}^{\infty}$ denote a sequence of r.v.s., all defined on a suitable probability space $(\mathcal{C}, \mathcal{B}, P)$ (sample space, σ-algebra, probability measure).

- **Definition:** Convergence in probability. We say that

$$X_n \xrightarrow{P} X \text{ as } n \to \infty$$

(X_n converges to X in probability) if

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0 \ \forall \varepsilon > 0$$

- **Definition:** Convergence in distribution. If F is a distribution function (CDF) we say that

$$X_n \xrightarrow{D} F \quad \text{as} \quad n \to \infty$$

(X_n converges to F in distribution) if

$$P(X_n \leq x) \to F(x) \quad \text{as} \quad n \to \infty \text{ for all } x \in C(F)$$

where $C(F)$ denotes the set of continuity points of F. If X has distribution function F, we also write

$$X_n \xrightarrow{D} X \quad \text{as} \quad n \to \infty$$

- **Properties:** As $n \to \infty$

1. $X_n \xrightarrow{P} X \Rightarrow aX_n \xrightarrow{P} aX$
2. $X_n \xrightarrow{P} X \Rightarrow g\left(X_n\right) \xrightarrow{P} g\left(X\right)$ if g is continuous
3. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$
4. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then

$$X_n + Y_n \xrightarrow{P} X + Y \text{ and } X_n Y_n \xrightarrow{P} XY \tag{1.1}$$

- **Example:** Let X be symmetric, i.e. $-X \sim X$, and define

$$X_n = (-1)^n X$$

Then

$$X_n \xrightarrow{D} X$$

(meaning that X_n converges to the distribution of X), since $F_{X_n} = F_X$ for all n, but unless X_n is constant,

$$X_n \nrightarrow X \text{ in probability}$$

- However, we have the following properties

  1. $X_n \xrightarrow{D} c \Rightarrow X_n \xrightarrow{P} c$
  2. $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 0$ then $X_n + Y_n \xrightarrow{D} X$
  3. $X_n \xrightarrow{D} X \Rightarrow g(X_n) \xrightarrow{D} g(X)$ if g is continuous
  4. **Slutsky's Theorem:** If $X_n \xrightarrow{D} X$ and $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$ then

$$A_n + B_n X_n \xrightarrow{P} a + bX$$

- **Example:** Let X_n and Y_n be two sequences such that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} X$ The following examples show that we do not in general have a result similar to (1.1) for convergence in distribution.

  1. Suppose that X is symmetric (see above), and let $X_n = X$ and $Y_n = -X$ for all n. Then
$$X_n + Y_n = X - X = 0$$
  so clearly $X_n + Y_n$ converges i distribution to 0 as $n \to \infty$.
  2. Now suppose that for each n, X_n and Y_n are independent and identically distributed with CDF $F(x) = P(X < x)$ for all x. Now
$$X_n + Y_n \xrightarrow{D} F_{X_1+Y_1}$$
  where $F_{X_1+Y_1}(x) = P(X_1 + Y_1 \leq x)$ for all x, corresponding to the convolution of X_1 and Y_1. Hence, the assumption that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} X$ is not enough to determine the limiting distribution of $X_n + Y_n$, which in fact depends on the sequence of joint distribution of X_n and Y_n.

- **Statistical setup**:

  Let $X_1, X_2, \ldots$ be a sequence of i.i.d. variables. Assume
$$\mu = \mathrm{E}(X_i) \quad \text{and} \quad \sigma^2 = \mathrm{Var}(X_i)$$
  Define for $n = 1, 2, \ldots$
$$T_n = \sum_{i=1}^{n} X_i \text{ and } \bar{X}_n = \frac{1}{n} T_n$$
  Then
$$\mathrm{E}(\bar{X}_n) = \mu \quad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

- **The (Weak) Law of Large Numbers** (LLN) says

$$\bar{X}_n \xrightarrow{P} \mu$$

Proof: Use Chebyshev's inequality

$$P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2/n}{\varepsilon^2} \to 0 \text{ as } n \to \infty$$

- Convergence to the standard normal distribution

$$P(X_n \leq x) \to \Phi(x) \quad \text{as} \quad n \to \infty,$$

for all $x \in \mathbb{R}$, where

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt.$$

- Now we define

$$Z_n = \sqrt{n}(\bar{X}_n - \mu)$$

for which

$$\mathrm{E}(Z_n) = 0 \quad \mathrm{Var}(Z_n) = \sigma^2$$

- **The Central Limit Theorem** (CLT) (see James, p. 265 or HMC p. 307) says

$$Z_n \xrightarrow{D} \mathrm{N}(0, \sigma^2) \quad \text{as} \quad n \to \infty$$

Practical use

$$\sqrt{n}(\bar{X}_n - \mu) \sim \mathrm{N}(0, \sigma^2) \quad \text{approx.}$$

which implies

$$\bar{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approx.}$$

Rule: The approximate normal distribution shares with $\bar{X}_n$ its mean and variance.

Example Bernoulli trials. Assume that the X_i are i.i.d. Bernoulli variables,

$$P(X_i = 1) = \mu = 1 - P(X_i = 0)$$

Hence we use μ as probability parameter, which is also the mean of X_i,

$$\mu = \mathrm{E}(X_i) \quad \text{and} \quad \sigma^2 = \mathrm{Var}(X_i) = \mu(1 - \mu)$$

Then

$$T_n = \sum_{i=1}^{n} X_i = \#\text{of 1s in a sample of } n$$

In fact $T_n \sim \mathrm{Bi}(n, \mu)$ (binomial distribution). Then, by the LLN

$$\bar{X}_n \xrightarrow{P} \mu$$

and by the CLT

$$Z_n \xrightarrow{D} \mathrm{N}(0, \mu(1-\mu))$$

so that

$$T_n \sim \mathrm{N}(n\mu, n\mu(1-\mu)) \quad \text{approx.}$$

- **Proofs based on the cumulant generating function.** Let X_i have cumulant generating function (CGF) $\kappa(s) = \log \mathrm{E}\left(e^{sX_i}\right)$. Note that $\overline{X}_n$ has CGF $n\kappa\left(s/n\right)$, which converges to $s\mu$, which is the CGF of the constant μ. This proves LLN. For $\sqrt{n}\left(\overline{X}_n - \mu\right)$ we have the CGF $n\kappa\left(s/\sqrt{n}\right) - s\mu\sqrt{n} = \frac{1}{2}\sigma^2 s^2 + O(n^{-1/2})$ which converges to $\mathrm{N}(0, \sigma^2)$.

- **Empirical variance:** Define

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \end{aligned}$$

Now, by the LLN

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathrm{E}(X_i^2) = \sigma^2 + \mu^2 \text{ as } n \to \infty$$

and

$$\bar{X}_n^2 \xrightarrow{P} \mu^2 \text{ as } n \to \infty$$

so by the properties above

$$S_n^2 \xrightarrow{P} \sigma^2 \text{ as } n \to \infty$$

and for that matter we also have

$$S_n \xrightarrow{P} \sigma \text{ as } n \to \infty$$

- **The Δ-method**: If the sequence X_n satisfies

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} \mathrm{N}(0, \sigma^2) \text{ as } n \to \infty$$

and if $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ and $\dot{g}(\theta) \neq 0$, then

$$\sqrt{n}\left[g\left(X_n\right) - g\left(\theta\right)\right] \xrightarrow{D} \mathrm{N}(0, \sigma^2\dot{g}^2(\theta)) \text{ as } n \to \infty$$

If $\dot{g}(\theta) = 0$, the asymptotic distribution is degenerate at zero. Note that

$$g\left(X_n\right) \sim \mathrm{N}(g\left(\theta\right), \sigma^2\dot{g}^2(\theta)/n), \text{ approx.}$$

so that $g\left(X_n\right)$ has asymptotic mean $g\left(\theta\right)$ and asymptotic variance $\sigma^2\dot{g}^2(\theta)/n$.

- **Proof** (sketch): By Taylor-expansion to first order, we obtain

$$\begin{aligned}\sqrt{n}\left[g\left(X_n\right)-g\left(\theta\right)\right] &\approx \dot{g}(\theta)\sqrt{n}\left(X_n-\theta\right)\\ &\xrightarrow{D} \dot{g}(\theta)\mathrm{N}(0,\sigma^2)\\ &= \mathrm{N}(0,\sigma^2\dot{g}^2(\theta))\end{aligned}$$

- **Definition:** A sequence $\{X_n\}_{n=1}^{\infty}$ is called *bounded in probability* if for any $\varepsilon > 0$ there exists $b_\varepsilon > 0$ such that

$$P\left(|X_n| \leq b_\varepsilon\right) \geq 1-\varepsilon \text{ for } n \text{ large enough.}$$

- Properties:

  1. If $X_n \xrightarrow{D} X$ then $\{X_n\}_{n=1}^{\infty}$ is bounded in probability.
  2. If $\{X_n\}_{n=1}^{\infty}$ is bounded in probability then

$$Y_n \xrightarrow{P} 0 \Rightarrow X_n Y_n \xrightarrow{P} 0$$

- The o and o_P notation. Recall that $a_n = o(b_n)$ for $b_n \to 0$ as $n \to \infty$ is defined by

$$\frac{a_n}{b_n} \to 0 \text{ as } n \to \infty$$

  This notation is used in connection with Taylor-expansions, e.g.

$$g(y) = g(x) + \dot{g}(x)(y-x) + o\left(|y-x|\right)$$

- o in probability, denoted o_P, is defined by

$$Y_n = o_P(X_n) \Leftrightarrow \frac{Y_n}{X_n} \xrightarrow{P} 0 \text{ as } n \to \infty$$

- Similarly O in probability, denoted O_P, is defined by

$$Y_n = O_P(X_n) \Leftrightarrow \frac{Y_n}{X_n} \text{ is bounded in probability.}$$

- **Theorem:** If $\{X_n\}_{n=1}^{\infty}$ is bounded in probability, and

$$Y_n = o_P(X_n)$$

  then

$$Y_n \xrightarrow{P} 0 \text{ as } n \to \infty$$

- Proof of Δ-method revisited. Use Taylor expansion with remainder term

$$g\left(X_n\right) - g\left(\theta\right) = \dot{g}(\theta)\left(X_n-\theta\right) + o_P(|X_n-\theta|)$$

  Then

$$\sqrt{n}\left[g\left(X_n\right)-g\left(\theta\right)\right] = \dot{g}(\theta)\sqrt{n}\left(X_n-\theta\right) + o_P(\sqrt{n}\,|X_n-\theta|)$$

  Since $\sqrt{n}\left(X_n-\theta\right) \xrightarrow{D} \mathrm{N}(0,\sigma^2)$ we find that $\sqrt{n}\,|X_n-\theta|$ is bounded in probability. Hence

$$o_P(\sqrt{n}\,|X_n-\theta|) \xrightarrow{P} 0 \text{ as } n \to \infty$$

  which implies the result.

2 The log likelihood function and its derivatives

2.1 Likelihood and log likelihood

- Likelihood and log likelihood: Let $X_1, X_2, \ldots, X_n$ be i.i.d. with either
  - probability density function $f_\theta(x)$ (continuous case);
  - probability mass function $f_\theta(x)$ (discrete case).

  θ is a real parameter with domain Ω (nonempty interval). θ is unknown, but we assume that the true distribution of $X_1, X_2, \ldots, X_n$ corresponds to $f_{\theta_0}(x)$ for some $\theta_0 \in \Omega$.

- **Regularity conditions:**
  1. The parameter θ is *identifiable*, i.e. if $f_{\theta_1}(x) = f_{\theta_2}(x)$ for almost all $x \in \mathbb{R}$ then $\theta_1 = \theta_2$.
  2. The support of $f_\theta(x)$ is the same for all $\theta \in \Omega$.
  3. The true parameter value θ_0 belongs to the interior of Ω.
  4. $f_\theta(x)$ is twice continuously differentiable with respect to θ for almost all x.
  5. $\frac{\partial}{\partial\theta}$ and $\int$ can be interchanged (continuous case), or $\frac{\partial}{\partial\theta}$ and $\sum$ can be interchanged (discrete case).

- The likelihood function is a stochastic function $L_n : \Omega \to [0, \infty)$ defined by

$$L_n(\theta) = f(X_1, X_2, \ldots, X_n; \theta) \text{ for } \theta \in \Omega,$$

  where

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

  is the joint probability density/mass function for $X_1, X_2, \ldots, X_n$.

  The **log likelihood** function is the stochastic function $\ell_n : \Omega \to \mathbb{R}$ defined by

$$\begin{aligned} \ell_n(\theta) &= \log L_n(\theta) = \log f(X_1, X_2, \ldots, X_n; \theta) \\ &= \sum_{i=1}^{n} \log f_\theta(X_i). \end{aligned}$$

  Strictly speaking, $\ell_n(\theta)$ takes the value $-\infty$ for $L_n(\theta) = 0$, but this is not a problem, because the region where $f(x_1, x_2, \ldots, x_n; \theta) = 0$ has probability zero.

Example - Bernoulli trials

$$\begin{aligned} f_\mu(x) &= \mu^x(1-\mu)^{1-x} \text{ for } x = 0, 1 \\ \ell_n(\mu) &= \sum_{i=1}^{n} \log\{\mu^{X_i}(1-\mu)^{(1-X_i)}\} \\ &= \sum_{i=1}^{n} \{X_i \log\mu + (1-X_i)\log(1-\mu)\} \quad 0 < \mu < 1 \\ &= T_n \log\frac{\mu}{1-\mu} + n\log(1-\mu). \end{aligned}$$

- **Parameter transformation.**

  Now suppose we work with ψ defined by $\theta = g(\psi)$ instead of θ, assuming that g is 1-1. Then the log likelihood for ψ is

$$\begin{aligned} \widetilde{\ell}_n(\psi) &= \sum_{i=1}^{n} \log f_{g(\psi)}(X_i) \\ &= \ell_n(g(\psi)) \end{aligned}$$

  so we just insert $\theta = g(\psi)$ in ℓ to obtain the new log likelihood.

  Example - Bernoulli trials:

  Let $\mu = \frac{e^\psi}{1+e^\psi}$, then $\psi = \log\frac{\mu}{1-\mu}$

$$\begin{aligned} \widetilde{\ell}_n(\psi) &= T_n\psi + n\log\frac{1}{1+e^\psi} \\ &= T_n\psi - n\log(1+e^\psi) \end{aligned}$$

- **Data transformation.** Consider a 1-1 transformation $Y_i = h(X_i)$, which is used instead of X_i.

  Continuous case We assume now that h is differentiable. Then Y_i has probability density function

$$f_\theta(h^{-1}(y))\frac{dx}{dy}(y),$$

  so the new log likelihood is

$$\begin{aligned} \widetilde{\ell}_n(\theta) &= \widetilde{\ell}(\theta, Y_1, Y_2, \ldots, Y_n) \\ &= \sum_{i=1}^{n} \log\{f_\theta(h^{-1}(Y_i))\frac{dx_i}{dy_i}(Y_i)\} \\ &= \sum_{i=1}^{n} \log\{f_\theta(X_i)\} + \sum_{i=1}^{n} \log\frac{dx_i}{dy_i}(Y_i) \\ &= \ell(\theta; X_1, X_2, \ldots, X_n) + \text{const.}, \end{aligned}$$

where "const." does not depend on θ. We use only $\dot{\ell}_n(\theta)$ and $\ddot{\ell}_n(\theta)$ and differences between log likelihood values (or likelihood ratios, see below) etc., so the constant may be disregarded. Hence: **Data transformation does not alter the likelihood**.

The same conclusion is obtained in the **discrete case**, because the probability mass function of Y is $f_\theta(h^{-1}(y)) = f_\theta(x)$, so $\widetilde{\ell}_n(\theta) = \ell_n(\theta)$.

Example - Bernoulli trials: Let $Y_i = 1 - X_i$, then $f_\mu(y) = \mu^{1-y}(1-\mu)^y$ for $y = 0, 1$.

$$\begin{aligned}\widetilde{\ell}_n(\mu) &= \sum_{i=1}^{n}\{(1-Y_i)\log\mu + Y_i\log(1-\mu)\} \\ &= \sum_{i=1}^{n}\{X_i\log\mu + (1-X_i)\log(1-\mu)\} \\ &= \ell_n(\mu)\end{aligned}$$

For the next result we assume Regularity conditions 1. (θ identifiable) and 2. (the $f_\theta(x)$ have common support).

Jensen's inequality If g is a strictly convex function, and X is a random variable with $\mathrm{E}(|X|) < \infty$ such that the distribution of X is not degenerate, then $g(\mathrm{E}(X)) < \mathrm{E}\,[g(X)]$. If instead g is strictly concave, then $g(\mathrm{E}(X)) > \mathrm{E}\,[g(X)]$.

Theorem

$$P_{\theta_0}\left(L_n(\theta_0) > L_n(\theta)\right) \to 1 \quad \text{as} \quad n \to \infty$$

for any fixed $\theta \neq \theta_0$.

