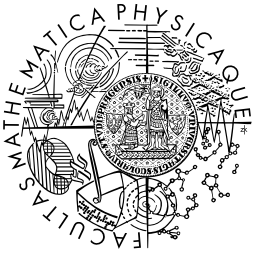

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MATHEMATICAL STATISTICS 3

NMST 424

Course notes

Last updated: April 30, 2024

I would like to thank Stanislav Nagy for finding many typos and misprints. I would like to thank all the students and colleagues who have helped me in improving the text.

Contents

Last update: April 30, 2024

1 Clippings from the asymptotic theory

1.1 The convergence of random vectors

Let $\boldsymbol{X}$ be a k-dimensional random vector (with the cumulative distribution function $F_{\boldsymbol{X}}$) and $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ be a sequence of k-dimensional random vectors (with the cumulative distribution functions $F_{\boldsymbol{X}_n}$).

Definition. We say that $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *in distribution* to $\boldsymbol{X}$), if

$$\lim_{n\to\infty} F_{\boldsymbol{X}_n}(\mathbf{x}) = F_{\boldsymbol{X}}(\mathbf{x})$$

for each point $\mathbf{x}$ of the continuity of $F_{\boldsymbol{X}}$.

Example 1. Let U be a random variable with a uniform distribution on the interval $(0,1)$. Put $X_n = U/n$. Show that $X_n \xrightarrow[n\to\infty]{d} 0$. But at the same time $F_{X_n}(0)$ does not converge to $F_X(0)$, where F_X is the cumulative distribution corresponding to the random variable that is equal to zero almost surely.

Let d be a metric in $\mathbb{R}^k$, e.g. the Euclidean metric $d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$.

Definition. We say that

- $\boldsymbol{X}_n \xrightarrow[n\to\infty]{P} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *in probability* to $\boldsymbol{X}$), if

$$\forall \varepsilon > 0 \ \lim_{n\to\infty} \mathsf{P}\left[\omega : d\big(\boldsymbol{X}_n(\omega), \boldsymbol{X}(\omega)\big) > \varepsilon\right] = 0;$$

- $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$ (i.e. $\boldsymbol{X}_n$ converges *almost surely* to $\boldsymbol{X}$), if

$$\mathsf{P}\left[\omega : \lim_{n\to\infty} d\big(\boldsymbol{X}_n(\omega), \boldsymbol{X}(\omega)\big) = 0\right] = 1.$$

Remark 1. For random vectors the convergence in probability and almost surely can be defined also component-wise. That is let $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})^{\mathsf{T}}$ and $\boldsymbol{X} = (X_1, \ldots, X_k)^{\mathsf{T}}$. Then

$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{P} \boldsymbol{X} \ (\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}) \quad \text{if} \quad X_{nj} \xrightarrow[n\to\infty]{P} X_j \ (X_{nj} \xrightarrow[n\to\infty]{\text{a.s.}} X_j), \quad \forall j = 1, \ldots, k.$$

But this is not true for the convergence in distribution for which we have the Cramér-Wold theorem that states

$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X} \iff \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{X}, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^k.$$

Theorem 1. *(**Continuous Mapping Theorem, CMT**) Let* $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$ *be continuous in each point of an open set* $C \subset \mathbb{R}^k$ *such that* $\mathsf{P}(\boldsymbol{X} \in C) = 1$*. Then*

(i) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{a.s.} \boldsymbol{X} \Rightarrow \mathbf{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{a.s.} \mathbf{g}(\boldsymbol{X})$*;*

(ii) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{P} \boldsymbol{X} \Rightarrow \mathbf{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{P} \mathbf{g}(\boldsymbol{X})$*;*

(iii) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X} \Rightarrow \mathbf{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{d} \mathbf{g}(\boldsymbol{X})$*.*

Proof. (i) *Almost sure convergence.*

$$\begin{aligned}
&\mathsf{P}\Big[\omega : \lim_{n\to\infty} d\big(\mathbf{g}(\boldsymbol{X}_n(\omega)), \mathbf{g}(\boldsymbol{X}(\omega))\big) = 0\Big] \\
&\quad \geq \mathsf{P}\Big[\omega : \lim_{n\to\infty} d\big(\mathbf{g}(\boldsymbol{X}_n(\omega)), \mathbf{g}(\boldsymbol{X}(\omega))\big) = 0, \boldsymbol{X}(\omega) \in C\Big] \\
&\quad \geq \mathsf{P}\Big[\omega : \lim_{n\to\infty} d\big(\boldsymbol{X}_n(\omega), \boldsymbol{X}(\omega)\big) = 0, \boldsymbol{X}(\omega) \in C\Big] = 1,
\end{aligned}$$

as C is an open set and $\mathsf{P}(\boldsymbol{X} \in C) = 1$.

(ii) *Convergence in probability.* Let $\varepsilon > 0$. Then for each $\delta > 0$

$$\begin{aligned}
&\mathsf{P}\Big[\omega : d\big(\mathbf{g}(\boldsymbol{X}_n(\omega)), \mathbf{g}(\boldsymbol{X}(\omega))\big) > \varepsilon\Big] \\
&\quad \leq \mathsf{P}\Big[d\big(\mathbf{g}(\boldsymbol{X}_n), \mathbf{g}(\boldsymbol{X})\big) > \varepsilon, d(\boldsymbol{X}_n, \boldsymbol{X}) \leq \delta\Big] + \mathsf{P}\Big[d(\boldsymbol{X}_n, \boldsymbol{X}) > \delta\Big] \\
&\quad \leq \mathsf{P}\Big[\boldsymbol{X} \in B^\delta\Big] + \underbrace{\mathsf{P}\Big[d(\boldsymbol{X}_n, \boldsymbol{X}) > \delta\Big]}_{\to 0, \forall \delta > 0},
\end{aligned}$$

where $B^\delta = \big\{\mathbf{x} \in \mathbb{R}^k; \exists \mathbf{y} \in \mathbb{R}^k : d(\mathbf{x}, \mathbf{y}) \leq \delta,\, d\big(\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{y})\big) > \varepsilon\big\}$. Further

$$\begin{aligned}
\mathsf{P}\left[\boldsymbol{X} \in B^\delta\right] &= \mathsf{P}\left[\boldsymbol{X} \in B^\delta, \boldsymbol{X} \in C\right] + \mathsf{P}\left[\boldsymbol{X} \in B^\delta, \boldsymbol{X} \notin C\right] \\
&= \mathsf{P}\left[\boldsymbol{X} \in B^\delta \cap C\right] + 0
\end{aligned}$$

and $\mathsf{P}\left[\boldsymbol{X} \in B^\delta \cap C\right]$ can be made arbitrarily small as $B^\delta \cap C \to \emptyset$ for $\delta \searrow 0$.

(iii) See for instance the proof of Theorem 13.6 in Lachout [2004].

□

Theorem 2. *(**Cramér-Slutsky, CS**) Let* $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$*,* $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{P} \mathbf{c}$*, then*

(i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X} + \mathbf{c}$*;*

(ii) $\boldsymbol{Y}_n \boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \mathbf{c}\, \boldsymbol{X}$*,*

where $\boldsymbol{Y}_n$ can be a sequence of random variables or vectors or matrices of appropriate dimensions ($\mathbb{R}$ or $\mathbb{R}^k$ or $\mathbb{R}^{m\times k}$) and analogously $\mathbf{c}$ can be either a number or a vector or a matrix of an appropriate dimension.

Proof. Note that it is sufficient to prove

$$(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow[n\to\infty]{d} (\boldsymbol{X}, \mathbf{c}). \tag{1}$$

Then the statement of the theorem follows from Continuous Mapping Theorem (Theorem 1).

To prove (1) note that

$$d\big((\boldsymbol{X}_n, \boldsymbol{Y}_n), (\boldsymbol{X}_n, \mathbf{c})\big) = d\big(\boldsymbol{Y}_n, \mathbf{c}\big) \xrightarrow[n\to\infty]{P} 0.$$

Thus by Theorem 13.7 in Lachout [2004] or Theorem 2.7 (iv) of van der Vaart [2000] it is sufficient to show that $(\boldsymbol{X}_n, \mathbf{c}) \xrightarrow[n\to\infty]{d} (\boldsymbol{X}, \mathbf{c})$. But this follows immediately with the help of the Cramér-Wold theorem. □

Definition 1. Let $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ be a sequence of random vectors and $\{r_n\}_{n=1}^{\infty}$ a sequence of positive constants. We write that

(i) $\boldsymbol{X}_n = o_P\big(\frac{1}{r_n}\big)$, if $(r_n\, \boldsymbol{X}_n) \xrightarrow[n\to\infty]{P} \mathbf{0}_k$, where $\mathbf{0}_k = (0, \dots, 0)^{\mathsf{T}}$ is a zero point in $\mathbb{R}^k$;

(ii) $\boldsymbol{X}_n = O_P\big(\frac{1}{r_n}\big)$, if

$$\forall \varepsilon > 0\ \exists K < \infty\ \exists n_0 \in \mathbb{N}\ \sup_{n\geq n_0} \mathsf{P}\left(r_n \left\|\boldsymbol{X}_n\right\| > K\right) < \varepsilon,$$

where $\|\cdot\|$ stands for instance for the Euclidean norm.

When $\boldsymbol{X}_n = O_P(1)$ then some authors say that $\{\boldsymbol{X}_n\}$ is (asymptotically) *bounded* in probability*. When $\boldsymbol{X}_n = o_P(1)$ then it is often said that $\{\boldsymbol{X}_n\}$ is (asymptotically) *negligible* in probability.

Remark 2. Note that

(i) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ implies $\boldsymbol{X}_n = O_P(1)$ (Prohorov's theorem, Portmanteau theorem, see e.g. Chapters 2.1 van der Vaart [2000]);

(ii) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{P} \mathbf{0}$ implies $\boldsymbol{X}_n = o_P(1)$;

(iii) $(r_n\, \boldsymbol{X}_n) \xrightarrow[n\to\infty]{P} \boldsymbol{X}$ or $(r_n\, \boldsymbol{X}_n) \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ implies $\boldsymbol{X}_n = O_P\big(\frac{1}{r_n}\big)$.

(iv) If $r_n \to \infty$ and $\boldsymbol{X}_n = O_P\big(\frac{1}{r_n}\big)$, then $\boldsymbol{X}_n = o_P(1)$.

* *omezená v pravděpodobnosti*

Proof of (iv). Note that it is sufficient to prove that for each $\varepsilon > 0$ and each $\eta > 0$ for all sufficiently large n it holds that $\mathsf{P}\left(\|\boldsymbol{X}_n\| > \varepsilon\right) < \eta$.

Note that $\boldsymbol{X}_n = O_P\left(\frac{1}{r_n}\right)$ implies there exists a finite constant K and $n_0 \in \mathbb{N}$ such that

$$\sup_{n \geq n_0} \mathsf{P}\left(r_n \|\boldsymbol{X}_n\| > K\right) < \varepsilon.$$

The statement now follows from the fact that

$$\mathsf{P}\left(\|\boldsymbol{X}_n\| > \varepsilon\right) = \mathsf{P}\left(r_n \|\boldsymbol{X}_n\| > \varepsilon\, r_n\right) < \eta$$

for all n such that $\varepsilon\, r_n > K$. □

Suppose that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are independent and identically distributed random vectors with a finite variance matrix. Then the law of large numbers implies

$$\overline{\boldsymbol{X}}_n = \mathsf{E}\ \boldsymbol{X}_1 + o_P(1).$$

With the help of the central limit theorem one can be even more specific about the remainder term and show that

$$\overline{\boldsymbol{X}}_n = \mathsf{E}\ \boldsymbol{X}_1 + O_P\left(\tfrac{1}{\sqrt{n}}\right).$$

Remark 3. Further note that the calculus with the random quantities $o_P(1)$ and $O_P(1)$ is analogous to the calculus with the (deterministic) quantities $o(1)$ and $O(1)$ in mathematical analysis. Thus, among others it holds that

(i) $o_P(1) + o_P(1) = o_P(1)$;

(ii) $o_P(1)\, O_P(1) = o_P(1)$;

(iii) $O_P(1)\, O_P(1) = O_P(1)$;

(iv) $o_P(1) + O_P(1) = O_P(1)$;

Proof of (ii). Let $\{\boldsymbol{X}_n\}, \{\boldsymbol{Y}_n\}$ be such that $\boldsymbol{X}_n = O_P(1), \boldsymbol{Y}_n = o_P(1)$ and $\boldsymbol{Y}_n \boldsymbol{X}_n$ makes sense. Let $\varepsilon > 0$ be given and consider for instance the Euclidean norm (for other norms the proof would go through up to a multiplicative constant in some of the arguments). Then one can find $K < \infty$ and $n_0 \in \mathbb{N}$ such that $\sup_{n \in \mathbb{N}_0} \mathsf{P}\left(\|\boldsymbol{X}_n\| > K\right) < \frac{\varepsilon}{2}$. Thus for all sufficiently large $n \in \mathbb{N}$

$$\begin{aligned}\mathsf{P}\left(\|\boldsymbol{Y}_n \boldsymbol{X}_n\| > \varepsilon\right) &\leq \mathsf{P}\left(\|\boldsymbol{Y}_n \boldsymbol{X}_n\| > \varepsilon, \|\boldsymbol{X}_n\| \leq K\right) + \mathsf{P}\left(\|\boldsymbol{X}_n\| > K\right) \\ &\leq \mathsf{P}\left(\|\boldsymbol{Y}_n\| > \tfrac{\varepsilon}{K}\right) + \tfrac{\varepsilon}{2} \leq \varepsilon,\end{aligned}$$

as $\boldsymbol{Y}_n = o_P(1)$.

We recommend the reader to prove the remaining statements as an exercise. □

For more details about the calculus with $o_P(1)$ and $O_P(1)$ see for instance Chapter 3.4 of Jiang [2010].

1.2 Δ-theorem

Let $\mathbf{T}_n = (T_{n1}, \ldots, T_{np})^\mathsf{T}$ be a p-dimensional random vector that converges to the constant $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\mathsf{T}$ and $\mathbf{g} = (g_1, \ldots, g_m)^\mathsf{T}$ be a function from (a subset of) $\mathbb{R}^p$ to $\mathbb{R}^m$. Denote the Jacobi matrix of the function $\mathbf{g}$ at the point $\mathbf{x}$ as $\mathbb{D}_\mathbf{g}(\mathbf{x})$, i.e.

$$\mathbb{D}_\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \nabla g_1(\mathbf{x}) \\ \vdots \\ \nabla g_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \ldots & \frac{\partial g_1(\mathbf{x})}{\partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\mathbf{x})}{\partial x_1} & \ldots & \frac{\partial g_m(\mathbf{x})}{\partial x_p} \end{pmatrix}.$$

Theorem 3. ***($\boldsymbol{\Delta}$-theorem)*** *Let $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = O_P(1)$. Further $\mathbf{g} : A \to \mathbb{R}^m$, where $A \subset \mathbb{R}^p$, $\boldsymbol{\mu}$ is an interior point of A and the first-order partial derivatives of $\mathbf{g}$ are continuous in a neighbourhood of $\boldsymbol{\mu}$. Then*

(i) $\sqrt{n}\,\big(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\mu})\big) - \mathbb{D}_\mathbf{g}(\boldsymbol{\mu})\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = o_P(1)$*;

(ii) moreover if $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \mathbb{\Sigma}\big)$*, then*

$$\sqrt{n}\,\big(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\mu})\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_m\big(\mathbf{0}_m, \mathbb{D}_\mathbf{g}(\boldsymbol{\mu})\,\mathbb{\Sigma}\,\mathbb{D}_\mathbf{g}^\mathsf{T}(\boldsymbol{\mu})\big). \tag{2}$$

Proof. Statement (i): For $j \in \{1, \ldots, m\}$ consider $g_j : A \to \mathbb{R}$ (the j-th coordinate of the function $\mathbf{g}$). From the assumptions of the theorem there exists a neighbourhood $\mathcal{U}_\delta(\boldsymbol{\mu})$ of the point $\boldsymbol{\mu}$ such that the function g_j has continuous partial derivatives in this neighbourhood. Further $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = O_P(1)$ implies $\mathbf{T}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\mu}$ (see for instance Remark 2(iv)), which yields that $\mathsf{P}\left(\mathbf{T}_n \in \mathcal{U}_\delta(\boldsymbol{\mu})\right) \xrightarrow[n\to\infty]{} 1$. Thus without loss of generality one can assume that $\mathbf{T}_n \in \mathcal{U}_\delta(\boldsymbol{\mu})$. Using this together with the mean value theorem there exists $\boldsymbol{\mu}_n^{j*}$ which lies between $\mathbf{T}_n$ and $\boldsymbol{\mu}$ such that

$$\begin{aligned} \sqrt{n}\,\big(g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu})\big) &= \nabla g_j(\boldsymbol{\mu}_n^{j*})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) \\ &= \nabla g_j(\boldsymbol{\mu})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) + \big[\nabla g_j(\boldsymbol{\mu}_n^{j*}) - \nabla g_j(\boldsymbol{\mu})\big]\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}). \end{aligned} \tag{3}$$

Further $\mathbf{T}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\mu}$ implies that $\boldsymbol{\mu}_n^{j*} \xrightarrow[n\to\infty]{P} \boldsymbol{\mu}$. Now the continuity of the partial derivatives of g_j in $\mathcal{U}_\delta(\boldsymbol{\mu})$ and CMT (Theorem 1) imply that

$$\nabla g_j(\boldsymbol{\mu}_n^{j*}) - \nabla g_j(\boldsymbol{\mu}) = o_P(1),$$

* $\mathbb{D}_\mathbf{g}(\boldsymbol{\mu})\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu})$ is sometimes called also the asymptotic linear approximation of $\sqrt{n}\,\big(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\mu})\big)$.

which together with $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = O_P(1)$ gives

$$\left[\nabla g_j(\boldsymbol{\mu}_n^{j*}) - \nabla g_j(\boldsymbol{\mu})\right] \sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = o_P(1). \tag{4}$$

Now combining (3) and (4) yields that for each $j = 1, \ldots, m$

$$\sqrt{n}\,\big(g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu})\big) = \nabla g_j(\boldsymbol{\mu})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) + o_P(1),$$

which implies the first statement of the theorem.

Statement (i) - the proof done at the lecture: For $j \in \{1, \ldots, m\}$ consider $g_j : A \to \mathbb{R}$ (the j-th coordinate of the function $\mathbf{g}$). Define the following function $h_j : A \to \mathbb{R}$ as

$$h_j(\mathbf{x}) = \begin{cases} \frac{g_j(\mathbf{x}) - g_j(\boldsymbol{\mu}) - \nabla g_j(\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})}{\|\mathbf{x}-\boldsymbol{\mu}\|}, & \mathbf{x} \neq \boldsymbol{\mu}, \\ 0, & \mathbf{x} = \boldsymbol{\mu}. \end{cases}$$

Note that h_j is continuous in $\boldsymbol{\mu}$.

Further $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = O_P(1)$ implies $\mathbf{T}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\mu}$ (see for instance Remark 2(iv)). Now using CMT (Theorem 1) implies $h_j(\mathbf{T}_n) = o_P(1)$. Thus

$$g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu}) - \nabla g_j(\boldsymbol{\mu})(\mathbf{T}_n - \boldsymbol{\mu}) = o_P(1)\,\|\mathbf{T}_n - \boldsymbol{\mu}\|,$$

which together with $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = O_P(1)$ and Remark 3(ii) gives

$$\sqrt{n}\,\big(g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu})\big) - \nabla g_j(\boldsymbol{\mu})(\mathbf{T}_n - \boldsymbol{\mu}) = o_P(1)\,\big\|\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu})\big\| = o_P(1).$$

Thus one can conclude that

$$\sqrt{n}\,\big(g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu})\big) = \nabla g_j(\boldsymbol{\mu})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) + o_P(1),$$

which implies the first statement of the theorem.

Statement (ii): By the first statement of the theorem one gets

$$\sqrt{n}\,\big(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\big) = \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\sqrt{n}\,(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) + o_P(1)$$

Now for the term $\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu})$ one can use the second statement of CS (Theorem 2) with $\boldsymbol{Y}_n = \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})$ and $\boldsymbol{X}_n = \sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu})$. Further, using now the first statement of CS with $\mathbf{c} = \mathbf{0}_m$ one can see that adding the term $o_P(1)$ does not alter the asymptotic distribution of $\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu})$. □

The end of class 1 (20. 2. 2024)

In the most common applications of Δ-theorem one often takes $\mathbf{T}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed. Then $\boldsymbol{\mu} = \mathsf{E}\,\boldsymbol{X}_1$ and the standard central limit theorem gives the asymptotic normality

$$\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = \sqrt{n}\,(\overline{\boldsymbol{X}}_n - \mathsf{E}\,\boldsymbol{X}_1) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{\Sigma}),$$

where $\mathbb{\Sigma} = \mathsf{var}(\boldsymbol{X}_i)$.

Note that then Theorem 3(i) implies that

$$\sqrt{n}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\right) = \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\sqrt{n}\,(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) + o_P(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{Z}_i + o_P(1),$$

where

$$\mathbf{Z}_i = \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,(\boldsymbol{X}_i - \boldsymbol{\mu}), \quad i \in \{1, \ldots, n\}$$

are independently distributed random vectors. Note that then the central limit theorem together with the Cramér-Slutsky theorem (Theorem 2(i)) implies that

$$\sqrt{n}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(\mathbf{0}, \mathsf{var}(\mathbf{Z}_1)\big).$$

Thus the asymptotic variance

$$\mathsf{avar}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n)\right) = \frac{1}{n}\,\mathsf{var}\left(\mathbf{Z}_1\right)$$

can be easily estimated as

$$\widehat{\mathsf{avar}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n)\right)} = \frac{1}{n}\,\mathbb{S}_Z^2,$$

where

$$\mathbb{S}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i^{\mathsf{T}}$$

is the sample variance matrice of the 'estimated' $\mathbf{Z}_i$ given by

$$\hat{\mathbf{Z}}_i = \mathbb{D}_{\mathbf{g}}(\overline{\boldsymbol{X}}_n)\,(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n),$$

as $\sum_{i=1}^{n}\hat{\mathbf{Z}}_i = \mathbf{0}$.

Remark 4. Instead of the continuity of the partial derivatives in a neighbourhood of $\boldsymbol{\mu}$, it would be sufficient to assume the existence of the total differential of the function $\mathbf{g}$ at the point $\boldsymbol{\mu}$ (see the alternative proof of (i) done at the lecture).

Sometimes instead of (2) we write shortly $\mathbf{g}(\mathbf{T}_n) \overset{\text{as}}{\approx} \mathsf{N}_m\big(\mathbf{g}(\boldsymbol{\mu}), \frac{1}{n}\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\mathbb{\Sigma}\,\mathbb{D}_{\mathbf{g}}^{\mathsf{T}}(\boldsymbol{\mu})\big)$. The quantity $\frac{1}{n}\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\mathbb{\Sigma}\,\mathbb{D}_{\mathbf{g}}^{\mathsf{T}}(\boldsymbol{\mu})$ is then called the **asymptotic variance** matrix of $\mathbf{g}(\mathbf{T}_n)$ and it is denoted as $\mathsf{avar}\left(\mathbf{g}(\mathbf{T}_n)\right)$. Note that the asymptotic variance has to be understood as the **variance of the asymptotic distribution**, but not as a limiting variance.

As the following three examples show for a sequence of random variables $\{Y_n\}$ the asymptotic variance $\mathsf{avar}(Y_n)$ may exist even if $\mathsf{var}(Y_n)$ does not exist for any $n \in \mathbb{N}$. Further even if $\mathsf{var}(Y_n)$ exists, then it **does not hold that** $\mathsf{var}(Y_n)/\,\mathsf{avar}(Y_n) \to 1$ as $n \to \infty$.

Example 2. Let $X \sim \mathsf{N}(0,1)$ and $\{\varepsilon_n\}$ be a sequence of random variables independent with X such that

$$\mathsf{P}(\varepsilon_n = -\sqrt{n}) = \tfrac{1}{2n}, \qquad \mathsf{P}(\varepsilon_n = 0) = 1 - \tfrac{1}{n}, \qquad \mathsf{P}(\varepsilon_n = \sqrt{n}) = \tfrac{1}{2n}.$$

Define $Y_n = X + \varepsilon_n$ and show that $Y_n \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1)$. Thus $\mathsf{avar}(Y_n) = 1$. On the other hand $\mathsf{var}(Y_n) = 2$ for each $n \in \mathbb{N}$.

Example 3. A random sample $X_1, \ldots, X_n$ from a zero-mean distribution with finite and positive variance. Find the asymptotic distribution of $Y_n = \overline{X}_n \exp\{\overline{X}_n^3\}$. Further compare $\mathsf{var}(Y_n)$ and $\mathsf{avar}(Y_n)$ when X_1 is distributed as $\mathsf{N}(0,1)$.

Example 4. Suppose you have a random sample $X_1, \ldots, X_n$ from a Bernoulli distribution with parameter p_X and you are interested in estimating the logarithm of the odd, i.e. $\theta_X = \log\left(\frac{p_X}{1-p_X}\right)$. Compare the variance and the asymptotic variance of $\widehat{\theta}_X = \log\left(\frac{\overline{X}_n}{1-\overline{X}_n}\right)$.

Example 5. Suppose you have two independent random samples from Bernoulli distribution. Derive the asymptotic distribution of the logarithm of odds-ratio.

Example 6. Suppose we observe independent identically distributed random vectors

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

and denote $\rho = \frac{\mathsf{cov}(X_1,Y_1)}{\sqrt{\mathsf{var}(X_1)\,\mathsf{var}(Y_1)}}$ the (Pearson's) correlation coefficient. Consider the sample correlation coefficient given by

$$\widehat{\rho}_n = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sum_{i=1}^n (Y_i - \overline{Y}_n)^2}}.$$

With the help of Theorem 3(i) derive (the asymptotic representation)

$$\sqrt{n}\,(\widehat{\rho}_n - \rho) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\widetilde{X}_i \widetilde{Y}_i - \tfrac{\rho}{2}\,\widetilde{X}_i^2 - \tfrac{\rho}{2}\,\widetilde{Y}_i^2\right] + o_P(1),$$

where $\widetilde{X}_i = \frac{X_i - \mathsf{E}\,X_1}{\sqrt{\mathsf{var}(X_1)}}$ and $\widetilde{Y}_i = \frac{Y_i - \mathsf{E}\,Y_1}{\sqrt{\mathsf{var}(Y_1)}}$ are standardized versions of X_i and Y_i. Conclude that

$$\sqrt{n}\,(\widehat{\rho}_n - \rho) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \mathsf{var}(Z_i)\big),$$

where $Z_i = \widetilde{X}_i \widetilde{Y}_i - \frac{\rho}{2}\,\widetilde{X}_i^2 - \frac{\rho}{2}\,\widetilde{Y}_i^2$. Derive the asymptotic distribution under the independence of X_i and Y_i and suggest a test of independence.

Further show that if $(X_i, Y_i)^\mathsf{T}$ follows the bivariate normal distribution then

$$\sqrt{n}\,(\widehat{\rho}_n - \rho) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, (1-\rho^2)^2\big).$$

Find the (asymptotic) variance stabilising transformation for $\widehat{\rho}_n$ (see Chapter 1.4) and derive the confidence interval for ρ.

Example 7. Consider a random sample from the Bernoulli distribution with the parameter p_X. Derive the asymptotic distribution of the estimator of $\theta_X = p_X(1-p_X)$ (variance of the Bernoulli distribution) given by $\widehat{\theta}_n = \frac{n}{n-1}\overline{X}_n(1-\overline{X}_n)$.

Example 8. Suppose that we observe $X_1, \ldots, X_n$ of a moving average sequence of order 1 given by

$$X_t = Y_t + \theta\, Y_{t-1}, \quad t \in \mathbb{Z},$$

where $\{Y_t, t \in \mathbb{Z}\}$ is a white noise sequence such that $\mathsf{E}\, Y_t = 0$ and $\mathsf{var}(Y_t) = \sigma^2$.

Using the fact that the autocorrelation function at lag 1 satisfies

$$r(1) = \frac{\theta}{1+\theta^2}$$

derive the estimator of θ and find its asymptotic distribution.

Hint. Note that by Bartlett's formula

$$\sqrt{n}\left(\widehat{r}_n(1) - r(1)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \sigma^2(\theta)\big),$$

where

$$\sigma^2(\theta) = 1 - 3\big(\tfrac{\theta}{1+\theta^2}\big)^2 + 4\big(\tfrac{\theta}{1+\theta^2}\big)^4.$$

The end of class 2 (23. 2. 2024)

1.3 Moment estimators

Suppose that the random vector $\boldsymbol{X}$ has a density $f(\mathbf{x};\boldsymbol{\theta})$ with respect to a σ-finite measure μ and that the density is known up to unknown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \Theta$. Let $\boldsymbol{\theta}_X$ be the true value* of this unknown parameter. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from this distribution and $t_1, \ldots, t_p$ be given real functions. For instance if the observations are one-dimensional one can take $t_j(x) = x^j$, $j \in \{1, \ldots, p\}$. For $j \in \{1, \ldots, p\}$ define the function $\tau_j : \Theta \to \mathbb{R}$ as

$$\tau_j(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\, t_j(\boldsymbol{X}_1) = \int t_j(\mathbf{x}) f(\mathbf{x};\boldsymbol{\theta})\, \mathrm{d}\mu(\mathbf{x}), \quad j \in \{1, \ldots, p\}.$$

Then the moment estimator† $\widehat{\boldsymbol{\theta}}_n$ of the parameter $\boldsymbol{\theta}$ is a solution to the estimating equations

$$\frac{1}{n}\sum_{i=1}^n t_1(\boldsymbol{X}_i) = \tau_1\big(\widehat{\boldsymbol{\theta}}_n\big), \ldots, \frac{1}{n}\sum_{i=1}^n t_p(\boldsymbol{X}_i) = \tau_p\big(\widehat{\boldsymbol{\theta}}_n\big).$$

Example 9. Let $X_1, \ldots, X_n$ be a random sample from the Beta distribution with the density $f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\, \mathbb{I}\{x \in (0,1)\}$. Consider $t_1(x) = x$ and $t_2(x) = x^2$. Then

$$\mathsf{E}\, X_1 = \frac{\alpha}{\alpha+\beta}, \quad \mathsf{E}\, X_1^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

* *skutečná hodnota* † *Momentový odhad*

Thus the estimating equations are

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\widehat{\alpha}}{\widehat{\alpha}+\widehat{\beta}}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\widehat{\alpha}(\widehat{\alpha}+1)}{(\widehat{\alpha}+\widehat{\beta})(\widehat{\alpha}+\widehat{\beta}+1)}.$$

Now denote $\boldsymbol{\tau}(\alpha,\beta) = (\tau_1(\alpha,\beta), \tau_2(\alpha,\beta))^{\mathsf{T}}$, where

$$\tau_1(\alpha,\beta) = \frac{\alpha}{\alpha+\beta}, \ \tau_2(\alpha,\beta) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Thus one can rewrite the estimating equations as

$$\boldsymbol{\tau}(\widehat{\alpha},\widehat{\beta}) = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2\Big)^{\mathsf{T}}.$$

Now provided that the inverse function $\boldsymbol{\tau}^{-1}$ exists one can write

$$\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} = \boldsymbol{\tau}^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2\Big)^{\mathsf{T}}.$$

and use Δ-theorem to derive the asymptotic distribution of the estimator $\binom{\widehat{\alpha}}{\widehat{\beta}}$.

Now in the general situation put

$$\mathbf{T}_n = \Big(\frac{1}{n}\sum_{i=1}^{n} t_1(\boldsymbol{X}_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} t_p(\boldsymbol{X}_i)\Big)^{\mathsf{T}} \tag{5}$$

and define the mapping $\boldsymbol{\tau}: \Theta \mapsto \mathbb{R}^p$ as $\boldsymbol{\tau}(\boldsymbol{\theta}) = \big(\tau_1(\boldsymbol{\theta}), \ldots, \tau_p(\boldsymbol{\theta})\big)^{\mathsf{T}}$. Note that provided there exists an inverse mapping $\boldsymbol{\tau}^{-1}$ one can write

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = \sqrt{n}\left(\boldsymbol{\tau}^{-1}(\mathbf{T}_n) - \boldsymbol{\tau}^{-1}\big(\boldsymbol{\tau}(\boldsymbol{\theta}_X)\big)\right). \tag{6}$$

Thus the asymptotic normality of the moment estimator $\widehat{\boldsymbol{\theta}}_n$ would follow by the $\boldsymbol{\Delta}$-theorem (Theorem 3) with $\mathbf{g} = \boldsymbol{\tau}^{-1}$. This is formalized in the following theorem.

Theorem 4. *Let $\boldsymbol{\theta}_X$ be an interior point of Θ and $\max_{j\in\{1,\ldots,p\}} \mathsf{var}_{\boldsymbol{\theta}_X}(t_j(\boldsymbol{X}_1)) < \infty$. Further let the function $\boldsymbol{\tau}$ be one-to-one and have continuous first-order partial derivatives in a neighbourhood of $\boldsymbol{\theta}_X$. Finally let the Jacobi matrix $\mathbb{D}_{\boldsymbol{\tau}}(\boldsymbol{\theta}_X)$ be regular. Then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \mathbb{D}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\theta}_X)\, \mathbb{\Sigma}(\boldsymbol{\theta}_X)\, [\mathbb{D}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\theta}_X)]^{\mathsf{T}}\big),$$

where $\mathbb{\Sigma}(\boldsymbol{\theta}_X) = \mathsf{var}_{\boldsymbol{\theta}_X}\big(t_1(\boldsymbol{X}_1), \ldots, t_p(\boldsymbol{X}_1)\big)$.