Proof (See HMC p. 322). The inequality $L_n(\theta_0) > L_n(\theta)$ is equivalent to

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0$$

Now by the Law of Large Numbers

$$R_n(\theta) \xrightarrow{P} \mathrm{E}_{\theta_0}\left\{\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right\} = m,$$

say. We note that for $\theta \neq \theta_0$ the distribution of $f_\theta(X_i)/f_{\theta_0}(X_i)$ is not degenerate, because (in the continuous case) we have

$$\mathrm{E}_{\theta_0}\left\{\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right\} = \int\frac{f_\theta(x)}{f_{\theta_0}(x)}f_{\theta_0}(x)dx = \int f_\theta(x)dx = 1$$

and $f_\theta(x)/f_{\theta_0}(x) = 1$ almost surely would imply $\theta = \theta_0$ by the identifiability of θ. The discrete case is similar.

We hence apply Jensen's inequality to the strictly convex function $g(x) = -\log x$, which yields

$$m = \mathrm{E}_{\theta_0}\left\{\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right\} < \log\mathrm{E}_{\theta_0}\left\{\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right\} = \log 1 = 0$$

Taking $\varepsilon = -m > 0$, the Law of Large Numbers implies that

$$P_{\theta_0}\left(R_n(\theta) \geq 0\right) = P_{\theta_0}\left(R_n(\theta) \geq m + \varepsilon\right) \leq P_{\theta_0}\left(|R_n(\theta) - m| \geq \varepsilon\right) \to 0 \text{ as } n \to \infty$$

Hence $P_{\theta_0}\{R_n(\theta) < 0\} \to 1$ as $n \to \infty$ which implies the desired conclusion.

Since $L_n(\theta_0) > L_n(\theta)$ with high probability for n large, we conclude that $L_n(\theta)$ will tend to have its maximum near θ_0, the true value of θ. This motivates the idea of maximum likelihood estimation, to be introduced below.

2.2 The score function and the Fisher information function

We now assume that the function $\theta \mapsto f_\theta(x)$ is twice continuously differentiable.

- Define the **score function** (random function) by

$$\begin{aligned} U_n(\theta) &= U_n(\theta, X_1, X_2, \ldots, X_n) \\ &= \dot{\ell}_n(\theta) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i) \end{aligned}$$

  This function is also known as the *efficient score*.

- Define the **Fisher information function** also called *expected information* by

$$I_n(\theta) = \mathrm{Var}_\theta\{U_n(\theta)\}$$

  Note that $I_n(\theta)$ is a function from Ω into $[0, \infty)$. $I_n(\theta)$ is known as the *intrinsic accuracy* in the physics literature.

- **Properties** (under regularity conditions):

  1. $\mathrm{E}_\theta\{U_n(\theta)\} = 0$ (first Bartlett identity)
  2. $\mathrm{Var}_\theta\{U_n(\theta)\} = \mathrm{E}_\theta\{U_n^2(\theta)\}$
  3. $I_n(\theta) = -\mathrm{E}_\theta\{\ddot{\ell}_n(\theta)\} = -\mathrm{E}_\theta\{\dot{U}_n(\theta)\}$ (second Bartlett identity)

  Example Bernoulli trials Let $\theta = \log \frac{\mu}{1-\mu}$ (actually θ is the ψ from above). Then

$$\begin{aligned} \ell_n(\theta) &= T_n\theta - n\log(1 + e^\theta) \\ U_n(\theta) &= T_n - n\frac{e^\theta}{1 + e^\theta} \\ \mathrm{E}_\theta\{U_n(\theta)\} &= \mathrm{E}_\theta(T_n) - n\frac{e^\theta}{1 + e^\theta} = 0 \end{aligned}$$

  because $\mathrm{E}_\theta(T_n) = n\mu = n\frac{e^\theta}{1+e^\theta}$.

$$\begin{aligned} I_n(\theta) &= \mathrm{Var}_\theta\left(T_n - n\frac{e^\theta}{1+e^\theta}\right) \\ &= \mathrm{Var}_\theta(T_n) \\ &= n\mu(1-\mu) \\ &= n\frac{e^\theta}{(1+e^\theta)^2} \end{aligned}$$

- **Regularity conditions:** (see Cox and Hinkley (1974), p.281) Assume that $\frac{\partial}{\partial\theta}$ and $\int$ can be interchanged (or $\frac{\partial}{\partial\theta}$ and $\sum$ in the discrete case). We know

$$\int f_\theta(x)dx = 1 \quad \text{for} \quad \theta \in \Omega$$

so that for θ in the interior of Ω

$$\frac{\partial}{\partial\theta}\int f_\theta(x)dx = 0$$

By the regularity condition,

$$\int \frac{\partial}{\partial\theta} f_\theta(x)dx = 0$$

or

$$\int \frac{\partial}{\partial\theta}\log f_\theta(x) f_\theta(x)dx = 0$$

because

$$\frac{\partial}{\partial\theta}\log f_\theta(x) = \frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}.$$

Hence

$$\mathrm{E}_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\right\} = 0$$

The proof in the discrete case is similar. Now

$$\begin{aligned} \mathrm{E}_\theta\{U_n(\theta)\} &= \mathrm{E}_\theta\left[\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_\theta(X_i)\right] \\ &= \sum_{i=1}^n \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right] = 0 \end{aligned} \tag{2.1}$$

Hence, by the shortcut formula,

$$\mathrm{Var}_\theta\left[U_\theta\right] = \mathrm{E}_\theta\left[U_n^2(\theta)\right] \tag{2.2}$$

Differentiating (2.1) once more we obtain

$$\begin{aligned} 0 &= \int \frac{\partial^2}{\partial\theta^2} \log f_\theta(x) f_\theta(x) + \left(\frac{\partial}{\partial\theta} \log f_\theta(x) \right)^2 f_\theta(x) dx \\ &= \int \frac{\partial^2}{\partial\theta^2} \log f_\theta(x) f_\theta(x) dx + \int \left(\frac{\partial}{\partial\theta} \log f_\theta(x) \right)^2 f_\theta(x) dx \\ &= \mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_1) \right\} + \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial\theta} \log f_\theta(X_1) \right]^2 \right\} \end{aligned}$$

Hence

$$\begin{aligned} I_n(\theta) &= \sum_{i=1}^n \mathrm{Var}_\theta \left[\frac{\partial}{\partial\theta} \log f_\theta(X_i) \right] \\ &= \sum_{i=1}^n \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial\theta} \log f_\theta(X_i) \right]^2 \right\} \\ &= -\sum_{i=1}^n \mathrm{E}_\theta \left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i) \right] \\ &= -\mathrm{E}_\theta \sum_{i=1}^n \left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i) \right] \\ &= -\mathrm{E}_\theta \left[\ddot{\ell}_n(\theta) \right] \end{aligned} \tag{2.3}$$

Note that $I_n(\theta) = ni(\theta)$, where $i(\theta) = \mathrm{E}_\theta\{(\frac{\partial}{\partial\theta} \log f_\theta(X_i))^2\}$. So the information of the sample is n times the information of a single observation.

Example - Bernoulli trials

$$\begin{aligned} I_n(\theta) &= \mathrm{Var}_\theta\{U_n(\theta)\} \\ &= \mathrm{Var}_\theta\{T_n\} \\ &= n\mu(1-\mu) \end{aligned}$$

Maximum information for $\mu = \frac{1}{2}$.

$$\ddot{\ell}_n(\theta) = -n \frac{e^\theta}{(1+e^\theta)^2} = -n\mu(1-\mu)$$

so

$$I_n(\theta) = -\mathrm{E}_\theta\{\ddot{\ell}_n(\theta)\}.$$

2.3 Observed information

- **Definition:** The observed information for θ (a stochastic function) is defined by

$$J_n(\theta) = -\ddot{\ell}_n(\theta).$$

By (2.3) we have

$$I_n(\theta) = \mathrm{E}_\theta\{J_n(\theta)\}.$$

Moreover, since

$$J_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i),$$

we have, by the Law of Large Numbers

$$\frac{1}{n} J_n(\theta) \xrightarrow{P} i(\theta) = -\mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i) \right\}.$$

- **Reparametrization** Let $\theta = g(\psi)$ and assume that g 1-1 and differentiable, then ψ has observed information $\tilde{J}_n(\psi)$, where

$$\begin{aligned}
\tilde{J}(\psi) &= -\tilde{\ddot{\ell}}_n(\psi) = -\frac{\partial^2}{\partial\psi^2}\ell_n(g(\psi)) \\
&= -\frac{\partial}{\partial\psi}\left\{\dot{\ell}_n(g(\psi))\frac{\partial\theta}{\partial\psi}\right\} \\
&= -\ddot{\ell}_n(g(\psi))\left(\frac{\partial\theta}{\partial\psi}\right)^2 - \dot{\ell}_n(g(\psi))\frac{\partial^2\theta}{\partial\psi^2}
\end{aligned}$$

Hence

$$\tilde{J}_n(\psi) = J_n(g(\psi))\left(\frac{\partial\theta}{\partial\psi}\right)^2 - U_n(g(\psi))\frac{\partial^2\theta}{\partial\psi^2}$$

and

$$\tilde{I}_n(\psi) = I_n(g(\psi))\left(\frac{\partial\theta}{\partial\psi}\right)^2$$

because $\mathrm{E}_\theta\{U_n(\theta)\} = 0$.

- **Example - Bernoulli trials.** Find $J_n(\mu)$ for $\theta = \log\frac{\mu}{1-\mu} = g(\mu)$. Recall that $U_n(\theta) = T_n - n\frac{e^\theta}{1+e^\theta}$ and $J_n(\theta) = n\frac{e^\theta}{(1+e^\theta)^2} = n\mu(1-\mu)$. Hence

$$\begin{aligned}
\theta &= g(\mu) = \log\mu - \log(1-\mu) \\
\frac{\partial\theta}{\partial\mu} &= \frac{1}{\mu} + \frac{1}{1-\mu} = \frac{1}{\mu(1-\mu)} \\
\frac{\partial^2\theta}{\partial\mu^2} &= \frac{2\mu-1}{\mu^2(1-\mu)^2} \\
\tilde{J}_n(\mu) &= n\mu(1-\mu)\frac{1}{\mu^2(1-\mu)^2} - (T_n - n\mu)\frac{2\mu-1}{\mu^2(1-\mu)^2}
\end{aligned}$$

and

$$\tilde{I}_n(\mu) = \frac{n}{\mu(1-\mu)}$$

which now has minimum for $\mu = \frac{1}{2}$.

- **Example - Uniform distribution.** Consider the uniform distribution on $(0, \theta)$ with PDF

$$f_\theta(x) = \theta^{-1} 1_{(0,\theta)}(x)$$

This is an example, where the regularity conditions are not satisfied, because the support depends on θ. This means that although

$$\frac{\partial}{\partial\theta} \int f_\theta(x) dx = 0,$$

the left-hand side

$$\begin{aligned} \frac{\partial}{\partial\theta} \int f_\theta(x) dx &= \frac{\partial}{\partial\theta} \int_0^\theta \theta^{-1} dx \\ &= \theta^{-1} + \int_0^\theta \frac{\partial}{\partial\theta} \theta^{-1} dx \\ &= \theta^{-1} - \int_0^\theta \theta^{-2} dx \end{aligned}$$

contains the extra term θ^{-1} due to an application of the chain rule. Let $X_{(n)} = \max\{X_1, \ldots, X_n\}$. The likelihood is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n \left[\theta^{-1} 1_{(0,\theta)}(X_i)\right] \\ &= \theta^{-n} 1_{(0,\theta)}(X_{(n)}) \end{aligned}$$

Note that $P_\theta\left(X_{(n)} < \theta\right) = 1$, so that $L_n(\theta) = \theta^{-n}$ for $\theta \geq X_{(n)}$. The likelihood is hence decreasing for $\theta \geq X_{(n)}$, and zero to the left of $X_{(n)}$, and the maximum likelihood estimator is $\widehat{\theta}_n = X_{(n)}$. Let us now go through the standard calculations, and see what goes wrong, if anything. The log likelihood is

$$\ell_n(\theta) = -n \log\theta \text{ for } \theta \geq X_{(n)}$$

The score function is

$$U_n(\theta) = -n/\theta \text{ for } \theta \geq X_{(n)}$$

with mean

$$\mathrm{E}_\theta\left(U_n(\theta)\right) = -n/\theta$$

which is not zero, so the first Bartlett identity is not satisfied. Moreover,

$$I_n\left(\theta\right) = \mathrm{Var}_\theta\left(U_n(\theta)\right) = 0$$

which is disturbing, because zero Fisher information in principle implies that the sample contains no information about the parameter θ. The observed information, however, is

$$J_n(\theta) = -n/\theta^2$$

which is negative, and $\mathrm{E}_\theta(J(\theta)) = -n/\theta^2$, so the second Bartlett identity also is not satisfied. The good news is that the maximum likelihood estimator $\widehat{\theta}_n = X_{(n)}$ is a reasonably good estimate for θ. Note, however, that $\widehat{\theta}_n$ does not satisfy the likelihood equation $U_n(\theta) = 0$, and neither $I_n(\theta)$ nor $J_n(\theta)$ seem to express the information in the sample about θ in any reasonable way.

2.4 The Cramér-Rao inequality

Theorem: If $\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2, \ldots, X_n)$ is an *unbiased* estimator of θ, then

$$\mathrm{Var}_\theta(\tilde{\theta}_n) \geq I_n^{-1}(\theta).$$

Proof: (See Silvey, 1975 p. 36) Let $f_\theta(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n; \theta)$. By unbiasedness, $\mathrm{E}_\theta(\tilde{\theta}_n) = \theta$, or

$$\int \tilde{\theta}_n(\boldsymbol{x}) f_\theta(\boldsymbol{x}) d\boldsymbol{x} = \theta$$

Differentiating with respect to θ and interchanging $\frac{\partial}{\partial\theta}$ and $\int$ (given regularity conditions) we have

$$\int \tilde{\theta}_n(\boldsymbol{x}) \frac{\partial}{\partial\theta} f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 1$$

or

$$\int \tilde{\theta}_n(\boldsymbol{x}) U_n(\theta) f_\theta(\boldsymbol{x}) d\boldsymbol{x} = 1, \qquad U_n(\theta) = \frac{\frac{\partial}{\partial\theta} f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x})}.$$

Hence

$$\begin{aligned} 1 &= \mathrm{E}_\theta(\tilde{\theta}_n U_n(\theta)) \\ &= \mathrm{Cov}_\theta(\tilde{\theta}_n, U_n(\theta)), \end{aligned}$$

because $\mathrm{E}_\theta\{U_n(\theta)\} = 0$. By the Cauchy-Schwarz inequality we obtain

$$1 = \mathrm{Cov}_\theta^2(\tilde{\theta}_n, U_n(\theta)) \leq \mathrm{Var}_\theta(\tilde{\theta}_n)\mathrm{Var}_\theta(U_n(\theta)).$$

Since $\mathrm{Var}_\theta(U_n(\theta)) = I_n(\theta)$, the inequality follows.

The quantity $I_n^{-1}(\theta)$ is called the **Cramér-Rao Lower Bound**. An unbiased estimator $\tilde{\theta}_n$ with $\mathrm{Var}_\theta(\tilde{\theta}_n) = I_n^{-1}(\theta)$ is called an *efficient* estimator. If $\tilde{\theta}_n$ is unbiased, but not necessarily efficient, we call

$$\mathrm{Eff}_\theta(\tilde{\theta}_n) = \frac{I_n^{-1}(\theta)}{\mathrm{Var}_\theta(\tilde{\theta}_n)}$$

the **efficiency** of $\tilde{\theta}_n$. An efficient estimator is hence the same as an estimator with efficiency 1 for all θ If the estimator is biased, the bias should be taken into account (see Cox and Hinkley (1974), p. 254), by defining the mean square error (MSE) as follows

$$\mathrm{MSE}_\theta(\tilde{\theta}_n) = \mathrm{E}\left[\left(\tilde{\theta}_n - \theta\right)^2\right].$$

Example - Poisson distribution Po(θ) with PMF

$$f_\theta(x) = \frac{\theta^x}{x!}e^{-\theta} \quad x = 0, 1, 2, \ldots$$

The log likelihood

$$\ell_n(\theta) = \sum_{i=1}^{n} X_i \log\theta - n\theta - \sum_{i=1}^{n} \log\left(X_i!\right)$$

The score function

$$U_n(\theta) = \sum_{i=1}^{n} X_i\theta^{-1} - n$$

The Fisher information

$$I_n(\theta) = \mathrm{Var}_\theta(U_n(\theta)) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}$$

The maximum likelihood estimator defined by $U_n(\hat{\theta}) = 0$, is $\hat{\theta}_n = \bar{X}_n$, and since $\mathrm{E}_\theta\left(\bar{X}_n\right) = \theta$, $\hat{\theta}_n$ is unbiased. Its variance is

$$\mathrm{Var}_\theta(\hat{\theta}_n) = \frac{1}{n}\theta = \frac{1}{I_n(\theta)}$$

so the estimator is efficient (the Cramér-Rao lower bound is attained).

Example - Geometric distribution has PMF

$$f_\theta(x) = \theta(1-\theta)^x, \qquad x = 0, 1, \ldots \qquad 0 < \theta < 1.$$

Define $\tilde{\theta}_1$ (for $n = 1$) by

$$\tilde{\theta}_1(x) = \begin{cases} 1 & \text{for} \quad x = 0 \\ 0 & \text{for} \quad x > 0 \end{cases}$$

Then $\tilde{\theta}_1$ is unbiased,

$$\mathrm{E}_\theta(\tilde{\theta}_1) = 1\theta + \sum_{x=1}^{\infty} 0\theta(1-\theta)^x = \theta$$

and

$$\mathrm{E}_\theta(\tilde{\theta}_1^2) = \mathrm{E}_\theta(\tilde{\theta}_1) = \theta$$

Hence $\mathrm{Var}_\theta(\tilde{\theta}) = \theta - \theta^2 = \theta(1-\theta)$. The log likelihood is

$$\ell_1(\theta) = \log\theta + X_1\log(1-\theta)$$

The score function is

$$U_1(\theta) = \theta^{-1} - X_1/(1-\theta)$$

The Fisher information is

$$I_1(\theta) = \mathrm{Var}_\theta(U_1(\theta)) = \frac{1-\theta}{\theta^2(1-\theta)^2} = \frac{1}{\theta^2(1-\theta)}.$$

Hence, by the Cramér-Rao inequality

$$\mathrm{Var}_\theta(\tilde{\theta}_1) \geq \frac{1}{I_1(\theta)}$$

or

$$\theta(1-\theta) \geq \theta^2(1-\theta)$$

which is indeed satisfied for $0 < \theta < 1$. The efficiency of $\tilde{\theta}_1$ is

$$\mathrm{Eff}_\theta(\tilde{\theta}_1) = \frac{I_1^{-1}(\theta)}{\mathrm{Var}_\theta(\tilde{\theta}_1)} = \frac{\theta^2(1-\theta)}{\theta(1-\theta)} = \theta$$

Hence the efficiency may be anywhere between 0 and 1, depending on the value of θ. Similar conclusions may be reached for $n > 1$ as well.

Note: In the course ST802: Estimating Functions we learn about estimating functions, which are random functions that correspond to unbiased estimating equations, which in turn define a large variety of estimators, including maximum likelihood estimators and unbiased estimators. We note that not all maximum likelihood estimators are unbiased, and not all unbiased estimators are maximum likelihood estimators. One of the main topics of ST802 is to show that maximum likelihood estimators are optimal, in a suitable sense, among all estimating function estimators.

3 Asymptotic likelihood theory

3.1 Asymptotic normality of the score function

Main result:

$$\frac{U_n(\theta)}{\sqrt{n}} \xrightarrow{D} \mathrm{N}(0, i(\theta)) \text{ as } n \to \infty, \tag{3.1}$$

where

$$i(\theta) = \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f_\theta(X_1) \right]^2 \right\}.$$

We may also write, for n large

$$U_n(\theta) \sim \mathrm{N}(0, ni(\theta)) = \mathrm{N}(0, I_n(\theta)) \text{ approx.}$$

Proof: Note that

$$U_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i).$$

Since this is a sum of i.i.d. random variables (for any given θ), and since

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right]=0$$

and

$$\mathrm{Var}_\theta\{\frac{\partial}{\partial\theta}\log f_\theta(X_i)\}=i(\theta),$$

the Central Limit Theorem gives

$$\sqrt{n}\left\{\frac{U_n(\theta)}{n}-0\right\}\xrightarrow{D}\mathrm{N}(0,i(\theta)),$$

as desired. This result will be used in the next section.

3.2 The maximum likelihood estimator

The maximum likelihood estimator $\hat{\theta}_n$ is defined as the value of θ that maximizes $\ell_n(\theta)$ in Ω, i.e. satisfies

$$\ell_n(\hat{\theta}_n)\geq\ell_n(\theta)\qquad\text{for any}\quad\theta\quad\text{in}\quad\Omega.$$

In most cases of interest $\hat{\theta}_n$ is a local maximum in the interior of Ω, and satisfies $\dot{\ell}_n(\hat{\theta}_n)=0$ that is

$$U_n(\hat{\theta}_n)=0,$$

which we call the **likelihood equation**. Often it is the random variable $J_n(\hat{\theta}_n)$ rather than the random function $J_n(\theta)$ that is called the **observed information**. Note that if $\hat{\theta}_n$ is a local maximum we have $J_n(\hat{\theta}_n)>0$. There may be problems with $\hat{\theta}_n$ on the boundary of Ω, but this is more common in the discrete case. Note that if $\theta=g(\psi)$, where g is 1-1 and differentiable, the likelihood equation becomes

$$U_n(g(\hat{\psi}_n))\frac{\partial\theta}{\partial\psi}=0$$

so that $\hat{\psi}_n=g^{-1}(\hat{\theta}_n)$.

Example - Bernoulli trials Let $\theta=\log\frac{\mu}{1-\mu}$, then the score function is given by

$$U_n(\theta)=T_n-n\frac{e^\theta}{1+e^\theta},$$

which gives the likelihood equation

$$T_n/n=\bar{X}_n=\frac{e^\theta}{1+e^\theta}$$

with solution $\hat{\theta}_n=\log\frac{\bar{X}_n}{1-\bar{X}_n}$. Since $\mu=\frac{e^\theta}{1+e^\theta}$ we obtain $\hat{\mu}_n=\bar{X}_n$. If $T_n=0$ or $T_n=n$, the likelihood equation $U_n(\theta)=0$ has no solution. However, $\hat{\mu}_n=\bar{X}_n$ is valid even if $T_n=0$ or n, and maximizes the likelihood.

We shall now show that the maximum likelihood estimator has the following two properties:

1. $\hat{\theta}_n$ is *consistent*, that is

$$\hat{\theta}_n \xrightarrow{P} \theta \text{ as } n \to \infty,$$

where $\xrightarrow{P}$ denotes convergence in probability under P_θ. By definition this means

$$\forall \epsilon > 0 \quad P_\theta(|\hat{\theta}_n - \theta| \geq \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

We also say that $\hat{\theta}_n$ is asymptotically unbiased, although this terminology is somewhat imprecise.

2. $\hat{\theta}_n$ is asymptotically normal and asymptotically efficient,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}\left(0, \frac{1}{i(\theta)}\right)$$

Since $i(\theta) = I_n(\theta)/n$ we have

$$\mathrm{Var}_\theta(\hat{\theta}_n) = I_n^{-1}(\theta), \text{ approx.}$$

which is the basis for the claim of asymptotic efficiency.

These properties show that $\hat{\theta}_n$ is essentially the best available estimator for θ when the sample size is large. The performance of $\hat{\theta}_n$ is generally good also for small samples, and $\hat{\theta}_n$ is widely used, although often we need to investigate the behavior of $\hat{\theta}_n$ for small samples via simulation.

3.3 Exponential families

Before proving properties 1. and 2. in the general case, we consider exponential families, where the proof of these properties is much simpler. Let X have probability density/mass function

$$f_\theta(x) = a(x)e^{\theta x - \kappa(\theta)}, \quad x \in \mathbb{R}$$

where θ is the canonical parameter with domain Ω.

Theorem $\mathrm{E}_\theta(X) = \dot{\kappa}(\theta)$ and $\mathrm{Var}_\theta(X) = \ddot{\kappa}(\theta)$.

Proof (Continuous case). Since $\int f_\theta(x)dx = 1$ we have

$$M(\theta) = e^{\kappa(\theta)} = \int a(x)e^{\theta x}dx$$

It may be shown that in this case $\int$ and $\frac{\partial}{\partial\theta}$ can be interchanged, so

$$\dot{M}(\theta) = \int xa(x)e^{\theta x}dx$$

and

$$\ddot{M}(\theta) = \int x^2a(x)e^{\theta x}dx.$$

Hence

$$\frac{\dot{M}(\theta)}{M(\theta)} = \int xa(x)e^{\theta x - \kappa(\theta)}dx = \mathrm{E}_\theta(X)$$

and

$$\frac{\ddot{M}(\theta)}{M(\theta)} = \int x^2 a(x)e^{\theta x - \kappa(\theta)}dx = \mathrm{E}_\theta(X^2)$$

Since $\kappa(\theta) = \log M(\theta)$, we find

$$\dot{\kappa}(\theta) = \frac{\dot{M}(\theta)}{M(\theta)} = \mathrm{E}_\theta(X)$$

and

$$\begin{aligned}
\ddot{\kappa}(\theta) &= \frac{\ddot{M}(\theta)M(\theta) - \dot{M}^2(\theta)}{M^2(\theta)} \\
&= \frac{\ddot{M}(\theta)}{M(\theta)} - \left[\frac{\dot{M}(\theta)}{M(\theta)}\right]^2 \\
&= \mathrm{E}_\theta(X^2) - \mathrm{E}_\theta^2(X) \\
&= \mathrm{Var}_\theta(X)
\end{aligned}$$

Now look at the log likelihood for $X_1, X_2, \ldots, X_n$ i.i.d.

$$\ell_n(\theta) = \sum_{i=1}^{n} \log a(X_i) + \theta T_n - n\kappa(\theta),$$

and score function

$$U_n(\theta) = T_n - n\dot{\kappa}(\theta).$$

The likelihood equation is

$$\frac{1}{n}T_n = \dot{\kappa}(\theta)$$

or

$$\bar{X}_n = \dot{\kappa}(\theta)$$

and

$$\begin{aligned}
I_n(\theta) &= \mathrm{Var}_\theta(U_n(\theta)) \\
&= \mathrm{Var}_\theta(T_n) \\
&= n\ddot{\kappa}(\theta)
\end{aligned}$$

Note that $-J_n(\theta) = \ddot{\ell}_n(\theta) = -n\ddot{\kappa}(\theta)$, confirming that $I_n(\theta) = \mathrm{E}_\theta(J_n(\theta))$. Now define $\tau(\theta) = \dot{\kappa}(\theta)$. Since $\ddot{\kappa}(\theta) = \mathrm{Var}_\theta(X) > 0$ we obtain

$$\dot{\tau}(\theta) = \ddot{\kappa}(\theta) > 0,$$

so that τ is strictly increasing, and the solution to

$$\bar{X}_n = \tau(\theta)$$

is unique, if it exists, and is

$$\hat{\theta}_n = \tau^{-1}(\bar{X}_n)$$

where τ^{-1} is increasing and differentiable.

We now look at the two asymptotic properties of $\hat{\theta}_n$.

1. **Consistency.** By the Law of Large Numbers

$$\bar{X}_n \xrightarrow{P} \mathrm{E}_\theta(X_1) = \tau(\theta) \quad \text{as} \quad n \to \infty$$

and since τ^{-1} is continuous,

$$\hat{\theta}_n = \tau^{-1}(\bar{X}_n) \xrightarrow{P} \tau^{-1}(\tau(\theta)) = \theta \quad \text{as} \quad n \to \infty$$

Hence $\hat{\theta}_n$ is consistent. Before continuing, we recall the Δ-method. Assume that

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathrm{N}(0, \sigma^2) \quad \text{as} \quad n \to \infty$$

Let $\psi_n = g(\bar{X}_n)$ where g is differentiable. Then

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{D} \mathrm{N}(0, \sigma^2 \dot{g}^2(\mu)), \text{ as} \quad n \to \infty$$

which follows from the expansion

$$g(\bar{X}_n) - g(\mu)) \approx \dot{g}(\mu)(\bar{X}_n - \mu).$$

2. **Efficiency and asymptotic normality.** By the CLT we have

$$\sqrt{n}(\bar{X}_n - \tau(\theta)) \xrightarrow{D} \mathrm{N}(0, \ddot{\kappa}(\theta)) \quad \text{as} \quad n \to \infty$$

because $\ddot{\kappa}(\theta) = \mathrm{Var}_\theta(X_1)$. Since τ^{-1} is differentiable, with derivative

$$\frac{\partial}{\partial x}\tau^{-1}(x) = \frac{1}{\dot{\tau}(\tau^{-1}(x))} = \frac{1}{\ddot{\kappa}(\tau^{-1}(x))}$$

we obtain

$$\begin{aligned} \sqrt{n}(\tau^{-1}(\bar{X}_n) - \tau^{-1}(\tau(\theta))) &\xrightarrow{D} \mathrm{N}\left(0, \frac{\ddot{\kappa}(\theta)}{\ddot{\kappa}(\theta)^2}\right) \\ &= \mathrm{N}\left(0, \frac{1}{\ddot{\kappa}(\theta)}\right) = \mathrm{N}\left(0, \frac{1}{i(\theta)}\right), \end{aligned}$$

and in particular $\hat{\theta}_n$ is asymptotically efficient. We now pass from the canonical parameter θ to a general parameter ψ defined by $\theta = g(\psi)$ where g is 1-1 and differentiable. Then, $\hat{\theta}_n = g(\hat{\psi}_n)$. $\hat{\psi}_n$ is consistent, because g is continuous. The expected information for ψ is $(n = 1)$

$$\begin{aligned} \tilde{\imath}(\psi) &= i(\theta)\left(\frac{\partial\theta}{\partial\psi}\right)^2 \\ &= i(g(\psi))\dot{g}^2(\psi). \end{aligned}$$

By the Δ-method applied to $g^{-1}(\hat{\theta}_n)$ we obtain

$$\begin{aligned}\sqrt{n}(g^{-1}(\hat{\theta}_n) - g^{-1}(\theta)) &\xrightarrow{D} \mathrm{N}\left(0, \frac{1}{i(\theta)\dot{g}^2(\theta)}\right) \\ &= \mathrm{N}\left(0, \frac{1}{\tilde{i}(\psi)}\right) \quad \text{as} \quad n \to \infty,\end{aligned}$$

so

$$\sqrt{n}(\hat{\psi}_n - \psi) \xrightarrow{D} \mathrm{N}\left(0, \frac{1}{\tilde{i}(\psi)}\right) \quad \text{as} \quad n \to \infty$$

Hence we have asymptotic normality and efficiency of $\hat{\psi}_n$. Let us apply this to the parameter

$$\mu = \mathrm{E}_\theta(X) = \tau(\theta)$$
$$\hat{\mu}_n = \tau(\hat{\theta}_n) = \bar{X}_n$$

Since $\mathrm{E}_\theta(X) = \tau(\theta)$, $\hat{\mu}_n$ is unbiased. Now $\dot{\tau}(\theta) = \ddot{\kappa}(\theta)$, so

$$\begin{aligned}\sqrt{n}(\hat{\mu}_n - \mu) &\xrightarrow{D} \mathrm{N}\left(0, \frac{\dot{\tau}^2(\theta)}{\ddot{\kappa}(\theta)}\right) \\ = \mathrm{N}(0, \ddot{\kappa}(\theta)) \quad &\text{as} \quad n \to \infty\end{aligned}$$

Since $\mathrm{Var}_\theta(\hat{\mu}_n) = \frac{1}{n}\ddot{\kappa}(\theta)$, the Cramér-Rao lower bound is attained.

Example - Normal ($X_i \sim \mathrm{N}(\mu, 1)$, i.i.d.) Look at the PDF

$$\begin{aligned}f_\mu(x) &= (2\pi)^{-1/2} e^{-\frac{1}{2}(x-\mu)^2} \\ &= (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2 - \frac{1}{2}\mu^2 + x\mu\right) \\ &= (2\pi)^{-1/2} e^{-\frac{1}{2}x^2} \exp\left(x\mu - \frac{1}{2}\mu^2\right)\end{aligned}$$

We can identify this as a natural exponential family with $\theta = \mu$ and

$$\begin{aligned}\kappa(\theta) &= \frac{1}{2}\theta^2 \\ \dot{\kappa}(\theta) &= \theta = \mu \\ \ddot{\kappa}(\theta) &= 1\end{aligned}$$

$$\begin{aligned}\ell_n(\mu) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (X_i - \mu)^2 \\ &= \text{const.} + T_n\mu - \frac{n}{2}\mu^2 \\ U_n(\mu) &= \sum_{i=1}^n (X_i - \mu) = T_n - n\mu \\ \hat{\mu}_n &= \frac{1}{n}T_n = \bar{X}_n \sim \mathrm{N}(\mu, 1/n) \\ I_n(\mu) &= J_n(\mu) = \mathrm{Var}_\mu(U_n(\mu)) = n \\ i(\mu) &= 1\end{aligned}$$

so the asymptotic distribution

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{D} \mathrm{N}(0, 1)$$

is exact for all $n = 1, 2, \ldots$. Note that $\hat{\mu}_n$ is unbiased, i.e. $\mathrm{E}_\mu(\hat{\mu}_n) = \mu$ and $\mathrm{Var}_\mu(\hat{\mu}_n) = 1/I_n(\mu)$ (CR lower bound is attained).

Example - Exponential distribution, parameter θ

$$f_\theta(x) = \theta e^{-\theta x}, \quad x > 0$$

$$\begin{aligned} \ell_n(\theta) &= n \log \theta - \theta T_n \\ U_n(\theta) &= \frac{n}{\theta} - T_n \\ I_n(\theta) &= \frac{n}{\theta^2} \\ \hat{\theta}_n &= \frac{n}{T_n}. \end{aligned}$$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}(0, \theta^2)$$

Example - Poisson $\mathrm{Po}(\theta)$

$$\hat{\theta}_n = \bar{X}_n$$

$$i(\theta) = 1/\theta$$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}(0, \theta)$$

Example - Binomial distribution

$$\hat{\mu}_n = \bar{X}_n$$

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{D} \mathrm{N}(0, \mu(1-\mu)) \quad \text{as} \quad n \to \infty$$

Example - Geometric distribution with PMF

$$f_\theta(x) = \theta(1-\theta)^x \text{ for } x = 1, 2, \ldots$$

has log likelihood, score function etc.

$$\begin{aligned} \ell_n(\theta) &= n \log \theta + \sum_{i=1}^{n} X_i \log(1-\theta) \\ U_n(\theta) &= \frac{n}{\theta} - \frac{T_n}{1-\theta} \\ \hat{\theta}_n &= \frac{n}{T_n + n} \\ i(\theta) &= \frac{1}{\theta^2(1-\theta)} \end{aligned}$$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}(0, \theta^2(1-\theta)) \quad \text{as} \quad n \to \infty$$

3.4 Consistency of the maximum likelihood estimator

In the following we let θ_0 denote the true value of θ, and we assume that the distributions $f_\theta(x)$ have common support.