Proof. By the assumptions of the theorem and the inverse function theorem (Theorem A13) there exists an open neighbourhood U containing $\boldsymbol{\theta}_X$ and an open neighbourhood V containing $\boldsymbol{\tau}(\boldsymbol{\theta}_X)$ such that $\boldsymbol{\tau} : U \to V$ is a differentiable bijection with a differentiable inverse $\boldsymbol{\tau}^{-1} : V \to U$. Further note that $\mathbf{T}_n$ defined in (5) satisfies $\mathsf{P}\left(\mathbf{T}_n \in V\right) \xrightarrow[n\to\infty]{} 1$. Thus one can use (6) and apply the Δ-theorem (Theorem 3) with $\mathbf{g} = \boldsymbol{\tau}^{-1}$, $\boldsymbol{\mu} = \boldsymbol{\tau}(\boldsymbol{\theta}_X)$ and $A = V$ to get

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\Big(\mathbf{0}, \mathbb{D}_{\boldsymbol{\tau}^{-1}}\left(\boldsymbol{\tau}(\boldsymbol{\theta}_X)\right) \mathbb{\Sigma}(\boldsymbol{\theta}_X) \left[\mathbb{D}_{\boldsymbol{\tau}^{-1}}\left(\boldsymbol{\tau}(\boldsymbol{\theta}_X)\right)\right]^{\mathsf{T}}\Big).$$

The statement of the theorem now follows from the identity

$$\mathbb{D}_{\boldsymbol{\tau}^{-1}}\left(\boldsymbol{\tau}(\boldsymbol{\theta}_X)\right) = \mathbb{D}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\theta}_X).$$

□

The asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ is usually estimated as

$$\tfrac{1}{n}\, \mathbb{D}_{\boldsymbol{\tau}}^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right) \widehat{\mathbb{\Sigma}}_n \left[\mathbb{D}_{\boldsymbol{\tau}}^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\right]^{\mathsf{T}},$$

where as $\widehat{\mathbb{\Sigma}}_n$ one can take either $\mathbb{\Sigma}(\widehat{\boldsymbol{\theta}}_n)$ or the empirical variance matrix

$$\widehat{\mathbb{\Sigma}}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{Z}_i - \overline{\mathbf{Z}}_n\right)\left(\mathbf{Z}_i - \overline{\mathbf{Z}}_n\right)^{\mathsf{T}},$$

with $\mathbf{Z}_i = \left(t_1(\boldsymbol{X}_i), \ldots, t_p(\boldsymbol{X}_i)\right)^{\mathsf{T}}$.

Confidence intervals for θ_{Xj}

Let θ_{Xj} stand for the j-th component of the true value of the parameter $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^{\mathsf{T}}$.

Put $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\theta}_{n1}, \ldots, \widehat{\theta}_{np}\right)^{\mathsf{T}}$ and $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^{\mathsf{T}}$. By Theorem 4 we know that

$$\sqrt{n}\left(\widehat{\theta}_{nj} - \theta_{Xj}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, v_{jj}(\boldsymbol{\theta}_X)\big), \qquad j \in \{1, \ldots, p\},$$

where $v_{jj}(\boldsymbol{\theta}_X)$ is the j-th diagonal element of the asymptotic variance matrix

$$\mathbb{V} = \mathbb{D}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\theta}_X)\, \mathbb{\Sigma}(\boldsymbol{\theta}_X) \left[\mathbb{D}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\theta}_X)\right]^{\mathsf{T}}. \tag{7}$$

Thus the (asymptotic two-sided) confidence interval for θ_{Xj} is given by

$$\left(\widehat{\theta}_{nj} - u_{1-\frac{\alpha}{2}} \sqrt{\tfrac{\widehat{v}_{jj}}{n}},\ \widehat{\theta}_{nj} + u_{1-\frac{\alpha}{2}} \sqrt{\tfrac{\widehat{v}_{jj}}{n}}\right),$$

where $\widehat{v}_{jj}$ is the j-th diagonal element of the estimated variance matrix

$$\widehat{\mathbb{V}}_n = \mathbb{D}_{\boldsymbol{\tau}}^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right) \widehat{\mathbb{\Sigma}}_n \left[\mathbb{D}_{\boldsymbol{\tau}}^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\right]^{\mathsf{T}}.$$

Applications of moment estimators

As maximum likelihood estimators are preferred over moment estimators, the use of moment estimators is limited. Nevertheless the moment estimators can be of interest in the following situations:

- the calculation of the maximum likelihood estimate is computationally too prohibitive due to a very complex model or a huge amount of data;
- moment estimates can be used as the starting values for the numerical algorithms that search for maximum likelihood estimates.

The choice of the functions $t_1, \dots, t_p$

The most common choice $t_j(x) = x^j$, where $j \in \{1, \dots, p\}$ for the univariate observations is not necessary the most appropriate one. The idea is to choose the functions $t_1, \dots, t_p$ so that the asymptotic variance matrix (7) is in some sense 'minimized'. But this is usually a too difficult problem. Nevertheless one should at least check that the vector function $\boldsymbol{\tau} : \Theta \to \mathbb{R}^p$ is one-to-one, otherwise the parameter $\boldsymbol{\theta}_X$ might not be identifiable with the given $t_1, \dots, t_p$.

Now the continuity of $\boldsymbol{\tau}$ guarantees the consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$. To guarantee also the asymptotic normality one needs that the Jacobi matrix $\mathbb{D}_{\boldsymbol{\tau}}(\boldsymbol{\theta})$ is regular for each $\boldsymbol{\theta} \in \Theta$.

To be more specific, consider the one-dimensional parameter θ and for a given function t introduce

$$\tau(\theta) = \mathsf{E}_\theta \, t(\boldsymbol{X}_1).$$

Then we need that $\tau : \Theta \to \mathbb{R}$ is a one-to-one function. Otherwise it might happen that with probability going to one the estimating function

$$\tau(\widehat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} t(\boldsymbol{X}_i).$$

has more roots (whose values are in the parameter space Θ) and we do not know which of the root is the appropriate (consistent) one.

Example 10. Let $X_1, \dots, X_n$ be independent identically distributed random variables from the discrete distribution given as

$$\mathsf{P}(X_1 = -1) = p, \quad \mathsf{P}(X_1 = 0) = 1 - p - p^2, \quad \mathsf{P}(X_1 = 2) = p^2,$$

where $p \in \Theta = (0, \frac{-1+\sqrt{5}}{2})$.

Now the standard choice $t(x) = x$ yields that $\tau(p) = \mathsf{E}_p \, X_1 = 2p^2 - p$. Note that the estimating equation given by

$$2\widehat{p}_n^2 - \widehat{p}_n = \overline{X}_n$$

has two roots

$$\widehat{p}_n^{(1,2)} = \tfrac{1}{4} \pm \sqrt{\tfrac{\overline{X}_n}{2} + \tfrac{1}{16}}.$$

Show that if the true value of the parameter $p_X \in (0, \frac{1}{2})$, then

$$\widehat{p}_n^{(1)} \xrightarrow[n\to\infty]{P} \tfrac{1}{4} - |p_X - \tfrac{1}{4}|, \quad \widehat{p}_n^{(2)} \xrightarrow[n\to\infty]{P} \tfrac{1}{4} + |p_X - \tfrac{1}{4}|.$$

Thus except for the $p_X = \frac{1}{4}$ the roots $\widehat{p}_n^{(1)}$ and $\widehat{p}_n^{(2)}$ converge in distribution to different limits and only one of these limits is the true value of the parameter p_X. Note also $p_X = \frac{1}{4}$, then both the roots are consistent, but as $\tau'(\frac{1}{4}) = 0$ neither of the roots is asymptotically normal.

Show that taking $t(x) = x^2$ or simply $t(x) = \mathbb{I}\{x = -1\}$ does not introduce such problematic issues.

1.4 Confidence intervals and asymptotic variance-stabilising transformation

In this section* we are interested in constructing a confidence interval for (one-dimensional) parameter θ_X. Suppose we have an estimator $\widehat{\theta}_n$ of parameter θ_X such that

$$\sqrt{n}\,\big(\widehat{\theta}_n - \theta_X\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \sigma^2(\theta_X)\big), \tag{8}$$

where $\sigma^2(\cdot)$ is a function continuous in the true value of the parameter (θ_X).

Standard asymptotic confidence interval of 'Wald' type

This interval is based on the fact that

$$\frac{\sqrt{n}\,\big(\widehat{\theta}_n - \theta_X\big)}{\sigma(\widehat{\theta}_n)} \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, 1\big)$$

and thus

$$\Big(\widehat{\theta}_n - \tfrac{u_{1-\alpha/2}\,\sigma(\widehat{\theta}_n)}{\sqrt{n}}, \widehat{\theta}_n + \tfrac{u_{1-\alpha/2}\,\sigma(\widehat{\theta}_n)}{\sqrt{n}}\Big) \tag{9}$$

is a confidence interval for parameter θ_X with the asymptotic coverage $1 - \alpha$.

The advantage of the confidence interval (9) is that it is easy to calculate. On the other hand the simulations show that for small sample size and/or if $|\sigma'(\theta)|$ is large then the actual coverage of this confidence interval can be much smaller than $1 - \alpha$.

* Not presented at the lecture. It is assumed that this is known from the bachelor degree.

Implicit (asymptotic) confidence interval of 'Wilson' type

This interval is based directly on (8) and it is given implicitly by

$$\left\{\theta : \left|\frac{\sqrt{n}\,(\widehat{\theta}_n - \theta)}{\sigma(\theta)}\right| < u_{1-\alpha/2}\right\}. \tag{10}$$

Note that (10) can be viewed as the set of θ for which we do not reject the null hypothesis

$$H_0 : \theta_X = \theta \quad \text{against the alternative} \quad H_1 : \theta_X \neq \theta$$

with the critical region

$$\left|\frac{\sqrt{n}\,(\widehat{\theta}_n - \theta)}{\sigma(\theta)}\right| \geq u_{1-\alpha/2}.$$

In fact the set given by (10) does not have to be necessarily an interval. But usually the function $\theta \mapsto \frac{\sqrt{n}\,(\widehat{\theta}_n - \theta)}{\sigma(\theta)}$ is not increasing which guarantees that the set (10) is indeed an interval.

It was observed that usually the actual coverage of this implicit confidence interval is closer to $1-\alpha$ than for the standard asymptotic confidence interval (9). In particular if one is interested in two-sided intervals then the implicit confidence interval (10) works surprisingly well even for very small samples. Its disadvantage is that in general one does not have an explicit formula for this interval and often it has to be found with the help of methods of numerical mathematics.

Confidence interval based on the transformation stabilizing the asymptotic variance

Put $g(\theta) = \int \frac{1}{\sigma(\theta)}\,\mathrm{d}\theta$. Then with the help of (8) and Δ-theorem it holds

$$\sqrt{n}\,\big(g(\widehat{\theta}_n) - g(\theta_X)\big) \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1).$$

Thus the set $\left(g(\widehat{\theta}_n) - \frac{u_{1-\alpha/2}}{\sqrt{n}}, g(\widehat{\theta}_n) + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)$ is a confidence set for $g(\theta_X)$. Now as g is an increasing function (note that $g'(\theta) > 0$) one can conclude that

$$\left(g^{-1}\left(g(\widehat{\theta}_n) - \tfrac{u_{1-\alpha/2}}{\sqrt{n}}\right), g^{-1}\left(g(\widehat{\theta}_n) + \tfrac{u_{1-\alpha/2}}{\sqrt{n}}\right)\right) \tag{11}$$

is a confidence interval for the parameter θ_X with the asymptotic coverage $1-\alpha$.

The actual coverage of this confidence interval is also usually closer to $1-\alpha$ than for the standard confidence interval (9). On the other hand when one is interested in two-sided confidence interval then the implicit confidence interval (10) usually works better. But the advantage of (11) is that one usually has an explicit formula for the confidence interval (provided that g and g^{-1} can be explicitly calculated). The confidence interval (11) is also usually a better choice than the the implicit confidence interval when one is interested in one-sided confidence intervals.

Example 11. A random sample from Poisson distribution. Find the transformation that stabilises the asymptotic variance of $\overline{X}_n$ and based on this transformation derive the asymptotic confidence intervals for λ.

Example 12. Fisher's Z-transformation and various confidence intervals for the correlation coefficient.

Example 13. Consider a random sample from Bernoulli distribution. Find the asymptotic variance-stabilizing transformation for $\overline{X}_n$ and construct the confidence interval based on this transformation.

Literature: van der Vaart [2000] – Chapters 2.1, 2.2, 3.1, 3.2 and 4.1. In particular Theorems 2.3, 2.4, 2.8 and 3.1.

The expected end of class 3 (27. 2. 2024)

2 Maximum likelihood methods

Suppose we have a random sample of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ being distributed as the generic vector $\boldsymbol{X} = (X_1, \ldots, X_k)^\mathsf{T}$ that has a density $f(\mathbf{x}; \boldsymbol{\theta})$ with respect to a σ-finite measure μ and that the density is known up to an unknown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \Theta$. Let $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$ be the true value of the parameter.

Define the *likelihood function* as

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\theta})$$

and the *log-likelihood function* as

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

The *maximum likelihood estimator* of parameter $\boldsymbol{\theta}_X$ is defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}) \quad \text{or alternatively as} \quad \widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}). \tag{12}$$

The (exact) distribution of $\widehat{\boldsymbol{\theta}}_n$ is usually too difficult or even impossible to calculate. Thus to make the inference about $\boldsymbol{\theta}_X$ we need to derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_n$.

2.1 Asymptotic normality of maximum likelihood estimator

Regularity assumptions

Let $I(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}} \left[\frac{\partial \log f(\boldsymbol{X}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, \frac{\partial \log f(\boldsymbol{X}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} \right]$ be the Fisher information matrix.

[R0] For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ it holds that $f(\mathbf{x}; \boldsymbol{\theta}_1) = f(\mathbf{x}; \boldsymbol{\theta}_2)$ μ-almost everywhere if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. (*Identifiability*)

[R1] The number of parameters p in the model is constant.

[R2] The support set $S = \left\{ \mathbf{x} \in \mathbb{R}^k : f(\mathbf{x}; \boldsymbol{\theta}) > 0 \right\}$ does not depend on the value of the parameter $\boldsymbol{\theta}$.

[R3] (The true value of the parameter) $\boldsymbol{\theta}_X$ is an interior point of the parameter space Θ.

[R4] The density $f(\mathbf{x}; \boldsymbol{\theta})$ is three-times differentiable with respect to $\boldsymbol{\theta}$ on an open neighbourhood U of $\boldsymbol{\theta}_X$ (for μ-almost all $\mathbf{x}$). Further there exists a function $M(\mathbf{x})$ such that for each $j, k, l \in \{1, \ldots, p\}$

$$\sup_{\boldsymbol{\theta} \in U} \left| \frac{\partial^3 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j \, \partial \theta_k \, \partial \theta_l} \right| \leq M(\mathbf{x}),$$

for μ-almost all $\mathbf{x}$ and

$$\mathsf{E}_{\boldsymbol{\theta}_X} M(\boldsymbol{X}_1) < \infty.$$

[R5] The Fisher information matrix $I(\boldsymbol{\theta}_X)$ is finite and positive definite.

[R6] The order of differentiation and integration can be interchanged in expressions such as

$$\frac{\partial}{\partial \theta_j} \int h(\mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x}) = \int \frac{\partial}{\partial \theta_j} h(\mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x}),$$

where $h(\mathbf{x}; \boldsymbol{\theta})$ is either $f(\mathbf{x}; \boldsymbol{\theta})$ or $\partial f(\mathbf{x}; \boldsymbol{\theta}) / \partial \theta_k$ and $j, k \in \{1, \ldots, p\}$.

Note that thanks to assumption **[R6]** one can calculate the Fisher information matrix as

$$I(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2 \log f(\boldsymbol{X}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^\mathsf{T}} \right],$$

see for instance Lemma 5.3 of Lehmann and Casella [1998] or Theorem 7.27 of Anděl [2007].

Example 14. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $\mathsf{N}(\mu_1 + \mu_2, 1)$. Then the identifiability assumption **[R0]** is not satisfied for the vector parameter $\boldsymbol{\theta} = (\mu_1, \mu_2)^\mathsf{T}$.

Example 15. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution $\mathsf{U}(0, \theta)$. Note that assumption **[R2]** is not satisfied.

Show that the maximum likelihood estimator of θ is $\widehat{\theta}_n = \max_{1 \le i \le n} \{X_i\}$. Derive the asymptotic distribution of $n\,(\widehat{\theta}_n - \theta)$.

Remark 5. Note that in particular assumption **[R4]** is rather strict. There are ways how to derive the asymptotic normality of the maximum likelihood estimator under less strict assumptions but that would require concepts that are out of the scope of this course.

The *score function* of the i-th observation $\boldsymbol{X}_i$ for the parameter $\boldsymbol{\theta}$ is defined as

$$\mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{\partial \log f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

The random vector

$$\mathbf{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is called *the score statistic*.

We search for the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ as a solution of the system of the likelihood equations

$$\mathbf{U}_n(\widehat{\boldsymbol{\theta}}_n) \stackrel{!}{=} \mathbf{0}_p. \tag{13}$$

Further define *the observed (empirical) information matrix* as

$$I_n(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial \mathbf{U}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_i;\boldsymbol{\theta}),$$

where

$$I(\boldsymbol{X}_i;\boldsymbol{\theta}) = -\frac{\partial \mathbf{U}(\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = -\frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^\mathsf{T}}$$

is the contribution of the i-th observation to the information matrix.

In what follows it will be useful to prove that $I_n\big(\widehat{\boldsymbol{\theta}}_n\big) \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_X) = \mathsf{E}\, I(\boldsymbol{X}_1;\boldsymbol{\theta})$ (provided that $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$). The following technical lemma is a generalization of this result that will be convenient in the proofs of the several theorems that will follow.

Lemma 1. *Suppose that assumptions* [R0]-[R6] *hold. Let ε_n be a sequence of positive numbers going to zero. Then*

$$\max_{j,k\in\{1,\ldots,p\}} \sup_{\boldsymbol{\theta}\in U_{\varepsilon_n}} \Big|\big(I_n(\boldsymbol{\theta}) - I(\boldsymbol{\theta}_X)\big)_{jk}\Big| = o_P(1),$$

where

$$U_{\varepsilon_n} = \big\{\boldsymbol{\theta}\in\Theta:\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_X\| \le \varepsilon_n\big\}$$

and $\big(I_n(\boldsymbol{\theta}) - I(\boldsymbol{\theta}_X)\big)_{jk}$ stands for the (j,k)-element of the difference of the matrices $I_n(\boldsymbol{\theta}) - I(\boldsymbol{\theta}_X)$.

Proof. Using assumption [R4] and the law of large numbers one can bound

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in U_{\varepsilon_n}} \Big|\big(I_n(\boldsymbol{\theta}) - I(\boldsymbol{\theta}_X)\big)_{jk}\Big| &\le \sup_{\boldsymbol{\theta}\in U_{\varepsilon_n}} \Big|\big(I_n(\boldsymbol{\theta}) - I_n(\boldsymbol{\theta}_X)\big)_{jk}\Big| + \Big|\big(I_n(\boldsymbol{\theta}_X) - I(\boldsymbol{\theta}_X)\big)_{jk}\Big| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{p} M(\boldsymbol{X}_i)\,\varepsilon_n + o_P(1) = O_P(1)\,o(1) + o_P(1) = o_P(1),\end{aligned}$$

which implies the statement of the lemma. □

The end of class 4 (1. 3. 2024)

Corollary 1. *Let the assumptions of Lemma 1 be satisfied. Further let $\widehat{\mathbf{t}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$. Then for each $j,k\in\{1,\ldots,p\}$*

$$\Big|\big(I_n(\widehat{\mathbf{t}}_n) - I(\boldsymbol{\theta}_X)\big)_{jk}\Big| = o_P(1).$$

Proof. Note that $\widehat{\mathbf{t}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ implies that there exists a sequence of positive constants $\{\varepsilon_n\}$ going to zero such that

$$\mathsf{P}\left(\widehat{\mathbf{t}}_n \in U_{\varepsilon_n}\right) \xrightarrow[n\to\infty]{} 1.$$

The corollary now follows from Lemma 1 and from the fact that one can bound

$$\begin{aligned}\left|\big(I_n(\widehat{\mathbf{t}}_n) - I(\boldsymbol{\theta}_X)\big)_{jk}\right| &= \left|\big(I_n(\widehat{\mathbf{t}}_n) - I(\boldsymbol{\theta}_X)\big)_{jk}\right| \mathbb{I}\{\widehat{\mathbf{t}}_n \in U_{\varepsilon_n}\} + \left|\big(I_n(\widehat{\mathbf{t}}_n) - I(\boldsymbol{\theta}_X)\big)_{jk}\right| \mathbb{I}\{\widehat{\mathbf{t}}_n \notin U_{\varepsilon_n}\} \\ &\leq \sup_{\boldsymbol{\theta} \in U_{\varepsilon_n}} \left|\big(I_n(\boldsymbol{\theta}) - I(\boldsymbol{\theta}_X)\big)_{jk}\right| + \left|\big(I_n(\widehat{\mathbf{t}}_n) - I(\boldsymbol{\theta}_X)\big)_{jk}\right| \mathbb{I}\{\widehat{\mathbf{t}}_n \notin U_{\varepsilon_n}\}.\end{aligned}$$

□

Theorem 5. *Suppose that assumptions* **[R0]**-**[R6]** *hold.*

(i) Then with probability tending to one as $n \to \infty$ there exists a consistent solution $\widehat{\boldsymbol{\theta}}_n$ of the likelihood equations (13).[*]

(ii) Any consistent solution $\widehat{\boldsymbol{\theta}}_n$ of the likelihood equations (13) satisfies,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = \big[I(\boldsymbol{\theta}_X)\big]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) + o_P(1), \tag{14}$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, I^{-1}(\boldsymbol{\theta}_X)\big). \tag{15}$$

Proof of (i). First, we need to prove the existence of the consistent root $\widehat{\boldsymbol{\theta}}_n$ of the likelihood equations. This can be deduced from a more general Theorem 9. An alternative approach can be found in the proof of Theorem 5.1 of Lehmann and Casella [1998, Chapter 6].

Proof of (ii). Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a consistent solution of the likelihood equations. Then by the mean value theorem (applied to each component of $\mathbf{U}_n(\boldsymbol{\theta})$) one gets that

$$\mathbf{0}_p = \mathbf{U}_n\big(\widehat{\boldsymbol{\theta}}_n\big) = \mathbf{U}_n(\boldsymbol{\theta}_X) - n\, I_n^*\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right),$$

where I_n^* is a matrix with the elements

$$i_{n,jk}^* = \frac{1}{n}\sum_{i=1}^{n} \frac{-\partial^2 \log f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\partial\theta_j\, \partial\theta_k}\Big|_{\boldsymbol{\theta}=\widehat{\mathbf{t}}_n^{(j)}}, \qquad j,k \in \{1,\ldots,p\},$$

with $\widehat{\mathbf{t}}_n^{(j)}$ being between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_X$. Thus the consistency of $\widehat{\boldsymbol{\theta}}_n$ implies that $\widehat{\mathbf{t}}_n^{(j)} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ and one can use Corollary 1 to show that

$$I_n^* \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_X). \tag{16}$$

Thus with probability going to one there exists $[I_n^*]^{-1}$ and one can write

$$n\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = [I_n^*]^{-1}\, \mathbf{U}_n(\boldsymbol{\theta}_X).$$

[*] Thus defining the estimator as an appropriately chosen root of the likelihood equations (provided that the likelihood equations has at least one root) and zero otherwise yields a consistent estimator of θ_X.

Now the central limit theorem for independent identically distributed random vectors implies that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, I(\boldsymbol{\theta}_X)\big). \tag{17}$$

Note that (17) yields that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) = O_P(1)$. Thus using (16) and CMT (Theorem 1) implies that

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) &= [I_n^*]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) \\
&= \left[I^{-1}(\boldsymbol{\theta}_X) + o_P(1)\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) \\
&= I^{-1}(\boldsymbol{\theta}_X) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}_X) + o_P(1).
\end{aligned}$$

Now (15) follows by CS (Theorem 2) and (17).

□

Remark 6. While the proof of consistency is for $p = 1$ relatively simple [see e.g. Theorem 22 of Nagy], for $p > 1$ it is much more involved. The reason is that while the border of the neighbourbood in $\mathbb{R}$ is a two-point set, in $\mathbb{R}^p$ $(p > 1)$ it is an uncountable set.

Remark 7. Note that strictly speaking Theorem 5 does not guarantee the asymptotic normality of the maximum likelihood estimator but of an appropriately chosen root of the likelihood equations (13). As illustrated in Example 19 it may happen that the maximum likelihood estimator defined by (12) is not a consistent estimator of $\boldsymbol{\theta}_X$ even if all the regularity assumptions **[R0]**-**[R6]** are satisfied. It may also happen that the maximum likelihood estimator does not exist (see the example on page 21). That is why some authors define the maximum likelihood estimator in regular families as an appropriately chosen root of the likelihood equations.

Fortunately for many models commonly used in applications the log-likelihood function $\ell_n(\boldsymbol{\theta})$ is (almost surely) convex. Then the maximum likelihood estimator is the only solution to the likelihood equations and Theorem 5 guarantees that this estimator is asymptotically normal. If $\ell_n(\boldsymbol{\theta})$ is not convex, there might be more roots to the likelihood equations and the choice of an appropriate (consistent) root of the estimating equations is more delicate both from the theoretical as well as the numerical point of view. Other available consistent estimators (e.g. moment estimators) can be very useful as for instance the starting points of the numerical algorithms that search for the root of the likelihood equations.

Example 16. Let $X_1, \ldots, X_n$ be a random sample from Bernoulli distribution* $\mathsf{Be}(p)$. Note that if either $\sum_{i=1}^{n} X_i = 0$ or $\sum_{i=1}^{n} X_i = n$ then there is no root of the likelihood equation. Nevertheless the probability of both events converges to zero as $n \to \infty$ whenever $p_X \in (0, 1)$.

* Alternativního rozdělení

Example 17. Let $X_1, \ldots, X_n$ be a random sample from the Pareto distribution with the density

$$f(x) = \frac{\beta\, \alpha^{\beta}}{x^{\beta+1}}\, \mathbb{I}\{x \geq \alpha\}, \qquad \beta > 0,\ \alpha > 0,$$

where both parameters are unknown.

(i) Find the maximum likelihood estimator of $\widehat{\boldsymbol{\theta}}_n = \big(\widehat{\alpha}_n, \widehat{\beta}_n\big)^{\mathsf{T}}$ of the parameter $\boldsymbol{\theta} = (\alpha, \beta)^{\mathsf{T}}$.

(ii) Derive the asymptotic distribution of $n\,\big(\widehat{\alpha}_n - \alpha\big)$.

(iii) Derive the asymptotic distribution of $\sqrt{n}\,\big(\widehat{\beta}_n - \beta\big)$.

Example 18. Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{N}(\mu, 1)$ where the parameter space for the parameter μ is restricted to $[0, \infty)$. Find the maximum likelihood estimator of μ and derive its asymptotic distribution. Do not forget to consider the special case $\mu = 0$.

Example 19. Let $X_1, \ldots, X_n$ be a random sample from the mixture of distributions $\mathsf{N}(0, 1)$ and $\mathsf{N}\big(\theta, \exp\{-2/\theta^2\}\big)$ with equal weights and the parameter space given by $\Theta = (0, \infty)$. Define the estimator of the parameter θ as $\widehat{\theta}_n^{(ML)} = \arg\max_{\theta \in \Theta} \ell_n(\theta)$. Then it can be shown that $\widehat{\theta}_n^{(ML)} \xrightarrow[n\to\infty]{P} 0$, thus $\widehat{\theta}_n^{(ML)}$ is not consistent estimator.

Nevertheless note that the assumptions **[R0]**-**[R6]** are met. Thus by Theorem 5 there exists a different root $(\widehat{\theta}_n)$ of the likelihood equation such that this estimator satisfies (14) and (15).

Example 20. Let $X_1, \ldots, X_n$ be a random sample from the mixture of distributions $\mathsf{N}(0, 1)$ and $\mathsf{N}(\mu, \sigma^2)$ with equal weights and the parameter space for the parameter $\boldsymbol{\theta} = (\mu, \sigma)^{\mathsf{T}}$ is given by $\Theta = \mathbb{R} \times (0, \infty)$. Show that

$$\sup_{(\mu,\sigma^2)^{\mathsf{T}} \in \Theta} \ell_n(\mu, \sigma^2) = \infty$$

and that the maximum likelihood estimator does not exist. But similarly as in Example 19 Theorem 5 still holds.

2.2 Asymptotic efficiency of maximum likelihood estimators

Recall the Rao-Cramér inequality. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from the regular family of densities $\mathcal{F} = \big\{f(\mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\big\}$, and $\mathbf{T}_n$ be an *unbiased* estimator of $\boldsymbol{\theta}_X$ (based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$). Then

$$\mathsf{var}\,\big(\mathbf{T}_n\big) - \frac{1}{n}\, I^{-1}(\boldsymbol{\theta}_X) \geq 0.$$

By Theorem 5 we have that (under appropriate regularity assumptions)

$$\mathsf{avar}\,\big(\widehat{\boldsymbol{\theta}}_n\big) = \frac{1}{n}\, I^{-1}(\boldsymbol{\theta}_X).$$

Thus the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ attains the lower bound in Rao-Cramér inequality.

Remark 8. Note that strictly speaking comparing with the Rao-Cramér bound is not fair. Generally, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ is not unbiased. Further, Rao-Cramér inequality speaks about the bound on the variance, but we compare the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ with this bound. Nevertheless it can be shown that in regular models there exists a lower bound for the asymptotic variances of the estimators that are asymptotically normal with zero mean and in some (natural) sense regular (see Example 21 below). And this bound is indeed given by $\frac{1}{n} I^{-1}(\boldsymbol{\theta}_X)$. See also Serfling [1980, Chapter 4.1.3] and the references therein.

Example 21. Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{N}(\theta, 1)$, where $\theta \in \mathbb{R}$. Define the estimator of θ as

$$\widehat{\theta}_n^{(S)} = \begin{cases} 0, & \text{if} \quad |\overline{X}_n| \leq n^{-1/4}, \\ \overline{X}_n, & \text{if} \quad |\overline{X}_n| > n^{-1/4}. \end{cases}$$

This estimator is called also *Hodges* or *shrinkage* estimator. Show that if $\theta_X \neq 0$ then $\sqrt{n}\,(\widehat{\theta}_n^{(S)} - \theta_X) \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1)$ and if $\theta_X = 0$ then even $n^r(\widehat{\theta}_n^{(S)} - \theta_X) \xrightarrow[n\to\infty]{P} 0$ for each $r \in \mathbb{N}$. Thus from the point-wise asymptotic point of view, the estimator $\widehat{\theta}_n^{(S)}$ is better than the standard maximum likelihood estimator that is given by the sample mean $\overline{X}_n$.

But on the other hand consider the following sequence of the true values of the parameter $\theta_X^{(n)} = n^{-1/4}$. Then show that for an arbitrarily large value of K

$$\liminf_{n\to\infty} \mathsf{P}\left(\sqrt{n}\,(\widehat{\theta}_n^{(S)} - \theta_X^{(n)}) \geq K\right) \geq \frac{1}{2}\,.$$

Thus the sequence $\sqrt{n}\,(\widehat{\theta}_n^{(S)} - \theta_X^{(n)})$ is not tight and so it does not converge in distribution. Such a non-uniform behaviour of the estimator $\widehat{\theta}_n^{(S)}$ is usually considered as undesirable. Thus the aim of the regularity assumptions on the estimators is to avoid such estimators that from the point-wise view can be considered as superior (*superefficient*) to the maximum likelihood estimators.*

The end of class 5 (5. 3. 2024)

2.3 Estimation of the asymptotic variance matrix

To do the inference about the parameter $\boldsymbol{\theta}_X$ we need to have a consistent estimator of $I(\boldsymbol{\theta}_X)$. Usually, we use one of the following estimators

$$I\big(\widehat{\boldsymbol{\theta}}_n\big) \quad \text{or} \quad I_n\big(\widehat{\boldsymbol{\theta}}_n\big) \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}\big(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n\big)\, \mathbf{U}^{\mathsf{T}}(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n).$$

The consistency of $I\big(\widehat{\boldsymbol{\theta}}_n\big)$ follows by CMT (Theorem 1), provided (the matrix function) $I(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}_X$, which follows by assumption [R4].

* Note that the issue of superefficiency is behind the claimed 'oracle'-properties of some regularized estimators (e.g. adaptive LASSO), see Leeb and Pötscher [2008] and the references therein.

The consistency of $I_n(\widehat{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_X)$ follows from Corollary 1 and Theorem 5.

On the other hand the consistency of $\frac{1}{n}\sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n)\mathbf{U}^{\mathsf{T}}(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n)$ does not automatically follow from assumptions [R0]-[R6]. It can be proved analogously as Corollary 1 provided the following assumption holds.

[R7] There exists an open neighbourhood U of $\boldsymbol{\theta}_X$ such that for each j,k in $\{1,\dots,p\}$ there exists a function $M_{jkl}(\mathbf{x})$ such that

$$\sup_{\boldsymbol{\theta}\in U}\left|\frac{\partial^2 \log f(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k}\right| \le M_{jk}(\mathbf{x})$$

for μ-almost all $\mathbf{x}$ and

$$\mathsf{E}_{\boldsymbol{\theta}_X} M_{jk}^2(\boldsymbol{X}_i) < \infty.$$

Literature: Anděl [2007] Chapter 7.6.5, Lehmann and Casella [1998] Chapter 6.5, Kulich [2014].