Recall that we have shown the following theorem above.

Theorem

$$P_{\theta_0}\left(\ell_n(\theta_0) > \ell_n(\theta)\right) \to 1 \quad \text{as} \quad n \to \infty$$

for any fixed $\theta \neq \theta_0$.

Theorem Consistency (Lehmann (1998) p. 413) With probability tending to 1 as $n \to \infty$, the likelihood equation has a solution $\hat{\theta}_n$ which is consistent.

Proof Let $\delta > 0$ be such that $\theta_0 - \delta$ and $\theta_0 + \delta$ are both in Ω, and define

$$A_n = \{x : \ell_n(\theta_0) > \ell_n(\theta_0 - \delta) \quad \text{and} \quad \ell_n(\theta_0) > \ell_n(\theta_0 + \delta)\}.$$

Then by the previous theorem we may conclude that $P_{\theta_0}(A_n) \to 1$ as $n \to \infty$. In fact, first note that for events A and B, we have

$$P(A^c \cap B^c) = 1 - P(A \cup B) \geq 1 - P(A) - P(B)$$

This implies that

$$\begin{aligned} P_{\theta_0}(A_n) &\geq 1 - P_{\theta_0}\left(\ell_n(\theta_0) \leq \ell_n(\theta_0 - \delta)\right) - P_{\theta_0}\left(\ell_n(\theta_0) \leq \ell_n(\theta_0 + \delta)\right) \\ &\to 1 \text{ as } n \to \infty. \end{aligned}$$

Hence, for any $x \in A_n$ there exists a $\hat{\theta}_n(\delta) \in (\theta_0 - \delta, \theta_0 + \delta)$ such that $\dot{\ell}_n(\hat{\theta}_n(\delta)) = 0$ and $\hat{\theta}_n(\delta)$ is a local maximum for $\ell_n(\theta)$. Hence

$$P_{\theta_0}\left(|\hat{\theta}_n(\delta) - \theta_0| < \delta\right) \geq P_{\theta_0}(A_n) \to 1 \quad \text{as} \quad n \to \infty.$$

We finally need to determine a sequence which does not depend on δ. Let $\hat{\theta}_n$ be the root closest to θ_0. Then, for any $\delta > 0$

$$P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| < \delta\right) \to 1 \quad \text{as } n \to \infty$$

which proves the consistency.

We note that $\hat{\theta}_n$ is *not* necessarily the maximum likelihood estimator, but we shall work with $\hat{\theta}_n$ from now on. However, $\hat{\theta}_n$ is in fact a stationary point of ℓ_n.

Corollary: If $\hat{\theta}_n$ is unique, then it is consistent. Provided that $\hat{\theta}_n$ is a solution to the likelihood equation, this follows from the above proof.

3.5 Efficiency and asymptotic normality

Now assume that there exists a function $M(x)$ such that

$$\left|\frac{\partial^3}{\partial \theta^3} \log f_\theta(x)\right| < M(x) \quad \text{for all} \quad x,$$

and $\mathrm{E}_\theta(M(X)) < \infty$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, 1/i(\theta_0)) \quad \text{as} \quad n \to \infty$$

Proof: Expand $\dot{\ell}_n(\hat{\theta}_n)$ around θ_0

$$\dot{\ell}_n(\hat{\theta}_n) = \dot{\ell}_n(\theta_0) + (\hat{\theta}_n - \theta_0)\ddot{\ell}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \dddot{\ell}(\theta_n^*),$$

where θ_n^* lies between θ_0 and $\hat{\theta}_n$. The left-hand side is zero, so

$$\begin{aligned} 0 &= U_n(\theta_0) - (\hat{\theta}_n - \theta_0) J_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \dddot{\ell}(\theta_n^*) \\ &= U_n(\theta_0) - (\hat{\theta}_n - \theta_0)\left[J_n(\theta_0) - \frac{1}{2}(\hat{\theta}_n - \theta_0) \dddot{\ell}(\theta_n^*) \right] \end{aligned}$$

Hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2} U_n(\theta_0)}{\frac{1}{n} J_n(\theta_0) - \frac{1}{2}(\hat{\theta}_n - \theta_0)\frac{1}{n}\dddot{\ell}(\theta_n^*)}$$

We now use the following facts:

1. $n^{-1/2} U_n(\theta_0) \xrightarrow{D} \mathrm{N}(0, i(\theta_0))$ as $n \to \infty$ (shown above).
2. $\frac{1}{n} J_n(\theta_0) \xrightarrow{P} i(\theta_0)$ as $n \to \infty$ by the Law of Large Numbers.
3. $\frac{1}{n}\dddot{\ell}(\theta_n^*)$ is bounded in probability

$$\begin{aligned} \left| \frac{1}{n} \dddot{\ell}(\theta_n^*) \right| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f_{\theta_n^*}(X_i) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} M(X_i) \xrightarrow{P} \mathrm{E}_{\theta_0}\left[M(X_1) \right]. \end{aligned}$$

Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ has the same asymptotic distribution as

$$\frac{U_n(\theta_0)}{\sqrt{n} i(\theta_0)} \quad \text{which is} \quad \mathrm{N}(0, 1/i(\theta_0))$$

We have used that $\hat{\theta}_n \xrightarrow{P} \theta_0$ so that $\hat{\theta}_n - \theta_0 \xrightarrow{P} 0$, which is hence bounded in probability. Note also that

$$\mathrm{Var}\left(\frac{U_n(\theta_0)}{\sqrt{n} i(\theta_0)} \right) = \frac{n i(\theta_0)}{n i^2(\theta_0)} = \frac{1}{i(\theta_0)}$$

3.6 The Weibull distribution

Consider the Weibull distribution with PDF

$$f_\theta(x) = \theta x^{\theta-1} e^{-x^\theta} \text{ for } x > 0, \tag{3.2}$$

where $\theta > 0$ is a parameter. It is easy to show that (3.2) is a PDF, by the substitution $z = x^\theta$. Let $T_n = \sum_{i=1}^n \log X_i$, where $X_1, \ldots, X_n$ are i.i.d. from the Weibull distribution. Then we may write the log likelihood as follows:

$$\ell_n(\theta) = n \log \theta + (\theta - 1) T_n - \sum_{i=1}^n e^{\theta \log X_i}.$$

The score function is

$$U_n(\theta) = \frac{n}{\theta} + T_n - \sum_{i=1}^n (\log X_i)\, e^{\theta \log X_i}$$

and the observed information is

$$J_n(\theta) = \frac{n}{\theta^2} + \sum_{i=1}^n (\log X_i)^2\, e^{\theta \log X_i} > 0 \tag{3.3}$$

so the log likelihood is concave. Hence there is at most one root of the likelihood equation, and this root is the maximum likelihood estimator. First note that $U_n(\theta)$ goes to ∞ as θ goes to zero. Also, note that we may write $U_n(\theta)$ as follows:

$$\begin{aligned} U_n(\theta) &= \frac{n}{\theta} + \sum_{i=1}^n \log X_i - \sum_{i=1}^n (\log X_i)\, e^{\theta \log X_i} \\ &= \frac{n}{\theta} + \sum_{i=1}^n \log X_i \left(1 - e^{\theta \log X_i}\right). \end{aligned}$$

In the special case where $X_i = 1$ for all i, the last term is zero and we obtain $U_n(\theta) = \frac{n}{\theta}$, which goes to 0 as θ goes to ∞. So in this case (which has probability zero) the likelihood equation has no root and the maximum likelihood estimate is $\theta = \infty$, which is outside of the parameter domain for θ. Except for this special case, we observe that the terms $\log X_i \left(1 - e^{\theta \log X_i}\right)$ are all negative, because $\log X_i$ and $1 - e^{\theta \log X_i}$ have opposite signs. In this case, the asymptotic value of $U_n(\theta)$ as θ goes to ∞ is either $-\infty$ (if at least one $X_i > 1$) or goes to a negative constant (if all $X_i \leq 1$ and at least one $X_i < 1$). Hence, with probability 1, there is a unique root $\widehat{\theta}_n$ of the likelihood equation $U_n(\theta) = 0$.

Now let us find the unit Fisher information. Using (3.3) we find using the substitution $y = x^\theta$

that

$$\begin{aligned} i(\theta) &= \frac{1}{\theta^2} + \int_0^\infty (\log x)^2 \, x^\theta \theta x^{\theta-1} e^{-x^\theta} \, dx \\ &= \frac{1}{\theta^2} + \int_0^\infty \left(\log y^{1/\theta}\right)^2 y e^{-y} dy \\ &= \frac{1}{\theta^2} + \frac{1}{\theta^2} \int_0^\infty (\log y)^2 \, y e^{-y} dy \\ &= \frac{1}{\theta^2} \left[\frac{1}{6}\pi^2 + (\gamma - 1)^2\right] \\ &= \frac{1}{\theta^2} 1.9781\ldots, \end{aligned}$$

where $\gamma = 0.577221\ldots$ is Euler's constant, using for example Maple. The maximum likelihood estimator $\widehat{\theta}_n$ is consistent and asymptotically normal and efficient,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{D} \mathrm{N}(0, 1/i(\theta)).$$

3.7 Location models

Example: Information for a location family (HMC p. 329–330). Consider the location model with PDF

$$f_\theta(x) = f(x - \theta) \text{ for } x \in \mathbb{R},$$

where $\theta \in \mathbb{R}$ and f is a given PDF. It is useful to let $h(x) = -\log f(x)$, or $f(x) = e^{-h(x)}$. For a random sample $X_1, \ldots, X_n$ from the location model, we obtain the log likelihood

$$\ell_n(\theta) = -\sum_{i=1}^n h(X_i - \theta).$$

The score function is

$$U_n(\theta) = \sum_{i=1}^n \dot{h}(X_i - \theta)$$

where dots denote derivatives, and the observed information is

$$J_n(\theta) = \sum_{i=1}^n \ddot{h}(X_i - \theta).$$

Hence, we note that if h is strictly convex, so that $\ddot{h}(x) > 0$ for all $x \in \mathbb{R}$, we obtain $J_n(\theta) > 0$, and so the log likelihood is strictly concave, and there is at most one root of the likelihood equation. In general, h may of course not be strictly convex, in which case that discussion of maximum likelihood is more involved.

We note that the first Bartlett identity takes the form

$$\begin{aligned} \mathrm{E}_\theta(\dot{h}(X_1 - \theta)) &= \int_{-\infty}^\infty \dot{h}(x - \theta) e^{-h(x-\theta)} \, dx \\ &= \int_{-\infty}^\infty \dot{h}(z) e^{-h(z)} \, dz = 0, \end{aligned}$$

where we have used the substitution $z = x - \theta$. Note that the equation is trivially satisfied if h is an even function ($h(x) = h(-x)$) and $\dot{h}$ hence an odd function ($\dot{h}(x) = -\dot{h}(-x)$). Now let us find the unit Fisher information (also called the Intrinsic Accuracy)

$$\begin{aligned} i(\theta) &= \mathrm{E}_\theta(\ddot{h}(X_1 - \theta)) \\ &= \int_{-\infty}^{\infty} \ddot{h}(x-\theta) e^{-h(x-\theta)}\, dx \\ &= \int_{-\infty}^{\infty} \ddot{h}(x) e^{-h(x)}\, dx. \end{aligned}$$

An alternative expression, obtained from the second Bartlett identity, is

$$\begin{aligned} i(\theta) &= \mathrm{E}_\theta(\dot{h}^2(X_1 - \theta)) \\ &= \int_{-\infty}^{\infty} \dot{h}^2(x-\theta) e^{-h(x-\theta)}\, dx \\ &= \int_{-\infty}^{\infty} \dot{h}^2(z) e^{-h(z)}\, dz. \end{aligned}$$

As an example we consider the Cauchy distribution with

$$f(x) = \frac{1}{\pi\,(1+x^2)},$$

for which

$$h(x) = \log \pi + \log(1 + x^2)$$

and

$$\dot{h}(x) = \frac{2x}{1+x^2}$$

In this case, h is not convex. In fact

$$\ddot{h}(x) = \frac{2\left(1-x^2\right)}{\left(1+x^2\right)^2},$$

which changes sign for $x = \pm 1$.

Using for example Maple we obtain

$$i(\theta) = \int_{-\infty}^{\infty} \frac{4z^2}{\pi\,(1+z^2)^3}\, dz = \frac{1}{2}.$$

The maximum likelihood estimator $\widehat{\theta}_n$ is consistent and asymptotically normal and efficient, but the likelihood equation $U_n(\theta) = 0$ may have other roots than $\widehat{\theta}_n$.

4 Vector parameters

4.1 The score vector and the Fisher information matrix

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with probability density/mass function $f_{\boldsymbol{\theta}}(x)$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top \in \Omega$, where Ω is a region in $\mathbb{R}^p$.

Example - $X_i \sim \mathrm{Ga}(\theta, \lambda)$, with density function

$$f_{\boldsymbol{\theta}}(x) = \frac{\theta^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\theta x} \quad \text{for } x > 0 \quad \text{where} \quad \boldsymbol{\theta} = (\theta, \lambda)^\top \in \mathbb{R}^2_+.$$

The likelihood function $L_n : \Omega \to [0, \infty)$ is now a random function of vector argument defined by

$$L(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}) = \prod_{i=1}^n f_{\boldsymbol{\theta}}(X_i) \quad \text{for } \boldsymbol{\theta} \in \Omega$$

The log likelihood function $\ell_n : \Omega \to \mathbb{R}$ is a random function of vector argument defined by

$$\ell(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_{\boldsymbol{\theta}}(X_i) \quad \text{for } \boldsymbol{\theta} \in \Omega$$

The **score vector** $\boldsymbol{U} : \Omega \to \mathbb{R}^p$ is a random vector function

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_p} \end{pmatrix} \quad p \times 1 \quad \text{vector.}$$

We sometimes use the gradient notation

$$\boldsymbol{U}(\boldsymbol{\theta}) = \nabla_\theta \ell(\boldsymbol{\theta})$$

Here and in the following, we often drop the subscript n, and instead $U_j(\boldsymbol{\theta})$ will denote the jth component of $\boldsymbol{U}(\boldsymbol{\theta})$ etc. The score vector satisfies the Bartlett identity

$$\mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{U}(\boldsymbol{\theta})\} = \mathbf{0},$$

that is $\mathrm{E}_{\boldsymbol{\theta}}\{\frac{\partial \ell}{\partial \theta_j}\} = 0$ for $j = 1, \ldots, p$.

Expected Information Matrix Definition

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta}) &= \mathrm{Var}_{\boldsymbol{\theta}}\{\boldsymbol{U}(\boldsymbol{\theta})\} \quad p \times p \quad \text{matrix} \\ &= \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}^\top(\boldsymbol{\theta})\} \\ I_{jk}(\boldsymbol{\theta}) &= \mathrm{Cov}_{\boldsymbol{\theta}}\{U_j(\boldsymbol{\theta}), U_k(\boldsymbol{\theta})\} \\ &= \mathrm{E}_{\boldsymbol{\theta}}\{U_j(\boldsymbol{\theta})U_k(\boldsymbol{\theta})\} \end{aligned}$$

Reparametrization $\boldsymbol{\theta} = g(\boldsymbol{\psi})$, g : 1-1 differentiable gives the score vector

$$\tilde{\boldsymbol{U}}(\boldsymbol{\psi}) = \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\psi}} \boldsymbol{U}(\boldsymbol{\theta})$$

and expected information matrix

$$\tilde{\boldsymbol{I}}(\boldsymbol{\psi}) = \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\psi}} \boldsymbol{I}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}^\top}$$

The observed Information Matrix

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\theta}) &= -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \quad p \times p \text{ matrix} \\ \boldsymbol{J}_{jk}(\boldsymbol{\theta}) &= -\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} \end{aligned}$$

Second Bartlett identity

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{J}(\boldsymbol{\theta})\}$$

4.2 Cramér-Rao inequality (generalized)

Define $I^{jk}(\boldsymbol{\theta}) = \{\boldsymbol{I}^{-1}(\boldsymbol{\theta})\}_{jk}$. If $\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ_1, i.e.

$$\mathrm{E}_{\boldsymbol{\theta}}\{\tilde{\theta}_n\} = \theta_1,$$

then

$$\mathrm{Var}_{\boldsymbol{\theta}}\{\tilde{\theta}_n\} \geq I^{11}(\boldsymbol{\theta})$$

See Cox and Hinkley (1974) p. 256. The proof is based on a generalized version of the Cauchy-Schwarz inequality.