2.4 Asymptotic tests (without nuisance parameters)

Suppose we are interested in testing the null hypothesis

$$H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0 \text{ against the alternative } H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0.$$

Let $\widehat{I}_n$ be an estimate of the Fisher information matrix $I(\boldsymbol{\theta}_X)$ or $I(\boldsymbol{\theta}_0)$. Basically there are three tests that can be considered.

Likelihood ratio test is based on the test statistic

$$LR_n = 2\left(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)\right).$$

Wald test is based on the test statistic

$$W_n = n\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)^{\mathsf{T}} \widehat{I}_n \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right).$$

Rao score test is based on the test statistic

$$R_n = \frac{1}{n}\,\mathbf{U}_n^{\mathsf{T}}(\boldsymbol{\theta}_0)\,\widehat{I}_n^{-1}\,\mathbf{U}_n(\boldsymbol{\theta}_0). \tag{18}$$

Note that the advantage of the likelihood ratio test (LR_n) is that one does not need to estimate the Fisher information matrix. On the other hand the advantage of Rao score test (R_n) is that you do not need to calculate the maximal likelihood estimator $\widehat{\boldsymbol{\theta}}_n$. That is why in Rao score statistic (R_n) one uses usually either $I(\boldsymbol{\theta}_0)$ or $I_n(\boldsymbol{\theta}_0)$ as $\widehat{I}_n$. On the other hand usually (for historical reasons) $I(\widehat{\boldsymbol{\theta}}_n)$ or $I_n(\widehat{\boldsymbol{\theta}}_n)$ is used for Wald statistic (W_n).

The next theorem says that all the test statistics have the same asymptotic distribution under the null hypothesis.

Theorem 6. *Suppose that the null hypothesis holds, assumptions* [R0]-[R6] *are satisfied,* $\widehat{I}_n \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_0)$ *and* $\widehat{\boldsymbol{\theta}}_n$ *is a consistent solution of the likelihood equations. Then each of the test statistics* LR_n, W_n *and* R_n *converges in distribution to* χ^2*-distribution with* p *degrees of freedom.*

Proof. $\underline{R_n}$: Note that R_n can be rewritten as

$$R_n = \Big(\big[\widehat{I}_n\big]^{-\frac{1}{2}} \tfrac{1}{\sqrt{n}}\, \mathbf{U}_n(\boldsymbol{\theta}_0)\Big)^{\mathsf{T}} \Big(\big[\widehat{I}_n\big]^{-\frac{1}{2}} \tfrac{1}{\sqrt{n}}\, \mathbf{U}_n(\boldsymbol{\theta}_0)\Big).$$

Now by the asymptotic normality of the score statistic (17), consistency of $\widehat{I}_n$ and CS (Theorem 2) one gets that

$$\big[\widehat{I}_n\big]^{-\frac{1}{2}} \tfrac{1}{\sqrt{n}}\, \mathbf{U}_n(\boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{I}_p),$$

where $\mathbb{I}_p$ is an identity matrix of dimension $p \times p$. Now the statement follows by using CMT (Theorem 1) with $g(x_1, \ldots, x_p) = \sum_{j=1}^p x_j^2$.

$\underline{W_n}$: One can rewrite W_n as

$$W_n = \Big(\big[\widehat{I}_n\big]^{\frac{1}{2}} \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)\Big)^{\mathsf{T}} \Big(\big[\widehat{I}_n\big]^{\frac{1}{2}} \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)\Big).$$

Now the statement follows by analogous reasoning as for R_n, as by Theorem 5 and CS (Theorem 2) one gets

$$\big[\widehat{I}_n\big]^{\frac{1}{2}} \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{I}_p).$$

$\underline{LR_n}$: With the help of the second order Taylor expansion around $\widehat{\boldsymbol{\theta}}_n$ one gets:

$$\ell_n(\boldsymbol{\theta}_0) = \ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) + \underbrace{\mathbf{U}_n^{\mathsf{T}}\big(\widehat{\boldsymbol{\theta}}_n\big)}_{=\mathbf{0}_p^{\mathsf{T}}} \big(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n\big) - \frac{n}{2}\big(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n\big)^{\mathsf{T}} I_n(\boldsymbol{\theta}_n^*)\big(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n\big),$$

where $\boldsymbol{\theta}_n^*$ lies between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$. Applying Corollary 1 yields $I_n(\boldsymbol{\theta}_n^*) \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_0)$. Thus analogously as above one gets

$$LR_n = 2\big(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n(\boldsymbol{\theta}_0)\big) = \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)^{\mathsf{T}} I_n(\boldsymbol{\theta}_n^*) \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow[n\to\infty]{d} \chi_p^2.$$

□

Remark 9. Note that using the asymptotic representation (14) of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ and the derivations done in the proof of Theorem 6 one can show that the difference of each of the two test statistics (LR_n, W_n and R_n) converges under the null hypothesis to zero in probability.

Nevertheless, in simulations it is observed that the actual level (the probability of type one error) of the test for the Wald test (W_n) can be substantially different from the prescribed

level α. Unfortunately, usually the test is anti-conservative, i.e. the actual level is higher than the prescribed level α. This happens in particular for small samples and/or when the curvature of the log-likelihood $\ell_n(\boldsymbol{\theta})$ is relatively high (as measured for instance by $I(\boldsymbol{\theta})$). The latter happens often if $\boldsymbol{\theta}_0$ is close to the border of the parameter space Θ. That is why some authors recommend either the score test R_n or likelihood ratio test LR_n whose actual levels are usually very close to the prescribed level α even in small samples.

Example 22. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of K-variate random vectors from the multinomial distribution $\mathsf{Mult}_K(1, \mathbf{p})$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})^\mathsf{T}$ and $\mathbf{p} = (p_1, \ldots, p_K)^\mathsf{T}$. Suppose we are interested in testing the null hypothesis

$$H_0 : \mathbf{p}_X = \mathbf{p}^0, \qquad H_1 : \mathbf{p}_X \neq \mathbf{p}^0,$$

where $\mathbf{p}^0 = (p_1^0, \ldots, p_K^0)^\mathsf{T}$ is a given value of the parameter $\mathbf{p}$. For $k \in \{1, \ldots, K\}$ put $n_k = \sum_{i=1}^n X_{ik}$. Derive that

$$LR_n = 2 \sum_{k=1}^K n_k \log\left(\tfrac{n_k}{np_k^0}\right).$$

Further if one uses $I(\widehat{\boldsymbol{\theta}}_n)$ in the Wald test and $I(\boldsymbol{\theta}_0)$ in the Rao score test, then

$$W_n = \sum_{k=1}^K \frac{(n_k - np_k^0)^2}{n_k}, \qquad R_n = \sum_{k=1}^K \frac{(n_k - np_k^0)^2}{np_k^0}.$$

Show that each of the test statistics converges to χ^2-distribution with $K-1$ degrees of freedom.

Note that Rao score test (R_n) corresponds to the standard χ^2-test of goodness-of-fit in multinomial distribution.

Hint. One has to be careful as it is not possible to take $\boldsymbol{\theta} = (p_1, \ldots, p_K)^\mathsf{T}$, as $p_K = 1 - \sum_{k=1}^{K-1} p_k$ (which violates assumption **[R3]**, as the corresponding parameter space would not have any interior points). To avoid this problem one has to take for instance $\boldsymbol{\theta} = (p_1, \ldots, p_{K-1})^\mathsf{T}$.

2.5 Asymptotic confidence sets

Sometimes we are interested in the confidence set for the whole vector parameter $\boldsymbol{\theta}_X$. Then we usually use the following confidence set

$$\left\{\boldsymbol{\theta} \in \Theta : n\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\mathsf{T} \widehat{I}_n (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq \chi^2_p(1-\alpha)\right\},$$

where $\widehat{I}_n$ is a consistent estimator of $I(\boldsymbol{\theta}_X)$. Usually $I_n\big(\widehat{\boldsymbol{\theta}}_n\big)$ or $I\big(\widehat{\boldsymbol{\theta}}_n\big)$ are used as $\widehat{I}_n$. Then the resulting confidence set is an ellipsoid.

Confidence intervals for θ_{Xj}

In most of the applications we are interested in confidence intervals for components θ_{Xj} of the parameter $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$.

Put $\widehat{\boldsymbol{\theta}}_n = \big(\widehat{\theta}_{n1}, \ldots, \widehat{\theta}_{np}\big)^\mathsf{T}$ and $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$. By Theorem 5 we know that

$$\sqrt{n}\,\big(\widehat{\theta}_{nj} - \theta_{Xj}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, i^{jj}(\boldsymbol{\theta}_X)\big), \qquad j \in \{1, \ldots, p\},$$

where $i^{jj}(\boldsymbol{\theta}_X)$ is the j-th diagonal element of $I^{-1}(\boldsymbol{\theta}_X)$. Thus the asymptotic variance of $\widehat{\theta}_{jn}$ is given by $\mathsf{avar}\,\big(\widehat{\theta}_{nj}\big) = \frac{i^{jj}(\boldsymbol{\theta}_X)}{n}$, which can be estimated by $\widehat{\mathsf{avar}\,\big(\widehat{\theta}_{nj}\big)} = \frac{i_n^{jj}}{n}$, where i_n^{jj} is the j-th diagonal element of $\widehat{I}_n^{-1}$. Thus the two-sided (asymptotic) confidence interval for θ_{Xj} is given by

$$\left(\widehat{\theta}_{nj} - u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{i_n^{jj}}{n}},\ \widehat{\theta}_{nj} + u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{i_n^{jj}}{n}}\right). \tag{19}$$

Remark 10. The approaches presented in this section are based on the Wald test statistic. The approaches based on the other test statistics are also possible. For instance one can construct the confidence set for $\boldsymbol{\theta}_X$ as

$$\big\{\boldsymbol{\theta} \in \Theta :\ 2\,\big(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n(\boldsymbol{\theta})\big) \le \chi^2_p(1-\alpha)\big\}.$$

But such a confidence set is for $p > 1$ very difficult to calculate. Nevertheless, as we will see later there exists an approach to calculate the confidence interval for θ_{Xj} with the help of the profile likelihood.

2.6 Asymptotic tests with nuisance parameters

Denote $\boldsymbol{\tau}$ the first q $(1 \le q < p)$ components of the vector $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ the remaining $p - q$ components, i.e.

$$\boldsymbol{\theta} = (\boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T})^\mathsf{T} = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_p)^\mathsf{T}.$$

We want to test the null hypothesis that

$$H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0, \qquad H_1 : \boldsymbol{\tau}_X \neq \boldsymbol{\tau}_0$$

and the remaining parameters $\boldsymbol{\psi}$ are considered as nuisance*. In regression problems this corresponds to situation when one wants to test that a given regressor (interaction) has an effect on the response. Then one is testing that all the parameters corresponding to this regressor (interaction) are zero.

In what follows all the vectors and matrices appearing in the notation of maximum likelihood estimation theory are decomposed into the first q (part 1) and the remaining $p - q$

* *rušivé*

components (part 2), i.e.

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{\boldsymbol{\tau}}_n \\ \widehat{\boldsymbol{\psi}}_n \end{pmatrix}, \quad \mathbf{U}_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{U}_{1n}(\boldsymbol{\theta}) \\ \mathbf{U}_{2n}(\boldsymbol{\theta}) \end{pmatrix},$$

and

$$I(\boldsymbol{\theta}) = \begin{pmatrix} I_{11}(\boldsymbol{\theta}) & I_{12}(\boldsymbol{\theta}) \\ I_{21}(\boldsymbol{\theta}) & I_{22}(\boldsymbol{\theta}) \end{pmatrix}, \quad I_n(\boldsymbol{\theta}) = \begin{pmatrix} I_{11n}(\boldsymbol{\theta}) & I_{12n}(\boldsymbol{\theta}) \\ I_{21n}(\boldsymbol{\theta}) & I_{22n}(\boldsymbol{\theta}) \end{pmatrix}. \tag{20}$$

Lemma 2. *Let $\mathbb{J}$ be a symmetric non-singular matrix of order $p \times p$ that can be written in the block form as*

$$\mathbb{J} = \begin{pmatrix} \mathbb{J}_{11} & \mathbb{J}_{12} \\ \mathbb{J}_{21} & \mathbb{J}_{22} \end{pmatrix}.$$

Denote

$$\mathbb{J}_{11\cdot 2} = \mathbb{J}_{11} - \mathbb{J}_{12}\mathbb{J}_{22}^{-1}\mathbb{J}_{21}, \qquad \mathbb{J}_{22\cdot 1} = \mathbb{J}_{22} - \mathbb{J}_{21}\mathbb{J}_{11}^{-1}\mathbb{J}_{12}.$$

Then

$$\mathbb{J}^{-1} = \begin{pmatrix} \mathbb{J}^{11} & \mathbb{J}^{12} \\ \mathbb{J}^{21} & \mathbb{J}^{22} \end{pmatrix},$$

where

$$\mathbb{J}^{11} = \mathbb{J}_{11\cdot 2}^{-1}, \quad \mathbb{J}^{22} = \mathbb{J}_{22\cdot 1}^{-1}, \quad \mathbb{J}^{12} = -\mathbb{J}_{11\cdot 2}^{-1}\,\mathbb{J}_{12}\,\mathbb{J}_{22}^{-1}, \quad \mathbb{J}^{21} = -\mathbb{J}_{22\cdot 1}^{-1}\,\mathbb{J}_{21}\,\mathbb{J}_{11}^{-1}.$$

Proof. Calculate $\mathbb{J}^{-1}\,\mathbb{J}$. □

Suppose that the parametric space can be written as $\Theta = \Theta_{\boldsymbol{\tau}} \times \Theta_{\boldsymbol{\psi}}$, where $\Theta_{\boldsymbol{\tau}} \subset \mathbb{R}^q$ and $\Theta_{\boldsymbol{\psi}} \subset \mathbb{R}^{p-q}$.

Denote $\widetilde{\boldsymbol{\theta}}_n$ the estimator of $\boldsymbol{\theta}$ under the null hypothesis, i.e.

$$\widetilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \boldsymbol{\tau}_0 \\ \widetilde{\boldsymbol{\psi}}_n \end{pmatrix}, \quad \text{where} \quad \widetilde{\boldsymbol{\psi}}_n \text{ solves } \mathbf{U}_{2n}(\boldsymbol{\tau}_0, \widetilde{\boldsymbol{\psi}}_n) \stackrel{!}{=} \mathbf{0}_{p-q}.$$

The end of class 6 (8. 3. 2024)

Let $\widehat{I}_n^{11}$ be an estimate of the corresponding block $I^{11}(\boldsymbol{\theta}_X)$ in the inverse of Fisher information matrix $I^{-1}(\boldsymbol{\theta}_X)$. The three asymptotic tests of the null hypothesis $H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$ are as follows.

Likelihood ratio test is based on the test statistic

$$LR_n^* = 2\left(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big)\right). \tag{21}$$

Wald test is based on the test statistic

$$W_n^* = n\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^{\mathsf{T}}\left[\widehat{I}_n^{11}\right]^{-1}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right).$$

Rao score test is based on the test statistic

$$R_n^* = \frac{1}{n}\,\mathbf{U}_{1n}^{\mathsf{T}}\big(\widetilde{\boldsymbol{\theta}}_n\big)\,\widehat{I}_n^{11}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big). \tag{22}$$

Remark 11. As $\mathbf{U}_{2n}(\widetilde{\boldsymbol{\theta}}_n) = \mathbf{0}_{p-q}$, the test statistic of the Rao score test can be also written in a form

$$R_n^* = \frac{1}{n}\,\mathbf{U}_n^{\mathsf{T}}(\widetilde{\boldsymbol{\theta}}_n)\,\widehat{I}_n^{-1}\,\mathbf{U}_n(\widetilde{\boldsymbol{\theta}}_n),$$

which is a straightforward analogy of the test statistic (18) of the Rao score test in case of no nuisance parameters.

Similarly as in the previous section the advantage of the likelihood ratio test (LR_n^*) is that one does not need to estimate $I^{-1}(\boldsymbol{\theta}_X)$. On the other hand the advantage of Rao score test (R_n^*) is that it is sufficient to calculate the maximal likelihood estimator only under the null hypothesis.

The next theorem is an analogy to Theorem 6. It says that all the test statistics have the same asymptotic distribution under the null hypothesis.

Theorem 7. *Suppose that the null hypothesis holds, assumptions* **[R0]**-**[R6]** *are satisfied and* $\widehat{I}_n^{11} \xrightarrow[n\to\infty]{P} I^{11}(\boldsymbol{\theta}_X)$. *Further assume that both* $\widehat{\boldsymbol{\theta}}_n$ *and* $\widetilde{\boldsymbol{\theta}}_n$ *are consistent estimators of* $\boldsymbol{\theta}_X$. *Then each of the test statistics* LR_n^*, W_n^* *and* R_n^* *converges in distribution to* χ^2*-distribution with* q *degrees of freedom.*

Proof. First note if the null hypothesis holds then $\boldsymbol{\theta}_X = \left(\boldsymbol{\tau}_0^{\mathsf{T}}, \boldsymbol{\psi}_X^{\mathsf{T}}\right)^{\mathsf{T}}$, where $\boldsymbol{\psi}_X$ stands for the true value of $\boldsymbol{\psi}$.

$\underline{W_n^*}$: Note that by Theorem 5 $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, I^{-1}(\boldsymbol{\theta}_X)\big)$, which yields

$$\sqrt{n}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_q\big(\mathbf{0}_q, I^{11}(\boldsymbol{\theta}_X)\big).$$

Thus analogously as in the proof of Theorem 6 one can show that

$$\left[\widehat{I}_n^{11}\right]^{-\frac{1}{2}} \sqrt{n}(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0) \xrightarrow[n\to\infty]{d} \mathsf{N}_q(\mathbf{0}_q, \mathbb{I}_q),$$

which further with the CMT (Theorem 1) implies

$$W_n^* = \left\{\left[\widehat{I}_n^{11}\right]^{-\frac{1}{2}} \sqrt{n}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)\right\}^{\mathsf{T}} \left\{\left[\widehat{I}_n^{11}\right]^{-\frac{1}{2}} \sqrt{n}(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)\right\} \xrightarrow[n\to\infty]{d} \chi_q^2.$$

$\underline{R_n^*}$: By the mean value theorem (applied to each component of $\mathbf{U}_{1n}(\boldsymbol{\theta})$) one gets

$$\frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\widetilde{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) - I_{12n}^*\,\sqrt{n}\left(\widetilde{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_X\right), \tag{23}$$

where I_{12n}^* is the $(1,2)$-block of the observed Fisher matrix whose j-th row ($j \in \{1,\ldots,q\}$) is evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\widetilde{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_X$. As $\boldsymbol{\theta}_n^{j*} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$, Corollary 1 implies that

$$I_{12n}^* \xrightarrow[n\to\infty]{P} I_{12}(\boldsymbol{\theta}_X). \tag{24}$$

Further note that $\widetilde{\boldsymbol{\psi}}_n$ is a maximum likelihood estimator in the model

$$\mathcal{F}_0 = \big\{ f(\mathbf{x}; \boldsymbol{\tau}_0, \boldsymbol{\psi}); \boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}} \big\}.$$

As the null hypothesis holds, using Theorem 5 one gets

$$\sqrt{n}\,\big(\widetilde{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_X\big) = I_{22}^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) + o_P(1). \tag{25}$$

Combining (23), (24) and (25) yields

$$\frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) = \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) - I_{12}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) + o_P(1). \tag{26}$$

Now using (26) and the central limit theorem (for i.i.d. vectors), which implies that (written in a block form)

$$\frac{1}{\sqrt{n}}\,\mathbf{U}_n(\boldsymbol{\theta}_X) = \begin{pmatrix} \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) \\ \frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) \end{pmatrix} \xrightarrow[n\to\infty]{d} \mathsf{N}_p\left(\mathbf{0}_p, \begin{pmatrix} I_{11}(\boldsymbol{\theta}_X) & I_{12}(\boldsymbol{\theta}_X) \\ I_{21}(\boldsymbol{\theta}_X) & I_{22}(\boldsymbol{\theta}_X) \end{pmatrix}\right),$$

one gets

$$\begin{aligned} \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) &= \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) - I_{12}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X)\frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) + o_P(1) \\ &= \big(\mathbb{I}_q, -I_{12}(\boldsymbol{\theta}_X)\, I_{22}^{-1}(\boldsymbol{\theta}_X)\big) \begin{pmatrix} \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) \\ \frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) \end{pmatrix} + o_P(1) \xrightarrow[n\to\infty]{d} \mathsf{N}_q\big(0, K(\boldsymbol{\theta}_X)\big), \end{aligned}$$

where

$$\begin{aligned} K(\boldsymbol{\theta}_X) &= \big(\mathbb{I}_q, -I_{12}(\boldsymbol{\theta}_X)\, I_{22}^{-1}(\boldsymbol{\theta}_X)\big) \begin{pmatrix} I_{11}(\boldsymbol{\theta}_X) & I_{12}(\boldsymbol{\theta}_X) \\ I_{21}(\boldsymbol{\theta}_X) & I_{22}(\boldsymbol{\theta}_X) \end{pmatrix} \begin{pmatrix} \mathbb{I}_q \\ -I_{22}^{-1}(\boldsymbol{\theta}_X)\, I_{21}(\boldsymbol{\theta}_X) \end{pmatrix} \\ &= I_{11}(\boldsymbol{\theta}_X) - 2I_{12}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X) I_{21}(\boldsymbol{\theta}_X) + I_{12}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X) I_{22}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X) I_{21}(\boldsymbol{\theta}_X) \\ &= I_{11}(\boldsymbol{\theta}_X) - I_{12}(\boldsymbol{\theta}_X)\, I_{22}^{-1}(\boldsymbol{\theta}_X)\, I_{21}(\boldsymbol{\theta}_X) = I_{11\cdot 2}(\boldsymbol{\theta}_X) \overset{\text{Lemma 2}}{=} \big[I^{11}(\boldsymbol{\theta}_X)\big]^{-1}. \end{aligned}$$

Thus $\frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_q\Big(\mathbf{0}_q, \big[I^{11}(\boldsymbol{\theta}_X)\big]^{-1}\Big)$, which further with the help of CS (Theorem 2) and CMT (Theorem 1) implies the statement of the theorem for R_n^*.

$\underline{LR_n^*}$: By the second-order Taylor expansion around the point $\widehat{\boldsymbol{\theta}}_n$ one gets

$$\ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big) = \ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) + \underbrace{\mathbf{U}_n^{\mathsf{T}}\big(\widehat{\boldsymbol{\theta}}_n\big)}_{=\mathbf{0}_p^{\mathsf{T}}}\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big) - \frac{n}{2}\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big)^{\mathsf{T}} I_n(\boldsymbol{\theta}_n^*)\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big), \tag{27}$$

where $\boldsymbol{\theta}_n^*$ is between $\widetilde{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\theta}}_n$. Thus $\boldsymbol{\theta}_n^* \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ and Corollary 1 implies $I_n(\boldsymbol{\theta}_n^*) \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_X)$.

Further by Theorem 5

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = I^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\,\mathbf{U}_n(\boldsymbol{\theta}_X) + o_P(1),$$

which together with (25) implies

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) + \sqrt{n}\left(\boldsymbol{\theta}_X - \widetilde{\boldsymbol{\theta}}_n\right) \\
&= I^{-1}(\boldsymbol{\theta}_X)\frac{1}{\sqrt{n}}\,\mathbf{U}_n(\boldsymbol{\theta}_X) - \begin{pmatrix} \mathbf{0}_q \\ I_{22}^{-1}(\boldsymbol{\theta}_X)\frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) \end{pmatrix} + o_P(1) \\
&= \mathbb{A}(\boldsymbol{\theta}_X)\frac{1}{\sqrt{n}}\,\mathbf{U}_n(\boldsymbol{\theta}_X) + o_P(1),
\end{aligned}$$

where

$$\mathbb{A}(\boldsymbol{\theta}_X) = I^{-1}(\boldsymbol{\theta}_X) - \begin{pmatrix} \mathbf{0}_{q\times q} & \mathbf{0}_{q\times(p-q)} \\ \mathbf{0}_{(p-q)\times q} & I_{22}^{-1}(\boldsymbol{\theta}_X) \end{pmatrix}.$$

By the central limit theorem (for i.i.d. vectors) and the symmetry of matrix $\mathbb{A}(\boldsymbol{\theta}_X)$

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \mathbb{A}(\boldsymbol{\theta}_X)\,I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X)\big). \tag{28}$$

Now we will use Lemma A6 about the distribution of a quadratic form from Appendix.

Put

$$\mathbb{B} = I(\boldsymbol{\theta}_X) \qquad \text{and} \qquad \mathbb{V} = \mathbb{A}(\boldsymbol{\theta}_X)\,I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X).$$

Now $\mathbb{B}\mathbb{V} = I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X)\,I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X)$, where

$$\begin{aligned}
I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X) &= \begin{pmatrix} I_{11}(\boldsymbol{\theta}_X) & I_{12}(\boldsymbol{\theta}_X) \\ I_{21}(\boldsymbol{\theta}_X) & I_{22}(\boldsymbol{\theta}_X) \end{pmatrix}\left(I^{-1}(\boldsymbol{\theta}_X) - \begin{pmatrix} \mathbb{0}_{q\times q} & \mathbb{0}_{q\times(p-q)} \\ \mathbb{0}_{(p-q)\times q} & I_{22}^{-1}(\boldsymbol{\theta}_X) \end{pmatrix}\right) \\
&= \mathbb{I}_p - \underbrace{\begin{pmatrix} \mathbb{0}_{q\times q} & I_{12}(\boldsymbol{\theta}_X)I_{22}^{-1}(\boldsymbol{\theta}_X) \\ \mathbb{0}_{(p-q)\times q} & \mathbb{I}_{p-q} \end{pmatrix}}_{=:\mathbb{D}}.
\end{aligned}$$

Note that matrix $\mathbb{D}$ is idempotent, thus also $\mathbb{I}_p - \mathbb{D}$ and $\mathbb{B}\mathbb{V} = (\mathbb{I}_p - \mathbb{D})(\mathbb{I}_p - \mathbb{D})$ are idempotent. Now using (27), (28), CS (Theorem 2), Lemma A6 and CMT (Theorem 1) one gets

$$LR_n^* = 2\left(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big)\right) = \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\right)^{\mathsf{T}} I(\boldsymbol{\theta}_X)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\right) + o_P(1) \xrightarrow[n\to\infty]{d} \chi^2_{\mathrm{tr}(\mathbb{B}\mathbb{V})},$$

where $\mathrm{tr}(\mathbb{B}\mathbb{V}) = \mathrm{tr}(\mathbb{I}_p) - \mathrm{tr}(\mathbb{D}) = p - (p-q) = q$. □

Suppose that both $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta})$ and $\widetilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta_0} \ell_n(\boldsymbol{\theta})$ (where Θ_0 stands for the parameter space under the null hypothesis) are consistent estimator under the null hypothesis. Then the likelihood ratio test can be rewritten as

$$LR_n^* = 2\left(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big)\right) = 2\left(\sup_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Theta_0} \ell_n(\boldsymbol{\theta})\right). \tag{29}$$

So with the likelihood ratio test one does not need to bother with the parametrization of the parametric spaces Θ and Θ_0 so that it fits into the framework of testing $H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$. The degrees of freedom of the asymptotic distribution are determined as the difference of the dimensions of the parametric spaces Θ and Θ_0.

Example 23. The following data gives the number of male children among the first 12 children of family size 13 in 6115 families taken from hospital records in the 19th century Saxony. The 13th child is ignored to assuage the effect of families non-randomly stopping when a desired gender is reached. Test the null hypothesis that the gender of the babies can

Nr. of boys	0	1	2	3	4	5	6	7	8	9	10	11	12
Nr. of families	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

be viewed as realisations of independent random variables having the same probability of a baby boy for each family.

Hint. Let X_i stand for the number of boys in the i-th family ($i \in \{1, \dots, n\}$, where n stands for the sample size). Then the counts in the table can be represented by

$$n_k = \sum_{i=1}^{n} \mathbb{I}\{X_i = k\}, \quad k \in \{0, 1, \dots, 12\}$$

and the table can be viewed as a realisation of a random vector $(n_0, n_1, \dots, n_{12})^\mathsf{T}$ that follows multinomial distribution $\mathsf{Mult}_{13}(n, \boldsymbol{\pi})$.

Note that under the null hypothesis X_i follows the binomial distribution, thus

$$\pi_k = \mathsf{P}(X_i = k) = \binom{12}{k} p^k (1-p)^{12-k}, \quad k \in \{0, 1, \dots, 12\},$$

where $p \in (0, 1)$ is the probability of baby boy.

Thus to parametrize the problem (so that it fits into the framework of this section) put $\psi = p$ and get

$$\pi_0 = (1-\psi)^{12}, \quad \pi_k = \binom{12}{k} \psi^k (1-\psi)^{12-k} + \tau_k, \quad k \in \{1, \dots, 11\},$$

and $\pi_{12} = 1 - \sum_{k=0}^{11} \pi_k$. The hypotheses can now be written as

$$H_0 : (\tau_1, \dots, \tau_{11})^\mathsf{T} = \mathbf{0}_{11}, \qquad H_1 : (\tau_1, \dots, \tau_{11})^\mathsf{T} \neq \mathbf{0}_{11}.$$

Nevertheless it would be rather tedious to derive either the Wald statistic (W_n^*) or Rao score statistic (R_n^*) as one needs to calculate the score statistic and (empirical) Fisher information matrix.

On the other hand using (29) it is straightforward to calculate the likelihood ratio test LR_n^* as

$$\sup_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) = \sum_{k=0}^{12} n_k \log\left(\tfrac{n_k}{n}\right)$$

and

$$\sup_{\boldsymbol{\theta}\in\Theta_0} \ell_n(\boldsymbol{\theta}) = \sum_{k=0}^{12} n_k \log \widetilde{\pi}_k, \quad \text{where} \quad \widetilde{\pi}_k = \binom{12}{k} (\widetilde{\psi}_n)^k (1-\widetilde{\psi}_n)^{12-k}, \quad \text{with} \quad \widetilde{\psi}_n = \sum_{k=1}^{12} \tfrac{k\, n_k}{12\, n}.$$

By Theorem 7 under the null hypothesis the test statistic LR_n^* converges in distribution to χ^2-distribution with 11 degrees of freedom.

Another approach to test the hypothesis of interest would be (to forget about the test statistics LR_n^*, W_n^*, R_n^* and) to use the standard χ^2-test of goodness-of-fit in multinomial distribution with estimated parameters. The test statistics would be

$$X^2 = \sum_{k=0}^{12} \frac{(n_k - n\,\widetilde{\pi}_k)^2}{n\,\widetilde{\pi}_k} \tag{30}$$

and under the null hypothesis it has also asymptotically χ^2-distribution with 11 degrees of freedom. In fact it can be proved* that the test statistic X^2 given by (30) corresponds to the test statistic of the Rao score test (R_n^*) with $I^{11}(\widetilde{\boldsymbol{\theta}}_n)$ taken as $\widehat{I}_n^{11}$.

Example 24. Breusch-Pagan test of heteroscedasticity.

Example 25. Suppose that you observe independent identically distributed random vectors $\binom{Y_1}{\boldsymbol{X}_1}, \ldots, \binom{Y_n}{\boldsymbol{X}_n}$ such that

$$\mathsf{P}\left(Y_1 = 1 \,|\, \boldsymbol{X}_1\right) = \frac{\exp\{\alpha + \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\alpha + \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}\}}, \qquad \mathsf{P}\left(Y_1 = 0 \,|\, \boldsymbol{X}_1\right) = \frac{1}{1+\exp\{\alpha + \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}\}},$$

where the distribution of $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1p})^\mathsf{T}$ does not depend on the unknown parameters α a $\boldsymbol{\beta}$.

(i) Derive a test for the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ against the alternative that $H_1 : \boldsymbol{\beta} \neq \mathbf{0}_p$.

(ii) Find the confidence set for the parameter $\boldsymbol{\beta}$.

Literature: Anděl [2007] Chapter 8.6, Kulich [2014], Zvára [2008] pp. 122–128.

The end of class 7 (12. 3. 2024)

* More precisely, it is said so in the textbooks but I have not managed to find the derivation.

2.7 Profile likelihood*

Let $\boldsymbol{\theta}$ be divided into $\boldsymbol{\tau}$ containing the first q components ($1 \leq q < p$) and $\boldsymbol{\psi}$ containing the remaining $p - q$ components, i.e.

$$\boldsymbol{\theta} = (\boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T})^\mathsf{T} = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_p)^\mathsf{T}.$$

Write the likelihood of the parameter $\boldsymbol{\theta}$ as $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\tau}, \boldsymbol{\psi})$ and analogously for log-likelihood, score function, Fisher information matrix, ...

In this subsection we will assume that there exists $\widehat{\boldsymbol{\theta}}_n$ which is a unique maximum of the funciton $\ell_n(\boldsymbol{\theta})$ and also a consistent estimator of $\boldsymbol{\theta}_X$. Similarly for $\boldsymbol{\tau}_X = \boldsymbol{\tau}$ let $\widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})$ be a unique maximum of $\ell_n(\boldsymbol{\tau}, \boldsymbol{\psi})$ and a consistent estimator of $\boldsymbol{\psi}_X$.

The profile likelihood and the profile log-likelihood for the parameter $\boldsymbol{\tau}$ are defined subsequently as

$$L_n^{(p)}(\boldsymbol{\tau}) = \max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} L_n(\boldsymbol{\tau}, \boldsymbol{\psi}), \qquad \ell_n^{(p)}(\boldsymbol{\tau}) = \log L_n^{(p)}(\boldsymbol{\tau}) = \max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} \ell_n(\boldsymbol{\tau}, \boldsymbol{\psi}).$$

In the following we will show that one can work with the profile likelihood as with the 'standard' likelihood.

First of all put

$$\widehat{\boldsymbol{\tau}}_n^{(p)} = \operatorname*{arg\,max}_{\boldsymbol{\tau} \in \Theta_{\boldsymbol{\tau}}} \ell_n^{(p)}(\boldsymbol{\tau}).$$

Note that

$$\ell_n^{(p)}\big(\widehat{\boldsymbol{\tau}}_n^{(p)}\big) = \max_{\boldsymbol{\tau} \in \Theta_{\boldsymbol{\tau}}} \ell_n^{(p)}(\boldsymbol{\tau}) = \max_{\boldsymbol{\tau} \in \Theta_{\boldsymbol{\tau}}} \max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} \ell_n(\boldsymbol{\tau}, \boldsymbol{\psi}) = \max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}) = \ell_n(\widehat{\boldsymbol{\theta}}_n).$$

As we assume that $\widehat{\boldsymbol{\theta}}_n$ is a unique maximizer of $\ell_n(\boldsymbol{\theta})$, this implies that

$$\widehat{\boldsymbol{\tau}}_n^{(p)} = \widehat{\boldsymbol{\tau}}_n,$$

where $\widehat{\boldsymbol{\tau}}_n$ stands for the first q-coordinates of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$.

Further denote

$$\widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau}) = \operatorname*{arg\,max}_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} \ell_n(\boldsymbol{\tau}, \boldsymbol{\psi}), \qquad \widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau}) = \big(\boldsymbol{\tau}^\mathsf{T}, \widetilde{\boldsymbol{\psi}}_n^\mathsf{T}(\boldsymbol{\tau})\big)^\mathsf{T},$$

and define the profile score statistic and profile (empirical) information matrix as

$$\mathbf{U}_n^{(p)}(\boldsymbol{\tau}) = \frac{\partial \ell_n^{(p)}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}}, \qquad I_n^{(p)}(\boldsymbol{\tau}) = -\frac{1}{n} \frac{\partial \mathbf{U}_n^{(p)}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}}.$$

The following lemma shows how the quantities $\mathbf{U}_n^{(p)}(\boldsymbol{\tau})$ and $I_n^{(p)}(\boldsymbol{\tau})$ are related with $\mathbf{U}_n(\boldsymbol{\theta})$ and $I_n(\boldsymbol{\theta})$.

---

* *Profilová věrohodnost.*

Lemma 3. *Suppose that assumptions* [R0]-[R6] *are satisfied. Then (with probability tending to one) on a neighbourbood of* $\boldsymbol{\tau}_X$

$$\mathbf{U}_n^{(p)}(\boldsymbol{\tau}) = \mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big), \quad I_n^{(p)}(\boldsymbol{\tau}) = I_{11n}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big) - I_{12n}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big) I_{22n}^{-1}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big) I_{21n}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big),$$

where $I_{jkn}(\boldsymbol{\theta})$ *(for* $j,k \in \{1,2\}$*) were introduced in* (20).

Proof. $\underline{\mathbf{U}_n^{(p)}(\boldsymbol{\tau})}$: Let us calculate

$$\begin{aligned}\big[\mathbf{U}_n^{(p)}(\boldsymbol{\tau})\big]^\mathsf{T} &= \frac{\partial \ell_n^{(p)}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}} = \frac{\partial \ell_n\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big)}{\partial \boldsymbol{\tau}^\mathsf{T}} \\ &= \mathbf{U}_{1n}^\mathsf{T}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) + \mathbf{U}_{2n}^\mathsf{T}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) \frac{\partial \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}} = \mathbf{U}_{1n}^\mathsf{T}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big), \end{aligned} \tag{31}$$

where the last equality follows from the fact that $\widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau}) = \arg\max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}} \ell_n^{(p)}(\boldsymbol{\tau}, \boldsymbol{\psi})$, which implies that $\mathbf{U}_{2n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) = \mathbf{0}_{p-q}$.

$\underline{I_n^{(p)}(\boldsymbol{\tau})}$: Note that with the help of (31)

$$\begin{aligned} I_n^{(p)}(\boldsymbol{\tau}) &= -\frac{1}{n}\frac{\partial \mathbf{U}_n^{(p)}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}} = -\frac{1}{n}\frac{\partial \mathbf{U}_{1n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big)}{\partial \boldsymbol{\tau}^\mathsf{T}} \\ &= I_{11,n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) + I_{12,n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) \frac{\partial \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}}. \end{aligned} \tag{32}$$

Further by differentiating both sides of the identity

$$\mathbf{U}_{2n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) = \mathbf{0}_{p-q}$$

with respect to $\boldsymbol{\tau}^\mathsf{T}$ one gets

$$I_{21,n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) + I_{22,n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big) \frac{\partial \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}} = \mathbb{0}_{(p-q)\times q},$$

which implies that

$$\frac{\partial \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}^\mathsf{T}} = -I_{22,n}^{-1}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big)\, I_{21,n}\big(\boldsymbol{\tau}, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau})\big). \tag{33}$$

Now combining (32) and (33) implies the statement of the theorem for $I_n^{(p)}(\boldsymbol{\tau})$. □

Tests based on profile likelihood

Define the (profile) test statistics of the null hypothesis $H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$ as

$$\begin{aligned} LR_n^{(p)} &= 2\left(\ell_n^{(p)}\big(\widehat{\boldsymbol{\tau}}_n\big) - \ell_n^{(p)}(\boldsymbol{\tau}_0)\right), \\ W_n^{(p)} &= n\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^\mathsf{T} \widehat{I}_n^{(p)} \left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right), \\ R_n^{(p)} &= \frac{1}{n}\left[\mathbf{U}_n^{(p)}(\boldsymbol{\tau}_0)\right]^\mathsf{T} \left[\widehat{I}_n^{(p)}\right]^{-1} \mathbf{U}_n^{(p)}(\boldsymbol{\tau}_0), \end{aligned}$$

where one can use for instance $I_n^{(p)}(\boldsymbol{\tau}_0)$ or $I_n^{(p)}(\widehat{\boldsymbol{\tau}}_n)$ as $\widehat{I}_n^{(p)}$.

Theorem 8. *Suppose that the null hypothesis holds and assumptions* **[R0]**-**[R6]** *are satisfied. Then each of the test statistics* $LR_n^{(p)}$, $W_n^{(p)}$ *and* $R_n^{(p)}$ *converges in distribution to* χ^2*-distribution with* q *degrees of freedom.*

Proof. $\underline{LR_n^{(p)}}$: Note that

$$\ell_n^{(p)}(\widehat{\boldsymbol{\tau}}_n) = \ell_n(\widehat{\boldsymbol{\tau}}_n, \widehat{\boldsymbol{\psi}}_n) = \ell_n\big(\widehat{\boldsymbol{\theta}}_n\big)$$

and further

$$\ell_n^{(p)}(\boldsymbol{\tau}_0) = \max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}} \ell_n(\boldsymbol{\tau}_0, \boldsymbol{\psi}) = \ell_n(\boldsymbol{\tau}_0, \widetilde{\boldsymbol{\psi}}_n) = \ell_n(\widetilde{\boldsymbol{\theta}}_n).$$

Thus $LR_n^{(p)} = LR_n^*$, where LR_n^* is the test statistic of the likelihood ratio test in the presence of nuisance parameters given by (21). Thus the statement of the theorem follows by Theorem 7.

$\underline{W_n^{(p)}}$: Follows from Theorem 7 and the fact that by Lemmas 1, 2 and 3

$$\widehat{I}_n^{(p)} \xrightarrow[n\to\infty]{P} I_{11}(\boldsymbol{\theta}_X) - I_{12}(\boldsymbol{\theta}_X)\, I_{22}^{-1}(\boldsymbol{\theta}_X)\, I_{21}(\boldsymbol{\theta}_X) = \big[I^{11}(\boldsymbol{\theta}_X)\big]^{-1}. \tag{34}$$

$\underline{R_n^{(p)}}$: By Lemma 3 one has $\mathbf{U}_n^{(p)}(\boldsymbol{\tau}) = \mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n(\boldsymbol{\tau})\big)$. Thus $R_n^{(p)} = R_n^*$ with $\widehat{I}_n^{11} = \big[\widehat{I}_n^{(p)}\big]^{-1}$, where R_n^* is Rao score test statistic in the presence of nuisance parameters defined in (22). The statement of the theorem now follows by (34) and Theorem 7. □

Confidence interval for θ_{Xj}

One of the applications of the profile likelihood is to construct a confidence interval for θ_{Xj}. Let $\tau = \theta_j$ and $\boldsymbol{\psi}$ contains the remaining coordinates of the parameter $\boldsymbol{\theta}$. Then the set

$$\Big\{\theta_j : 2\left(\ell_n^{(p)}(\widehat{\theta}_{nj}) - \ell_n^{(p)}(\theta_j)\right) \le \chi_1^2(1-\alpha)\Big\}$$

is the asymptotic confidence interval for θ_{Xj}. Although this confidence interval is more difficult to calculate than the Wald-type confidence interval given by (19), the simulations show that it has better finite sample properties. In R-software these intervals for GLM models are calculated by the function `confint`.

Example 26. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with density

$$f(x) = \frac{1}{\Gamma(\beta)}\, \lambda^\beta\, x^{\beta-1}\, \exp\{-\lambda\, x\}\, \mathbb{I}\{x > 0\}.$$

Suppose we are interested in parameter β and parameter λ is nuisance. Derive the profile likelihood for parameter β and the Rao score test of the null hypothesis $H_0 : \beta_X = \beta_0$ against $H_1 : \beta_X \neq \beta_0$ that is based on the profile likelihood.

Solution: The likelihood and log-likelihood are given by

$$L_n(\beta,\lambda) = \prod_{i=1}^{n} \frac{1}{\Gamma(\beta)} \lambda^{\beta} X_i^{\beta-1} \mathrm{e}^{-\lambda X_i},$$

$$\ell_n(\beta,\lambda) = -n\log\Gamma(\beta) + n\beta\log\lambda + (\beta-1)\sum_{i=1}^{n}\log X_i - \lambda\sum_{i=1}^{n} X_i.$$

For a given β we can find $\widetilde{\lambda}_n(\beta)$ by

$$\begin{aligned} \frac{\partial \ell_n(\beta,\lambda)}{\partial\lambda} = \frac{n\beta}{\lambda} - \sum_{i=1}^{n} X_i &\stackrel{!}{=} 0 \\ \widetilde{\lambda}_n(\beta) &= \frac{\beta}{\overline{X}_n}. \end{aligned}$$

Thus the profile log-likelihood is

$$\ell_n^{(p)}(\beta) = -n\log\Gamma(\beta) + n\beta\log\left(\tfrac{\beta}{\overline{X}_n}\right) + (\beta-1)\sum_{i=1}^{n}\log X_i - n\beta$$

and its corresponding score function

$$U_n^{(p)}(\beta) = -\frac{n\,\Gamma'(\beta)}{\Gamma(\beta)} + n\log\left(\tfrac{\beta}{\overline{X}_n}\right) + n + \sum_{i=1}^{n}\log X_i - n.$$

Statistic of Rao score test of the null hypothesis $H_0 : \beta_X = \beta_0$ against $H_1 : \beta_X \neq \beta_0$ is now given by

$$R_n^{(p)} = \frac{\left[U_n^{(p)}(\beta_0)\right]^2}{n\, I_n^{(p)}(\beta_0)},$$

where

$$I_n^{(p)}(\beta) = -\frac{1}{n}\frac{\partial U_n^{(p)}(\beta)}{\partial\beta} = \left[\frac{\Gamma''(\beta)}{\Gamma(\beta)} - \left(\frac{\Gamma'(\beta)}{\Gamma(\beta)}\right)^2 - \frac{1}{\beta}\right].$$

Example 27. Box-Cox transformation. See Zvára [2008] pp. 149–151.

Remark 12. Although we have shown that one can work with the profile likelihood as with the standard likelihood not all the properties are shared. For instance for standard score statistic one has $\mathsf{E}\ \mathbf{U}_n(\boldsymbol{\theta}_X) = \mathbf{0}_p$. But this is not guaranteed for profile score statistic as by Lemma 3

$$\mathsf{E}\ \mathbf{U}_n^{(p)}(\boldsymbol{\tau}_X) = \mathsf{E}\ \mathbf{U}_{1n}(\boldsymbol{\tau}_X, \widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau}_X))$$

and the expectation on the right-hand side of the previous equation is typically not zero due to the random argument $\widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau}_X)$ (for illustration think of $\mathsf{E}\ U_n^{(p)}(\beta_X)$ in Example 26). From the proof of Theorem 7 we only know that $\frac{1}{\sqrt{n}}\mathbf{U}_n^{(p)}(\boldsymbol{\tau}_X)$ converges in distribution to a zero-mean Gaussian distribution.

Note also that we have avoided defining the profile Fisher information matrix. The thing is that the only definition that makes sense would be $I^{(p)}(\boldsymbol{\tau}_X) = \left[I^{11}(\boldsymbol{\tau}_X, \boldsymbol{\psi}_X)\right]^{-1}$. But this is not nice as it depends on the nuisance parameter $\boldsymbol{\psi}_X$. Further, it does not hold that $I^{(p)}(\boldsymbol{\tau}_X)$ is the expectation of $I_n^{(p)}(\boldsymbol{\tau}_X)$. It only holds that

$$I_n^{(p)}(\boldsymbol{\tau}_X) \xrightarrow[n\to\infty]{P} I^{(p)}(\boldsymbol{\tau}_X).$$

2.8 Some notes on maximum likelihood in case of not i.i.d. random vectors

Let observations $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ have a joint density $f_n(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\theta})$ that is known up to the unknown parameter $\boldsymbol{\theta}$ from the parametric space Θ. Analogously as in 'i.i.d case' one can define the *likelihood function* as

$$L_n(\boldsymbol{\theta}) = f_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\theta}),$$

the *log-likelihood function* as

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}),$$

and *the score statistic as*

$$\mathbf{U}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

The *maximum likelihood estimator* (of parameter $\boldsymbol{\theta}_X$) is then defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}) \quad \text{or more generally as} \quad \mathbf{U}_n(\widehat{\boldsymbol{\theta}}_n) \overset{!}{=} \mathbf{0}_p.$$

Finally *the observed (empirical) Fisher information matrix* as

$$I_n(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}}.$$

The role of the theoretical Fisher information matrix $I(\boldsymbol{\theta})$ in 'i.i.d' settings is now taken by the *limit 'average' Fisher information matrix*

$$\bar{I}(\boldsymbol{\theta}) = \lim_{n\to\infty} \frac{1}{n}\, \mathsf{E}\left[\frac{-\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}}\right].$$

In 'nice (regular) models' (see also Remark 13 below) it holds that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\left(\mathbf{0}, \bar{I}^{-1}(\boldsymbol{\theta}_X)\right).$$

The most straightforward estimator of $\bar{I}(\boldsymbol{\theta}_X)$ is $I_n\left(\widehat{\boldsymbol{\theta}}_n\right)$ and thus the estimator of the asymptotic variance matrix of $\widehat{\boldsymbol{\theta}}_n$ is

$$\widehat{\mathsf{avar}\left(\widehat{\boldsymbol{\theta}}_n\right)} = \frac{1}{n}\, I_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right) = \left[\frac{-\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}\right]^{-1}.$$

That is why some authors prefer to define the empirical Fisher information without $\frac{1}{n}$ simply as

$$\widetilde{I}_n(\boldsymbol{\theta}) = \frac{-\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}}$$

and they speak about it as the Fisher information of all observations.

Remark 13. An inspection of the proof of Theorem 9 (for Z-estimators) reveals that we need to show the analogy of Lemma 1 with $I(\boldsymbol{\theta}_X)$ replaced with $\bar{I}(\boldsymbol{\theta}_X)$ and that

$$\frac{1}{\sqrt{n}}\, \mathbf{U}_n(\boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \bar{I}(\boldsymbol{\theta}_X)\big).$$

Example 28. Suppose we have K independent samples, that is for each $k \in \{1, \ldots, K\}$ the random variables $\boldsymbol{X}_{ki}, i \in \{1, \ldots, n_k\}$ are independent and identically distributed with density $f_k(\mathbf{x}; \boldsymbol{\theta})$ (with respect to a σ-finite measure μ). Further let all the random variables be independent and let $\lim_{n\to\infty} \frac{n_k}{n} = w_k$, where $n = n_1 + \ldots + n_K$. Then

$$\begin{aligned}
L_n(\boldsymbol{\theta}) &= \prod_{k=1}^{K} \prod_{i=1}^{n_k} f_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}), \\
\ell_n(\boldsymbol{\theta}) &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \log f_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}), \\
\mathbf{U}_n(\boldsymbol{\theta}) &= \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\partial \log f_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \\
I_n(\boldsymbol{\theta}) &= -\frac{1}{n} \frac{\partial \mathbf{U}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = -\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\partial^2 \log f_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}}, \\
\bar{I}(\boldsymbol{\theta}) &= \lim_{n\to\infty} \mathsf{E}\, I_n(\boldsymbol{\theta}) = \lim_{n\to\infty} \sum_{k=1}^{K} \underbrace{\frac{n_k}{n}}_{\to w_k} I^{(k)}(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, I^{(k)}(\boldsymbol{\theta}),
\end{aligned}$$

where $I^{(k)}(\boldsymbol{\theta})$ is Fisher information matrix of $\boldsymbol{X}_{ki}$ (i.e. for the density $f_k(\mathbf{x}; \boldsymbol{\theta})$).

In standard applications $\boldsymbol{\theta} = \big(\boldsymbol{\theta}_1^\mathsf{T}, \ldots, \boldsymbol{\theta}_K^\mathsf{T}\big)^\mathsf{T}$, and the density $f_k(\mathbf{x}; \boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}_k$, i.e. $f_k(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta}_k)$. And we are usually interested in testing the null hypothesis that all the distributions are the same, that is

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \ldots = \boldsymbol{\theta}_K \qquad H_1 : \exists_{k,j \in \{1,\ldots,K\}} \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_j.$$

See also Example 32.

The end of class 8 (15. 3. 2024)

Random vs. fixed design

Sometimes in regression it is useful do distinguish random design and fixed design.

In **random design** we assume that the values of the covariates are realisations of random vectors. Thus (in the most simple situation) we assume that we observe independent and identically distributed random vectors

$$\begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \boldsymbol{X}_n \end{pmatrix}, \tag{35}$$

where the conditional distribution of $Y_i|\boldsymbol{X}_i$ is known up to the unknown parameter $\boldsymbol{\theta}$ and the distribution of $\boldsymbol{X}_i$ does not depend on $\boldsymbol{\theta}$. Put $f(y_i|\mathbf{x}_i;\boldsymbol{\theta})$ for the conditional density of $Y_i|\boldsymbol{X}_i = \mathbf{x}_i$ and $f_{\boldsymbol{X}}(\mathbf{x})$ for the density of $\boldsymbol{X}_i$. Then the likelihood and the log-likelihood (for the parameter $\boldsymbol{\theta}$) are given by

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \prod_{i=1}^n f_{Y,\boldsymbol{X}}(Y_i, \boldsymbol{X}_i;\boldsymbol{\theta}) = \prod_{i=1}^n f(Y_i|\boldsymbol{X}_i;\boldsymbol{\theta}) f_{\boldsymbol{X}}(\boldsymbol{X}_i) \\ \ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^n \log f(Y_i|\boldsymbol{X}_i;\boldsymbol{\theta}) + \sum_{i=1}^n \log f_{\boldsymbol{X}}(\boldsymbol{X}_i). \end{aligned} \tag{36}$$

In **fixed design** it is assumed that the values of the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed when planning the experiment (before measuring the response). Now we observe $Y_1, \ldots, Y_n$ independent (but not identically distributed) random variables with the densities $f(y_1|\mathbf{x}_1;\boldsymbol{\theta})$, $\ldots$, $f(y_n|\mathbf{x}_n;\boldsymbol{\theta})$. Then the log-likelihood is given by

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(Y_i|\mathbf{x}_i;\boldsymbol{\theta}). \tag{37}$$

Comparing the log-likelihoods in (36) and (37) one can see that (once the data are observed) they differ only by $\sum_{i=1}^n \log f_{\boldsymbol{X}}(\boldsymbol{X}_i)$ which does not depend on $\boldsymbol{\theta}$. Thus in terms of (likelihood based) inference for a given dataset both approaches are equivalent. The only difference is that the theory for the fixed design is more difficult.

Example 29. *Poisson regression.*
Random design approach: We assume that we observe independent identically distributed random vectors (35) and that $Y_i|\boldsymbol{X}_i \sim \mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, where $\lambda(\mathbf{x}) = \exp\{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}\}$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$. Then (provided assumptions [R0]-[R6] are satisfied)

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, I^{-1}(\boldsymbol{\beta}_X)\big), \text{ where } I(\boldsymbol{\beta}_X) = \mathsf{E}\left[\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T} \exp\left\{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X\right\}\right].$$

Fixed design approach: We assume that we observe independent random variables $Y_1, \ldots, Y_n$ and we have the known constants $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that $Y_i \sim \mathsf{Po}\big(\lambda(\mathbf{x}_i)\big)$, where $\lambda(\mathbf{x}) =$

$\exp\{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}\}$. Then it can be shown (that under mild assumptions on $\mathbf{x}_1, \ldots, \mathbf{x}_n$)

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \bar{I}^{-1}(\boldsymbol{\beta}_X)\big), \text{ where } \bar{I}(\boldsymbol{\beta}_X) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\,\mathbf{x}_i^\mathsf{T} \exp\left\{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_X\right\}.$$

Note that in practice both $I(\boldsymbol{\beta}_X)$ and $\bar{I}(\boldsymbol{\beta}_X)$ would be estimated by

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\,\boldsymbol{X}_i^\mathsf{T} \exp\left\{\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n\right\} \qquad \text{or} \qquad \widehat{\bar{I}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\,\mathbf{x}_i^\mathsf{T} \exp\left\{\mathbf{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n\right\}.$$

Thus for observed data the estimators coincide. The only difference is in notation in which you distinguish whether you think of the observed values of the covariates as the realizations of the random vectors or as fixed constants.

Example 30. Note that alternatively one can view the K-sample problem described in Example 28 also within i.i.d framework. Consider the data as a realization of the random sample $\binom{\mathbf{Z}_1}{J_1}, \ldots, \binom{\mathbf{Z}_n}{J_n}$, where J_i takes values in $\{1, \ldots, K\}$ and the conditional distribution of $\mathbf{Z}_i$ given $J_i = j$ is given by the density $f_j(\mathbf{x}; \boldsymbol{\theta})$.

Example 31. Maximum likelihood estimation in AR(1) process.

Example 32. Suppose that X_{ki}, $k \in \{1, \ldots, K\}$, $i = 1, \ldots, n_K$ be independent random variables such that X_{ki} follows Bernoulli distribution with parameter p_k. We are interested in testing the hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_K \qquad H_1 : \exists_{k,j\in\{1,\ldots,K\}}\, p_k \neq p_j.$$

Note that one can easily construct a likelihood ratio test.

Alternatively one can view the data as $K \times 2$ contingency table and use the χ^2-test of independence. It can be proved that this test is in fact the Rao-score test for this problem.

Literature: Hoadley [1971].

The end of class 9 (19. 3. 2024)

2.9 Conditional and marginal likelihood*

In some models the number of parameters is increasing as the sample size increases. Formally let $\boldsymbol{\theta}^{(n)} = (\theta_1, \ldots, \theta_{p_n})^\mathsf{T}$, where p_n is a non-decreasing function of n. Let $\boldsymbol{\theta}^{(n)}$ be divided into $\boldsymbol{\tau}$ containing the first q components (with q being fixed) and $\boldsymbol{\psi}^{(n)}$ containing the remaining $(p_n - q)$ components.

Example 33. *Strongly stratified sample.* Let Y_{ij}, $i \in \{1, \ldots, N\}$, $j \in \{1, 2\}$ be independent random variables such that $Y_{ij} \sim \mathsf{N}(\mu_i, \sigma^2)$. Derive the maximum likelihood estimator of σ^2. Is this estimator consistent as $N \to \infty$?

* *Podmíněná a marginální věrohodnost.*

Solution. The joint density of all the observations $\boldsymbol{Y} = \big(Y_{ij}, i \in \{1,\dots,N\},\ j \in \{1,2\}\big)$ is

$$f(\mathbf{y};\sigma^2,\mu_1,\dots,\mu_N) = \prod_{i=1}^{N}\prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big\{-\tfrac{(Y_{ij}-\mu_i)^2}{2\sigma^2}\big\} \tag{38}$$

and thus the log-likelihood is given by

$$\ell_n(\sigma^2,\mu_1,\dots,\mu_N) = -N\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{2}(Y_{ij}-\mu_i)^2 - N\log(2\pi).$$

Differentiating with respect to $\mu_1,\dots,\mu_N$ and σ^2 one easily finds that

$$\widehat{\mu}_i = \frac{Y_{i1}+Y_{i2}}{2}, \quad i \in \{1,\dots,N\}$$

and

$$\begin{aligned}\widehat{\sigma}^2_N &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2}(Y_{ij}-\widehat{\mu}_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}\Big[\big(Y_{i1}-\tfrac{Y_{i1}+Y_{i2}}{2}\big)^2 + \big(Y_{i2}-\tfrac{Y_{i1}+Y_{i2}}{2}\big)^2\Big]\\ &= \frac{1}{2N}\sum_{i=1}^{N}\Big[\big(\tfrac{Y_{i1}-Y_{i2}}{2}\big)^2 + \big(\tfrac{Y_{i2}-Y_{i1}}{2}\big)^2\Big] = \frac{1}{4N}\sum_{i=1}^{N}(Y_{i1}-Y_{i2})^2.\end{aligned}$$

Thus

$$\widehat{\sigma}^2_N \xrightarrow[N\to\infty]{P} \tfrac{1}{4}\,\mathsf{E}\,(Y_{i1}-Y_{i2})^2 = \tfrac{1}{4}\,\mathsf{var}(Y_{i1}-Y_{i2}) = \tfrac{2\sigma^2}{4} = \tfrac{\sigma^2}{2} \neq \sigma^2.$$

Note that in Example 33 each observation carries information on σ^2, but the maximum likelihood estimator of σ^2 is not even consistent. The problem is that the dimension of nuisance parameters $\boldsymbol{\psi}^{(N)} = (\mu_1,\dots,\mu_N)^{\mathsf{T}}$ is increasing to infinity (too quickly). Marginal and conditional likelihoods are two attempts to modify the likelihood so that it yields a consistent (and hopefully also asymptotically normal) estimator of the parameter $\boldsymbol{\tau}$.

Suppose that one can use data $\mathbb{X}$ to calculate $\mathbf{V}$ whose distribution depends only on parameter $\boldsymbol{\tau}$ (and not on $\boldsymbol{\psi}^{(n)}$). Then *the marginal (log-)likelihood* of parameter $\boldsymbol{\tau}$ is defined as

$$L_n^{(M)}(\boldsymbol{\tau}) = f(\mathbf{V};\boldsymbol{\tau}), \qquad \ell_n^{(M)}(\boldsymbol{\tau}) = \log\big(L_n^{(M)}(\boldsymbol{\tau})\big),$$

where $f(\mathbf{v};\boldsymbol{\tau})$ is the joint density of $\mathbf{V}$ with respect to a σ-finite measure μ.

Suppose that one can use data $\mathbb{X}$ to calculate $\mathbf{V}$ and $\mathbf{W}$ such that the conditional distribution of $\mathbf{V}$ given $\mathbf{W}$ depends only on parameter $\boldsymbol{\tau}$ (and not on $\boldsymbol{\psi}^{(n)}$). Then *the conditional (log-)likelihood* of parameter $\boldsymbol{\tau}$ is defined as

$$L_n^{(C)}(\boldsymbol{\tau}) = f(\mathbf{V}\,|\,\mathbf{W};\boldsymbol{\tau}), \qquad \ell_n^{(C)}(\boldsymbol{\tau}) = \log\big(L_n^{(C)}(\boldsymbol{\tau})\big),$$

where $f(\mathbf{v}|\mathbf{w};\boldsymbol{\tau})$ is the conditional density of $\mathbf{V}$ given $\mathbf{W} = \mathbf{w}$ with respect to a σ-finite measure μ.

Remark 14. (i) If $\mathbf{V}$ is independent of $\mathbf{W}$, then $f(\mathbf{V}|\mathbf{W};\boldsymbol{\tau}) = f(\mathbf{V};\boldsymbol{\tau})$ and thus $L_n^{(M)}(\boldsymbol{\tau}) = L_n^{(C)}(\boldsymbol{\tau})$.

(ii) '*Automatic calculation* of $\ell_n^{(C)}(\boldsymbol{\tau})$':

$$\ell_n^{(C)}(\boldsymbol{\tau}) = \log\left(\frac{f(\mathbf{V},\mathbf{W};\boldsymbol{\tau},\boldsymbol{\psi}^{(n)})}{f(\mathbf{W};\boldsymbol{\tau},\boldsymbol{\psi}^{(n)})}\right) = \ell_{n,\mathbf{V},\mathbf{W}}(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)}) - \ell_{n,\mathbf{W}}(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)}),$$

where $\ell_{n,\mathbf{V},\mathbf{W}}(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)})$ is the log-likelihod of $(\mathbf{V},\mathbf{W})$ and $\ell_{n,\mathbf{W}}(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)})$ is the log-likelihod of $\mathbf{W}$. Note that using this approach we do not need to derive the conditional distribution of $\mathbf{V}$ given $\mathbf{W}$.

(iii) It can be shown that (under certain regularity assumptions) one can work with $L_n^{(M)}(\boldsymbol{\tau})$ and $L_n^{(C)}(\boldsymbol{\tau})$ as with 'standard' likelihoods.

The question of interest is how to find $\mathbf{V}$ and $\mathbf{W}$ so that we do not loose too many information about $\boldsymbol{\tau}$. To the best of my knowledge for marginal likelihood there are only ad-hoc approaches.

For conditional likelihood one can use the theory of sufficient statistics. Suppose that for each fixed value of $\boldsymbol{\tau}$ the statistic $\mathbf{S}_n(\mathbb{X})$ is sufficient for $\boldsymbol{\psi}^{(n)}$. Thus the conditional distribution of $\mathbb{X}$ given $\mathbf{S}_n(\mathbb{X})$ does not depend on $\boldsymbol{\psi}^{(n)}$. This implies that when constructing the conditional likelihood $L_n^{(C)}(\boldsymbol{\tau})$ one can take $\mathbf{S}_n(\mathbb{X})$ as $\mathbf{W}$ and $\mathbb{X}$ as $\mathbf{V}$.

Exponential family

Let the dataset $\mathbb{X}$ have the density (with respect to a σ-finite measure μ) of the form

$$f(\mathbf{x};\boldsymbol{\tau},\boldsymbol{\psi}^{(n)}) = \exp\left\{\sum_{j=1}^{q} Q_j(\boldsymbol{\tau})\, T_j(\mathbf{x}) + \sum_{j=1}^{p_n-q} R_j(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)})\, S_j(\mathbf{x})\right\} a\big(\boldsymbol{\tau},\boldsymbol{\psi}^{(n)}\big)\, h(\mathbf{x}), \qquad (39)$$

where $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_q)^{\mathsf{T}}$ and $\boldsymbol{\psi}^{(n)} = \big(\psi_1^{(n)},\ldots,\psi_{p_n-q}^{(n)}\big)^{\mathsf{T}}$. Put $\mathbf{S}_n(\mathbb{X}) = \big(S_1(\mathbb{X}),\ldots,S_{p_n-q}(\mathbb{X})\big)^{\mathsf{T}}$ and note that for a fixed value of $\boldsymbol{\tau}$ the statistic $\mathbf{S}_n(\mathbb{X})$ is sufficient for $\boldsymbol{\psi}^{(n)}$. Thus one can put $\mathbf{W} = \mathbf{S}_n(\mathbb{X})$ and $\mathbf{V} = \mathbb{X}$.

Example 33. *Strongly stratified sample (cont.).* Derive the marginal and conditional likelihood.

Marginal likelihood. For $i \in \{1,\ldots,N\}$ consider $V_i = \frac{Y_{i1}-Y_{i2}}{\sqrt{2}}$. Then $V_i \sim \mathsf{N}(0,\sigma^2)$. Thus the marginal likehood is the likelihood of $V_1,\ldots,V_n$ and is given by

$$L_n^{(M)}(\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big\{-\tfrac{V_i^2}{2\sigma^2}\big\}.$$

Further the marginal log-likehood is given by

$$\ell_n^{(M)}(\sigma^2) = -\tfrac{N}{2}\,\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N} V_i^2 - \tfrac{N}{2}\,\log 2\pi = -\tfrac{N}{2}\,\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{i=1}^{N}(Y_{i1}-Y_{i2})^2 - \tfrac{N}{2}\,\log 2\pi.$$

With this marginal log-likelihood one can work in the 'standard' way. That is one can for instance derive the maximum (marginal) likelihood estimator

$$\widehat{\sigma}_N^{2(M)} = \frac{1}{2N}\sum_{i=1}^{N}(Y_{i1}-Y_{i2})^2.$$

It is straightforward to show that this estimator is consistent and that

$$\sqrt{n}\,\big(\widehat{\sigma}_N^{2(M)} - \sigma^2\big) \xrightarrow[N\to\infty]{d} \mathsf{N}\big(0, 2\sigma^4\big),$$

where the asymptotic variance $2\sigma^2$ can be calculated as $\mathsf{var}\big((Y_{i1}-Y_{i2})^2\big)$ or as one over the Fisher information that corresponds to $\ell_N^{(M)}(\sigma^2)$.