Asymptotic Normality of the score function Recall that the score function

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_i)$$

is a sum of i.i.d. variables with mean zero,

$$\mathrm{E}_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_i) \right\} = \mathbf{0}$$

and variance equal to the unit Fisher information matrix

$$\mathrm{Var}_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_i) \right\} = \boldsymbol{i}(\boldsymbol{\theta}).$$

Recall that the total Fisher information matrix is $\boldsymbol{I}(\boldsymbol{\theta}) = n\boldsymbol{i}(\boldsymbol{\theta})$.

By the Multivariate Central Limit Theorem

$$\frac{1}{\sqrt{n}} \boldsymbol{U}(\theta) \xrightarrow{D} \mathrm{N}_p(0, \boldsymbol{i}(\boldsymbol{\theta}))$$

Note that since

$$\frac{1}{n}\boldsymbol{J}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f_{\boldsymbol{\theta}}(X_i)\right\}$$

then by the Law of Large Numbers

$$\frac{1}{n}\boldsymbol{J}(\boldsymbol{\theta}) \xrightarrow{P} -\mathrm{E}_{\boldsymbol{\theta}}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f_{\boldsymbol{\theta}}(X_i)\right\} = \boldsymbol{i}(\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE) The MLE $\hat{\boldsymbol{\theta}}_n \in \Omega$ is defined by $L(\hat{\boldsymbol{\theta}}_n) \geq L(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Omega$. In general $\widehat{\boldsymbol{\theta}}_n$ satisfies the likelihood equation

$$\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$$

or the p equations with p unknowns,

$$\begin{pmatrix} U_1(\boldsymbol{\theta}) = 0 \\ \vdots \\ U_p(\boldsymbol{\theta}) = 0 \end{pmatrix}$$

See picture of likelihood contours.

4.3 Consistency and asymptotic normality of the maximum likelihood estimator

Consistency ($\boldsymbol{\theta}_0$ = true value)

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0 \quad \text{as} \quad n \to \infty$$

that is

$$P_{\theta_0}\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| > \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty \quad \forall \epsilon > 0$$

Asymptotic normality, asymptotic efficiency

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \mathrm{N}_p(0, \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)) \quad \text{as} \quad n \to \infty$$

By the Cramér-Rao inequality, $\boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)$ is the "best" obtainable variance for an unbiased estimator; hence $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient (see ST802 notes).

Proof (sketch): From Lehmann (1998), p. 429–434. Recall from the one-parameter case that for any $\theta \neq \theta_0$

$$P_{\theta_0}\left(L_n(\theta_0) > L_n(\theta)\right) \to 1 \text{ as } n \to \infty.$$

Let Q_a denote the sphere, center $\boldsymbol{\theta}_0$, radius $a > 0$ for a small, there is high probability that

$$\ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0) \quad \text{for} \quad \boldsymbol{\theta} \in Q_a,$$

and hence $\ell(\boldsymbol{\theta})$ has a local maximum in the interior of Q_a, and this maximum satisfies the likelihood equation $\boldsymbol{U}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Hence we have shown that with probability tending to 1 there exists a root of $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ near $\boldsymbol{\theta}_0$, proving consistency.

Asymptotic normality: Expand $\boldsymbol{U}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$,

$$\boldsymbol{U}(\hat{\boldsymbol{\theta}}_n) \approx \boldsymbol{U}(\boldsymbol{\theta}_0) - \boldsymbol{J}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

where $\boldsymbol{\theta}_n^*$ is on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$. Now $\boldsymbol{U}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, so

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \left\{ \frac{1}{n} \boldsymbol{J}(\boldsymbol{\theta}_n^*) \right\}^{-1} \frac{1}{\sqrt{n}} \boldsymbol{U}(\boldsymbol{\theta}_0)$$

We have $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$, so $\boldsymbol{\theta}_n^* \to \boldsymbol{\theta}_0$ and hence $\frac{1}{n}\boldsymbol{J}(\boldsymbol{\theta}_n^*) \to \boldsymbol{i}(\boldsymbol{\theta}_0)$. By the asymptotic normality of $\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\theta}_0)$ we obtain

$$\boldsymbol{i}^{-1}(\boldsymbol{\theta}_0) \frac{1}{\sqrt{n}} \boldsymbol{U}(\boldsymbol{\theta}_0) \to \mathrm{N}_p(0, \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{i}(\boldsymbol{\theta}_0) \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)) = \mathrm{N}_p(0, \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)).$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \mathrm{N}(0, \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)) \quad \text{as} \quad n \to \infty.$$

As before, we interpret this as saying that, for n large, we have

$$\hat{\boldsymbol{\theta}}_n \sim \mathrm{N}_p(\boldsymbol{\theta}_0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)),$$

where $\boldsymbol{I}(\boldsymbol{\theta}_0) = n\boldsymbol{i}(\boldsymbol{\theta}_0)$ is the total Fisher information matrix.

Example - Normal distribution (HMC p. 354) $X_i \sim \mathrm{N}(\mu, \tau)$, $\boldsymbol{\theta} = (\mu, \tau) \in \mathbb{R} \times \mathbb{R}_+$ (note that HMC use σ as parameter, whereas we use $\tau = \sigma^2$). The log likelihood for a sample of size n is

$$\ell(\theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \tau - \frac{1}{2\tau} \sum_{i=1}^{n} (X_i - \mu)^2$$

Straightforward calculation give

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} (X_i - \mu)/\tau$$

$$U_2(\boldsymbol{\theta}) = -\frac{n}{2\tau} + \sum_{i=1}^{n} (X_i - \mu)^2/(2\tau^2)$$

and also

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\tau} & \sum_{i=1}^{n}(X_i - \mu)/\tau^2 \\ \sum_{i=1}^{n}(X_i - \mu)/\tau^2 & \sum_{i=1}^{n}(X_i - \mu)^2/\tau^3 - \frac{n}{2\tau^2} \end{pmatrix}$$

We hence obtain

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\tau} & 0 \\ 0 & \frac{n}{2\tau^2} \end{pmatrix}$$

and

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\tau}{n} & 0 \\ 0 & \frac{2\tau^2}{n} \end{pmatrix}$$

The maximum likelihood estimates are

$$\hat{\mu}_n = \bar{X}_n \quad \text{and} \quad \widehat{\tau}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

The Cramér-Rao lower bound for μ is τ/n, which is attained by $\hat{\mu}_n$. $\widehat{\tau}_n$ is not unbiased, but

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

is unbiased, $\mathrm{E}_\theta(S_n^2) = \tau$. The variance of S_n^2 is $\frac{2\tau^2}{n-1}$, which is bigger than the Cramér-Rao lower bound $\frac{2\tau^2}{n}$, but for n large, the difference is small. By the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\mu}_n \\ \widehat{\tau}_n\end{pmatrix} - \begin{pmatrix}\mu \\ \tau\end{pmatrix}\right\} \xrightarrow{D} \mathrm{N}_2\left(0, \begin{pmatrix}\tau & 0 \\ 0 & 2\tau^2\end{pmatrix}\right)$$

The exact distributions of $\hat{\mu}_n$ and $\widehat{\tau}_n$ are

$$\hat{\mu}_n \sim \mathrm{N}(\mu, \tau/n) \quad \text{and} \quad \widehat{\tau}_n \sim \frac{\tau}{n}\chi^2(n-1)$$

Example gamma distribution (continued) - $X_i \sim \mathrm{Ga}(\theta, \lambda)$, with density function

$$f_{\boldsymbol{\theta}}(x) = \frac{\theta^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\theta x} \quad \text{for } x > 0 \quad \text{where} \quad \boldsymbol{\theta} = (\theta, \lambda)^\top \in \mathbb{R}_+^2.$$

Log likelihood

$$\ell(\theta, \lambda) = n\lambda \log\theta - n\log\Gamma(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log X_i - \theta\sum_{i=1}^{n}X_i$$

In order to handle the derivative of the gamma function, we introduce the digamma function

$$\psi(\lambda) = \frac{d}{d\lambda}\log\Gamma(\lambda)$$

and the trigamma function

$$\psi_1(\lambda) = \frac{d^2}{d\lambda^2}\log\Gamma(\lambda)$$

Now the components of the score function are

$$U_1(\theta, \lambda) = \frac{n\lambda}{\theta} - \sum_{i=1}^{n}X_i$$

and

$$U_2(\theta, \lambda) = n\log\theta - n\psi(\lambda) + \sum_{i=1}^{n}\log X_i$$

The likelihood equations are hence equivalent to

$$\frac{\lambda}{\theta} = \overline{X}_n \tag{4.1}$$

$$\psi(\lambda) - \log\lambda = \overline{L}_n - \log\overline{X}_n, \tag{4.2}$$

where $\overline{L}_n$ denotes the average of $\log X_i$. The observed information matrix is

$$\boldsymbol{J}(\theta,\lambda) = n\begin{bmatrix} \frac{\lambda}{\theta^2} & \frac{1}{\theta} \\ \frac{1}{\theta} & \psi_1(\lambda) \end{bmatrix}$$

This matrix is non-random, and so $\boldsymbol{I}(\theta,\lambda) = \boldsymbol{J}(\theta,\lambda)$, and since $\boldsymbol{I}(\theta,\lambda)$ is positive-definite, it follows that the log likelihood is strictly concave. In particular, it follows that $\psi_1(\lambda) > 0$, and furthermore, the equation (4.2) has a unique solution.

4.4 Parameter orthogonality

Consider a statistical model parametrized by the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$. In case the Fisher information matrix is diagonal,

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} I_{11}(\boldsymbol{\theta}) & 0 \\ 0 & I_{22}(\boldsymbol{\theta}) \end{bmatrix},$$

the parameters θ_1 and θ_2 are said to be *orthogonal*. In this case the inverse Fisher information matrix is also diagonal,

$$\boldsymbol{I}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} 1/I_{11}(\boldsymbol{\theta}) & 0 \\ 0 & 1/I_{22}(\boldsymbol{\theta}) \end{bmatrix}.$$

It follows that the maximum likelihood estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are asymptotically independent, with asymptotic normal distributions

$$\widehat{\theta}_j \mathrel{\dot\sim} \mathrm{N}(\theta_j, 1/I_{jj}(\boldsymbol{\theta})). \tag{4.3}$$

Here we note that (4.3) implies, for example, that the asymptotic distribution of $\widehat{\theta}_1$ is the same, whether or not the second parameter θ_2 is considered known or not. This follows because $I_{11}(\boldsymbol{\theta})$ is the Fisher information for θ_1 when θ_2 is known, making the asymptotic distribution of $\widehat{\theta}_1$ to be $\widehat{\theta}_1 \mathrel{\dot\sim}\mathrm{N}(\theta_1, 1/I_{11}(\boldsymbol{\theta}))$.

The notion of parameter orthogonality may be generalized in various ways. If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ consists of p parameters, we say that $\theta_1, \ldots, \theta_p$ are orthogonal parameters if the Fisher information matrix for $\boldsymbol{\theta}$ is diagonal. If $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\top$ is a p-dimensional parameter consisting of two components $\boldsymbol{\theta}_1$ (q-dim) and $\boldsymbol{\theta}_2$ ($(p-q)$-dim), then the parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be orthogonal if the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$, when partitioned in blocks corresponding to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, is block diagonal. In these cases, the consequences are roughly speaking the same as above, namely that the components of $\boldsymbol{\theta}$ are asymptotically independent, and that the asymptotic distribution of one component does not depend on whether or not the other component or the remaining elements are known or not.

A further option is to use the observed information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ to define orthogonality. We can hence talk about the parameters θ_1 and θ_2 being *observed orthogonal*. If we want to distinguish the original concept of orthogonality, we talk about θ_1 and θ_2 being *expected orthogonal parameters*.

4.5 Exponential dispersion models

Consider the distribution with PDF or PMF defined by

$$f(x;\theta,\lambda) = a(x;\lambda)e^{\lambda[x\theta-\kappa(\theta)]}. \tag{4.4}$$

Note that (4.4) is a natural exponential family for each value of λ. The gamma and normal distributions are of this form. In particular we find that the mean and variance are

$$\begin{aligned} \mathrm{E}(X) &= \dot{\kappa}(\theta) && (4.5)\\ \mathrm{Var}(X) &= \lambda^{-1}\ddot{\kappa}(\theta) && (4.6) \end{aligned}$$

Taking $n = 1$ in the calculations, we obtain

$$\begin{aligned} \ell(\theta,\lambda) &= \log a(X;\lambda) + \lambda\left[X\theta - \kappa(\theta)\right] \\ &= c(X;\lambda) + \lambda\left[X\theta - \kappa(\theta)\right], \end{aligned}$$

say. The components of the score function are

$$\begin{aligned} U_1(\theta,\lambda) &= \lambda\left[X - \dot{\kappa}(\theta)\right] \\ U_2(\theta,\lambda) &= \dot{c}(X;\lambda) + X\theta - \kappa(\theta) \end{aligned}$$

where a dot denotes derivative with respect to λ. Note that (4.5) follows from the first Bartlett identity, and that (4.6) follows from the second Bartlett identity.

The observed information matrix is

$$\boldsymbol{J}(\theta,\lambda) = \begin{bmatrix} \lambda\ddot{\kappa}(\theta) & -\left[X - \dot{\kappa}(\theta)\right] \\ -\left[X - \dot{\kappa}(\theta)\right] & -\ddot{c}(X;\lambda) \end{bmatrix}.$$

Since $X - \dot{\kappa}(\theta)$ has mean zero, we obtain the following Fisher information matrix

$$\boldsymbol{I}(\theta,\lambda) = \begin{bmatrix} \lambda\ddot{\kappa}(\theta) & 0 \\ 0 & -\mathrm{E}_{\theta,\lambda}\left[\ddot{c}(X;\lambda)\right] \end{bmatrix}.$$

This is an example where θ and λ are orthogonal parameters, meaning that the Fisher information matrix is diagonal.

4.6 Linear regression

Let $Y_1,\ldots,Y_n$ be independent and assume that

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \tau) \text{ for } i = 1,\ldots,n,$$

where $x_1,\ldots,x_n$ are constants satisfying

$$x_1 + \cdots + x_n = 0. \tag{4.7}$$

This is the standard linear regression model with x being centered, so that $\overline{x} = 0$. Let us go through the likelihood calculations for this model.

The log likelihood for the three parameters is

$$\ell(\alpha, \beta, \tau) = -\frac{n}{2}\log(2\pi\tau) - \frac{1}{2\tau}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2 .$$

We now introduce the notation

$$\begin{aligned} S_Y &= Y_1 + \cdots + Y_n, \\ S_{xY} &= \sum_{i=1}^{n} x_i Y_i \\ S_{xx} &= \sum_{i=1}^{n} x_i^2. \end{aligned}$$

We assume that $S_{xx} > 0$, in other words that not all x_i are identical.

The first component of the score function is

$$\frac{\partial \ell}{\partial \alpha} = \frac{1}{\tau}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i) = \frac{1}{\tau}(S_Y - n\alpha),$$

where we have used (4.7). The solution to the first likelihood equation is hence

$$\widehat{\alpha} = \overline{Y}_n \tag{4.8}$$

with distribution $\mathrm{N}(\alpha, \tau/n)$.

The next component of the score function is

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\tau}\sum_{i=1}^{n} x_i (Y_i - \alpha - \beta x_i) = \frac{1}{\tau}(S_{xY} - \beta S_{xx}),$$

where once again we have used (4.7). The solution to the second likelihood equation is

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}. \tag{4.9}$$

with distribution $\mathrm{N}(\beta, \tau/S_{xx})$.

The third component of the score function is

$$\begin{aligned} \frac{\partial \ell}{\partial \tau} &= -\frac{n}{2\tau} + \frac{1}{2\tau^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2 \\ &= -\frac{n}{2\tau} + \frac{1}{2\tau^2}\left(S_{YY} + \alpha^2 + \beta^2 S_{xx} - 2\alpha S_Y - 2\beta S_{xY}\right), \end{aligned}$$

where we have used (4.7) once more. Inserting the solutions (4.8) and (4.9), the solution to the third likelihood equation is

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

We know from the theory of linear models that an unbiased estimator may be obtained as follows:

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

with distribution

$$\widetilde{\tau} \sim \frac{\tau}{n-2}\chi^2(n-2),$$

which has mean τ (being unbiased) and variance

$$\mathrm{Var}\,(\widetilde{\tau}) = \frac{2\tau^2}{n-2}. \tag{4.10}$$

We shall now show that the Fisher information matrix is diagonal, as follows:

$$\boldsymbol{I}(\alpha,\beta,\tau) = \begin{bmatrix} \frac{n}{\tau} & 0 & 0 \\ 0 & \frac{S_{xx}}{\tau} & 0 \\ 0 & 0 & \frac{n}{2\tau^2} \end{bmatrix},$$

making the three parameters orthogonal. The calculation of the entries of $\boldsymbol{I}(\alpha,\beta,\tau)$ goes as follows. The first two diagonal elements of the second derivative matrix are

$$\frac{\partial^2\ell}{\partial\alpha^2} = -\frac{n}{\tau}$$

$$\frac{\partial^2\ell}{\partial\beta^2} = -\frac{S_{xx}}{\tau}$$

which immediately give the first two diagonal elements of $\boldsymbol{I}(\alpha,\beta,\tau)$. The third diagonal element is

$$\frac{\partial^2\ell}{\partial\tau^2} = \frac{n}{2\tau^2} - \frac{1}{\tau^3}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2. \tag{4.11}$$

Since

$$\mathrm{E}\left[(Y_i - \alpha - \beta x_i)^2\right] = \tau,$$

it follows that the mean of (4.11) is

$$\mathrm{E}\left(\frac{\partial^2\ell}{\partial\tau^2}\right) = \frac{n}{2\tau^2} - \frac{n\tau}{\tau^3} = -\frac{n}{2\tau^2},$$

giving the third diagonal element of $\boldsymbol{I}(\alpha,\beta,\tau)$. We also need to show that the three mixed derivatives have mean zero. First note that

$$\frac{\partial^2\ell}{\partial\alpha\partial\beta} = 0.$$

Also note that the two mixed derivatives with respect to τ are

$$\frac{\partial^2\ell}{\partial\tau\partial\alpha} = -\frac{1}{\tau^2}(S_Y - n\alpha)$$

and

$$\frac{\partial^2 \ell}{\partial \tau \partial \beta} = -\frac{1}{\tau^2}\left(S_{xY} - \beta S_{xx}\right),$$

both of which have mean zero. This completes the calculation of the Fisher information matrix.