Conditional likelihood. Note that the joint density (38) of $\boldsymbol{Y} = \big(Y_{ij}, i \in \{1,\dots,N\},\ j \in \{1,2\}\big)$ can be written as

$$f(\mathbf{y};\sigma^2,\dots) = \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{2} y_{ij}^2 + \sum_{i=1}^{N}\frac{\mu_i}{\sigma^2}(y_{i1}+y_{i2}) - \sum_{i=1}^{N}\frac{\mu_i^2}{\sigma^2}\right\}\frac{1}{(2\pi\sigma^2)^N}. \tag{40}$$

Now the above density can be written in the form (39) with $\tau = -\frac{1}{\sigma^2}$, $\psi_i = \mu_i$, $T(\mathbf{y}) = \sum_{i=1}^{N}\sum_{j=1}^{2} y_{ij}^2$ and $S_i(\mathbf{y}) = y_{i1} + y_{i2}$. Thus by Remark 14(ii) the conditional log-likelihood of $\boldsymbol{Y}$ given $\mathbf{S}(\boldsymbol{Y}) = \big(S_1(\boldsymbol{Y}),\dots,S_N(\boldsymbol{Y})\big)^{\mathsf{T}}$ can be calculated as

$$\ell_n^{(C)}(\sigma^2) = \ell_n(\sigma^2,\mu_1,\dots,\mu_N) - \ell_{n,\mathbf{S}(\boldsymbol{Y})}(\sigma^2,\mu_1,\dots,\mu_N).$$

Here $\ell_n(\sigma^2,\mu_1,\dots,\mu_N)$ is the (standard) log-likehood of $\boldsymbol{Y}$ which can be with the help of (40) rewritten as

$$\ell_n(\sigma^2,\dots) = -N\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_{i1}^2+Y_{i2}^2) + \sum_{i=1}^{N}\frac{\mu_i}{\sigma^2}(Y_{i1}+Y_{i2}) - \sum_{i=1}^{N}\frac{\mu_i^2}{\sigma^2} - N\log 2\pi \tag{41}$$

and $\ell_{n,\mathbf{S}(\boldsymbol{Y})}(\sigma^2,\mu_1,\dots,\mu_N)$ is the log-likelihood of $\mathbf{S}(\boldsymbol{Y}) = (Y_{11}+Y_{12},\dots,Y_{N1}+Y_{N2})^{\mathsf{T}}$. As the components of $\mathbf{S}(\boldsymbol{Y})$ are independent random variables with the distribution $\mathsf{N}(2\mu_i, 2\sigma^2)$ $(i \in \{1,\dots,N\})$, the log-likelihood $\ell_{n,\mathbf{S}(\boldsymbol{Y})}(\sigma^2,\dots)$ is given by

$$\begin{aligned}
\ell_{n,\mathbf{S}(\boldsymbol{Y})}(\sigma^2,\dots) &= \log\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi 2\sigma^2}}\exp\big\{-\tfrac{(Y_{i1}+Y_{i2}-2\mu_i)^2}{2\cdot 2\sigma^2}\big\}\right)\\
&= -\frac{N}{2}\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{i=1}^{N}(Y_{i1}^2+Y_{i2}^2+2Y_{i1}Y_{i2})\\
&\quad + \sum_{i=1}^{N}\frac{4\mu_i}{4\sigma^2}(Y_{i1}+Y_{i2}) - \sum_{i=1}^{N}\frac{4\mu_i^2}{4\sigma^2} + \tfrac{N}{2}\log(4\pi).
\end{aligned} \tag{42}$$

Thus comparing (41) and (42) one gets

$$\ell_n^{(C)}(\sigma^2) = \ell_n(\sigma^2,\ldots) - \ell_{n,\mathbf{S}(\boldsymbol{Y})}(\sigma^2,\ldots) = -\tfrac{N}{2}\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{i=1}^{N}(Y_{i1}^2 + Y_{i2}^2 - 2Y_{i1}Y_{i2}) - \tfrac{N}{4}\log\pi$$
$$= -\tfrac{N}{2}\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{i=1}^{N}(Y_{i1} - Y_{i2})^2 - \tfrac{N}{4}\log\pi = \ell_N^{(M)}(\sigma^2) + \tfrac{N}{4}\log\pi,$$

where the (irrelevant) difference between the $\ell_n^{(C)}(\sigma^2)$ and $\ell_N^{(M)}(\sigma^2)$ comes from the fact that for the conditional likelihood we use the conditional distribution of $\boldsymbol{Y}$ giwen $\mathbf{W} = \mathbf{S}(\boldsymbol{Y})$ instead of the conditional distribution of $(\mathbf{V}, \mathbf{W})$ given $\mathbf{W} = \mathbf{S}(\boldsymbol{Y})$. Note also that in the latter case one would get directly the marginal likelihood of $\mathbf{V}$, as $\mathbf{V}$ is independent of $\mathbf{W}$.

Example 34. Let Y_{ij}, $i \in \{1,\ldots,N\}$, $j \in \{1,2\}$ be independent random variables such that $Y_{i1} \sim \mathsf{Exp}(\psi_i)$ and $Y_{i2} \sim \mathsf{Exp}(\tau\,\psi_i)$ where $\tau > 0$ and ψ_i are unknown parameters. Show that the distribution of $V_i = \frac{Y_{i2}}{Y_{i1}}$ depends only on parameter τ (and not on ψ_i). Derive the marginal likelihood of τ that is based on $\mathbf{V} = (V_1,\ldots,V_N)^\mathsf{T}$.

The end of class 10 (22. 3. 2024)

Example 35. Let Y_{ij}, $i \in \{1,\ldots,I\}$, $j \in \{0,1\}$ be independent, $Y_{ij} \sim \mathsf{Bi}(n_{ij}, p_{ij})$, where

$$\log\left(\tfrac{p_{ij}}{1-p_{ij}}\right) = \psi_i + \tau\,\mathbb{I}\{j=1\}.$$

Suppose we are interested in testing the null hypothesis $H_0 : \tau = 0$ against the alternative $H_1 : \tau \neq 0$.

Note that the standard tests based on the maximum likelihood as described in Chapter 2.6 require that I is fixed and all the sample sizes n_{ij} tend to infinity. This implies that using conditional likelihood is reasonable in situations when (some) n_{ij} are small.

The Rao score test based on the conditional likelihood in this situation coincides with Cochran-Mantel-Haenszel test and its test statistic is given by

$$R_n^{(C)} = \frac{\left(\sum_{i=1}^{I} Y_{i1} - \mathsf{E}_{H_0}[Y_{i1}\,|\,Y_{i+}]\right)^2}{\sum_{i=1}^{I}\mathsf{var}_{H_0}[Y_{i1}\,|\,Y_{i+}]} = \frac{\left(\sum_{i=1}^{I} Y_{i1} - Y_{i+}\frac{n_{i1}}{n_{i+}}\right)^2}{\sum_{i=1}^{I} Y_{i+}\frac{n_{i1}n_{i0}}{n_{i+}^2}\frac{n_{i+}-Y_{i+}}{n_{i+}-1}}, \tag{43}$$

where $Y_{i+} = Y_{i0} + Y_{i1}$ and $n_{i+} = n_{i0} + n_{i1}$. Under the null hypothesis $R_n^{(C)} \xrightarrow[n\to\infty]{d} \chi_1^2$, where $n = \sum_{i=1}^{I}\sum_{j=0}^{1} n_{ij}$.

Example 36. Consider in Example 35 the special case $I = 1$. Thus the model simplifies to comparing two binomial distributions. Let $Y_0 \sim \mathsf{Bi}(n_0, p_0)$ and $Y_1 \sim \mathsf{Bi}(n_1, p_1)$. Note that the standard approaches of testing the null hypothesis $H_0 : p_0 = p_1$ against the alternative $H_1 : p_0 \neq p_1$ are asymptotic.

Conditional approach offers an exact inference. Analogously as in Example 35 introduce the parametrization

$$\log\left(\tfrac{p_j}{1-p_j}\right) = \psi + \tau\,\mathbb{I}\{j=1\}, \qquad j = 0, 1.$$

Note that in this parametrization τ is the logarithm of odds-ratio.

Put $Y_+ = Y_0 + Y_1$ and $y_+ = y_0 + y_1$. Then

$$\mathsf{P}_\tau\left(Y_1 = k \,|\, Y_+ = y_+\right) = \frac{\binom{n_1}{k}\binom{n_0}{y_+-k}\mathrm{e}^{\tau\,k}}{\sum_{l\in\mathcal{K}}\binom{n_1}{l}\binom{n_0}{y_+-l}\mathrm{e}^{\tau\,l}}, \quad k \in \mathcal{K}, \tag{44}$$

where $\mathcal{K} = \big\{\max\{0, y_+ - n_0\}, \ldots, \min\{y_+, n_1\}\big\}$.

Thus the p-value of the 'exact' test of the null hypothesis $H_0 : \tau = \tau_0$ against $H_1 : \tau \neq \tau_0$ would be

$$p(\tau_0) = 2\min\big\{\mathsf{P}_{\tau_0}(Y_1 \leq y_1 \,|\, Y_+ = y_1 + y_2), \mathsf{P}_{\tau_0}(Y_1 \geq y_1 \,|\, Y_+ = y_1 + y_2)\big\}, \tag{45}$$

where y_0 and y_1 are the observed values of Y_0 and Y_1 respectively.

By the inversion of the test one can define the 'exact' confidence interval for τ as the set of those values for which we do not reject the null hypothesis, i.e.

$$CI = (\widehat{\tau}_L, \widehat{\tau}_U) = \{\tau \in \mathbb{R} : p(\tau) > \alpha\}.$$

The confidence interval for odds-ratio calculated by the function `fisher.test()` is now given by $\left(\mathrm{e}^{\widehat{\tau}_L}, \mathrm{e}^{\widehat{\tau}_U}\right)$.

The special case presents testing the null hypothesis $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. Then (44) simplifies to

$$\mathsf{P}_0(Y_1 = k \,|\, Y_+ = y_+) = \frac{\binom{n_1}{k}\binom{n_0}{y_+-k}}{\sum_{l\in\mathcal{K}}\binom{n_1}{l}\binom{n_0}{y_+-l}} = \frac{\binom{n_1}{k}\binom{n_0}{y_+-k}}{\binom{n_1+n_0}{y_+}}, \quad k \in \mathcal{K}.$$

This corresponds to *Fisher's exact test* sometimes known also as *Fisher's factorial test*. Be careful that the p-value of the test as implemented in `fisher.test()` is not calculated by (45) but as

$$\widetilde{p} = \sum_{k\in\mathcal{K}_-} \mathsf{P}_0(Y_1 = k \,|\, Y_+ = y_+),$$

where

$$\mathcal{K}_- = \big\{k \in \mathcal{K} : \mathsf{P}_0(Y_1 = k \,|\, Y_+ = y_+) \leq \mathsf{P}_0(Y_1 = y_1 \,|\, Y_+ = y_+)\big\},$$

which sometimes slightly differs from $p(0)$ as defined in (45).

Note that Fisher's exact test presents an alternative to the χ^2-square test of independence in the 2×2 contingency table

	Sample 1	Sample 2
Success	y_0	y_1
Failure	$n_0 - y_0$	$n_1 - y_1$

,

which is an asymptotic test.

Example 37. Consider in Example 35 the special case $n_{i0} = n_{i1} = 1$ for each $i \in \{1, \dots, I\}$. Introduce

$$N_{jk} = \sum_{i=1}^{I} \mathbb{I}\{Y_{i0} = j, Y_{i1} = k\}, \qquad j, k \in \{0, 1\}.$$

Then the test statistic (43) simplifies to

$$R_n^{(C)} = \frac{(N_{01} - N_{10})^2}{N_{01} + N_{10}},$$

which is known as McNemar's test.

Example 38. Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be two independent random samples from the Poisson distributions. Let λ_X be the true value of the parameter for the first sample and λ_Y for the second sample. Note that $\mathbf{S} = (S_1, S_2)^\mathsf{T} = \left(\sum_{i=1}^{n_1} X_i, \sum_{i=1}^{n_2} Y_i\right)^\mathsf{T}$ is a sufficient statistic for the parameter $\boldsymbol{\theta} = (\lambda_X, \lambda_Y)^\mathsf{T}$. Derive the conditional distribution of S_1 given $S_1 + S_2$. Use this result to find an exact test of

$$H_0 : \lambda_X = \lambda_Y, \qquad H_1 : \lambda_X \neq \lambda_Y.$$

Further derive an 'exact' confidence interval for the ratio $\frac{\lambda_X}{\lambda_Y}$.

Literature: Pawitan [2001] Chapters 10.1–10.5.

The end of class 11 (26. 3. 2024)

3 M- and Z-estimators

M- and Z-estimators present a very general class of estimators that include most of the commonly used estimators.*

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be a random sample from a distribution F from the model $\mathcal{F}$ and one is interested in estimating some quantity (p-dimensional parameter) of this distribution, say $\boldsymbol{\theta}(F)$.

* *One should be careful as the terminology may vary. But (among others) minimum contrast estimators, pseudo-likelihood estimators, quasi-likelihood estimators and estimating equations can be usually viewed as either M-estimators or Z-estimators.*

M-estimator

Let ρ be a function defined on $S_{\boldsymbol{X}} \times \Theta$, where $S_{\boldsymbol{X}}$ is the support of F. Futher denote Θ the parameter space, i.e. $\Theta = \{\boldsymbol{\theta}(F), F \in \mathcal{F}\}$. The **$M$-estimator**[*] is defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \rho(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

Examples of M-estimators

Note that in *parametric models* $\{f(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ the maximum likelihood (ML) estimator

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \log f(\boldsymbol{X}_i; \boldsymbol{\theta})$$

can be viewed as the M-estimator with

$$\rho_{ML}(\mathbf{x}; \boldsymbol{\theta}) = -\log f(\mathbf{x}; \boldsymbol{\theta}).$$

In *regression problems* one observes $\mathbf{Z}_1 = \binom{Y_1}{\boldsymbol{X}_1}, \ldots, \mathbf{Z}_n = \binom{Y_n}{\boldsymbol{X}_n}$. Then the least squares (LS) estimator of regression parameters

$$\widehat{\boldsymbol{\beta}}_n^{(LS)} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^{\mathsf{T}}\mathbf{b})^2$$

can be viewed as the M-estimator with

$$\rho_{LS}(\mathbf{z}; \boldsymbol{\beta}) = \rho_{LS}(\mathbf{x}, y; \boldsymbol{\beta}) = \big(y - \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}\big)^2.$$

Similarly the least absolute deviation (LAD) estimator

$$\widehat{\boldsymbol{\beta}}_n^{(LAD)} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} |Y_i - \boldsymbol{X}_i^{\mathsf{T}}\mathbf{b}|$$

can be viewed as the M-estimator with

$$\rho_{LAD}(\mathbf{z}; \boldsymbol{\beta}) = \rho(\mathbf{x}, y; \boldsymbol{\beta}) = \big|y - \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}\big|.$$

Z-estimator

Often the maximizing value in the definition of M-estimator is sought by setting a derivative (or the set of partial derivatives if $\boldsymbol{\theta}$ is multidimensional) equal to zero. Thus we search for $\widehat{\boldsymbol{\theta}}_n$ as the point that solves the set of estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}\big(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n\big) = \mathbf{0}_p, \qquad \text{where} \qquad \boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \rho(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{46}$$

[*] *M-odhad*

Note that

$$\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}) = \big(\psi_1(\mathbf{x};\boldsymbol{\theta}),\ldots,\psi_p(\mathbf{x};\boldsymbol{\theta})\big)^\mathsf{T} = \Big(\frac{\partial\rho(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_1},\ldots,\frac{\partial\rho(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_p}\Big)^\mathsf{T}.$$

Generally let $\boldsymbol{\psi}$ be a p-dimensional vector function (not necessarily a derivative of the function ρ) defined on $S_{\boldsymbol{X}} \times \Theta$. Then we define the Z-estimator* as the solution of the system of equations (46).

Note that the maximum likelihood (ML) and the least squares (LS) estimators can be also viewed as Z-estimators with

$$\boldsymbol{\psi}_{ML}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}, \qquad \boldsymbol{\psi}_{LS}(\mathbf{x},y;\boldsymbol{\beta}) = \big(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}\big)\,\mathbf{x}.$$

Literature: van der Vaart [2000] – Chapter 5.1.

3.1 Identifiability of parameters† via M- and/or Z-estimators

When using M- or Z-estimators one should check the potential of these estimators to identify the parameters of interest. Note that by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^n \rho(\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathsf{E}\ \rho(\boldsymbol{X}_1;\boldsymbol{\theta}) + o_P(1), \qquad \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathsf{E}\ \boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}) + o_P(1).$$

Thus the M-estimator $\widehat{\boldsymbol{\theta}}_n$ identifies (at the population level) the quantity

$$\boldsymbol{\theta}_X = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathsf{E}\ \rho(\boldsymbol{X}_1;\boldsymbol{\theta})$$

and analogously Z-estimator identifies $\boldsymbol{\theta}_X$ such that

$$\mathsf{E}\ \boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X) = \mathbf{0}_p.$$

Example 39. Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be i.i.d. observations from a distribution with a density $f(\mathbf{x})$ (with respect to a σ-finite measure μ). By assuming that f belongs to a parametric family of densities $\mathcal{F} = \big\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\big\}$ we are estimating (identifying) $\boldsymbol{\theta}_X$ such that

$$\boldsymbol{\theta}_X = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{E}\ \log f(\boldsymbol{X}_1;\boldsymbol{\theta})$$

Provided that the true density $f(\mathbf{x})$ has the support $S_{\boldsymbol{X}}$ that is the same as the support of $f(\mathbf{x};\boldsymbol{\theta})$ for each $\boldsymbol{\theta}\in\Theta$, this can be further rewritten as

$$\boldsymbol{\theta}_X = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{E}\ \log\Big[\tfrac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1)}\Big].$$

---

* *Z-odhad* † *Identifikovatelnost parametrů.*

Now by Jensen's inequality

$$\mathsf{E}\,\log\left[\tfrac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1)}\right] \le \log\left\{\mathsf{E}\,\left[\tfrac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1)}\right]\right\} = \log\left\{\int_{S_{\boldsymbol{X}}} \tfrac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x})}\, f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})\right\} = \log\{1\} = 0.$$

Suppose that our (parametric) assumption **is right** and there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $f(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_0)$. Then $\mathsf{E}\,\log\left[\frac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1;\boldsymbol{\theta}_0)}\right]$ is maximised for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and thus $\boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ (i.e. the maximum likelihood method identifies the true value of the parameter).

Suppose that our (parametric) assumption **is not right** and that $f \notin \mathcal{F}$. Then

$$\begin{aligned}\boldsymbol{\theta}_X &= \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{E}\,\log\left[\tfrac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1)}\right] = \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{S_{\boldsymbol{X}}} \log\left[\tfrac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x})}\right] f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{S_{\boldsymbol{X}}} \log\left[\tfrac{f(\mathbf{x})}{f(\mathbf{x};\boldsymbol{\theta})}\right] f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}).\end{aligned}$$

The integral $\int_{S_{\boldsymbol{X}}} \log\left[\frac{f(\mathbf{x})}{f(\mathbf{x};\boldsymbol{\theta})}\right] f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})$ is called the *Kullback–Leibler divergence* from $f(\mathbf{x};\boldsymbol{\theta})$ to $f(\mathbf{x})$ (it measures how $f(\mathbf{x};\boldsymbol{\theta})$ diverges from $f(\mathbf{x})$). Thus $\boldsymbol{\theta}_X$ is the point of parameter space Θ for which the Kullback–Leibler divergence from $\mathcal{F}$ to f is minimised.

3.2 Asymptotic distribution of Z-estimators

Analogously as for the maximum likelihood estimator the basic asymptotic results will be formulated for Z-estimators. In order to do that put $\mathbf{Z}(\boldsymbol{\theta}) = \mathsf{E}\,\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta})$ and $\mathbb{D}_{\boldsymbol{\psi}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}$ (the Jacobi matrix of $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$).

To state the theorem about asymptotic normality we will need the following regularity assumptions. These assumptions are analogous to assumptions [R0]-[R6] for the maximum likelihood estimators.

[Z0] *Identifiability.* $\boldsymbol{\theta}_X$ satisfies $\mathbf{Z}(\boldsymbol{\theta}_X) = \mathbf{0}_p$.

[Z1] The number of parameters p in the model is *constant*.

[Z2] (The true value of the parameter) $\boldsymbol{\theta}_X$ is an *interior point* of the parameter space Θ.

[Z3] For μ-almost all $\mathbf{x}$ each component of the function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ is *differentiable* with respect to $\boldsymbol{\theta}$.

[Z4] There exists $\alpha > 0$, an open neighbourhood U of $\boldsymbol{\theta}_X$ and a function $M(\mathbf{x})$ so that for each $j,k \in \{1,\ldots,p\}$ and for each $\boldsymbol{\theta} \in U$

$$\left|\frac{\partial\psi_j(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_k} - \frac{\partial\psi_j(\mathbf{x};\boldsymbol{\theta}_X)}{\partial\theta_k}\right| \le M(\mathbf{x})\,\|\boldsymbol{\theta} - \boldsymbol{\theta}_X\|^\alpha$$

for μ-almost all $\mathbf{x}$ and $\mathsf{E}\,M(\boldsymbol{X}_1) < \infty$.

[Z5] The matrix

$$\mathbb{\Gamma}(\boldsymbol{\theta}_X) = \mathsf{E}\, \mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_1; \boldsymbol{\theta}_X) \tag{47}$$

is finite and *regular*.

[Z6] The *variance matrix*

$$\mathbb{\Sigma}(\boldsymbol{\theta}_X) = \mathsf{var}\left(\boldsymbol{\psi}(\boldsymbol{X}_1; \boldsymbol{\theta}_X)\right) = \mathsf{E}\left[\boldsymbol{\psi}(\boldsymbol{X}_1; \boldsymbol{\theta}_X)\, \boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_1; \boldsymbol{\theta}_X)\right] \tag{48}$$

is finite.

The end of class 12 (2. 4. 2024)

Introduce

$$\mathbb{\Gamma}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

The following technical lemma says that if $\boldsymbol{\theta}$ is 'close' to $\boldsymbol{\theta}_X$, then $\mathbb{\Gamma}_n(\boldsymbol{\theta})$ is close to $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$. This result will be useful for the proof of the consistency and asymptotic normality of Z-estimators. Note that it is an analogy of Lemma 1.

Lemma 4. *Suppose that assumptions* [Z1]-[Z5] *are satisfied. Let* $\{\varepsilon_n\}$ *be a sequence of positive numbers going to zero. Then*

$$\max_{j,k \in \{1,\ldots,p\}} \sup_{\boldsymbol{\theta} \in U_{\varepsilon_n}} \left| \left(\mathbb{\Gamma}_n(\boldsymbol{\theta}) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)\right)_{jk} \right| = o_P(1),$$

where

$$U_{\varepsilon_n} = \left\{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_X\| \leq \varepsilon_n\right\}$$

and $\left(\mathbb{\Gamma}_n(\boldsymbol{\theta}) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)\right)_{jk}$ *stands for the* (j,k)*-element of the difference of the matrices* $\mathbb{\Gamma}_n(\boldsymbol{\theta}) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)$.

Proof. Using assumption [Z4] and the law of large numbers one can bound

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in U_{\varepsilon_n}} \left| \left(\mathbb{\Gamma}_n(\boldsymbol{\theta}) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)\right)_{jk} \right| &\leq \sup_{\boldsymbol{\theta} \in U_{\varepsilon_n}} \left| \left(\mathbb{\Gamma}_n(\boldsymbol{\theta}) - \mathbb{\Gamma}_n(\boldsymbol{\theta}_X)\right)_{jk} \right| + \left| \left(\mathbb{\Gamma}_n(\boldsymbol{\theta}_X) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)\right)_{jk} \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} M(\boldsymbol{X}_i)\, \varepsilon_n^{\alpha} + o_P(1) = O_P(1)\, o(1) + o_P(1) = o_P(1),
\end{aligned}$$

which implies the statement of the lemma. □

Suppose now that $\widehat{\mathbf{t}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$. Note that the above lemma (together with the reasoning of Corollary 1) one gets that for each $j, k \in \{1, \ldots, p\}$:

$$\left| \left(\mathbb{\Gamma}_n(\widehat{\mathbf{t}}_n) - \mathbb{\Gamma}(\boldsymbol{\theta}_X)\right)_{jk} \right| = o_P(1). \tag{49}$$

Theorem 9. *Suppose that assumptions* [Z0]-[Z6] *are satisfied.*

(i) Then with probability going to one there exists a consistent solution $\widehat{\boldsymbol{\theta}}_n$ to the estimating equations (46).

(ii) Further, if $\widehat{\boldsymbol{\theta}}_n$ is a consistent root of the estimating equations (46), then

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = -\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) + o_P(1), \tag{50}$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\Big(\mathbf{0}_p, \mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\mathbb{\Sigma}(\boldsymbol{\theta}_X)\left[\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\right]^{\mathsf{T}}\Big), \tag{51}$$

where the matrices $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$ and $\mathbb{\Sigma}(\boldsymbol{\theta}_X)$ are defined in (47) and (48) respectively.

Proof. Consistency: Introduce the vector function

$$h_n(\boldsymbol{\theta}) = \boldsymbol{\theta} - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbf{Z}_n(\boldsymbol{\theta}),$$

where

$$\mathbf{Z}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}).$$

In what follows we will show that with probability going to one (as $n\to\infty$) the mapping h_n is a contraction on $U_{\varepsilon_n} = \{\boldsymbol{\theta}\in\Theta : \|\boldsymbol{\theta}-\boldsymbol{\theta}_X\|\le\varepsilon_n\}$, where $\{\varepsilon_n\}$ is a sequence of positive numbers going to zero such that $\varepsilon_n\sqrt{n}\xrightarrow[n\to\infty]{}\infty$. Having proved that then by the Banach fixed point theorem (Theorem A14) there exists a unique fixed point $\widehat{\boldsymbol{\theta}}_n\in U_{\varepsilon_n}$ such that $h_n(\widehat{\boldsymbol{\theta}}_n)=\widehat{\boldsymbol{\theta}}_n$ and thus also $\mathbf{Z}_n(\widehat{\boldsymbol{\theta}}_n)=\mathbf{0}_p$. This implies the existence of a consistent root of the estimating equations (46).

Showing that h_n is a contraction on U_{ε_n}. Let $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in U_{\varepsilon_n}$ then

$$\begin{aligned}\left\|h_n(\boldsymbol{\theta}_1)-h_n(\boldsymbol{\theta}_2)\right\| &= \left\|(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2) - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\left(\mathbf{Z}_n(\boldsymbol{\theta}_1)-\mathbf{Z}_n(\boldsymbol{\theta}_2)\right)\right\| \\ &= \left\|\left(\mathbb{I}_p - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbb{\Gamma}_n^*\right)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\right\|,\end{aligned} \tag{52}$$

where $\mathbb{\Gamma}_n^*$ is $(p\times p)$-matrix whose j-th row is the j-th row of the matrix

$$\mathbb{\Gamma}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_i;\boldsymbol{\theta})$$

evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\boldsymbol{\theta}_1$ a $\boldsymbol{\theta}_2$. Note that $\boldsymbol{\theta}_n^{j*}\in U_{\varepsilon_n}$. Now by Lemma 4 and assumption **[Z5]**

$$a_n = \max_{j,k\in\{1,\ldots,p\}}\sup_{\boldsymbol{\theta}\in U_{\varepsilon_n}}\left|\left(\mathbb{I}_p - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbb{\Gamma}_n(\boldsymbol{\theta})\right)_{jk}\right| = o_P(1). \tag{53}$$

So with the help of (52) and (53) it holds that uniformly in $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U_{\varepsilon_n}$

$$\big\|h_n(\boldsymbol{\theta}_1) - h_n(\boldsymbol{\theta}_2)\big\| \leq o_P(1)\,\big\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\big\|, \tag{54}$$

which implies that there exists $q \in (0,1)$ such that

$$\mathsf{P}\left(\forall_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in U_{\varepsilon_n}}\ \big\|h_n(\boldsymbol{\theta}_1) - h_n(\boldsymbol{\theta}_2)\big\| \leq q\big\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\big\|\right) \xrightarrow[n\to\infty]{} 1.$$

Thus to show that h_n is a contraction on U_{ε_n} it remains to prove that (with probability going to one) $h_n : U_{\varepsilon_n} \to U_{\varepsilon_n}$. Note that for each $\boldsymbol{\theta} \in U_{\varepsilon_n}$ the inequality (54) implies

$$h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}_X) = o_P(1)\,\varepsilon_n, \tag{55}$$

where the $o_P(1)$ term does not depend on $\boldsymbol{\theta}$. Further

$$h_n(\boldsymbol{\theta}_X) = \boldsymbol{\theta}_X - \big[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\big]^{-1}\mathbf{Z}_n(\boldsymbol{\theta}_X) = \boldsymbol{\theta}_X + O_P\big(\tfrac{1}{\sqrt{n}}\big), \tag{56}$$

where we have used that by the central limit theorem $\mathbf{Z}_n(\boldsymbol{\theta}_X) = O_P\big(\frac{1}{\sqrt{n}}\big)$. Now combining (55) and (56) yields that

$$h_n(\boldsymbol{\theta}) = o_P(1)\,\varepsilon_n + h_n(\boldsymbol{\theta}_X) = o_P(1)\,\varepsilon_n + \boldsymbol{\theta}_X + O_P\big(\tfrac{1}{\sqrt{n}}\big),$$

which further together with the assumption $\varepsilon_n\sqrt{n} \xrightarrow[n\to\infty]{} \infty$ gives

$$h_n(\boldsymbol{\theta}) - \boldsymbol{\theta}_X = \varepsilon_n\Big(o_P(1) + O_P\big(\tfrac{1}{\varepsilon_n\sqrt{n}}\big)\Big) = \varepsilon_n o_P(1).$$

This finally implies that $\mathsf{P}\big(\forall \boldsymbol{\theta} \in U_{\varepsilon_n} : h_n(\boldsymbol{\theta}) \in U_{\varepsilon_n}\big) \xrightarrow[n\to\infty]{} 1$, which was to be proved.

Asymptotic normality: This is proved analogously as in Theorem 5. Let $\widehat{\boldsymbol{\theta}}_n$ be a consistent root of the estimating equations. Then by the mean value theorem applied to each component of $\mathbf{Z}_n\big(\widehat{\boldsymbol{\theta}}_n\big)$ one gets

$$\mathbf{0}_p = \mathbf{Z}_n\big(\widehat{\boldsymbol{\theta}}_n\big) = \mathbf{Z}_n(\boldsymbol{\theta}_X) + \mathbb{\Gamma}_n^*\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big),$$

where similarly as in the proof of consistency $\mathbb{\Gamma}_n^*$ is $(p \times p)$-matrix whose j-th row is the j-th row of the matrix $\mathbb{\Gamma}_n(\boldsymbol{\theta})$ evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\widehat{\boldsymbol{\theta}}_n$ a $\boldsymbol{\theta}_X$. Thus $\boldsymbol{\theta}_n^{j*} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ as $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_X$. So one can use (49) to conclude that $\mathbb{\Gamma}_n^* \xrightarrow[n\to\infty]{P} \mathbb{\Gamma}(\boldsymbol{\theta}_X)$. Now with the help of CS (Theorem 2) one can write

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big) = -[\mathbb{\Gamma}_n^*]^{-1}\sqrt{n}\,\mathbf{Z}_n(\boldsymbol{\theta}_X) = -\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) + o_P(1),$$

which with the help of the central limit theorem (for i.i.d. random vectors) and CS (Theorem 2) implies the second statement of the theorem. □

Remark 15. If there exists a real function $\rho(\mathbf{x};\boldsymbol{\theta})$ such that $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial\rho(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, then the matrix $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$ is symmetric and one can simply write $\mathbb{\Gamma}(\boldsymbol{\theta}_X)^{-1}$ instead of $[\mathbb{\Gamma}(\boldsymbol{\theta}_X)^{-1}]^{\mathsf{T}}$ in (51).

Asymptotic variance estimations

Note that by Theorem 9 one has

$$\widehat{\boldsymbol{\theta}}_n \overset{\text{as}}{\approx} \mathsf{N}_p\big(\boldsymbol{\theta}_X, \tfrac{1}{n}\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\mathbb{\Sigma}(\boldsymbol{\theta}_X)\,\big[\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\big]^{\mathsf{T}}\big).$$

Thus the most straightforward estimate of the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ is the 'sandwich estimator' given by

$$\widehat{\mathsf{avar}\,(\widehat{\boldsymbol{\theta}}_n)} = \frac{1}{n}\,\widehat{\mathbb{\Gamma}}_n^{-1}\,\widehat{\mathbb{\Sigma}}_n\,\big[\widehat{\mathbb{\Gamma}}_n^{-1}\big]^{\mathsf{T}}, \tag{57}$$

where

$$\widehat{\mathbb{\Gamma}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{D}_{\boldsymbol{\psi}}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) \qquad \text{and} \qquad \widehat{\mathbb{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big)\,\boldsymbol{\psi}^{\mathsf{T}}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big).$$

Note that Lemma 4 together with the consistency of $\widehat{\boldsymbol{\theta}}_n$ implies that

$$\widehat{\mathbb{\Gamma}}_n \xrightarrow[n\to\infty]{P} \mathbb{\Gamma}(\boldsymbol{\theta}_X).$$

It is more tedious to give some general assumptions so that it also holds

$$\widehat{\mathbb{\Sigma}}_n \xrightarrow[n\to\infty]{P} \mathbb{\Sigma}(\boldsymbol{\theta}_X).$$

To derive such assumptions rewrite

$$\begin{aligned}
\widehat{\mathbb{\Sigma}}_n = {} & \frac{1}{n}\sum_{i=1}^{n}\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]\,\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]^{\mathsf{T}} \\
& + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\,\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]^{\mathsf{T}} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]\,\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) \\
& + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\,\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X).
\end{aligned} \tag{58}$$

Now by the law of large numbers the last summand in (58) converges in probability to $\mathbb{\Sigma}(\boldsymbol{\theta}_X)$, thus it is sufficient to show that the remaining terms are of order $o_P(1)$. With the help of assumption **[Z4]** this can be done for instance by assuming that for each $j,k \in \{1,\dots,p\}$

$$\mathsf{E}\; M_{jk}^2(\boldsymbol{X}_1) < \infty \quad \text{and} \quad \mathsf{E}\left|\tfrac{\partial\psi_j(\boldsymbol{X}_1;\boldsymbol{\theta}_X)}{\partial\theta_k}\right|^2 < \infty.$$

Confidence sets and confidence intervals

Suppose that $\widehat{\mathbb{V}}_n$ is a consistent estimator of $\mathbb{V} = \mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\mathbb{\Sigma}(\boldsymbol{\theta}_X)\,[\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)]^{\mathsf{T}}$.

Then by the Cramér-Slutsky theorem the confidence set (ellipsoid) for the parameter $\boldsymbol{\theta}_X$ is given by

$$\Big\{\boldsymbol{\theta} \in \Theta : \, n\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^{\mathsf{T}} \widehat{\mathbb{V}}_n^{-1} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq \chi^2_p(1-\alpha)\Big\}.$$

The 'Wald-type' (asymptotic) confidence interval for θ_{Xj} (the j-th coordinate of $\boldsymbol{\theta}_X$) is given by

$$\Big[\widehat{\theta}_{nj} - \tfrac{u_{1-\alpha/2}\sqrt{\widehat{v}_{n,jj}}}{\sqrt{n}}, \, \widehat{\theta}_{nj} + \tfrac{u_{1-\alpha/2}\sqrt{\widehat{v}_{n,jj}}}{\sqrt{n}}\Big],$$

where $\widehat{\theta}_{nj}$ is the j-th coordinate of $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{v}_{n,jj}$ is the j-th diagonal element of the matrix $\widehat{\mathbb{V}}_n$.
Literature: Sen et al. [2010] Chapter 8.2.

3.3 Likelihood under model misspecification

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample with a density f (with respect to a σ-finite measure μ). Then the maximum likelihood estimator can be viewed as the M-estimator with $\rho(\mathbf{x}; \boldsymbol{\theta}) = -\log f(\mathbf{x}; \boldsymbol{\theta})$ or Z-estimator with $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. From Example 39 we know that when assuming $f \in \mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$, the method of the maximum likelihood identifies the parameter

$$\boldsymbol{\theta}_X = \arg\min_{\boldsymbol{\theta} \in \Theta} \int_{S_{\boldsymbol{X}}} \log\left[\tfrac{f(\mathbf{x})}{f(\mathbf{x};\boldsymbol{\theta})}\right] f(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x}).$$

Further by Theorem 9 we also know that (with probability going to one there exists a consistent solution $\widehat{\boldsymbol{\theta}}_n$ of (46) which satisfies)

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \mathbb{I}^{-1}(\boldsymbol{\theta}_X)\mathbb{\Sigma}(\boldsymbol{\theta}_X)\big[\mathbb{I}^{-1}(\boldsymbol{\theta}_X)\big]^{\mathsf{T}}\big).$$

Suppose that our parametric assumption is right and $f \in \mathcal{F}$, i.e. there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $f(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}_0)$. Then the identified parameter is equal to $\boldsymbol{\theta}_0$, i.e. $\boldsymbol{\theta}_X = \boldsymbol{\theta}_0$. Further it is easy to see that $\mathbb{I}(\boldsymbol{\theta}_X) = I(\boldsymbol{\theta}_X) = \mathbb{\Sigma}(\boldsymbol{\theta}_X)$, where $I(\boldsymbol{\theta}_X)$ is the Fisher information matrix. Thus

$$\mathbb{I}^{-1}(\boldsymbol{\theta}_X)\mathbb{\Sigma}(\boldsymbol{\theta}_X)\mathbb{I}^{-1}(\boldsymbol{\theta}_X) = I^{-1}(\boldsymbol{\theta}_X).$$

So one can view Theorem 5 as a special case of Theorem 9. Further, when doing the inference about $\boldsymbol{\theta}_X$ it is sufficient to estimate the Fisher information matrix.

Often in practice we are not completely sure that $f \in \mathcal{F}$. If we are not sure about the parametric assumption then it is safer to view the estimator $\widehat{\boldsymbol{\theta}}_n$ as an Z-estimator with $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. The asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ can now be estimated with the help of 'sandwich estimator' (57) where

$$\begin{aligned}
\widehat{\mathbb{\Sigma}}_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n)\, \mathbf{U}^{\mathsf{T}}(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n), \text{ where } \mathbf{U}(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \\
\widehat{\mathbb{I}}_n &= \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n), \text{ where } I(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}.
\end{aligned}$$

This type of variance estimator is calculated for GLM models by the function `sandwich` (from the package with the same name).

The end of class 13 (5. 4. 2024)

Example 40. *Misspecified normal linear model.* Let $\binom{Y_1}{\boldsymbol{X}_1}, \dots, \binom{Y_n}{\boldsymbol{X}_n}$ be independent and identically distributed random vectors, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\mathsf{T}$. Note that if one assumes that $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \sim \mathsf{N}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}, \sigma^2)$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$, then the maximum likelihood estimation of $\boldsymbol{\beta}$ corresponds to the method of the least squares given by $\rho_{LS}(\mathbf{x}, y; \boldsymbol{\beta}) = \big(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}\big)^2$.

Show that without the assumption $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \sim \mathsf{N}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}, \sigma^2)$ the method of the least squares identifies the parameter

$$\boldsymbol{\beta}_X = \big[\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]^{-1}\ \mathsf{E}\ Y_1\boldsymbol{X}_1$$

and it holds that $\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V})$, where

$$\mathbb{V} = \big[\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]^{-1}\big[\mathsf{E}\ \sigma^2(\boldsymbol{X}_1)\,\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]\big[\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]^{-1},$$

with $\sigma^2(\boldsymbol{X}_1) = \mathsf{E}\,\big[(Y_1 - \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X)^2|\boldsymbol{X}_1\big]$.

Note that provided $\mathsf{E}\,\big[Y_1|\boldsymbol{X}_1\big] = \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0$ for some $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, then $\boldsymbol{\beta}_X = \boldsymbol{\beta}_0$ and $\sigma^2(\boldsymbol{X}_1) = \mathsf{var}(Y_1|\boldsymbol{X}_1)$.

Example 41. *Misspecified Poisson regression.* Let $\binom{Y_1}{\boldsymbol{X}_1}, \dots, \binom{Y_n}{\boldsymbol{X}_n}$ be independent and identically distributed random vectors, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\mathsf{T}$. Assume that the conditional distribution of Y_i given $\boldsymbol{X}_i$ is Poisson, i.e. $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \sim \mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, where $\lambda(\mathbf{x}) = \mathrm{e}^{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}}$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\mathsf{T}$. The score statistic for the maximum likelihood estimation is given by

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{X}_i\big(Y_i - \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}}\big).$$

Thus one can view the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n$ as the Z-estimator with

$$\boldsymbol{\psi}(\mathbf{x}, y; \boldsymbol{\beta}) = \mathbf{x}\big(y - \mathrm{e}^{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}}\big) \tag{59}$$

and $\boldsymbol{\beta}_X$ solves the system of equations

$$\mathsf{E}\ \boldsymbol{X}_1\big(Y_1 - \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X}\big) = \mathbf{0}_p.$$

Suppose now that $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \not\sim \mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, but one can still assume that there exists $\boldsymbol{\beta}_0$ such that $\mathsf{E}\,[Y_1|\boldsymbol{X}_1] = \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0}$. Then

$$\mathsf{E}\ \boldsymbol{X}_1\big(Y_1 - \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0}\big) = \mathsf{E}\,\Big\{\mathsf{E}\,\big[\boldsymbol{X}_1\big(Y_1 - \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0}\big)\big|\boldsymbol{X}_1\big]\Big\} = \mathsf{E}\,\big[\boldsymbol{X}_1\big(\mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0} - \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_0}\big)\big] = \mathbf{0}_p.$$

Thus $\boldsymbol{\beta}_X$ identifies $\boldsymbol{\beta}_0$ which describes the effect of the covariates on the expected mean value.

The above calculation implies that when we are not sure that the conditional distribution $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big)$ is $\mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, but we are willing to assume that $\mathsf{E}\,[Y_i|\boldsymbol{X}_i] = \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0}$ for some $\boldsymbol{\beta}_0 \in \mathbb{R}^p$,

then we can still use the score function (59) which identifies the parameter $\boldsymbol{\beta}_0$. By Theorem 9 we know that the estimator $\widehat{\boldsymbol{\beta}}_n$ is asymptotically normal with the matrices $\mathbb{\Gamma}(\boldsymbol{\beta}_X)$ and $\mathbb{\Sigma}(\boldsymbol{\beta}_X)$ given by

$$\mathbb{\Sigma}(\boldsymbol{\beta}_X) = \mathsf{E}\ \boldsymbol{X}_1 \boldsymbol{X}_1^\mathsf{T} \big(Y_1 - \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X}\big)^2 \quad \text{and} \quad \mathbb{\Gamma}(\boldsymbol{\beta}_X) = \mathsf{E}\ \boldsymbol{X}_1\, \boldsymbol{X}_1^\mathsf{T} \mathrm{e}^{\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X}.$$

Thus the asymptotic variance of the estimator $\widehat{\boldsymbol{\beta}}_n$ can be estimated by

$$\widehat{\mathsf{avar}\,(\widehat{\boldsymbol{\beta}}_n)} = \frac{1}{n}\, \widehat{\mathbb{\Gamma}}_n^{-1}\, \widehat{\mathbb{\Sigma}}_n\, \widehat{\mathbb{\Gamma}}_n^{-1},$$

where

$$\widehat{\mathbb{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T} \big(Y_i - \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n}\big)^2 \quad \text{and} \quad \widehat{\mathbb{\Gamma}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\, \boldsymbol{X}_i^\mathsf{T} \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n}.$$

Literature: White [1980], White [1982].

3.4 Asymptotic normality of M-estimators defined by convex minimization

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from a distribution F and one is interested in estimating some quantity $\boldsymbol{\theta}_X$ (p-dimensional parameter) of this distribution such that this parameter can be identified as

$$\boldsymbol{\theta}_X = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \mathsf{E}\ \rho(\boldsymbol{X}_1; \boldsymbol{\theta}),$$

where for each fixed $\mathbf{x}$ the function $\rho(\mathbf{x}; \boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. Further suppose that the parameter space Θ is a subset of $\mathbb{R}^p$.

A straightforward estimator of the parameter $\boldsymbol{\theta}_X$ is given by

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

As we see later we do not need to assume that the function $\rho(\mathbf{x}; \boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ for all $(\mathbf{x}, \boldsymbol{\theta}) \in S_{\boldsymbol{X}} \times \Theta$. Nevertheless we need a function that plays the role of the function $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta})$ in the definition of Z-estimators (46). Note that the convexity in $\boldsymbol{\theta}$ guarantees that for each $\mathbf{x}$ the function $\rho(\mathbf{x}, \boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ for almost all $\boldsymbol{\theta} \in \Theta$. So suppose that there exists a function $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta})$ such that $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \rho(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ whenever this derivative exists. Moreover suppose that similarly as for Z-estimators it holds that

$$\mathsf{E}\ \boldsymbol{\psi}(\boldsymbol{X}_1; \boldsymbol{\theta}_X) = \mathbf{0}_p. \tag{60}$$

For formulating the main result it is useful to introduce the 'remainder function'

$$R(\mathbf{x}; \mathbf{t}) = \rho(\mathbf{x}; \boldsymbol{\theta}_X + \mathbf{t}) - \rho(\mathbf{x}; \boldsymbol{\theta}_X) - \mathbf{t}^\mathsf{T} \boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}_X) \tag{61}$$

and the asymptotic (expected) objective function

$$M(\boldsymbol{\theta}) = \mathsf{E}\ \rho(\boldsymbol{X}_1; \boldsymbol{\theta}). \tag{62}$$

Theorem 10. *Assume that the function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ satisfies (60) and the functions $R(\mathbf{x};\mathbf{t})$ and $M(\boldsymbol{\theta})$ are defined by (61) and (62) respectively. Further suppose that*

(i) there exists a positive definitive matrix $\mathbb{J}(\boldsymbol{\theta}_X)$ such that

$$M(\boldsymbol{\theta}_X + \mathbf{t}) = M(\boldsymbol{\theta}_X) + \tfrac{1}{2}\,\mathbf{t}^{\mathsf{T}}\mathbb{J}(\boldsymbol{\theta}_X)\,\mathbf{t} + o(\|\mathbf{t}\|^2), \;\; \textit{as } \mathbf{t} \to \mathbf{0}_p;$$

(ii) $\mathsf{var}\left(R(\boldsymbol{X}_1;\mathbf{t})\right) = o(\|\mathbf{t}\|^2)$ *as* $\mathbf{t} \to \mathbf{0}_p$*;*

(iii) there exists a finite variance matrix $\mathbb{\Sigma}(\boldsymbol{\theta}_X) = \mathsf{var}\left(\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\right)$.

Then

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = -\left[\mathbb{J}(\boldsymbol{\theta}_X)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) + o_P(1),$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\Big(\mathbf{0}_p, \left[\mathbb{J}(\boldsymbol{\theta}_X)\right]^{-1}\mathbb{\Sigma}(\boldsymbol{\theta}_X)\left[\mathbb{J}(\boldsymbol{\theta}_X)\right]^{-1}\Big).$$

Proof. See the proof of Theorem 2.1 of Hjort and Pollard [2011]. □

Comparison of Theorems 9 and 10

First of all note that Theorem 10 yields *the asymptotic normality of the argument of the minimum* of $\frac{1}{n}\sum_{i=1}^{n}\rho(\boldsymbol{X}_i;\boldsymbol{\theta})$. On the other hand Theorem 9 guarantees asymptotic normality only for a consistent (i.e. an appropriately chosen) root of the estimating equations (46). But in case that there are more roots to the estimating equations (46) it is generally impossible to decide which of the roots is the consistent one.

Further it is worth noting that Theorem 10 allows for $\rho(\mathbf{x};\boldsymbol{\theta})$ that are *'less differentiable'*. Note that to calculate the matrix $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$ one needs the differentiability of the function $\boldsymbol{\psi}$. On the other hand the matrix $\mathbb{J}(\boldsymbol{\theta}_X)$ can be computed as the Hessian matrix of the function $M(\boldsymbol{\theta}) = \mathsf{E}\,\rho(\boldsymbol{X}_1;\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}_X$. Thus the assumption about the smoothness of $\boldsymbol{\psi}$ (i.e. **[Z3]** and **[Z4]**) can be replaced with the assumption that function $M(\boldsymbol{\theta})$ is twice continuously differentiable on a neighbourhood of $\boldsymbol{\theta}_X$. So the lack of smoothness of $\boldsymbol{\psi}$ can be compensated with the assumptions on the distribution of $\boldsymbol{X}_1$ so that the function $M(\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$. See also the application of Theorem 10 to derive the asymptotic normality of the sample median given below.

Vaguely speaking the assumptions of Theorems 10 are milder than assumptions of Theorem 9. More formally if assumptions **[Z3]**-**[Z6]** hold then also assumptions (i) and (iii) are satisfied. Further a closer inspection of the proof of Theorem 2.1 of Hjort and Pollard [2011] shows that the remainder term $r_n(s)$ there can be handled with the help of assumptions **[Z3]** and **[Z4]** instead of assumption (ii) of Theorem 10.

On the other hand note that if the function $\boldsymbol{\psi}$ does not meet at least assumption **[Z3]** then it is not any straightforward method how to estimate the matrix $\mathbb{J}(\boldsymbol{\theta}_X)$ which is needed to estimate the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$.

Sample median

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with density $f(y)$ that is positive and continuous in a neighbourhood of median $F^{-1}(0.5)$.

It is well known (see also Lemma 5 and Remark 17 in Chapter 5) that the sample median $\widetilde{m}_n$ can be written as

$$\widetilde{m}_n = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |X_i - \theta|.$$

Thus one can view $\widetilde{m}_n$ as an M-estimator with $\rho(x;\theta) = |x - \theta|$. For theoretical reasons it is advantageous to consider

$$\rho(x;\theta) = |x - \theta| - |x|,$$

which does not require that $\mathsf{E}\ |X_1| < \infty$ in order to define $M(\theta) = \mathsf{E}\ \rho(X;\theta)$. Note that then (see also Lemma 5 in Chapter 5)

$$F^{-1}(0.5) = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\ \rho(X;\theta).$$

Now one can use Theorem 10 to derive the asymptotic distribution of $\widetilde{m}_n$. Introduce

$$\psi(x;\theta) = -\operatorname{sign}(x - \theta)$$

and note that $\psi(x;\theta) = \partial\rho(x;\theta)/\partial\theta$ for $\theta \neq x$. Further it is easy to check that

$$\mathsf{E}\ \psi(X_1; F^{-1}(0.5)) = 0.$$

Provided that also the other assumptions of Theorem 10 are satisfied it remains to calculate $\mathbb{\Sigma}(\theta_X)$ and $\mathbb{\Gamma}(\theta_X)$. As $\theta \in \mathbb{R}$ the matrix $\mathbb{\Sigma}(\theta_X)$ reduces to

$$\sigma_\psi^2 = \mathsf{var}\left(\psi(X_1; F^{-1}(0.5))\right) = 1.$$

Further as $M(\theta) = -\mathsf{E}\ \int_0^\theta \operatorname{sign}(X_1 - t)\,\mathrm{d}t$ one can interchange the derivative and the integral to get

$$\frac{\partial M(\theta)}{\partial \theta} = -\mathsf{E}\ \left[\operatorname{sign}(X_1 - \theta)\right] = -\mathsf{P}(X_1 > \theta) + \mathsf{P}(X_1 < \theta) = 2\,F(\theta) - 1.$$

This further implies that (the matrix) $\mathbb{J}(\theta_X)$ reduces to

$$\gamma = \frac{\partial^2 M(\theta)}{\partial \theta^2}\Big|_{\theta = F^{-1}(0.5)} = 2\,f\big(F^{-1}(0.5)\big).$$

Finally one gets

$$\sqrt{n}\left(\widetilde{m}_n - F^{-1}(0.5)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\Big(0, \frac{1}{4\,f^2\big(F^{-1}(0.5)\big)}\Big).$$

Note also to estimate the asymptotic variance of $\widetilde{m}_n$ one needs to estimate the quantity $f\big(F^{-1}(0.5)\big)$ which is far from being straightforward.*

Literature: Hjort and Pollard [2011] Section 2A.

4 M-estimators and Z-estimators in robust statistics†

In statistics the word 'robust' has basically two meanings.

(i) We say that a procedure is robust, if it stays (approximately/asymptotically) valid even when some of the assumptions (under which the procedure is derived) are not satisfied. For instance the standard ANOVA F-statistic is robust against the violation of the normality of the observations provided that the variances of all the observations are the same (and finite).

(ii) People interested in robust statistics say that a procedure is robust, if it is not 'too much' influenced by the outlying observations. In what follows we will concentrate on this meaning of the robustness.

One of the standard measures of robustness is the **breakdown point**. Vaguely speaking the breakdown point of an estimator is the smallest percentage of observations that one has to change so that the estimator produces a nonsense value (e.g. $\pm\infty$ for location or regression estimator; 0 or $+\infty$ when estimating the scale).

The end of class 14 (9. 4. 2024)

Let $\widehat{\boldsymbol{\theta}}_n$ be an M- or Z-estimator of a parameter $\boldsymbol{\theta}_X$. Note that thanks to Theorems 9 or 10 (under appropriate assumptions) one has the following representation

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X = \frac{1}{n}\sum_{i=1}^{n} IF(\boldsymbol{X}_i) + o_P\big(\tfrac{1}{\sqrt{n}}\big),$$

where $IF(\mathbf{x}) = -\mathbb{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}_X)$ is called *the influence function*. Thus if one can ignore the remainder term $o_P\big(\frac{1}{\sqrt{n}}\big)$, then changing $\boldsymbol{X}_i$ to $\boldsymbol{X}_i + \boldsymbol{\Delta}$ results that the estimates $\widehat{\boldsymbol{\theta}}_n$ changes (approximately) by

$$\frac{1}{n}\big[IF(\boldsymbol{X}_i + \boldsymbol{\Delta}) - IF(\boldsymbol{X}_i)\big].$$

Thus provided that $IF(\mathbf{x})$ is bounded then also this change is bounded (and of order $O(\frac{1}{n})$).

Note that the above reasoning was not completely correct as the term $o_P\big(\frac{1}{\sqrt{n}}\big)$ was ignored. Nevertheless it can be proved that (under some mild assumptions excluding 'singular' cases) if

* This can be estimated with the help of kernel smoothing methods or circumvented with the help of boostrap methods. Both methods are included in the course NMST545 Mathematical Statistics 4. † *Robustní statistika*

the function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ is bounded then the breakdown point of the associated $M(Z)$-estimator is $\frac{1}{2}$.

4.1 Robust estimation of location*

Suppose that we observe a random sample $X_1, \ldots, X_n$ from a distribution F and we are interested in characterising the location.

Note that for the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ it is sufficient to change only one observation to get an arbitrary value of $\overline{X}_n$.

On the other hand when considering the sample median $\widetilde{m}_n = \widehat{F}_n^{-1}(0.5)$ then one needs to change at least half of the observations so that one can for instance change the estimator to $\pm\infty$.

When deciding between a sample mean and a sample median one has to take into consideration that if the distribution F is not symmetric then $\overline{X}_n$ and $\widetilde{m}_n$ estimate different quantities. But when one can hope that the distribution F is symmetric, then both $\overline{X}_n$ and $\widetilde{m}_n$ estimate the centre of the symmetry and one can be interested which of the estimators is more appropriate. By the maximum likelihood theory we know that $\overline{X}_n$ is efficient if F is normal while $\widetilde{m}_n$ is asymptotically efficient if F is doubly exponential (i.e. it has a density $f(x) = \frac{1}{2\sigma}\exp\{-\frac{|x-\theta|}{\sigma}\}$).

In robust statistics it is usually assumed that most of our observations follow normal distributions but there are some outlying values. This can be formalised by assuming that the distribution function F of each of the observations satisfies

$$F(x) = (1-\eta)\,\Phi\big(\tfrac{x-\mu}{\sigma}\big) + \eta\, G(x), \tag{63}$$

where η is usually interpreted as probability of having an outlying observation and G is a distribution (hopefully symmetric around μ) of outlying observations. It was found that if η is 'small' then using sample median is too pessimistic (and inefficient). We will mention here several alternative options.

Before we proceed note that both the sample mean $\overline{X}_n$ and the sample median $\widetilde{m}_n$ can be viewed as M-estimators as

$$\overline{X}_n = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n (X_i-\theta)^2 \quad \text{and} \quad \widetilde{m}_n = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n |X_i-\theta|. \tag{64}$$

Huber estimator

This estimator is defined as

$$\widehat{\theta}_n^{(H)} = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \rho_H(X_i-\theta),$$

* *Robustní odhad polohy*

where

$$\rho_H(x) = \begin{cases} \frac{x^2}{2}, & |x| \le k, \\ k \cdot \left(|x| - \frac{k}{2}\right), & |x| > k \end{cases} \tag{65}$$

and k is a given constant. Note that the 'score function' $\psi_H(x) = \rho'_H(x)$ of the estimator is

$$\psi_H(x) = \rho'_H(x) = \begin{cases} x, & |x| \le k, \\ k \cdot \operatorname{sign}(x), & |x| > k. \end{cases} \tag{66}$$

Thus one can see that for $x \in (-k, k)$ the function ψ_H corresponds to a score function of a sample mean (which is $\psi(x) = x$) while for $x \in (-\infty, k) \cup (k, \infty)$ it corresponds to a score function of a sample median (which is $\psi(x) = \operatorname{sign}(x)$). Thus Huber estimator presents a compromise between a sample mean and a sample median. So it is not surprising that $\widehat{\theta}_n^{(H)}$ is usually a value between the sample median and the sample mean.

When using Huber estimator one has to keep in mind that the identified parameter is

$$\theta^{(H)} = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\ \rho_H(X_1 - \theta).$$

Thus if the distribution F **is not symmetric** then $\mathsf{E}\ X_1$ generally does not coincide with $F^{-1}(0.5)$ and $\theta^{(H)}$ lies between $\mathsf{E}\ X_1$ and $F^{-1}(0.5)$.

On the other hand if the distribution F **is symmetric**, then $\theta^{(H)}$ coincides with the centre of symmetry, i.e. with $F^{-1}(0.5)$ (the median of F) and also with $\mathsf{E}\ X_1$, if the expectation exists. It was observed that for the contamination model (63) with G symmetric, Huber estimator usually performs better than the sample mean as well as the sample median. This can be proved analytically by showing that for $\eta > 0$ and G heavy tailed, then usually

$$\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right) < \min\left\{\mathsf{var}(\overline{X}_n), \mathsf{avar}(\widetilde{m}_n)\right\},$$

where the asymptotic variance $\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right)$ is derived in Example 42.

The nice thing about Huber estimator is that its loss function $\rho(x;\theta) = \rho_H(x-\theta)$ is convex (in θ) thus $\widehat{\theta}_n^{(H)}$ is not too difficult to calculate and with the help of Theorem 10 one can derive its asymptotic distribution (see also Example 42).

Example 42. With the help of Theorem 10 one can show that (under appropriate regularity assumptions)

$$\sqrt{n}\left(\widehat{\theta}_n^{(H)} - \theta^{(H)}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\left(0, \tfrac{\sigma_\psi^2}{\gamma^2}\right),$$

where

$$\gamma = \frac{\partial^2\ \mathsf{E}\ \rho_H(X_1 - \theta)}{\partial\theta^2}\Big|_{\theta=\theta^{(H)}} = F(\theta^{(H)} + k) - F(\theta^{(H)} - k)$$

and

$$\sigma_\psi^2 = \mathsf{var}\left(\psi_H(X_1 - \theta^{(H)})\right) = \int_{\theta^{(H)}-k}^{\theta^{(H)}+k} (x - \theta^{(H)})^2\, \mathrm{d}F(x) + k^2\big(1 - F(\theta^{(H)} + k) + F(\theta^{(H)} - k)\big).$$

Thus $\mathsf{avar}(\widehat{\theta}_n^{(H)}) = \frac{\sigma_\psi^2}{n\gamma^2}$.

The choice of the constant k is usually done as follows. Suppose that $X_1, \ldots, X_n$ follows $\mathsf{N}(0,1)$. Then one takes the smallest k such that

The end of class 15 (12. 4. 2024)

$$\frac{\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right)}{\mathsf{var}\left(\overline{X}_n\right)} \leq 1 + \delta,$$

where δ stands for the efficiency loss of Huber estimator under normal distributions. For instance the common choices are $\delta = 0.05$ or $\delta = 0.1$ which corresponds approximately to $k = 1.37$ or $k = 1.03$.

Other robust M/Z-estimators of location

The other most common M/Z-estimators are the following.

(i) **Cauchy-pseudolikelihood**: $\rho(x;\theta) = \log(1 + (x-\theta)^2)$. The problem is that this function is not convex in θ and the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{2\,(X_i - \widehat{\theta}_n)}{1 + (X_i - \widehat{\theta}_n)^2}}_{\psi(X_i;\widehat{\theta}_n)} \stackrel{!}{=} 0$$

has usually more roots.

(ii) **Tukey's biweight:**

$$\psi(x) = \begin{cases} x\left(1 - \frac{x^2}{k^2}\right)^2, & |x| \leq k, \\ 0, & |x| > k. \end{cases}$$

But also here the corresponding loss function ρ ($\psi = \rho'$) is not convex.

4.2 Robust studentized M/Z-estimators of location

The problem is that the M/Z-estimators presented above (except for the sample mean and the sample median) are not scale equivariant (i.e. $\widehat{\theta}_n(c\,\boldsymbol{X}) \neq c\,\widehat{\theta}_n(\boldsymbol{X})$ for each $c \in \mathbb{R}$). That is why in practice M/Z-estimators are usually defined as

$$\widehat{\theta}_n = \operatorname*{arg\,min}_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \rho\big(\tfrac{X_i-\theta}{S_n}\big), \text{ or as } \sum_{i=1}^{n} \psi\big(\tfrac{X_i-\widehat{\theta}_n}{S_n}\big) \stackrel{!}{=} 0,$$

where S_n is an appropriate estimator of scale*, which satisfies $S_n(c\,\boldsymbol{X}) = |c|\,S_n(\boldsymbol{X})$ for each $c \in \mathbb{R}$. The most common estimators of scale are as follows.

* *odhad měřítka*

Sample standard deviation

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Note that in robust statistics S_n is rather rarely used as it is not robust (i.e. it is sensitive to outlying observations).

Interquartile range*

$$S_n = IQR = \widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25),$$

where $\widehat{F}_n$ is the empirical distribution function $\big($i.e. $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i \le x\}\big)$. Some people prefer to use

$$\widetilde{S}_n = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

as it is desired that $\widetilde{S}_n$ estimates σ, when $X_1, \ldots, X_n$ is a random sample from $\mathsf{N}(\mu, \sigma^2)$.

Note that the breakdown point of interquartile range is 0.25.

Median absolute deviation†

This measure is given as the median absolute deviation from the median, i.e.

$$MAD = \mathrm{med}_{1\le i\le n}\{|X_i - \widehat{F}_n^{-1}(0.5)|\},$$

or its modification

$$\widetilde{MAD} = \frac{MAD}{\Phi^{-1}(0.75)},$$

so that it estimates σ for random samples from $\mathsf{N}(\mu, \sigma^2)$.

Note that the breakdown point of this estimator is 0.50.

Remark 16. Note that due to the studentization the functions $\rho(x;\theta) = \rho\big(\frac{x-\theta}{S_n}\big)$ and $\psi(x;\theta) = \psi\big(\frac{x-\theta}{S_n}\big)$ (when viewed as functions of x and θ) are random. Thus one can use neither Theorem 9 nor Theorem 10 to derive the asymptotic distribution of studentized M/Z-estimators.

Nevertheless, if $S_n \xrightarrow[n\to\infty]{P} S(F)$ and the distribution F is **symmetric**, then (under some regularity assumptions) it can be shown that the asymptotic distribution of studentized Z/M-estimators is the same as the asymptotic distribution of M/Z-estimators with $\rho(x;\theta) = \rho\big(\frac{x-\theta}{S(F)}\big)$ and $\psi(x;\theta) = \psi\big(\frac{x-\theta}{S(F)}\big)$ for which one can (usually) use either Theorem 9 or Theorem 10.

---

* *mezikvartilové rozpětí* † *mediánová absolutní odchylka*

4.3 Robust estimation in linear models

Suppose we observe independent random vectors $\binom{Y_1}{\boldsymbol{X}_1}, \ldots, \binom{Y_n}{\boldsymbol{X}_n}$ each of them having the same distribution as the generic random vector $\binom{Y}{\boldsymbol{X}}$.

4.3.1 The least squares method

This method results in the estimator

$$\widehat{\boldsymbol{\beta}}_n^{(LS)} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^n \left(Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i Y_i\right).$$

Note that if $\boldsymbol{X}_{ik} \neq 0$ then by changing Y_i one can arrive at any arbitrary value of $\widehat{\beta}_{nk}$.

From Example 40 we know that the method of the least squares identifies the parameter

$$\boldsymbol{\beta}_X^{(LS)} = \left[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]^{-1}\mathsf{E}\ \boldsymbol{X}Y$$

and it holds that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{(LS)} - \boldsymbol{\beta}_X^{(LS)}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(0,\mathbb{V}),\ \text{where } \mathbb{V} = \left[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]^{-1}\left[\mathsf{E}\ \sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]\left[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]^{-1},$$

with $\sigma^2(\boldsymbol{X}_1) = \mathsf{E}\left[(Y_1 - \boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}_X)^2|\boldsymbol{X}_1\right]$. Further provided $\mathsf{E}\left[Y|\boldsymbol{X}\right] = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X^{(LS)} = \boldsymbol{\beta}_0$ and $\sigma^2(\boldsymbol{X}) = \mathsf{var}(Y|\boldsymbol{X})$.

Suppose now that the first component of $\boldsymbol{X}_i$ is 1 (i.e. the model includes an intercept) and denote by $\widetilde{\boldsymbol{X}}_i$ the remaining components of $\boldsymbol{X}_i$. That is $\boldsymbol{X}_i = \binom{1}{\widetilde{\boldsymbol{X}}_i}$. Further suppose that the following model holds

$$Y = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + \varepsilon,\ \text{where}\quad \varepsilon \perp \widetilde{\boldsymbol{X}}. \tag{67}$$

Then $\mathsf{E}\left[Y \mid \boldsymbol{X}\right] = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\widetilde{\boldsymbol{X}} + \mathsf{E}\ \varepsilon$ and the method of the least squares identifies the parameter

$$\boldsymbol{\beta}_X^{(LS)} = \begin{pmatrix} \beta_0 + \mathsf{E}\ \varepsilon \\ \boldsymbol{\beta} \end{pmatrix}. \tag{68}$$

Further the asymptotic variance matrix $\mathbb{V}$ simplifies to

$$\mathbb{V} = \sigma^2\left(\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1},\quad \text{where}\quad \sigma^2 = \mathsf{var}(\varepsilon). \tag{69}$$

The expected end of class 16 (19. 4. 2024)

4.3.2 Method of the least absolute deviation*

This method is usually considered as a robust alternative to the least squares methods. The estimate of the regression parameter is given by

$$\widehat{\boldsymbol{\beta}}_n^{(LAD)} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n |Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}|,$$

* *Metoda nejmenších absolutních odchylek, mediánová regrese*

As we will see later (see Chapter 5) the LAD method models $\mathsf{med}[Y \,|\, \boldsymbol{X}] = F^{-1}_{Y|\boldsymbol{X}}(0.5)$ as $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$. So if indeed $\mathsf{med}[Y \,|\, \boldsymbol{X}] = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X^{(LAD)} = \boldsymbol{\beta}_0$.