The inverse Fisher information matrix is now

$$\boldsymbol{I}^{-1}(\alpha, \beta, \tau) = \begin{bmatrix} \frac{\tau}{n} & 0 & 0 \\ 0 & \frac{\tau}{S_{xx}} & 0 \\ 0 & 0 & \frac{2\tau^2}{n} \end{bmatrix}.$$

It follows from the calculations above that $\widehat{\alpha}$ and $\widehat{\beta}$ are both minimum variance unbiased estimators. As regards the unbiased estimator $\widetilde{\tau}$, its variance (4.10) does not achieve the Cramér-Rao lower bound, but since $\widetilde{\tau}$ is a function of the sufficient statistic $(S_Y, S_{xY}, S_{YY})^\top$, it is a minimum variance unbiased estimator (see the next section).

4.7 Exercises

1. Find the Fisher information matrix in the regression model when $\overline{x}$ is not assumed to be zero.

2. Consider the unit logistic distribution (cf. HMC Example 6.1.2) with pdf $f(x) = e^{-x}/(1+e^{-x})^2$. Investigate the location-scale version of the logistic distribution, i.e. the family of PDFs $f((x-\mu)/\sigma)/\sigma$ for $\mu \in \mathbb{R}$, $\sigma > 0$, and develop the likelihood, score vector, information matrix, maximum likelihood estimation etc. Use the function $h = -\log f$ in the notation, as in Section 10.2 of the notes. Show that μ and σ are orthogonal parameters, i.e. the Fisher information matrix is diagonal. Show that the function h is convex, and hence show that the solution to the likelihood equations is unique. The following substitution may be useful in order to simplify the integrals: $z = (x-\mu)/\sigma$.

5 Sufficiency

See HMC, Sections 7.2–7.4.

5.1 Definition

Let us start with a motivating example.

Example: $X_i \sim \mathrm{N}(\mu, \tau)$ i.i.d. with log likelihood

$$\begin{aligned} \ell(\mu, \tau) &= -\frac{n}{2}\log(2\pi\tau) - \frac{1}{2\tau}\sum_{i=1}^{n}(X_i - \mu)^2 \\ &= -\frac{n}{2}\log(2\pi\tau) - \frac{1}{2\tau}\left(\sum_{i=1}^{n} X_i^2 + n\mu^2 - 2\mu\sum_{i=1}^{n} X_i\right) \end{aligned}$$

Hence the log likelihood is determined by

$$S_X = \sum_{i=1}^{n} X_i \text{ and } S_{XX} = \sum_{i=1}^{n} X_i^2$$

The statistic $(S_X, S_{XX})^\top$ is called a *sufficient statistic* for (μ, τ). If we write the log likelihood in terms of the sufficient statistic,

$$\ell(\mu, \tau) = -\frac{n}{2}\log(2\pi\tau) - \frac{1}{2\tau}\left(S_{XX} + n\mu^2 - 2\mu S_X\right)$$

then (in the present case), the data enter the log likelihood only via the sufficient statistic. We also note that the sufficient statistic (S_X, S_{XX}) has dimension 2, so it is a nice summary statistic, as compared to the full data $X_1, \ldots, X_n$, which form an n-dimensional vector.

Now suppose that τ is known to have the value 1, say. Then the log likelihood for μ is

$$\ell(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}S_{XX} - \frac{1}{2}n\mu^2 + \mu S_X$$

Then the constant $-\frac{n}{2}\log(2\pi) - \frac{1}{2}S_{XX}$ does not influence the shape of the likelihood, which is determined solely by S_X, which is now the sufficient statistic. Hence, the sufficient statistic depends on which parameters are considered unknown. In the present example, the dimensions of the parameter and the sufficient statistic are the same (two in the first case, and one in the second case), although this is not generally the case.

Let $X_1, \ldots, X_n$ be i.i.d. with PDF/PMF $f(x; \theta)$, for $\theta \in \Omega$. Let

$$Y_1 = u_1(X_1, \ldots, X_n)$$

be a statistic with PDF/PMF $f_{Y_1}(y; \theta)$.

Definition 7.2.1: The statistic Y_1 is called *sufficient* for the parameter θ if and only if

$$\frac{f(x_1; \theta) \cdots f(x_n; \theta)}{f_{Y_1}(u_1(x_1, \ldots, x_n); \theta)} = H(x_1, \ldots, x_n), \tag{5.1}$$

where $H(x_1, \ldots, x_n)$ is a function that does not depend on $\theta \in \Omega$.

Note that in the discrete case, the ratio at the left-hand side of (5.1) is the conditional probability for the event $\{X_1 = x_1, \ldots, X_n = x_n\}$ given $Y_1 = y_1$, provided that $y_1 = u_1(x_1, \ldots, x_n)$. In other words, the conditional PMF of $X_1, \ldots, X_n$ given Y_1 is

$$f(x_1, \ldots, x_n | y_1; \theta) = \frac{f(x_1; \theta) \cdots f(x_n; \theta)}{f_{Y_1}(u_1(x_1, \ldots, x_n); \theta)}$$

provided that $y_1 = u_1(x_1, \ldots, x_n)$, and zero otherwise.

In the continuous case the conditional PDF of $X_1, \ldots, X_n$ given Y_1 is proportional to the left-hand side of (5.1),

$$f(x_1, \ldots, x_n | y_1; \theta) \propto \frac{f(x_1; \theta) \cdots f(x_n; \theta)}{f_{Y_1}(u_1(x_1, \ldots, x_n); \theta)}$$

provided that $y_1 = u_1(x_1, \ldots, x_n)$, and zero otherwise. The proportionality constant is, roughly speaking, a Jacobian. Hence, we may interpret the definition of sufficiency as saying that the conditional distribution of $X_1, \ldots, X_n$ given Y_1 is the same for all values of $\theta \in \Omega$, i.e. is independent of θ.

Note that the sample $(X_1, \ldots, X_n)^\top$ (an n-dimensional statistic) is always sufficient, and hence, a sufficient statistic always exists.

Example (Gamma distribution) (continued) - $X_i \sim \text{Ga}(\theta, \lambda)$ i.i.d. Let us take $\lambda = 2$, corresponding to the density function

$$f(x;\theta) = \frac{\theta^2}{\Gamma(2)} x e^{-\theta x} \quad \text{for } x > 0$$

By using moment generating functions, we know that $Y_1 = X_1 + \cdots + X_n$ is $\text{Ga}(\theta, 2n)$, with PDF

$$f_{Y_1}(y_1;\theta) = \frac{\theta^{2n}}{\Gamma(2n)} y_1^{2n-1} e^{-\theta y_1} \quad \text{for } y_1 > 0$$

Now look at the ratio

$$\frac{\prod_{i=1}^{n} \frac{\theta^2}{\Gamma(2)} x_i e^{-\theta x_i}}{\frac{\theta^{2n}}{\Gamma(2n)} \left(x_1 + \cdots + x_n\right)^{2n-1} e^{-\theta(x_1 + \cdots + x_n)}} = \frac{\Gamma(2n)\left(x_1 \cdots x_n\right)}{\Gamma^n(2)\left(x_1 + \cdots + x_n\right)^{2n-1}},$$

which is independent of θ. Hence Y_1 is sufficient for θ. Note that again the dimension of the sufficient statistic and the parameter are the same in this example.

5.2 The Fisher-Neyman factorization criterion

How is the definition of sufficiency related to the idea that the log likelihood is fully determined by the sufficient statistic? This follows from a criterion due to Fisher, which Neyman later proved to be a characterization of sufficiency.

Theorem 7.2.1 (Neyman). The statistic $Y_1 = u_1(X_1, \ldots, X_n)$ is sufficient for θ if and only if

$$f(x_1;\theta) \cdots f(x_n;\theta) = k_1\left(u_1(x_1, \ldots, x_n);\theta\right) k_2\left(x_1, \ldots, x_n\right), \tag{5.2}$$

where $k_2\left(x_1, \ldots, x_n\right)$ does not depend on θ.

Note that the left-hand side of (5.2) is the likelihood, so that (5.2) may also be written as

$$L(\theta) \propto k_1\left(Y_1;\theta\right),$$

in the sense that the proportionality constant does not depend on θ, although it may depend on $X_1, \ldots, X_n$. Furthermore, the log likelihood takes the form

$$\ell(\theta) = \text{const.} \; + \log k_1\left(Y_1;\theta\right),$$

so it follows that the the score function

$$U(\theta) = \dot{\ell}(\theta) = \frac{\partial}{\partial\theta} \log k_1\left(Y_1;\theta\right)$$

depends on the data $X_1, \ldots, X_n$ only through $Y_1 = u_1(X_1, \ldots, X_n)$.

Proof: If Y_1 is sufficient for θ, it follows from (5.1) that

$$f(x_1;\theta)\cdots f(x_n;\theta) = f_{Y_1}(u_1(x_1,\ldots,x_n);\theta)H\left(x_1,\ldots,x_n\right),$$

which by the definition of sufficiency is of the form (5.2). Conversely, consider the discrete case, and assume that (5.2) is satisfied. Then

$$f_{Y_1}(y_1;\theta) = k_1\left(y_1;\theta\right) \sum_{y_1 = u_1(x_1,\ldots,x_n)} k_2\left(x_1,\ldots,x_n\right),$$

where the sum is over all $x_1, \ldots, x_n$ satisfying $y_1 = u_1(x_1, \ldots, x_n)$. It follows that

$$\begin{aligned}\frac{f(x_1;\theta)\cdots f(x_n;\theta)}{f_{Y_1}(u_1(x_1,\ldots,x_n);\theta)} &= \frac{k_1\left(u_1(x_1,\ldots,x_n);\theta\right)k_2\left(x_1,\ldots,x_n\right)}{k_1\left(u_1(x_1,\ldots,x_n);\theta\right)\sum_{y_1=u_1(x_1,\ldots,x_n)}k_2\left(x_1,\ldots,x_n\right)} \\ &= \frac{k_2\left(x_1,\ldots,x_n\right)}{\sum_{y_1=u_1(x_1,\ldots,x_n)}k_2\left(x_1,\ldots,x_n\right)},\end{aligned}$$

which by the assumption about k_2 does not depend on θ. Hence, Y_1 is sufficient for θ.

See HMC p. 384–385 for the proof in the continuous case.

Example 7.2.5. Consider $X_1, \ldots, X_n$ i.i.d. from the power distribution

$$f(x;\theta) = \theta x^{\theta-1} \text{ for } 0 < x < 1,$$

where $\theta > 0$. Consider the statistic $Y_1 = \prod_{i=1}^{n} X_i$. Then

$$f(x_1;\theta)\cdots f(x_n;\theta) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}$$

Since this is a function of the data through $y_1 = \prod_{i=1}^{n} x_i$ only, it follows from the factorization criterion that Y_1 is sufficient for θ.

Example (Weibull distribution). For the Weibull distribution of Section 3.6 we found the following score function:

$$U_n(\theta) = \frac{n}{\theta} + T_n - \sum_{i=1}^{n} \left(\log X_i\right) e^{\theta \log X_i}$$

which is clearly not a function of any statistic of small dimension. In this case, the full sample $(X_1, \ldots, X_n)^\top$ seems to be the best sufficient statistic we can have.

5.3 The Rao–Blackwell theorem

A useful result for sufficiency is obtained from a theorem due to Rao and Blackwell. Let us first review some basic properties of conditional expectations. If X and Y are random variables and X has expectation, then

$$\mathrm{E}\left[\mathrm{E}\left(X|Y\right)\right] = \mathrm{E}\left(X\right) \tag{5.3}$$

and if X has variance, then

$$\mathrm{Var}\left(X\right) = \mathrm{E}\left[\mathrm{Var}\left(X|Y\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(X|Y\right)\right].$$

It follows that

$$\mathrm{Var}\left(X\right) \geq \mathrm{Var}\left[\mathrm{E}\left(X|Y\right)\right] \tag{5.4}$$

The conditional mean is used in the following result.

Theorem (Rao-Blackwell). Let the statistic $Y_1 = u_1(X_1, \ldots, X_n)$ be sufficient for θ, and let $Y_2 = u_2(X_1, \ldots, X_n)$ be an unbiased estimator for θ. Then

$$\widetilde{\theta} = \mathrm{E}_\theta\left(Y_2|Y_1\right) \tag{5.5}$$

is also an unbiased estimator of θ, it is a function of Y_1, and $\mathrm{Var}_\theta\left(\widetilde{\theta}\right) \leq \mathrm{Var}_\theta\left(Y_2\right)$ for all $\theta \in \Omega$.

Proof. Since Y_1 is sufficient for θ, it follows from the discussion above, that the conditional distribution of Y_2 given Y_1 does not depend on θ. In particular $\mathrm{E}_\theta\left(Y_2|Y_1\right)$ does not depend on θ, and $\widetilde{\theta}$ is hence a statistic, i.e. a function of $X_1, \ldots, X_n$ that does not involve θ. Using (5.3) along with the unbiasedness of Y_2, we obtain

$$\mathrm{E}_\theta\left(\widetilde{\theta}\right) = \mathrm{E}_\theta\left[\mathrm{E}_\theta\left(Y_2|Y_1\right)\right] = \mathrm{E}_\theta\left(Y_2\right) = \theta$$

so that $\widetilde{\theta}$ is also unbiased. By using (5.4) we obtain

$$\mathrm{Var}_\theta\left(\widetilde{\theta}\right) = \mathrm{Var}_\theta\left[\mathrm{E}_\theta\left(Y_2|Y_1\right)\right] \leq \mathrm{Var}_\theta\left(Y_2\right),$$

as desired.

The operation (5.5) is called **Rao-Blackwellization**. This operation always improves upon a given unbiased estimator if possible, and gives an estimator that is a function of the sufficient statistic Y_1. However, if Y_2 is already a function of Y_1, then Rao-Blackwellization does not change Y_2.

Definition 7.1.1. A statistic $Y_2 = u_2(X_1, \ldots, X_n)$ is called a *Minimum Variance Unbiased Estimator* (MVUE) for θ if Y_2 is unbiased for θ, and if the variance of Y_2 is less than or equal to the variance of any other unbiased estimator for θ.

In order to find the MVUE, if it exists, the Rao-Blackwell theorem tells us that we should always look among the Rao-Blackwellized statistics, i.e. estimators that are a function of a sufficient statistic. There is also the question if the MVUE is unique, but since there may be more than one sufficient statistic, we cannot in general guarantee that there is a unique MVUE.

A separate question is if the maximum likelihood estimator could be an MVUE. The following result relates the maximum likelihood estimator to sufficiency.

Theorem 7.3.2. If the maximum likelihood estimator $\widehat{\theta}$ is uniquely determined from $X_1, \ldots, X_n$, and $Y_1 = u_1(X_1, \ldots, X_n)$ is a sufficient statistic, then $\widehat{\theta}$ is a function of Y_1.

Proof. From the definition of sufficiency, we find that the likelihood has the form

$$L(\theta) = f_{Y_1}(u_1(X_1, \ldots, X_n); \theta) H(X_1, \ldots, X_n)$$

The maximum likelihood estimator $\widehat{\theta}$ satisfies

$$L(\widehat{\theta}) \geq L(\theta)$$

for all θ, or, equivalently,

$$f_{Y_1}(Y_1; \widehat{\theta}) \geq f_{Y_1}(Y_1; \theta)$$

Hence, we can always determine from the value of the sufficient statistic Y_1 if $\widehat{\theta}$ is in fact a maximum likelihood estimator. If there is more than one maximum likelihood estimator, one could in principle select between these based on the value of a statistic that is not a function of Y_1. However, if $\widehat{\theta}$ is uniquely determined, then $\widehat{\theta}$ must be a function of Y_1.

Example 7.3.1. Let $X_1, \ldots, X_n$ be i.i.d. random variables from the exponential distribution with PDF

$$f(x; \theta) = \theta e^{-\theta x}, \quad x > 0$$

with parameter $\theta > 0$. Then

$$f(x_1; \theta) \cdots f(x_n; \theta) = \theta^n e^{-\theta(x_1 + \cdots + x_n)},$$

so at $Y_1 = X_1 + \cdots + X_n$ is sufficient for θ. The log likelihood is

$$\ell(\theta) = n \log \theta - \theta Y_1$$

and the score function is hence

$$U(\theta) = \frac{n}{\theta} - Y_1$$

which yields the maximum likelihood estimator

$$\hat{\theta}_n = \frac{n}{Y_1} = \frac{1}{\overline{X}_n}$$

Note that $X_i \sim \text{Ga}(\theta, 1)$, which implies that $Y_1 = X_1 + \cdots + X_n$ also has a gamma distribution,

$$Y_1 \sim \text{Ga}(\theta, n)$$

Hence, we may calculate the mean of $\hat{\theta}_n$ as follows:

$$\begin{aligned}
\mathrm{E}_\theta\left(\hat{\theta}_n\right) &= \mathrm{E}_\theta\left(\frac{n}{Y_1}\right) \\
&= n\mathrm{E}_\theta\left(\frac{1}{Y_1}\right) \\
&= n\int_0^\infty y^{-1}\frac{\theta^n}{\Gamma(n)}y^{n-1}e^{-\theta y}\,dy \\
&= \frac{n\theta^n}{(n-1)!}\int_0^\infty y^{n-1-1}e^{-\theta y}\,dy \\
&= \frac{n\theta^n}{(n-1)!}\frac{(n-2)!}{\theta^{n-1}} \\
&= \frac{n}{n-1}\theta
\end{aligned}$$

where we have used that $\Gamma(n) = (n-1)!$. It follows that the statistic

$$\frac{n-1}{n}\hat{\theta}_n = \frac{n-1}{Y_1}$$

is an unbiased estimator for θ. This estimator is a function of the sufficient statistic Y_1, and hence cannot be improved further by Rao-Blackwellization based on conditioning on Y_1. Later, we shall see that this estimator is in fact the UMVE for θ, but for now, all we can say is that it is the best estimator for θ based on Y_1.

In general, there is no unique sufficient statistic. For example, in the above example, both Y_1 and the full sample $(X_1, \ldots, X_n)^\top$ are sufficient statistics. We hence need a method for selecting the best sufficient statistic, in some sense, perhaps the statistic with the smallest dimension.

5.4 The Lehmann-Scheffé theorem

The following definition can help us to find the sufficient statistic that has the smallest dimension.

Definition 7.4.1. The family $\{f_{Y_1}(\cdot;\theta) : \theta \in \Omega\}$ is called *complete* if the condition

$$\mathrm{E}_\theta\left[u\left(Y_1\right)\right] = 0 \text{ for all } \theta \in \Omega$$

implies that $u(y) = 0$ except for a set which has probability zero with respect to $f_{Y_1}(\cdot;\theta)$ for all $\theta \in \Omega$. We shall also say that the statistic Y_1 is complete.

In the exponential family setting, completeness can often be determined by appeal to the properties of Laplace transforms (moment generating functions), see below. Here is a simple example.

Example 7.4.1. Assume that Y_1 is exponentially distributed, i.e.

$$f_{Y_1}(y;\theta) = \theta e^{-\theta y} \text{ for } y > 0.$$

where $\theta > 0$. Then the condition $\mathrm{E}_\theta\left[u\left(Y_1\right)\right] = 0$ for all $\theta \in \Omega$ means

$$\theta\int_0^\infty u(y)\,e^{-\theta y}\,dy = 0 \text{ for all } \theta > 0. \tag{5.6}$$

The integral on the left-hand side is the Laplace transform of the function $u(y)$, so (5.6) implies that $u(y) = 0$ almost everywhere on $\mathbb{R}_+$, which shows that the family of exponential distributions is complete. Note that the behaviour of $u(y)$ for $y < 0$ can be arbitrary, which is of no consequence, because $\mathbb{R}_-$ has probability zero with respect to any member of the family of exponential distributions. In general, we need only determine $u(y)$ on the support of the distribution.

Example. Let us consider the Bernoulli distribution with PMF

$$f_{Y_1}(y;\mu) = \mu^y(1-\mu)^{1-y} \text{ for } y = 0, 1.$$

Let u be a function defined on the support of Y_1, i.e.

$$u(y) = \begin{cases} u_0 & \text{for} \quad y = 0 \\ u_1 & \text{for} \quad y = 1 \end{cases}$$

Then

$$\mathrm{E}_\theta[u(Y_1)] = (1-\mu)u_0 + \mu u_1 = u_0 + \mu(u_1 - u_0) \tag{5.7}$$

Hence, $\mathrm{E}_\mu[u(Y_1)] = 0$ for all $\mu \in (0,1)$ implies that the linear function (5.7) is zero for all $\mu \in (0,1)$. This, in turn, implies that both coefficients u_0 and $u_1 - u_0$ are zero, i.e. $u_0 = u_1 = 0$. Hence, the function $u(y)$ is zero on the support $\{0,1\}$.

Example. Consider an i.i.d. sample $X_1, \ldots, X_n$ from the normal distribution $\mathrm{N}(\mu, \mu)$ for $\mu > 0$ with equal mean and variance. Then the statistic $(\overline{X}_n, S_n^2)$ is sufficient, but

$$\mathrm{E}_\mu\left(\overline{X}_n - S_n^2\right) = 0 \text{ for all } \mu > 0.$$

Hence the statistic $(\overline{X}_n, S_n^2)$ is not complete, because we have found a nontrivial function of it with mean zero for all parameter values.

The next result, due to Lehmann and Scheffé, links sufficiency with completeness to produce a unique MVUE estimator.

Theorem (Lehmann-Scheffé). Let $\widetilde{\theta}$ be an unbiased estimator of θ such that $\widetilde{\theta}$ is a function of a complete sufficient statistic Y_1. Then $\widetilde{\theta}$ is the unique MVUE of θ.

Proof. By assumption $\widetilde{\theta} = u(Y_1)$ and $\mathrm{E}_\theta\left(\widetilde{\theta}\right) = \theta$ for all $\theta \in \Omega$. Let $\widetilde{\theta}_1 = v(Y_1)$ be unbiased, so that $\mathrm{E}_\theta\left(\widetilde{\theta}_1\right) = \theta$ for all $\theta \in \Omega$. By Rao-Blackwell, there is no loss of generality in assuming that $\widetilde{\theta}_1$ is a function of Y_1, because this can only make its variance smaller. Then

$$\mathrm{E}_\theta\left(\widetilde{\theta} - \widetilde{\theta}_1\right) = \mathrm{E}_\theta[u(Y_1) - v(Y_1)] = 0 \text{ for all } \theta \in \Omega.$$

By the completeness of Y_1 we conclude that $u(Y_1) - v(Y_1) = 0$ almost surely for all $\theta \in \Omega$, i.e. $\widetilde{\theta}_1 = \widetilde{\theta}$ almost surely. Hence, $\widetilde{\theta}$ is the unique minimum variance unbiased estimator for θ.

Example (Exponential families). Let us consider a family of distributions with PDF/PMF

$$f(x;\theta) = a(x)\exp\left[\theta^\top u(x) - \kappa(\theta)\right] \tag{5.8}$$

where $u(x)$ is a k-dimensional statistic and $\Omega \subseteq \mathbb{R}^k$. The cumulant function κ is defined by

$$\kappa(\theta) = \log \int a(x)\exp\left[\theta^\top u(x)\right] dx \text{ for } \theta \in \Omega$$

thereby guaranteeing that (5.8) is a PDF, with a similar definition in the discrete case. A family of distributions of this form is called an *exponential family* with canonical parameter θ, canonical parameter domain Ω, and canonical statistic $u(X)$. If the canonical parameter domain Ω contains an open set, then the statistic $u(X)$ is complete. This may be shown by appeal to the uniqueness of the moment generating function.

By the reparametrization $\theta = g(\psi)$ we obtain the family

$$f(x;\psi) = a(x)\exp\left[g^\top(\psi)u(x) - \kappa(g(\psi))\right]$$

in which case the completeness of $u(X)$ follows if the domain for $g^\top(\psi)$ contains an open set.

It is interesting to consider the proof of completeness in the case of a natural exponential family

$$f(x;\theta) = a(x)\exp\left[\theta x - \kappa(\theta)\right] \tag{5.9}$$

Then the equation $\mathrm{E}_\theta\left(t(X)\right) = 0$ for all $\theta \in \Omega$ for some statistic t implies

$$\int t(x)a(x)\exp\left[\theta x - \kappa(\theta)\right]\,dx = 0 \text{ for all } \theta \in \Omega$$

or

$$\int t(x)a(x)e^{\theta x}\,dx = 0 \text{ for all } \theta \in \Omega$$

If Ω contains an open interval, then the uniqueness of the Laplace transform implies that $t(x)a(x) = 0$ nearly everywhere, which translates into $t(X) = 0$ almost surely, thereby implying that X is complete.

Now, consider an i.i.d. sample $X_1, \ldots, X_n$ from the natural exponential family (5.9) with joint PDF/PMF

$$\begin{aligned} f(\boldsymbol{x};\theta) &= \prod_{i=1}^n \left\{a(x_i)\exp\left[\theta x_i - \kappa(\theta)\right]\right\} \\ &= \prod_{i=1}^n a(x_i)\exp\left[\theta\left(x_1 + \cdots + x_n\right) - n\kappa(\theta)\right] \end{aligned}$$

Let us transform to the joint distribution of $Y_1, X_2, \ldots, X_n$, where $Y_1 = X_1 + \cdots + X_n$, giving

$$f(y_1, x_2, \ldots, x_n;\theta) = a\left(y_1 - \sum_{i=2}^n x_i\right)\prod_{i=2}^n a(x_i)\exp\left[\theta y_1 - n\kappa(\theta)\right]$$

By integrating/summing out $x_2, \ldots, x_n$ we obtain the marginal distribution for Y_1, of the form

$$f(y_1;\theta) = a_0(x)\exp\left[\theta y_1 - n\kappa(\theta)\right]$$

for some function $a_0(x)$. Hence, Y_1 also follows a natural exponential family, now with cumulant function $n\kappa(\theta)$. In particular, Y_1 is complete if Ω contains an open interval.

6 The likelihood ratio test and other large-sample tests

6.1 Standard errors

The most common asymptotic technique is perhaps the use of standard errors. From the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ we obtain

$$\hat{\boldsymbol{\theta}}_n \sim \mathrm{N}\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{i}^{-1}(\boldsymbol{\theta})\right) = \mathrm{N}(\theta, \boldsymbol{I}^{-1}(\boldsymbol{\theta})), \text{ approx.}$$

The estimated asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_n$ is hence $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}_n)$. This gives the following standard error for $\hat{\theta}_{jn}$

$$\mathrm{se}(\hat{\theta}_{jn}) = \sqrt{I^{jj}(\hat{\boldsymbol{\theta}}_n)}$$

where standard error means the estimated value of the standard deviation of the estimator. Since $\frac{1}{n}\boldsymbol{J}(\boldsymbol{\theta}) \xrightarrow{P} \boldsymbol{i}(\boldsymbol{\theta})$, we may use $\boldsymbol{J}(\hat{\boldsymbol{\theta}}_n)$ instead of $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_n)$, giving the alternative standard error

$$\mathrm{se}(\hat{\theta}_{jn}) = \sqrt{J^{jj}(\hat{\boldsymbol{\theta}}_n)}$$

When we write $\mathrm{se}(\hat{\boldsymbol{\theta}}_{jn})$, we may use either of these two possibilities. A $1-\alpha$ confidence interval for θ_j is given by the endpoints

$$\hat{\theta}_j \pm \mathrm{se}(\hat{\theta}_{jn}) z_{1-\frac{\alpha}{2}},$$

where $\Phi(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$. Similarly, a test for the hypothesis $\theta_j = \theta_j^0$, say, may be performed using

$$Z = \frac{\hat{\theta}_{jn} - \theta_j^0}{\mathrm{se}(\hat{\theta}_{jn})}$$

whose distribution is approximately $\mathrm{N}(0,1)$ for n large. This test is an example of a *Wald test*. Actually, the Wald test is based on the fact that Z^2 follows asymptotically a $\chi^2(1)$ distribution.

6.2 The likelihood ratio test

We now consider tests for composite hypotheses. Let $\Omega_0 \subseteq \Omega$ be a subset of Ω of dimension $q > p$, and consider the hypothesis $H_0 : \boldsymbol{\theta} \in \Omega_0$. We shall assume that, after a reparametrization, H_0 may be written in the form

$$H_0 : \theta_{q+1} = \cdots = \theta_p = 0$$

The alternative hypothesis is $H_A : \boldsymbol{\theta} \notin \Omega_0$.

Let $\hat{\boldsymbol{\theta}}_n$ denote the MLE of $\boldsymbol{\theta}$ in Ω, and let $\tilde{\boldsymbol{\theta}}_n = (\tilde{\theta}_1, \ldots, \tilde{\theta}_q, 0, \ldots, 0)^\top$ denote the MLE of $\boldsymbol{\theta}$ under H_0 (so that $\tilde{\boldsymbol{\theta}}_n \in \Omega_0$). Define the log likelihood ratio test by

$$R_n = 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\tilde{\boldsymbol{\theta}}_n)\}.$$

Note that $R_n \geq 0$. We reject H_0 if $R_n > c$, which gives a test with level $P_{H_0}(R_n > c)$. We normally chose a specified level α, so that c is determined by the equation $P_{H_0}(R_n > c) = \alpha$.

Theorem Under H_0 we have

$$R_n \xrightarrow{D} \chi^2(p-q) \quad \text{as} \quad n \to \infty.$$

Proof Expand $\ell(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}_n$, and use that $\boldsymbol{U}(\hat{\boldsymbol{\theta}}_n) = 0$

$$\begin{aligned} 2\{\ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}}_n)\} &\approx 2(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{U}(\hat{\boldsymbol{\theta}}_n) - (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{J}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \\ &= -\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^\top \left\{ \frac{1}{n} \boldsymbol{J}(\hat{\boldsymbol{\theta}}_n) \right\} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \\ &\approx -\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{i}(\boldsymbol{\theta}) \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \end{aligned}$$

because $\frac{1}{n}\boldsymbol{J}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} \boldsymbol{i}(\boldsymbol{\theta})$. Now we use

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \approx \boldsymbol{i}^{-1}(\boldsymbol{\theta}) \frac{\boldsymbol{U}(\boldsymbol{\theta})}{\sqrt{n}}$$

to obtain

$$\begin{aligned} 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta})\} &\approx \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top \boldsymbol{i}(\boldsymbol{\theta}) \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ &\approx \frac{1}{\sqrt{n}} \boldsymbol{U}(\boldsymbol{\theta})^\top \boldsymbol{i}^{-1}(\boldsymbol{\theta}) \frac{1}{\sqrt{n}} \boldsymbol{U}(\boldsymbol{\theta}) \end{aligned}$$

Let $\boldsymbol{\theta}_0 = (\theta_{10}, \ldots, \theta_{q0}, 0, \ldots, 0)^\top$ denote the true value of $\boldsymbol{\theta}$ under H_0. From matrix theory we know that there exists an upper triangular matrix $\boldsymbol{i}_0^{1/2}$, such that

$$\boldsymbol{i}(\boldsymbol{\theta}_0) = \boldsymbol{i}_0^{1/2} \boldsymbol{i}_0^{\top/2},$$

where $\boldsymbol{i}_0^{\top/2} = \left(\boldsymbol{i}_0^{1/2}\right)^\top$

Let $\boldsymbol{\psi} = \boldsymbol{i}_0^{1/2} \boldsymbol{\theta}$ denote a new parameter, which has score function

$$\tilde{U}(\boldsymbol{\psi}) = \boldsymbol{i}_0^{-1/2} \boldsymbol{U}(\boldsymbol{i}_0^{-1/2} \boldsymbol{\psi}),$$

where $\boldsymbol{i}_0^{-1/2}$ is the inverse of $\boldsymbol{i}_0^{1/2}$. The Fisher information matrix for $\boldsymbol{\psi}$ is

$$\tilde{i}(\boldsymbol{\psi}) = \boldsymbol{i}_0^{-1/2} i(\boldsymbol{\theta}) \boldsymbol{i}_0^{-\top/2}$$

If $\psi_0 = \boldsymbol{i}_0^{1/2} \boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\psi}$ then

$$\tilde{i}(\boldsymbol{\psi}_0) = \boldsymbol{i}_0^{-1/2} \boldsymbol{i}_0^{1/2} \boldsymbol{i}_0^{\top/2} \boldsymbol{i}_0^{-\top/2} = \boldsymbol{I}_p,$$

the $p \times p$ identity matrix. Hence, we obtain

$$\begin{aligned} 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\} &\approx \frac{1}{\sqrt{n}} \tilde{\boldsymbol{U}}^\top(\boldsymbol{\psi}_0) \frac{1}{\sqrt{n}} \tilde{\boldsymbol{U}}(\boldsymbol{\psi}_0) \\ &= \sum_{j=1}^{p} \frac{1}{n} \tilde{U}_j^2(\boldsymbol{\psi}_0). \end{aligned}$$

Since $\boldsymbol{i}_0^{1/2}$ is upper triangular, H_0 is equivalent to $\psi_{q+1} = \cdots = \psi_p = 0$. Hence, arguments similar to the above show that

$$2\{\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\} \approx \sum_{j=1}^{q} \frac{1}{n}\tilde{U}_j^2(\boldsymbol{\psi}_0).$$

We hence obtain the following approximation to the likelihood ratio test

$$\begin{aligned} R_n &= 2\{\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\} - 2\{\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\} \\ &\approx \sum_{j=q+1}^{p} \frac{1}{n}\tilde{U}_j^2(\boldsymbol{\psi}_0). \end{aligned}$$

Since $\tilde{\imath}(\boldsymbol{\psi}_0)$ is the identity matrix, we have that $\frac{1}{\sqrt{n}}\tilde{U}_1(\boldsymbol{\psi}_0), \ldots, \frac{1}{\sqrt{n}}\tilde{U}_p(\boldsymbol{\psi}_0)$ are asymptotically independent, and

$$\frac{1}{\sqrt{n}}\tilde{U}_j(\boldsymbol{\psi}_0) \xrightarrow{D} \mathrm{N}(0,1) \text{ as } n \to \infty.$$

Hence, $\frac{1}{n}\tilde{U}_{q+1}^2(\boldsymbol{\psi}_0), \ldots, \frac{1}{n}\tilde{U}_p^2(\boldsymbol{\psi}_0)$ are asymptotically independent and $\chi^2(1)$ distributed, and consequently

$$\sum_{j=q+1}^{p} \frac{1}{n}\tilde{U}_j^2(\boldsymbol{\psi}_0) \sim \chi^2(p-q), \text{ approx.}$$

Hence, $R_n \xrightarrow{D} \chi^2(p-q)$ which we had to prove.