The *asymptotic distribution* of $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$ can be heuristically derived by Theorem 10 as follows. The score function is given by

$$\boldsymbol{\psi}(\mathbf{x}, y; \mathbf{b}) = -\operatorname{sign}(y - \mathbf{x}^\mathsf{T}\mathbf{b})\,\mathbf{x}.$$

Now put $M(\mathbf{b}) = \mathsf{E}\,\big[|Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}| - |Y|\big]$, where the random vector $\binom{Y}{\boldsymbol{X}}$ has the same distribution as $\binom{Y_i}{\boldsymbol{X}_i}$. Then

$$\begin{aligned}\frac{\partial M(\mathbf{b})}{\partial \mathbf{b}} &= \mathsf{E}\,\big[\operatorname{sign}(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b})\,(-\boldsymbol{X})\big] = -\,\mathsf{E}\,\boldsymbol{X}\big[\mathbb{I}\{Y > \boldsymbol{X}^\mathsf{T}\mathbf{b}\} - \mathbb{I}\{Y < \boldsymbol{X}^\mathsf{T}\mathbf{b}\}\big] \\ &= -\,\mathsf{E}\,\boldsymbol{X}\big[1 - 2F_{Y|\boldsymbol{X}}\big(\boldsymbol{X}^\mathsf{T}\mathbf{b}\big)\big].\end{aligned}$$

Thus

$$\frac{\partial^2 M(\mathbf{b})}{\partial \mathbf{b}\,\partial \mathbf{b}^\mathsf{T}} = 2\,\mathsf{E}\,\boldsymbol{X} f_{Y|\boldsymbol{X}}\big(\boldsymbol{X}^\mathsf{T}\mathbf{b}\big)\boldsymbol{X}^\mathsf{T},$$

which finally implies that

$$\mathbb{J}(\boldsymbol{\beta}_X^{(LAD)}) = \frac{\partial^2 M(\mathbf{b})}{\partial \mathbf{b}\,\partial \mathbf{b}^\mathsf{T}}\Big|_{\mathbf{b}=\boldsymbol{\beta}_X^{(LAD)}} = 2\,\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(LAD)}\big)\big] = 2\,\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(0.5)\big)\big].$$

Further as

$$\mathbb{\Sigma}(\boldsymbol{\beta}_X^{(LAD)}) = \mathsf{var}\,\big(\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}_X^{(LAD)})\big) = \mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T},$$

one gets that under appropriate regularity assumptions

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n^{(LAD)} - \boldsymbol{\beta}_X^{(LAD)}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}),$$

where

$$\mathbb{V} = \frac{1}{4}\Big(\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(0.5)\big)\big]\Big)^{-1}\,\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big(\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(0.5)\big)\big]\Big)^{-1}.$$

Note that if model (67) holds, then $\mathsf{med}(Y_1 \,|\, \boldsymbol{X}_1) = \beta_0 + \widetilde{\boldsymbol{X}}_1^\mathsf{T}\boldsymbol{\beta} + F_\varepsilon^{-1}(0.5)$, where F_ε^{-1} is the quantile function of ε_1 and thus

$$\boldsymbol{\beta}_X^{(LAD)} = \begin{pmatrix} \beta_0 + F_\varepsilon^{-1}(0.5) \\ \boldsymbol{\beta} \end{pmatrix}.$$

Thus when compared with the method of the least squares (68) one can see, that if model (67) holds then both methods identify the same slope parameter $\boldsymbol{\beta}$. The only difference is in intercept.

Further if model (67) holds then

$$f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(0.5)\big) = f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big),$$

which implies that

$$\mathbb{\Gamma}(\boldsymbol{\beta}_X^{(LAD)}) = 2\, f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\ \mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$$

and

$$\mathbb{V} = \frac{1}{4\big[f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\big]^2}\ \big(\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\big)^{-1}. \tag{70}$$

Now when one compares (69) with (70), one can see that the least absolute deviation method is favourable if

$$\frac{1}{\big[4f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\big]^2} < \mathsf{var}(\varepsilon).$$

Regarding the robustness of the least absolute deviation estimator note that in this special situation (i.e. if model (67) holds) $Y_i - \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_X^{(LAD)} = \varepsilon_i - F_\varepsilon^{-1}(0.5)$ and the asymptotic representation (50) of $\widehat{\boldsymbol{\beta}}_n$ implies

$$\widehat{\boldsymbol{\beta}}_n^{(LAD)} - \boldsymbol{\beta}_X^{(LAD)} = \frac{1}{n}\sum_{i=1}^n \Big(\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\Big)^{-1}\boldsymbol{X}_i\,\frac{\operatorname{sign}\big(\varepsilon_i - F_\varepsilon^{-1}(0.5)\big)}{2f_\varepsilon(F_\varepsilon^{-1}(0.5))} + o_P\big(\tfrac{1}{\sqrt{n}}\big).$$

Thus one can expect that the change of Y_i (or equivalently the change of ε_i) has only a bounded effect on $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$. On the other hand note that the change of $\boldsymbol{X}_i$ has an unbounded effect on $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$. Thus LAD method is robust with respect to the response but not with respect to the covariates.

The end of class 17 (19. 4. 2024)

4.3.3 Huber estimator of regression

Analogously as Huber estimator of location is a compromise between a sample mean and a sample median, Huber estimator of regression is a compromise between LS and LAD. Put

$$\widehat{\boldsymbol{\beta}}_n^{(H)} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho_H\big(Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}\big),$$

where ρ_H is defined in (65). Generally, it is difficult to interpret what is being modelled with Huber estimator of regression (it is something between $\mathsf{E}\,(Y \mid \boldsymbol{X})$ and $\mathsf{med}(Y \mid \boldsymbol{X})$). Note that it identifies

$$\boldsymbol{\beta}_X^{(H)} = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \mathsf{E}\ \rho_H(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}).$$

Equivalently $\boldsymbol{\beta}_X^{(H)}$ solves

$$\mathsf{E}\ \big[\psi_H\big(Y - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)}\big)\,\boldsymbol{X}\big] \stackrel{!}{=} \mathbf{0}_p,$$

where ψ_H is defined in (66).

Analogously as in Example 42 one can derive that under appropriate assumptions

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n^{(H)} - \boldsymbol{\beta}_X^{(H)}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}), \quad \text{with} \quad \mathbb{V} = \mathbb{\Gamma}^{-1}(\boldsymbol{\beta}_X^{(H)})\,\mathbb{\Sigma}(\boldsymbol{\beta}_X^{(H)})\,\mathbb{\Gamma}^{-1}(\boldsymbol{\beta}_X^{(H)}),$$

where

$$\mathbb{\Gamma}\big(\boldsymbol{\beta}_X^{(H)}\big) = \mathsf{E}_{\boldsymbol{X}}\left[F_{Y|\boldsymbol{X}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)}+k) - F_{Y|\boldsymbol{X}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)}-k)\right]\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$$

and

$$\mathbb{\Sigma}\big(\boldsymbol{\beta}_X^{(H)}\big) = \mathsf{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\,\mathsf{var}\left(\psi(Y-\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)})|\boldsymbol{X}\right)\right].$$

If model (67) holds then $\boldsymbol{\beta}_X^{(H)} = \binom{\beta_{X0}^{(H)}}{\widetilde{\boldsymbol{\beta}}_X^{(H)}}$ solves

$$\mathsf{E}\left[\psi_H\big(\beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + \varepsilon - \beta_{X0}^{(H)} - \widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{\beta}}_X^{(H)}\big)\,\boldsymbol{X}\right] \stackrel{!}{=} \mathbf{0}_p.$$

Thus $\boldsymbol{\beta}_X$ identifies the following parameter

$$\boldsymbol{\beta}_X^{(H)} = \begin{pmatrix} \beta_0 + \theta^{(H)} \\ \boldsymbol{\beta} \end{pmatrix},$$

where $\theta^{(H)}$ solves $\mathsf{E}\,\psi_H(\varepsilon - \theta^{(H)}) \stackrel{!}{=} 0$. So if model (67) holds then the interpretation of the regression slope coefficient ($\boldsymbol{\beta}$) is the same for each of the methods described above (LS, LAD, Huber regression).

Further the asymptotic variance matrix simplifies to

$$\mathbb{V} = \frac{\sigma_\psi^2}{\gamma^2}\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}, \tag{71}$$

with

$$\gamma = F_\varepsilon(\theta^{(H)}+k) - F_\varepsilon(\theta^{(H)}-k)$$

and

$$\sigma_\psi^2 = \int_{\theta^{(H)}-k}^{\theta^{(H)}+k} (x-\theta^{(H)})^2\,\mathrm{d}F_\varepsilon(x) + k^2\big(1 - F_\varepsilon(\theta^{(H)}+k) + F_\varepsilon(\theta^{(H)}-k)\big).$$

Using (69), (70) and (71) one sees that to compare the efficiency of the estimators $\widehat{\boldsymbol{\beta}}_n^{(LS)}$, $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$ and $\widehat{\boldsymbol{\beta}}_n^{(H)}$ it is sufficient to compare $\mathsf{var}(\varepsilon)$, $\frac{1}{4f_\varepsilon^2(F_\varepsilon^{-1}(0.5))}$ and $\frac{\sigma_\psi^2}{\gamma^2}$.

Regarding the robustness properties the influence function is given by

$$IF(\mathbf{x}, y) = \left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}\frac{1}{\gamma}\,\psi_H(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}_X^{(H)})\,\mathbf{x},$$

thus the estimator is robust in response but not in the covariate.

4.3.4 Studentized Huber estimator of regression

Analogously as in Chapter 4.2 in practice *the studentized Huber estimator* is usually used. This estimator is defined as

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min}\,\frac{1}{n}\sum_{i=1}^n \rho_H\left(\frac{Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}}{S_n}\right),$$

where S_n is an estimator of scale of ε_i. For instance one can take MAD or IQR calculated from the residuals from LAD regression $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n^{(LAD)}$.

Inference:

- With the help of Theorem 10 one can show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ of the (non-Studentized) Huber estimator.
- If model (67) holds, then it can be shown, that the estimate of the scale influences only the asymptotic distribution of the estimate of the intercept and not of the slope.

Literature: Maronna et al. [2006] Chapters 2.1-2.2 and Chapters 4.1-4.4.

5 Quantile regression*

Generally speaking, while the least squares method aims at estimating (modelling) a conditional expectation, quantile regression aims at estimating (modelling) a conditional quantile. This is of interest if the covariate may have different effects on different quantiles of the response.

Applications of the quantile regression can be found in medicine (e.g. constructing reference charts), finance (e.g. estimating value at risk), economics (e.g. wage and income studies, modelling household electricity demand) and environment modelling (e.g. modelling flood height).

5.1 Identification of quantiles

For a given $\tau \in (0,1)$ consider the following loss function

$$\rho_\tau(x) = \tau\, x\, \mathbb{I}\{x > 0\} + (1-\tau)\,(-x)\,\mathbb{I}\{x \leq 0\}.$$

Note that for $x \neq 0$ one gets

$$\psi_\tau(x) = \rho'_\tau(x) = \tau\, \mathbb{I}\{x > 0\} - (1-\tau)\,\mathbb{I}\{x < 0\}.$$

For $x = 0$ put $\psi_\tau(0) = 0$.

Lemma 5. *Let the random variable X have a cumulative distribution function F. Then*

$$F^{-1}(\tau) = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\,\big[\rho_\tau(X-\theta) - \rho_\tau(X)\big]. \tag{72}$$

* *Kvantilová regrese.*

Proof. Put $M(\theta) = \mathsf{E}\,\big[\rho_\tau(X-\theta) - \rho_\tau(X)\big]$. One can calculate

$$\begin{aligned} M(\theta) &= -\,\mathsf{E}\int_0^\theta \psi_\tau(X-t)\,\mathrm{d}t = -\int_0^\theta \mathsf{E}\;\psi_\tau(X-t)\,\mathrm{d}t \\ &= -\int_0^\theta \tau\;\mathsf{P}(X>t) - (1-\tau)\,\mathsf{P}(X<t)\,\mathrm{d}t. \\ &= -\int_0^\theta \tau - \tau\,F(t) - (1-\tau)F(t)\,\mathrm{d}t. \\ &= -\tau\,\theta + \int_0^\theta F(t)\,\mathrm{d}t. \end{aligned}$$

Now for each $\theta < F^{-1}(\tau)$

$$\begin{aligned} M'(\theta_-) &= -\tau + F(\theta_-) \le -\tau + F(\theta) < 0, \\ M'(\theta_+) &= -\tau + F(\theta_+) = -\tau + F(\theta) < 0. \end{aligned}$$

As the function $M(\theta)$ is continuous, this implies that $M(\theta)$ is decreasing on $\big(-\infty, F^{-1}(\tau)\big)$.

Analogously for $\theta > F^{-1}(\tau)$

$$\begin{aligned} M'(\theta_-) &= -\tau + F(\theta_-) \ge -\tau + F\big(F^{-1}(\tau)\big) \ge 0, \\ M'(\theta_+) &= -\tau + F(\theta_+) \ge -\tau + F\big(F^{-1}(\tau)\big) \ge 0. \end{aligned}$$

Thus the function $M(\theta)$ is non-decreasing on $\big(F^{-1}(\tau), +\infty\big)$. This further implies that $F^{-1}(\tau)$ is the point of the global minimum of the function $M(\theta)$. □

Remark 17. Suppose we observe a random sample $X_1, \dots, X_n$. Let $\widehat{F}_n$ be the corresponding empirical distribution function. Then by

$$\frac{1}{n}\sum_{i=1}^n \rho_\tau(X_i-\theta) = \mathsf{E}_{\,\widehat{F}_n}\,\rho_\tau(Z-\theta),$$

where the random variable Z has the distribution given by the empirical distribution function $\widehat{F}_n$ and $\mathsf{E}_{\,\widehat{F}_n}$ stands for the expectation with respect to this distribution.

Thus by Lemma 5

$$\widehat{F}_n^{-1}(\tau) = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \rho_\tau(X_i-\theta).$$

Note that for $\tau = 0.5$ one gets the characterization of the sample median as in (64).

Further note that from the proof of Lemma 5 it follows that the $\arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \rho_\tau(X_i-\theta)$ is not unique if there exists a root of the function $-\tau + \widehat{F}_n(\theta)$. This happens if $n\,\tau = i_0 \in \mathbb{N}$ and $X_{(i_0)} < X_{(i_0+1)}$. Then $M(\theta)$ is minimised by any value from the interval $\big[X_{(i_0)}, X_{(i_0+1)}\big]$. In this situation $\widehat{F}_n^{-1}(\tau) = X_{(i_0)}$ is the left point of this interval.

The end of class 18 (23. 4. 2024)

5.2 Regression quantiles*

Suppose that one observes independent and identically distributed random vectors

$$\begin{pmatrix} \boldsymbol{X}_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{X}_n \\ Y_n \end{pmatrix}$$

being distributed as the generic vector $\binom{\boldsymbol{X}}{Y}$.

The τ-th regression quantile is defined as

$$\widehat{\boldsymbol{\beta}}_n(\tau) = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}).$$

At the population level the regression quantile identifies the parameter

$$\boldsymbol{\beta}_X(\tau) = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \mathsf{E}\left[\rho_\tau(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}) - \rho_\tau(Y)\right].$$

Note that thanks to (72)

$$\begin{aligned}\mathsf{E}\left[\rho_\tau(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}) - \rho_\tau(Y)\right] &= \mathsf{E}\left\{\mathsf{E}\left[\rho_\tau(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}) - \rho_\tau(Y)\,|\,\boldsymbol{X}\right]\right\} \\ &\geq \mathsf{E}\left\{\mathsf{E}\left[\rho_\tau\big(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau)\big) - \rho_\tau(Y)\,|\,\boldsymbol{X}\right]\right\} = \mathsf{E}\left[\rho_\tau\big(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau)\big) - \rho_\tau(Y)\right],\end{aligned}$$

where $F^{-1}_{Y|\boldsymbol{X}}(\tau)$ is the τ-th conditional quantile of Y given $\boldsymbol{X}$. Thus if the model for $F^{-1}_{Y|\boldsymbol{X}}(\tau)$ is correctly specified, that is $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X(\tau) = \boldsymbol{\beta}_0$.

Often in applications we have $\boldsymbol{X} = \big(1, \widetilde{\boldsymbol{X}}^\mathsf{T}\big)^\mathsf{T}$ and assume that

$$Y = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + \varepsilon, \text{ where } \quad \varepsilon \perp \widetilde{\boldsymbol{X}}. \tag{73}$$

Then $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + F^{-1}_\varepsilon(\tau)$, where $F^{-1}_\varepsilon(\tau)$ is the τ-th quantile of the random error ε. Thus provided model (73) holds

$$\boldsymbol{\beta}_X(\tau) = \begin{pmatrix} \beta_0 + F^{-1}_\varepsilon(\tau) \\ \boldsymbol{\beta} \end{pmatrix}.$$

Thus if model (73) holds, then for $\tau_1 \neq \tau_2$ the regression quantiles $\boldsymbol{\beta}_X(\tau_1)$ and $\boldsymbol{\beta}_X(\tau_2)$ differ only in the intercepts. That is the effect of the covariate is the same for all quantiles of the response. But this is not true in general. In fact the regression quantiles are interesting in situations where the effect of the covariate can be different for different quantiles of the response.

As also illustrated by the following simple examples, the regression quantiles gives us a more detailed idea about the effect of the covariate on the response. This can be of interest on its own or as a check that we do not simplify the situation too much by considering only the effect of the covariate on the conditional expectation.

---

* *Regresní kvantily*

Example 43. To illustrate consider one-dimensional covariate X_i which is generated from the uniform distribution on the interval $(0, 1)$ and the error term ε_i which has an exponential distribution with mean 1 and which is independent of X_i. Further consider the following two models

- *The homoscedastic model* given by

$$Y_i = 1 + 2\,X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

- *The heteroscedastic model* given by

$$Y_i = 1 + 2\,X_i + 2\,X_i\,\varepsilon_i, \quad i = 1, \ldots, n.$$

On Figure 43 one can find a random sample of size 1 000 from these models. The solid lines represent the fitted regression quantiles for $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ assuming that the conditional quantile is in the simple linear form

$$F^{-1}_{Y|X}(\tau) = \beta_1(\tau) + \beta_2(\tau)\,X.$$

The standard least square estimator is included for the reason of comparison.

Note that in the homoscedastic model all the fitted lines are approximately parallel. This is in agreement with the above finding that in the 'strict linear model' (73) the slope of the (theoretical) regression quantiles is the same (up to the random variations that decreases as the sample size increases).

On the other hand in the heteroscedastic model the slopes differ and in this simple example we see that the effect of the covariate is stronger on larger conditional quantiles.

Homework exercise. In the homoscedastic as well as heteroscedastic model find the theoretical conditional quantile $F^{-1}_{Y|X}(\tau)$ for different values of τ and compare it with the conditional expectation $\mathsf{E}\,[Y|X]$. Compare the results with the fitted lines on Figure 43.

Example 44. Let $Y_1, \ldots, Y_{n_1}$ be a random sample with the distribution function F and $Y_{n_1+1}, \ldots, Y_{n_1+n_2}$ be a random sample from the distribution function G .

Often it is assumed that $G(x) = F(x + \mu)$ for each $x \in \mathbb{R}$. Thus alternatively we can formulate the two-sample problem as a linear regression problem with

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{74}$$

where

$$x_i = \begin{cases} 0, & i = 1, \ldots, n_1, \\ 1, & i = n_1 + 1, \ldots, n_1 + n_2, \end{cases}$$

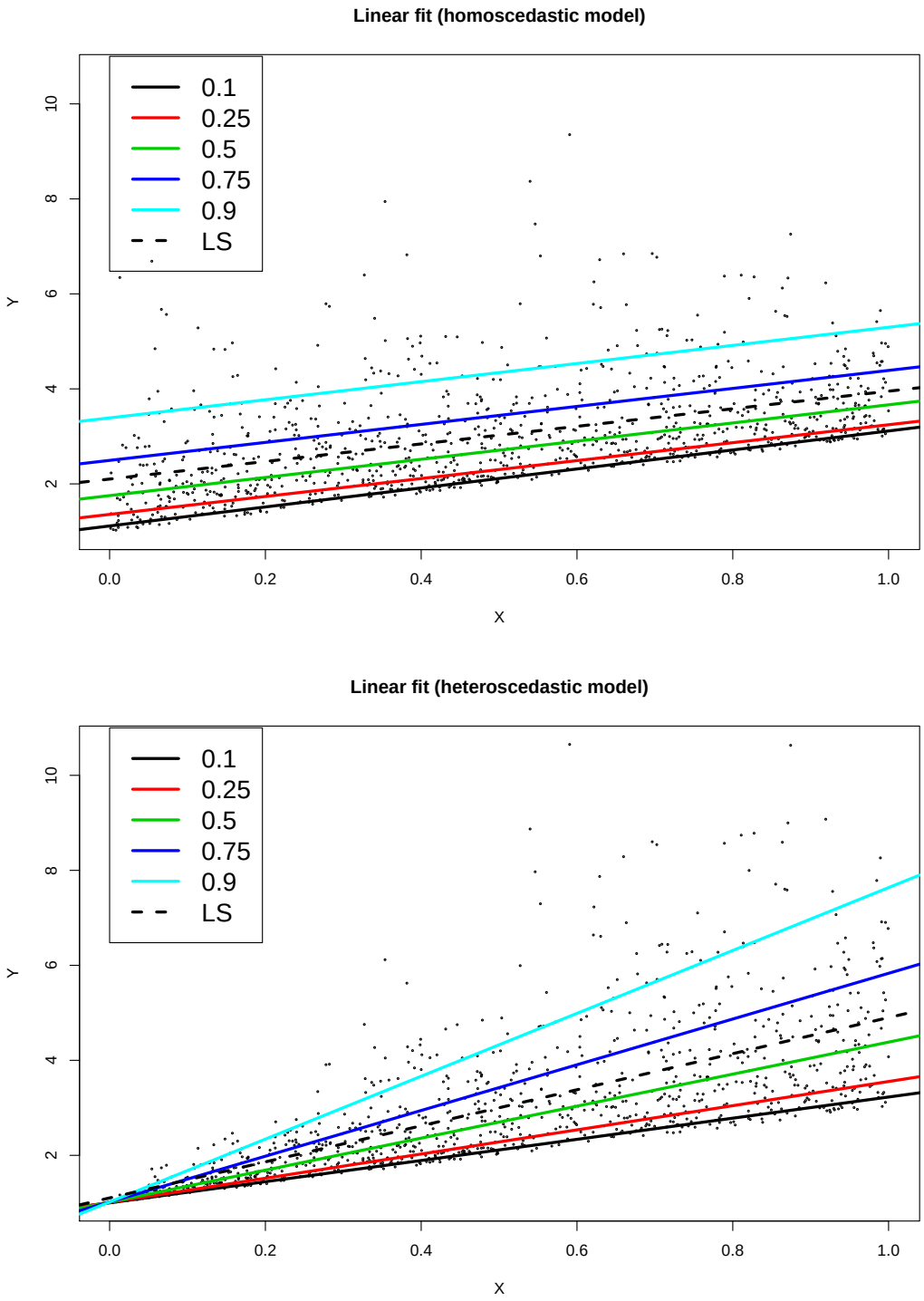

Figure 1: Fitted regression quantiles for $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ (solid lines with different colours) for homoscedastic model (the upper figure) and heteroscedastic model (the lower figure). The least squares fit is included for the reason of comparison (dashed line).

and the error term ε_i has a cumulative distribution function F. Usually we are interested in estimating β_1. By the LS method one gets

$$\widehat{\beta}_1 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} Y_i - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i \xrightarrow[n_1,n_2\to\infty]{\mathrm{P}} \underbrace{\mu_G - \mu_F}_{=:\mu} =: \beta_1^{LS},$$

where μ_F and μ_G stand for the expectation of an observation from the first and second sample respectively.

On the other hand let $n = n_1 + n_2$. Then the quantile regression yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(\tau) &= \arg\min_{b_0,b_1} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - b_0 - b_1 x_i) \\ &= \arg\min_{b_0,b_1} \frac{1}{n}\left(\sum_{i=1}^{n_1} \rho_\tau(Y_i - b_0) + \sum_{i=n_1+1}^{n_1+n_2} \rho_\tau(Y_i - b_0 - b_1)\right).\end{aligned}$$

The first sum is minimised by

$$\widehat{\beta_0}(\tau) = F_{n_1}^{-1}(\tau)$$

and the second sum by

$$\widehat{\beta_0(\tau) + \beta_1}(\tau) = G_{n_2}^{-1}(\tau).$$

Thus we get

$$\widehat{\beta}_1(\tau) = G_{n_2}^{-1}(\tau) - F_{n_1}^{-1}(\tau) \xrightarrow[n_1,n_2\to\infty]{\mathrm{P}} G^{-1}(\tau) - F^{-1}(\tau) := \beta_1(\tau).$$

Further if model (74) really holds, then $G^{-1}(\tau) = F^{-1}(\tau) + \mu$ and one gets $\beta_1(\tau) = \mu = \beta_1^{LS}$ for each $\tau \in (0,1)$.

Computing regression quantiles

The optimisation task

$$\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^{\mathsf{T}}\mathbf{b})$$

can be rewritten with the help of linear programming as minimisation of the objective function

$$\tau\sum_{i=1}^{n} r_i^+ + (1-\tau)\sum_{i=1}^{n} r_i^-,$$

subject to the following constrains

$$\begin{aligned}&\sum_{j=1}^{p} X_{ij}\, b_j + r_i^+ - r_i^- = Y_i, \qquad i = 1,\ldots,n,\\ &r_i^+ \geq 0,\ r_i^- \geq 0, \qquad i = 1,\ldots,n,\\ &b_j \in \mathbb{R}, \qquad j = 1,\ldots,p.\end{aligned}$$

Note that one can think of r_i^+ and r_i^- as the positive or negative part of the i-th residual, i.e.

$$r_i^+ = \left(Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}\right)_+, \quad r_i^- = \left(Y_i - \boldsymbol{X}_i^\mathsf{T}\mathbf{b}\right)_- .$$

This can be solved for instance with the help of *the simplex algorithm*.

5.3 Interpretation of the regression quantiles

Provided $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$ and the model is correctly specified then one can interpret $\widehat{\beta}_{nk}(\tau)$ $\left(\text{the } k\text{-th element of } \widehat{\boldsymbol{\beta}}_n(\tau)\right)$ as the estimated change of the conditional quantile of the response when the k-th element of the explanation variable increases by 1.

Intersection of the fitted regression quantiles

Note that it might happen that for a given value of the covariate $\mathbf{x}$ and given quantiles $0 < \tau_1 < \tau_2 < 1$

$$\widehat{F}^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_1) = \mathbf{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau_1) > \mathbf{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau_2) = \widehat{F}^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_2). \tag{75}$$

which is rather strange as we know that the theoretical quantiles for $\tau_1 < \tau_2$ must satisfy

$$F^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_1) \leq F^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_2).$$

Thus if one gets the inequality (75) (we also say that the regression quantiles cross) for $\mathbf{x}$ from the support of the covariate, it might indicate that the assumed linear model for the conditional quantile is not correct.

Transformed response

It is worth noting that if one models the conditional quantile of the transformed response, that is one assumes that $F^{-1}_{h(Y)|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$ for a given increasing transformation h, then

$$\tau = \mathsf{P}\left(h(Y) \leq \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} \,|\, \boldsymbol{X}\right) = \mathsf{P}\left(Y \leq h^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}) \,|\, \boldsymbol{X}\right),$$

which implies that $F^{-1}_{Y|\boldsymbol{X}}(\tau) = h^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})$. Analogously $F^{-1}_{Y|\boldsymbol{X}}(1-\tau) = h^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})$ for h decreasing. That is unlike for modelling of conditional expectation (through the least squares method), here we still have a link between $\boldsymbol{\beta}$ and the quantile of the original (not transformed) response $F^{-1}_{Y|\boldsymbol{X}}(\tau)$.

Thus from the practical point of view even if $\widehat{\boldsymbol{\beta}}_n(\tau)$ is estimated from the response-transformed data $\binom{\boldsymbol{X}_1}{h(Y_1)}, \ldots, \binom{\boldsymbol{X}_n}{h(Y_n)}$, one can still estimate the conditional quantile of the original (not transformed) data $\widehat{F}^{-1}_{Y|\boldsymbol{X}}(\tau) = h^{-1}\left(\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau)\right)$ (for h increasing). On the other hand if we estimate the conditional expectation of $\mathsf{E}\left[h(Y)|\boldsymbol{X}\right]$ as $\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n$, there is no general way how to use $\widehat{\boldsymbol{\beta}}_n$ to get an estimate of $\mathsf{E}\left[Y|\boldsymbol{X}\right]$.

A very common and popular transformation is log-transformation, i.e. $h(y) = \log y$. This results in $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \mathrm{e}^{\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}(\tau)}$ and $\mathrm{e}^{\beta_k(\tau)}$ measures how many times the conditional quantile $F^{-1}_{Y|\boldsymbol{X}}(\tau)$ changes when the k-th coordinate of the covariate is increased by adding one.

The end of class 19 (26. 4. 2024)

5.4 Inference for regression quantiles

Analogously as in Chapter 4.3.2 one can heuristically derive that under appropriate regularity assumption for fixed $\tau \in (0,1)$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}_X(\tau)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}),$$

where

$$\mathbb{V} = \Big(\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\right]\Big)^{-1} \tau(1-\tau)\,\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big(\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\right]\Big)^{-1}. \quad (76)$$

Note that if model (73) holds, then $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + F^{-1}_\varepsilon(\tau)$, where F^{-1}_ε is the quantile function of ε_1 and thus

$$\boldsymbol{\beta}_X(\tau) = \begin{pmatrix} \beta_0 + F^{-1}_\varepsilon(\tau) \\ \boldsymbol{\beta} \end{pmatrix}.$$

Further if model (73) holds then

$$f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big) = f_\varepsilon\big(F^{-1}_\varepsilon(\tau)\big),$$

which implies that

$$\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\right] = f_\varepsilon\big(F^{-1}_\varepsilon(\tau)\big)\,\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$$

and

$$\mathbb{V} = \frac{\tau(1-\tau)}{\left[f_\varepsilon\big(F^{-1}_\varepsilon(\tau)\big)\right]^2}\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}. \quad (77)$$

Estimation of asymptotic variance of $\widehat{\boldsymbol{\beta}}_n(\tau)$

Note that in general the asymptotic variance matrix (76) of $\widehat{\boldsymbol{\beta}}_n(\tau)$ is rather complicated and it is not clear how to estimate it. That is why nonparametric bootstrap is of interest.

If model (73) holds, then the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_n(\tau)$ simplifies considerably and one gets

$$\mathsf{avar}\left(\widehat{\boldsymbol{\beta}}_n(\tau)\right) = \tfrac{1}{n}\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}\frac{\tau(1-\tau)}{f^2_\varepsilon\big(F^{-1}_\varepsilon(\tau)\big)}.$$

The matrix $\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ can be estimated as $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}$. The difficulty is in estimating the sparsity function $s(\tau) = \frac{1}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)}$. In Chapter 4.10.1 of Koenker [2005] it is suggested that one can use the following estimate

$$\widehat{s}_n(\tau) = \frac{\widehat{F}^{-1}_{n\widehat{\varepsilon}}(\tau + h_n) - \widehat{F}^{-1}_{n\widehat{\varepsilon}}(\tau - h_n)}{2\,h_n},$$

where

$$\widehat{F}_{n\widehat{\varepsilon}}(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i - \boldsymbol{X}_i^\mathsf{T}\,\widehat{\boldsymbol{\beta}}_n(\tau) \le y\} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\widehat{\varepsilon}_i(\tau) \le y\}$$

is the empirical distribution function of the residuals and (the bandwidth) h_n is a sequence going to zero as $n \to \infty$. A possible choice of h_n (derived when assuming normal errors $\varepsilon_1, \ldots, \varepsilon_n$) is given by

$$h_n = n^{-1/3}\, u_{1-\alpha/2}^{2/3} \left[\frac{1.5\,\varphi^2(u_\tau)}{2\,u_\tau^2+1}\right]^{1/3},$$

where φ is the density of $\mathsf{N}(0,1)$. For details and other possible choices of h_n see Chapter 4.10.1 in Koenker [2005] and the references therein.

As estimating $\frac{1}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)}$ is rather delicate, also in this situation the nonparametric bootstrap* is of interest.

5.5 Asymptotic normality of sample quantiles†

Suppose that we have a random sample $X_1, \ldots, X_n$, where X_1 has a cumulative distribution function F. Note that for a given $\tau \in (0,1)$ thanks to Remark 17 one can view the sample quantile $\widehat{F}_n^{-1}(\tau)$ as the argument of minimum of a convex function. Thus analogously as in Chapter 3.4 one can derive that if $f(x)$ (the density of X_1) is positive and continuous in a neighbourhood of $F^{-1}(\tau)$, then

$$\sqrt{n}\left(\widehat{F}_n^{-1}(\tau) - F^{-1}(\tau)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \tfrac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}\big).$$

Literature: Koenker [2005], Sections 2.1, 2.4, 4.2. 4.10.