6.3 Wald and score tests

We shall now briefly consider two other types of test, which turn out to be asymptotically equivalent to the likelihood ratio test. For simplicity, we consider the simple hypothesis H_0 : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a given value of $\boldsymbol{\theta}$.

The first test is the *Wald test*, which is defined by the quadratic form

$$W_n = \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)^\top \boldsymbol{I}(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$$

whose asymptotic distribution under H_0 is $\chi^2(p)$. The second test is the Rao *score test*, which is also a quadratic form

$$S_n = \boldsymbol{U}^\top(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0).$$

The asymptotic distribution of S_n under H_0 is also $\chi^2(p)$. The three test statistics R_n, W_n and S_n are asymptotically equivalent. In all three cases, we reject the hypothesis H_0 if the test statistic is larger than $\chi^2_{1-\alpha}(p)$, the $1-\alpha$ quantile of the $\chi^2(p)$ distribution.

These tests may be generalized to the case where H_0 is composite, see e.g. Cox and Hinkley (1974), Chapter 9.

7 Maximum likelihood computation

7.1 Assumptions

We consider algorithms for calculating the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for a log likelihood $\ell(\boldsymbol{\theta})$. We assume the following Ω conditions. The parameter domain Ω is an open region (bounded or unbounded) of $\mathbb{R}^p$. The log likelihood $\ell(\boldsymbol{\theta})$ is twice differentiable in Ω.

Recall that the score function and observed information matrix are defined by

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\theta}) &= \dot{\ell}(\boldsymbol{\theta}) \\ \boldsymbol{J}(\boldsymbol{\theta}) &= -\dot{\boldsymbol{U}}(\boldsymbol{\theta}). \end{aligned}$$

The Fisher information

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}_\theta\left[\boldsymbol{J}(\boldsymbol{\theta})\right]$$

is positive-definite for any $\boldsymbol{\theta} \in \Omega$.

The overall objective is to find the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$, satisfying

$$\ell(\widehat{\boldsymbol{\theta}}) \geq \ell(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Omega.$$

In practice, the best we can hope for is to find a local maximum $\hat{\boldsymbol{\theta}}$ of ℓ; in particular $\hat{\boldsymbol{\theta}}$ is assumed to be a root of $\boldsymbol{U}$.

Our goal is to calculate $\hat{\boldsymbol{\theta}}$ with a given accuracy relative to the asymptotic standard errors

$$\mathrm{se}(\hat{\theta}_j) = \sqrt{I^{jj}(\hat{\boldsymbol{\theta}})}.$$

Hence, we use a convergence criterion of the form $10^{-d}\left|\boldsymbol{I}(\boldsymbol{\theta})\right|^{-1/2}$, where $|\cdot|$ denotes determinant, and d is the desired number of significant digits relative to a given $\mathrm{se}(\hat{\theta}_j)$. A good choice for d is 2 or 3. Sometimes the asymptotic standard error is calculated from $\boldsymbol{J}^{-1}(\boldsymbol{\theta})$, but this value is not suitable as a reference for the convergence criterion, because $\boldsymbol{J}(\boldsymbol{\theta})$ may not be positive-definite when $\boldsymbol{\theta}$ is far from $\hat{\boldsymbol{\theta}}$.

We consider methods that ensure convergence to a local maximum of ℓ. Our methods take the statistical nature of the problem into account by using $\boldsymbol{I}(\boldsymbol{\theta})$ instead of $\boldsymbol{J}(\boldsymbol{\theta})$. Systematic accounts of numerical optimization methods may be found in Dennis and Schnabel (1983), Smyth (2002) and Lange (2004).

7.2 Stabilized Newton methods

The best optimization methods for our purpose are the so-called stabilized Newton methods. Let $\boldsymbol{K}(\boldsymbol{\theta})$ be a given positive-definite information matrix, e.g. $\boldsymbol{I}(\boldsymbol{\theta})$ or $\boldsymbol{J}(\boldsymbol{\theta})$. Note that the requirement that $\boldsymbol{J}(\boldsymbol{\theta})$ be positive-definite is equivalent to ℓ being strictly concave. Hence, if ℓ is not strictly concave we must use $\boldsymbol{I}(\boldsymbol{\theta})$ instead of $\boldsymbol{J}(\boldsymbol{\theta})$.

A *stabilized Newton method* based on $\boldsymbol{K}(\boldsymbol{\theta})$ is an iterative method of the following form:

1. *Starting value:* Find a suitable starting value $\boldsymbol{\theta}_0$ and let $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

2. *Search direction:* For given $\boldsymbol{\theta}$, calculate $\boldsymbol{\delta} = \boldsymbol{K}^{-1}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ (the *stabilized Newton step*).

3. *Step length:* Compute a positive scalar α such that $\boldsymbol{\theta} + \alpha\boldsymbol{\delta}$ is inside Ω (*boundary check*) and such that
$$\ell(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}) > \ell(\boldsymbol{\theta}) \text{ (ascent check).} \tag{7.1}$$

4. *Convergence:* Stop when the *convergence criterion*
$$\|\alpha\boldsymbol{\delta}\| < 10^{-d} \left|\boldsymbol{I}(\boldsymbol{\theta})\right|^{-1/2} \tag{7.2}$$
is met (where $\|\cdot\|$ denotes Euclidean norm), or if the number of iterations exceeds a certain number *maxiter*.

5. *Update:* Otherwise update $\boldsymbol{\theta}$
$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \alpha\boldsymbol{\delta};$$
and return to Step 2, with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Starting with $\boldsymbol{\theta}_0$, the method calculates a sequence $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ that is designed to converge towards the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$. An algorithm satisfying (7.1) in each step is called an *ascent method.* In order to obtain an ascent method, it is important to use a positive-definite information matrix $\boldsymbol{K}(\boldsymbol{\theta})$, which assures that the search direction $\boldsymbol{\delta}$ points in an uphill direction, as shown below. In this way, a small enough step length α will guarantee that (7.1) is satisfied. For further information about stabilized Newton methods, see Bard (1974), Gill et al. (1981), Luenberger (1969) and Everitt (1987).

In practice it may be better to replace (7.2) by the criterion
$$\alpha^2 \boldsymbol{\delta}^\top \boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{\delta} < 10^{-2d} \tag{7.3}$$
based on the weighted norm $\|x\|_{I(\theta)} = \left(x^\top \boldsymbol{I}(\boldsymbol{\theta})x\right)^{1/2}$, say, which is slightly easier to handle than (7.2). Note, however, that either of (7.2) and (7.3) is satisfied for α small enough. Hence, the step length calculation should in practice be designed to always take a good step in the right direction, or in other words avoid making α so small that the iterations are halted prematurely.

7.3 The Newton-Raphson method

Assume now that ℓ is strictly concave, and take $\boldsymbol{K}(\boldsymbol{\theta}) = \boldsymbol{J}(\boldsymbol{\theta})$, which is now, by assumption, positive-definite for all $\boldsymbol{\theta}$. Taking $\alpha = 1$ gives the *Newton-Raphson method*
$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}), \tag{7.4}$$
where $\boldsymbol{\theta}^*$ is the updated value of $\boldsymbol{\theta}$, and $\boldsymbol{\delta} = \boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})$ is called the *Newton step.* Adding a step length calculation gives a *stabilized* Newton-Raphson method. The Newton-Raphson method derives from the Taylor-expansion
$$\boldsymbol{U}(\boldsymbol{\theta}^*) \approx \boldsymbol{U}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta})\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\right). \tag{7.5}$$
If $\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}$, then the left-hand side of (7.5) is zero, which motivates us to define $\boldsymbol{\theta}^*$ by (7.4), making the right-hand side of (7.5) zero.

Some properties of the Newton-Raphson method:

- The convergence of (7.4) is *quadratic* near the maximum, provided $\boldsymbol{U}(\boldsymbol{\theta})$ is Lipschitz continuous (Dennis and Schnabel, 1983, p. 22). Quadratic convergence means that, roughly speaking, the number of correct figures doubles in each iteration.

- The step length calculation is designed to avoid problems where the Newton step overshoots or undershoots the target because ℓ may not be quadratic away from the maximum.

- The step length calculation should be designed such that it does not interfere with the quadratic convergence of the algorithm near the maximum. As mentioned above, (7.2) or (7.3) are easily satisfied if α is chosen small enough, which risks halting the iterations prematurely.

7.4 Fisher's scoring method

Taking $\boldsymbol{K}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta})$ gives *Fisher's scoring method*

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}), \tag{7.6}$$

which is a widely applicable, and usually quite stable algorithm, when step length calculation is used. The main assumption for Fisher's scoring method is that $\boldsymbol{I}(\boldsymbol{\theta})$ should be positive-definite for all $\boldsymbol{\theta}$, which holds for any regular statistical model, making Fisher scoring the best general method for maximum likelihood computation.

Some properties of Fisher's scoring method:

- The convergence is usually *linear* near the maximum. Linear convergence means that, roughly speaking, the same number of correct figures are added in each iteration.

- Far from the maximum, the algorithm tends to make good steps in the right direction, making it robust to badly behaved log likelihoods or bad starting values. In particular, ℓ does not need to be strictly concave.

- As for the Newton-Raphson method, the step length calculation is important far from the maximum, but should be avoided near the maximum.

7.5 Step length calculation

As already mentioned above, the step length calculation is important in order to obtain an ascent method, which in turn helps avoiding divergence due to a poor starting value. The step length calculation may be implemented in many different ways, but a good method should strike a suitable balance between maintaining control at the beginning of the iterative process, while relying on the good convergence properties of the stabilized Newton methods near the maximum.

Let the function g be defined for $\alpha \geq 0$ as follows:

$$g(\alpha) = \ell(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}) - \ell(\boldsymbol{\theta}),$$

provided that $\boldsymbol{\theta} + \alpha\boldsymbol{\delta} \in \Omega$. We may then proceed with the step length calculation as follows:

1. *Boundary check:* If $\boldsymbol{\theta} + \alpha\boldsymbol{\delta} \notin \Omega$, then we repeatedly divide α by 2 until $\boldsymbol{\theta} + \alpha\boldsymbol{\delta} \in \Omega$.

2. *Quadratic interpolation*: If $g(\alpha) \leq 0$ replace α by

$$\frac{\alpha^2 \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta})}{2\left\{\alpha \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta}) - g(\alpha)\right\}}$$

Note here that setting $g(\alpha) = 0$ if $\boldsymbol{\theta} + \alpha\boldsymbol{\delta} \notin \Omega$ has the effect of halving the step length, so that 1. and 2. may be combined into a single step.

3. *Ascent check:* If $g(\alpha) > 0$ then exit the step length calculation with the current value of α, else return to Step 2.

Comments on the step length calculation.

- The effect of Step 2 is to locate approximately the maximum for g by a quadratic interpolation in the interval from 0 to α. Note that $g(0) = 0$ and $\dot{g}(0) = \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{U}^\top(\boldsymbol{\theta})\boldsymbol{K}^{-1}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta}) \geq 0$, where we have used the fact that $\boldsymbol{K}(\boldsymbol{\theta})$ is positive-definite, which in turn implies that $\boldsymbol{K}^{-1}(\boldsymbol{\theta})$ is positive-definite. Let $q(x) = ax^2 + bx$ be a quadratic function that agrees with g at 0 and α and has the same derivative as g at 0. The coefficients of q are then

$$\begin{aligned} a &= -\frac{\alpha \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta}) - g(\alpha)}{\alpha^2} < 0 \\ b &= \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta}) > 0, \end{aligned}$$

where we have used the fact that $g(\alpha) \leq 0$. The maximum for $q(x)$ is attained between 0 and $\alpha/2$ for the following value:

$$x = \frac{\alpha^2 \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta})}{2\left\{\alpha \boldsymbol{\delta}^\top \boldsymbol{U}(\boldsymbol{\theta}) - g(\alpha)\right\}}.$$

- Steps 2–3 guarantee an increase of the log likelihood value in each iteration. Note that the quadratic interpolation is skipped if $g(\alpha) > 0$, i.e. if the α obtained after the boundary check provides an increase of the log likelihood value.

7.6 Convergence and starting values

The algorithm will tend to converge towards a local maximum of ℓ somewhere near the starting value $\boldsymbol{\theta}_0$. To ensure that the iterations converge towards a statistically meaningful local maximum, it is hence useful to use a statistically meaningful starting value, for example based on a moment estimator.

If it is suspected that there are other local maxima than the one found, one may restart the algorithm from a new starting value several multiples of the criterion $|\boldsymbol{I}(\boldsymbol{\theta})|^{-1/2}$ away from the original root $\hat{\boldsymbol{\theta}}$.

References

[1] Andersen, E. B. (1980) Discrete Statistical Models With Social Science Applications. Amsterdam, North-Holland Publishing Company

[2] Arley, Niels and Buch, K. Rander (1950). Introduction to the Theory of Probability and Statistics. Wiley.

[3] Bain, L. J. and Engelhardt, M. (1987). Introduction to Probability and Mathematical Statistics. Duxbury Press, Boston.

[4] Casella, G. and Berger, R. L. (1990). Statistical Inference. Wadsworth, Belmont CA.

[5] Cox, D. R. and Hinkley, D. (1974) Theoretical Statistics. Chapman & Hall, London.

[6] Bard, Y. (1974). *Nonlinear Parameter Estimation.* New York, Academic Press.

[7] Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs, N.J., Prentice-Hall.

[8] Everitt, B. S. (1987). *Introduction to Optimization Methods and Their Application in Statistics.* London, Chapman and Hall.

[9] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Phil. Trans. Roy. Soc. London A 222, 309–368.

[10] Geyer, C. J. (2003). Maximum likelihood in R. Lecture notes. http://www.stat.umn.edu/geyer/5931/mle/mle.pdf

[11] Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization.* London, Academic Press.

[12] Hogg, R. V., McKean, J. W., and Craig, A. T. (2013). *Introduction to Mathematical Statistics.* Seventh Edn. Upper Saddle River, Pearson/Prentice Hall.

[13] James, B. (1981). Probabilidade: Um Curso em Nível Intermediário. IMPA, Rio de Janeiro.

[25] Knight, K. (2010) Mathematical Statistics. Taylor & Francis, London.

[15] Lange, K. (2004). *Optimization.* New York, Springer.

[25] Lehmann E. L. and Romano, J. P. (2005). Testing Statistical Hypotheses, 3rd Ed.. Springer, New York.

[17] Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation, 2nd Ed. Springer, New York.

[18] Luenberger, D. G. (1969). *Optimization by Vector Space Methods.* New York, Wiley.

[25] Mukhopadhyay, N. (2000) Probability and Statistical Inference. Taylor & Francis, London.

[20] Murison, B. (2000). Distribution Theory, Statistical Inference.
`http://turing.une.edu.au/~stat354/notes/notes.html`

[21] Rao, C. R. (1973). Linear Statistical Inference and Its Applications (2nd Edn.). Wiley, New York.

[22] Rice, J. A. (1988). Mathematical Statistics and Data Analysis. Wadsworth, Belmont CA.

[25] Rohatgi, V. K. and Saleh, A. K. M. E. (2011). An Introduction to Probability and Statistics. 2nd Ed. Wiley, Chichester.

[25] Roussas, G. G. (1997). A Course in Mathematical Statistics. Academic Press, San Diego.

[25] Roussas G. G. (2003) An Introduction to Probability and Statistical Inference. Academic Press, San Diego.

[26] Silvey, S. D. (1975). Statistical Inference. Chapman & Hall, London.

[27] Smyth, G. K. (2002). Optimization. In *Encyclopedia of Environmetrics* (Eds. A.H. El-Shaarawi and W.W. Piegorsch), Vol. 3, pp. 1481–1487. John Wiley & Sons, Chichester.

[28] Wackerly, D. D., Mendenhall, W. and Scheaffer, R. L. (2008). Mathematical Statistics with Applications. 7th Ed. Thomson-Brooks/Cole, Belmont CA.