6 EM-algorithm

It is an *iterative* algorithm to find the *maximum likelihood* estimator $\widehat{\boldsymbol{\theta}}_n$ in situations with missing data. It is also often used in situations when the model can be specified with the help of some unobserved variables and finding $\widehat{\boldsymbol{\theta}}_n$ would be (relatively) simple with the knowledge of those unobserved variables.

* Bootstrap and othere resampling methods are in detail explained in the course NMST545 Mathematical Statistics 4. † Not done at the lecture. It is assumed that it is known from the bachelor degree.

Example 45. Let $X_1, \ldots, X_n$ be a random sample from the distribution with the density

$$f(x; \boldsymbol{\pi}) = \sum_{j=1}^{G} \pi_j \, f_j(x),$$

where $f_1, \ldots, f_G$ are known densities and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^{\mathsf{T}}$ is a vector of unknown non-negative *mixing proportions* such that $\sum_{j=1}^{G} \pi_j = 1$. Find the maximum likelihood estimator of the parameter $\boldsymbol{\pi}$, i.e.

$$\widehat{\boldsymbol{\pi}}_n = \arg\max_{\boldsymbol{\pi} \in \Theta} \left(\prod_{i=1}^{n} f(X_i; \boldsymbol{\pi}) \right),$$

where $\Theta = \left\{ (\pi_1, \ldots, \pi_G)^{\mathsf{T}} \, : \, \pi_j \in [0,1], \ \sum_{j=1}^{G} \pi_j = 1 \right\}$.

Solution. A straightforward approach would be to maximize the log-likelihood

$$\ell_n(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{G} \pi_j f_j(X_i) \right).$$

Using for instance the parametrization $\pi_G = 1 - \sum_{j=1}^{G-1} \pi_j$, the system of score equations is given by

$$U_{jn}(\boldsymbol{\pi}) = \frac{\partial \ell_n(\boldsymbol{\pi})}{\partial \pi_j} = \sum_{i=1}^{n} \left[\frac{f_j(X_i)}{\sum_{l=1}^{G} \pi_l f_l(X_i)} - \frac{f_G(X_i)}{\sum_{l=1}^{G} \pi_l \, f_l(X_i)} \right] \stackrel{!}{=} 0, \quad j = 1, \ldots, G-1,$$

which requires some numerical routines.

Alternatively one can use the EM-algorithm, which runs as follows. Introduce $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^{\mathsf{T}} \sim \mathsf{Mult}_G(1; \boldsymbol{\pi})$, where

$$Z_{ij} = \begin{cases} 1, & X_i \text{ is generated from } f_j(x), \\ 0, & \text{otherwise.} \end{cases}$$

Note that one can think of our data as the realizations of the independent and identically distributed random vectors $\binom{X_1}{\mathbf{Z}_1}, \ldots, \binom{X_n}{\mathbf{Z}_n}$, where $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are missing.

Put $\mathbb{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$. The joint density of a random vector $\binom{X_i}{\mathbf{Z}_i}$ is given by

$$f_{X,\mathbf{Z}}(x, \mathbf{z}; \boldsymbol{\pi}) = f_{X|\mathbf{Z}}(x|\mathbf{z}; \boldsymbol{\pi}) \, f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\pi}) = \left(\sum_{j=1}^{G} z_j f_j(x) \right) \cdot \left(\prod_{j=1}^{G} \pi_j^{z_j} \right).$$

In the context of EM algorithm the random sample $\binom{X_1}{\mathbf{Z}_1}, \ldots, \binom{X_n}{\mathbf{Z}_n}$ is called *complete data*.

The corresponding log-likehood is called *complete log-likehood* and it is given by

$$\ell_n^C(\boldsymbol{\pi}) = \log\left\{\prod_{i=1}^n\left[\left(\sum_{j=1}^G Z_{ij} f_j(X_i)\right)\left(\prod_{j=1}^G \pi_j^{Z_{ij}}\right)\right]\right\}$$
$$= \sum_{i=1}^n\left[\log\left(\sum_{j=1}^G Z_{ij} f_j(X_i)\right)\right] + \sum_{i=1}^n\left[\sum_{j=1}^G Z_{ij}\log\pi_j\right].$$

If we knew $\mathbf{Z}_1,\dots,\mathbf{Z}_n$, then we would estimate simply $\widehat{\pi}_j = \frac{1}{n}\sum_{i=1}^n Z_{ij}, j = 1,\dots,G$. The EM algorithm runs in the following two steps:

(i) **E-step** (Expectation step): Let $\widehat{\boldsymbol{\pi}}^{(k)}$ be the current estimate of $\boldsymbol{\pi}$. In this step we calculate

$$Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}[\ell_n^C(\boldsymbol{\pi}) \,|\, \mathbb{X}],$$

where the expectation is taken with respect to the unobserved random vectors $\mathbf{Z}_1,\dots,\mathbf{Z}_n$. More precisely one has to take the expectation with respect to the conditional distribution of $\mathbf{Z}_1,\dots,\mathbf{Z}_n$ given $X_1,\dots,X_n$. As this distribution depends on the unknown parameter $\boldsymbol{\pi}$, this parameter is replaced with the current version of the estimate $\widehat{\boldsymbol{\pi}}^{(k)}$. This is indicated by $\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}$. Note that in this step one gets rid of the unobserved random vectors $\mathbf{Z}_1,\dots,\mathbf{Z}_n$.

(ii) **M-step** (Maximization step): The updated value of the estimate of $\boldsymbol{\pi}$ is calculated as

$$\widehat{\boldsymbol{\pi}}^{(k+1)} = \operatorname*{arg\,max}_{\boldsymbol{\pi}\in\Theta} Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big).$$

The end of class 20 (30. 4. 2024)

E-step in a detail:

$$Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^n \log\left(\sum_{j=1}^G Z_{ij} f_j(X_i)\right)\,\middle|\,\mathbb{X}\right] + \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^n\sum_{j=1}^G Z_{ij}\log\pi_j\,\middle|\,\mathbb{X}\right]. \quad (78)$$

Note that the first term on the right-hand side of the above equation does not depend on $\boldsymbol{\pi}$. Thus we do not need to calculate this term for M-step. To calculate the second term it is sufficient to calculate $\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[Z_{ij}\,|\,\mathbb{X}\right]$. To do that denote $\mathbf{e}_j = (0,\dots,0,1,0,\dots,0)^\mathsf{T}$ for the j-th canonical vector. Now with the help of Bayes theorem for densities (Theorem A15) one can calculate

$$\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[Z_{ij}\,|\,\mathbb{X}\right] = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[Z_{ij}\,|\,X_i\right] = \mathsf{P}_{\widehat{\boldsymbol{\pi}}^{(k)}}(Z_{ij}=1\,|\,X_i) = f_{\mathbf{Z}|X}(\mathbf{e}_j|X_i;\widehat{\boldsymbol{\pi}}^{(k)})$$
$$= \frac{f_{X|\mathbf{Z}}(X_i|\mathbf{e}_j;\widehat{\boldsymbol{\pi}}^{(k)}) f_{\mathbf{Z}}(\mathbf{e}_j;\widehat{\boldsymbol{\pi}}^{(k)})}{f_X(X_i;\widehat{\boldsymbol{\pi}}^{(k)})} = \frac{f_j(X_i)\,\widehat{\pi}_j^{(k)}}{\sum_{l=1}^G f_l(X_i)\,\widehat{\pi}_l^{(k)}} =: z_{ij}^{(k)}.$$

M-step in a detail: Note that with the help of the previous step and (78)

$$Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = const + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}^{(k)} \log \pi_j.$$

Analogously as when calculating the maximum likelihood estimator in a multinomial distribution one can show that the updated value of the estimate of $\boldsymbol{\pi}$ is given by

$$\widehat{\boldsymbol{\pi}}^{(k+1)} = \arg\max_{\boldsymbol{\pi}\in\Theta} Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^{(k)},$$

where $\mathbf{z}_i^{(k)} = \big(z_{i1}^{(k)}, \ldots, z_{iG}^{(k)}\big)^{\mathsf{T}}$ and so $\widehat{\pi}_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(k)}$ for $j \in \{1, \ldots, G\}$.

6.1 General description of the EM-algorithm

Denote the observed random variables as $\mathbb{Y}_{obs}$ and the unobserved (missing) random variables $\mathbb{Y}_{mis}$. Let $f(\mathbf{y}; \boldsymbol{\theta})$ be the joint density (with respect to a σ-finite measure μ) of $\mathbb{Y} = (\mathbb{Y}_{obs}, \mathbb{Y}_{mis})$ and denote $\ell_n^C(\boldsymbol{\theta})$ the *complete log-likelihood* of $\mathbb{Y}$. Our task is to maximize the *observed log-likelihood* $\ell_{obs}(\boldsymbol{\theta}) = \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta})$, where $f(\mathbf{y}_{obs}; \boldsymbol{\theta})$ is the density of $\mathbb{Y}_{obs}$. Note that

$$\begin{aligned}\ell_n^C(\boldsymbol{\theta}) &= \log f(\mathbb{Y}_{obs}, \mathbb{Y}_{mis}; \boldsymbol{\theta}) = \log\big(f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta})\, f(\mathbb{Y}_{obs}; \boldsymbol{\theta})\big)\\ &= \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta}) = \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \ell_{obs}(\boldsymbol{\theta}),\end{aligned}$$

where $f(\mathbf{y}_{mis}|\mathbf{y}_{obs}; \boldsymbol{\theta})$ stands for the conditional density of $\mathbb{Y}_{mis}$ given $\mathbb{Y}_{obs} = \mathbf{y}_{obs}$. Thus one can express the observed log-likelihood with the help of the complete log-likelihood as

$$\ell_{obs}(\boldsymbol{\theta}) = \ell_n^C(\boldsymbol{\theta}) - \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}). \tag{79}$$

Finally denote

$$Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \mathsf{E}_{\widetilde{\boldsymbol{\theta}}}\left[\ell_n^C(\boldsymbol{\theta}) \,|\, \mathbb{Y}_{obs}\right]. \tag{80}$$

EM-algorithm runs as follows:

Let $\widehat{\boldsymbol{\theta}}^{(k)}$ be the result of the k-th iteration of the EM-algorithm. The next iteration $\widehat{\boldsymbol{\theta}}^{(k+1)}$ is computed in two steps:

E-step: Calculate $Q\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\big)$.

M-step: Find $\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\big)$.

Note that at this moment it is not at all clear, if the EM-algorithm is a good idea. Remember that our task is to maximize the observed likelihood $\ell_{obs}(\boldsymbol{\theta})$. The following theorem is the first answer in this aspect.

Theorem 11. *Let the set* $\{y_{miss} : f(y_{miss}|\mathbb{Y}_{obs};\boldsymbol{\theta}) > 0\}$ *does not depend on* $\boldsymbol{\theta}$. *Further* $\ell_{obs}(\boldsymbol{\theta})$ *be the observed likelihood and* $\widehat{\boldsymbol{\theta}}^{(k)}$ *be the result of the* k*-th iteration of the EM-algorithm. Then*

$$\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right) \geq \ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right).$$

Proof. Note that the left-hand side of (79) does not depend on $\mathbb{Y}_{mis}$. Thus applying $\mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}[\,\cdot\,|\,\mathbb{Y}_{obs}]$ on both sides of (79) yields that

$$\begin{aligned}\ell_{obs}(\boldsymbol{\theta}) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}[\ell_n^C(\boldsymbol{\theta})\,|\,\mathbb{Y}_{obs}] - \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})\,\middle|\,\mathbb{Y}_{obs}\right] \\ &=: Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right). \end{aligned} \tag{81}$$

Now note that

$$\begin{aligned}\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right) &= Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right), \\ \ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) &= Q\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right).\end{aligned}$$

Thus to verify $\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right) \geq \ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$ it is sufficient to show that

$$Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \geq Q\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \quad \text{and also} \quad H\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \leq H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right). \tag{82}$$

Showing *the first inequality* in (82) is easy as from the M-step

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right),$$

which implies that $Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \geq Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right)$ for each $\boldsymbol{\theta}\in\Theta$.

To show *the second inequality* in (82) one gets with the help of Jensen's inequality that for each $\boldsymbol{\theta}\in\Theta$:

$$\begin{aligned} H\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})\,\middle|\,\mathbb{Y}_{obs}\right] \\ &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log\left(\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\right)\middle|\,\mathbb{Y}_{obs}\right] + \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)|\mathbb{Y}_{obs}\right] \\ &\overset{\text{Jensen}}{\leq} \log\left(\mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\middle|\,\mathbb{Y}_{obs}\right]\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \\ &= \log\left(\int \frac{f(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\cdot f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)\mathrm{d}\mu(\mathbf{y}_{mis})\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \\ &= \log(1) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) = H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right). \end{aligned} \tag{83}$$

□

6.2 Convergence of the EM-algorithm

Although from Theorem 11 we know that EM algorithm increases (more precisely does not decrease) the observed log-likelihood, it is still not clear whether the sequence $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{\infty}$ converges. And if it converges what is the limit.

To answer this question we need to introduce the following regularity assumptions.

- The parameter space Θ is a subset of $\mathbb{R}^p$.
- The set $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \ell_{obs}(\boldsymbol{\theta}) \geq \ell_{obs}(\boldsymbol{\theta}_0)\}$ is compact for any $\boldsymbol{\theta}_0 \in \Theta$ such that $\ell_{obs}(\boldsymbol{\theta}_0) > -\infty$.
- $\ell_{obs}(\boldsymbol{\theta})$ is continuous in Θ and differentiable in the interior of Θ.

Theorem 12. *Let the function $Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ defined in (80) be continuous both in $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$. Then all the limit points of any instance $\{\widehat{\boldsymbol{\theta}}^{(k)}\}$ are stationary points of $\ell_{obs}(\boldsymbol{\theta})$. Further $\{\ell_{obs}(\widehat{\boldsymbol{\theta}}^{(k)})\}$ converges monotonically to some value $\ell^* = \ell_{obs}(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is a stationary point of $\ell_{obs}(\boldsymbol{\theta})$.*

Proof. See Wu [1983]. □

Note that if $\boldsymbol{\theta}^*$ is a stationary point of $\ell_{obs}(\boldsymbol{\theta})$, then

$$\left.\frac{\partial \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbf{0}_p.$$

Thus by Theorem 12 the EM-algorithm finds a solution of the system of log-likelihood equations but in generally there is no guarantee that this is a global maximum of $\ell_{obs}(\boldsymbol{\theta})$.

Corollary 2. *Let the assumptions of Theorem 12 be satisfied. Further suppose that the function $\ell_{obs}(\boldsymbol{\theta})$ has a unique maximum $\widehat{\boldsymbol{\theta}}_n$ that is the only stationary point. Then $\widehat{\boldsymbol{\theta}}^{(k)} \to \widehat{\boldsymbol{\theta}}_n$ as $k \to \infty$.*

The expected end of class 21 (3. 5. 2024)

Example 46. Let $X_1, \ldots, X_n$ be a random sample from the distribution with the density

$$f(x) = w\,\tfrac{1}{\sigma_1}\,\varphi\big(\tfrac{x-\mu_1}{\sigma_1}\big) + (1-w)\,\tfrac{1}{\sigma_2}\,\varphi\big(\tfrac{x-\mu_2}{\sigma_2}\big),$$

where $w \in [0,1]$, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2, \sigma_2^2 \in (0, \infty)$ are unknown parameters and

$$\varphi(x) = \tfrac{1}{\sqrt{2\,\pi}}\,\exp\{-x^2/2\}$$

is the density of the standard normal distribution. Describe the EM algorithm to find the maximum likelihood estimates of the unknown parameters.

6.3 Rate of convergence of EM-algorithm

Note that in the M-step of the algorithm there might not be a unique value that maximizes $Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)$. Thus denote the set of maximizing points as $\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$, i.e.

$$\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) = \left\{\widetilde{\boldsymbol{\theta}} : Q\left(\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^{(k)}\right) = \max_{\boldsymbol{\theta}\in\Theta} Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)\right\}.$$

Then one needs to choose $\widehat{\boldsymbol{\theta}}^{(k+1)}$ as an element of the set $\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$. Thus let $\mathbf{M} : \Theta \to \Theta$ be a mapping such that

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \mathbf{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right).$$

Let $\widehat{\boldsymbol{\theta}}^{(k)} \to \boldsymbol{\theta}^*$ as $k \to \infty$. Note that then $\boldsymbol{\theta}^* = \mathbf{M}(\boldsymbol{\theta}^*)$. Assuming that $\mathbf{M}$ is sufficiently smooth one gets by the one term Taylor expansion around the point $\boldsymbol{\theta}^*$ the following approximation

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \mathbf{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) = \underbrace{\mathbf{M}(\boldsymbol{\theta}^*)}_{=\boldsymbol{\theta}^*} + \left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \left(\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right) + o\left(\left\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\|\right).$$

Thus

$$\widehat{\boldsymbol{\theta}}^{(k+1)} - \boldsymbol{\theta}^* = \left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \left(\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right) + o\left(\left\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\|\right) \tag{84}$$

and the Jacobi matrix $\left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ measures approximately the rate of convergence. It can be shown that

$$\left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = [I_n^C(\boldsymbol{\theta}^*)]^{-1} I_n^{mis}(\boldsymbol{\theta}^*), \tag{85}$$

where

$$I_n^C(\boldsymbol{\theta}) = -\,\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell_n^C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} \,\middle|\, \mathbb{Y}_{obs}\right]$$

can be considered as the empirical Fisher information matrix from the complete data and

$$I_n^{mis}(\boldsymbol{\theta}) = -\,\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} \middle| \mathbb{Y}_{obs}\right],$$

can be considered as the empirical Fisher information matrix of the contribution of the missing data (that is not explained by the observed data).

Note that by (84) and (85) in the presence of missing data the convergence is only linear. Further the bigger proportion of missing data the 'bigger' $I_n^{mis}(\boldsymbol{\theta})$ and the slower is the convergence.

The expected end of class 22 (7. 5. 2024)

6.4 The EM algorithm in exponential families

Let the complete data $\mathbb{Y}$ have a density with respect to a σ-finite measure μ given by

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\Big\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbf{y}) \Big\}\, b(\boldsymbol{\theta})\, c(\mathbf{y}) \tag{86}$$

and the standard choice of the parametric space is

$$\Theta = \left\{ \boldsymbol{\theta} : \int \exp\Big\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbf{y}) \Big\}\, c(\mathbf{y})\, \mathrm{d}\mu(\mathbf{y}) < \infty \right\}.$$

Note that $\mathbf{T}(\mathbb{Y}) = \big(T_1(\mathbb{Y}), \ldots, T_p(\mathbb{Y})\big)^{\mathsf{T}}$ is a sufficient statistic for $\boldsymbol{\theta}$.

The log-likelihood of the complete data is now given by

$$\ell_n^C(\boldsymbol{\theta}) = \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbb{Y}) + \log b(\boldsymbol{\theta}) + const.,$$

which yields that the function Q from the EM-algorithm is given by

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\ell_n^C(\boldsymbol{\theta}) | \mathbb{Y}_{obs}\right] = \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[T_j(\mathbb{Y}) \,\middle|\, \mathbb{Y}_{obs}\right] + \log b(\boldsymbol{\theta}) + const. \\ &= \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \widehat{T}_j^{(k)} + \log b(\boldsymbol{\theta}) + const., \end{aligned}$$

where we put $\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[T_j(\mathbb{Y}) \,\middle|\, \mathbb{Y}_{obs}\right]$.

The nice thing about exponential families is that in the E-step of the algorithm we do not need to calculate $Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)$ for each $\boldsymbol{\theta}$ separately but it is sufficient to calculate

$$\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[T_j(\mathbb{Y}) \,\middle|\, \mathbb{Y}_{obs}\right], \quad j = 1, \ldots, p,$$

and in the M-step we maximize

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \left\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \widehat{T}_j^{(k)} + \log b(\boldsymbol{\theta}) \right\}. \tag{87}$$

Interval censoring

Let $-\infty = d_0 < d_1 < \ldots < d_M = \infty$ be a division of $\mathbb{R}$. Further let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables whose exact values are not observed. Instead of each Y_i we only know that $Y_i \in (d_{q_i-1}, d_{q_i}]$, for some $q_i \in \{1, \ldots, M\}$. Thus we observed independent and identically distributed random variables $X_1, \ldots, X_n$ such that $X_i = q_i$ if $Y_i \in (d_{q_i-1}, d_{q_i}]$.

Suppose now that Y_i has a density $f(y;\boldsymbol{\theta})$ of the form

$$f(y;\boldsymbol{\theta}) = \exp\left\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, t_j(y) \right\} b_1(\boldsymbol{\theta})\, c_1(y).$$

Thus the joint density of the random sample $Y_1, \dots, Y_n$ is of the form (86) where

$$T_j(\mathbb{Y}) = \sum_{i=1}^{n} t_j(Y_i), \qquad j = 1, \dots, p.$$

Thus in the E-step of the EM-algorithm it is sufficient to calculate

$$\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}} \left[T_j(\mathbb{Y}) \,\middle|\, X_1, \dots, X_n \right] = \sum_{i=1}^{n} \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}} \left[t_j(Y_i) \,|\, X_i \right], \quad j = 1, \dots, p,$$

and the M-step is given by (87) where $b(\boldsymbol{\theta}) = b_1^n(\boldsymbol{\theta})$.

Example 47. Suppose that $Y_i \sim \mathsf{Exp}(\lambda)$, i.e. $f(y;\lambda) = \lambda \mathrm{e}^{-\lambda y} \mathbb{I}\{y > 0\}$. Thus $p = 1$, $t_1(y) = y$, $a_1(\lambda) = -\lambda$ and $b_1(\lambda) = \lambda$.

In the E-step one needs to calculate $\mathsf{E}_{\hat{\lambda}^{(k)}}[Y_i \,|\, X_i]$. Note that the conditional distribution of Y_i given that $Y_i \in (a, b]$ has a density $\frac{\lambda \mathrm{e}^{-\lambda y}}{\mathrm{e}^{-\lambda a} - \mathrm{e}^{-\lambda b}} \mathbb{I}\{y \in (a, b]\}$. Thus with the help of the integration by parts

$$\begin{aligned} \widehat{Y}_i^{(k)} := \mathsf{E}_{\hat{\lambda}^{(k)}}[Y_i \,|\, X_i = q_i] &= \frac{1}{\mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_{i-1}}} - \mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_i}}} \int_{d_{q_{i-1}}}^{d_{q_i}} x\, \hat{\lambda}^{(k)} \mathrm{e}^{-\hat{\lambda}^{(k)} x}\, \mathrm{d}x \\ &= \frac{d_{q_{i-1}}\, \mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_{i-1}}} - d_{q_i} \mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_i}}}{\mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_{i-1}}} - \mathrm{e}^{-\hat{\lambda}^{(k)} d_{q_i}}} + \frac{1}{\hat{\lambda}^{(k)}} \end{aligned}$$

and with the help of (87) one gets that

$$\hat{\lambda}^{(k+1)} = \arg\max_{\lambda > 0} \left\{ Q\left(\lambda, \hat{\lambda}^{(k)}\right) \right\} = \arg\max_{\lambda > 0} \left\{ -\lambda \sum_{i=1}^{n} \widehat{Y}_i^{(k)} + n \log \lambda \right\} = \frac{1}{\frac{1}{n} \sum \widehat{Y}_i^{(k)}}.$$

Literature: McLachlan and Krishnan [2008] Chapters 1.4.3, 1.5.1, 1.5.3, 2.4, 2.7, 3.2, 3.4.4, 3.5.3, 3.9 and 5.9.

7 Missing data*

For $i = 1, \dots, I$ let $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})^{\mathsf{T}}$ represent the data of the i-th subject that could be ideally observed. Let $\mathbf{R}_i = (R_{i1}, \dots, R_{in_i})^{\mathsf{T}}$, where

$$R_{ij} = \begin{cases} 1, & \text{if } Y_{ij} \text{ is observed}, \\ 0, & \text{otherwise.} \end{cases}$$

* *Chybějící data*

Let $\mathbb{Y}_{obs}$ represent Y_{ij} such that $R_{ij} = 1$ and $\mathbb{Y}_{mis}$ represent Y_{ij} such that $R_{ij} = 0$. Thus the available data are given by

$$(\mathbb{Y}_{obs}, \mathbf{R}_1, \ldots, \mathbf{R}_I) = (\mathbb{Y}_{obs}, \boldsymbol{R}),$$

where $\boldsymbol{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_I)$. Note that the complete data can be represented as

$$(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_I, \boldsymbol{R}) = (\mathbb{Y}_{obs}, \mathbb{Y}_{mis}\, \boldsymbol{R}) =: (\mathbb{Y}, \boldsymbol{R}).$$

Suppose that the distribution of $\mathbb{Y}$ depends on a parameter $\boldsymbol{\theta}$ (which we are interested in) and the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ depends on $\boldsymbol{\psi}$. Then the joint density of the complete data can be written as

$$f(\mathbf{y}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{r}|\mathbf{y}; \boldsymbol{\psi})\, f(\mathbf{y}; \boldsymbol{\theta}).$$

Now integrating the above density with respect to $\mathbf{y}_{mis}$ yields the density of the available data

$$f(\mathbf{y}_{obs}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\theta})\, f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\psi})\, \mathrm{d}\mu(\mathbf{y}_{mis}). \tag{88}$$

In what follows we will say that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are *separable* if $\boldsymbol{\theta} \in \Omega_\theta$, $\boldsymbol{\psi} \in \Omega_\psi$ and $(\boldsymbol{\theta}, \boldsymbol{\psi})^\mathsf{T} \in \Omega_\theta \times \Omega_\psi$.

7.1 Basic concepts for the mechanism of missing

Depending on what can be assumed about the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ we distinguish three situations.

Missing completely at random (MCAR). Suppose that $\boldsymbol{R}$ is independent of $\mathbb{Y}$, thus one can write $f(\mathbf{r}|\mathbf{y}; \boldsymbol{\psi}) = f(\mathbf{r}; \boldsymbol{\psi})$ and with the help of (88) one gets

$$f(\mathbf{y}_{obs}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_{obs}; \boldsymbol{\theta}) f(\mathbf{r}; \boldsymbol{\psi}),$$

which further implies that the observed log-likelihood is of the form

$$\ell_{obs}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \log f(\boldsymbol{R}; \boldsymbol{\psi}).$$

Note that if the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable then the second term on the right-hand side of the above equation does not depend on $\boldsymbol{\theta}$ and can be ignored when one is interested only in $\boldsymbol{\theta}$.

Example 48. Let $Y_1, \ldots, Y_n$ be a random sample from the exponential distribution $\mathsf{Exp}(\lambda)$. Let $R_1, \ldots, R_n$ be a random sample independent with $Y_1, \ldots, Y_n$ and R_i follows a Bernoulli distribution with a parameter p_i (e.g. $p_i = \frac{1}{1+i}$).

Missing at random (MAR). Suppose that the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ is the same as the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}_{obs}$. Thus one can write $f(\mathbf{r}|\mathbf{y};\boldsymbol{\psi}) = f(\mathbf{r}|\mathbf{y}_{obs};\boldsymbol{\psi})$ and with the help of (88)

$$f(\mathbf{y}_{obs},\mathbf{r};\boldsymbol{\theta},\boldsymbol{\psi}) = f(\mathbf{y}_{obs};\boldsymbol{\theta})f(\mathbf{r}|\mathbf{y}_{obs};\boldsymbol{\psi}),$$

which further implies that the observed log-likelihood is of the form

$$\ell_{obs}(\boldsymbol{\theta},\boldsymbol{\psi}) = \log f(\mathbb{Y}_{obs};\boldsymbol{\theta}) + \log f(\boldsymbol{R}|\mathbb{Y}_{obs};\boldsymbol{\psi}).$$

Note that although MAR is not so strict in assumptions as MCAR, also here the second term on the right-hand side of the above equation does not depend on $\boldsymbol{\theta}$ provided that $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable.

Example 49. Let $(\boldsymbol{X}_1^\mathsf{T}, Y_1, R_1)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^\mathsf{T}, Y_n, R_n)^\mathsf{T}$ be independent and identically distributed random vectors, where the covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are always completely observed. Let R_i stand for the indicator of missing of Y_i and

$$\mathsf{P}(R_i = 1 \,|\, \boldsymbol{X}_i, Y_i) = r(\boldsymbol{X}_i),$$

where $r(\mathbf{x})$ is a given (but possibly unknown) function.

Missing not at random (MNAR). In this concept neither the distribution of $\boldsymbol{R}$ is independent of $\mathbb{Y}$ nor the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}_{obs}$ is independent of $\mathbb{Y}_{mis}$. Thus the density of the observed data is generally given by (88). To proceed one has to make some other assumptions about the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ (i.e. about the density $f(\mathbf{r}|\mathbf{y}_{obs},\mathbf{y}_{mis};\boldsymbol{\psi})$).

Example 50. *Maximum likelihood estimator for the right-censored data from an exponential distribution.* Suppose that $Y_1, \ldots, Y_n$ is a random sample from the exponential distribution with the density $f(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}\mathbb{I}\{x > 0\}$. Nevertheless we observe Y_i only if $Y_i \leq C$, where C is a known constant (e.g. duration of the study). If $Y_i > C$ then we do not observe the value of Y_i (we only now that Y_i is greater than C).

Note that

$$f(\mathbf{y}_{obs},\mathbf{y}_{mis};\lambda) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda y_i}$$

and

$$f(\mathbf{r}|\mathbf{y}_{obs},\mathbf{y}_{mis}) = \prod_{i=1}^{n} \big[\mathbb{I}\{y_i \leq C\}\big]^{r_i} \big[\mathbb{I}\{y_i > C\}\big]^{1-r_i}.$$

Although this conditionally density depends on $\mathbf{y}_{mis}$ (thus we are in a situation of MNAR), we can proceed because this conditionally density is completely known.

Let n_0 be the number of fully observed Y_i (i.e. $n_0 = \sum_{i=1}^{n} \mathbb{I}\{Y_i \leq C\}$). For simplicity of notation assume that $Y_1, \ldots, Y_n$ are ordered in such a way that $Y_1, \ldots, Y_{n_0}$ are fully observed and $Y_{n_0+1}, \ldots, Y_n$ are censored (i.e. $Y_i > C$ for $i \in \{n_0 + 1, \ldots, n\}$). Thus the corresponding components of $\boldsymbol{R}$ are given by $R_i = 1$ for $i \in \{1, \ldots, n_0\}$ and zero otherwise.

Now with the help of (88) one can calculate

$$\begin{aligned} f(\boldsymbol{Y}_{obs}, \mathbf{R}; \lambda) &= \prod_{i=1}^{n_0} \lambda \mathrm{e}^{-\lambda Y_i} \int_C^\infty \cdots \int_C^\infty \prod_{i=n_0+1}^{n} \lambda \, \mathrm{e}^{-\lambda y_i} \mathrm{d}y_{n_0+1}, \ldots, \mathrm{d}y_n \\ &= \lambda^{n_0} \mathrm{e}^{-\lambda \sum_{i=1}^{n_0} Y_i} \left[\mathrm{e}^{-\lambda C}\right]^{n-n_0}. \end{aligned}$$

The corresponding log-likelihood of the observed data is

$$\ell_{obs}(\lambda) = n_0 \log \lambda - \lambda \sum_{i=1}^{n_0} Y_i - (n - n_0) C \lambda,$$

which is maximised at

$$\widehat{\lambda}_n = \frac{1}{\frac{1}{n_0} \sum_{i=1}^{n_0} Y_i + \frac{(n-n_0)C}{n_0}}.$$

Note that the above example is in fact rather exceptional as the missing mechanism is given by the design of the study and thus known.

The general problem of all the concepts is that if missing is not a part of the design of the study then no assumptions about the relationship of $\mathbb{Y}_{mis}$ and $\boldsymbol{R}$ can be verified as we do not observe $\mathbb{Y}_{mis}$.

The expected end of class 23 (10. 5. 2024)

7.2 Methods for dealing with missing data

Complete case analysis (CCA)

In the analysis we use only the subjects with the full record, i.e. only subjects for which no information is missing.

Advantages and disadvantages:

+ simplicity;

– the inference about $\boldsymbol{\theta}$ is 'biased' (i.e. the parameter $\boldsymbol{\theta}$ is generally not identified), if MCAR does not hold;

– even if MCAR holds, then this method may not provide an effective use of data.

Example 51. Suppose that we have five observations on each subject. Each observation is missing with probability 0.1 and the observations are missing independently on each other. Thus on average only 59 % ($0.9^5 \doteq 0.59$) of the records will be complete.

Available case analysis (ACA)

In each of the analyses one uses all the data that are available for this particular analysis.

Example 52. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from $\mathsf{N}((\mu_1, \mu_2, \mu_3)^\mathsf{T}, \mathbb{\Sigma}_{3\times 3})$. Then the covariance $\sigma_{ij} = \mathsf{cov}(X_{1i}, X_{1j})$ is estimated from all the vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ for which both the i-th and the j-th coordinate is observed.

Advantages and disadvantages:

+ simplicity;

+ more data can be used than with CCA;

− the inference about $\boldsymbol{\theta}$ is biased, if MCAR does not hold;

− it can result in estimates with strange features (e.g. there is no guarantee that the estimate of the variance matrix $\widehat{\mathbb{\Sigma}}$ in Example 52 is positive semidefinite).

Direct (ignorable) observed likelihood

The inference is based on $\log f(\mathbb{Y}_{obs}; \boldsymbol{\theta})$, that is the distribution of $\boldsymbol{R}$ is 'ignored'.
Advantages and disadvantages:

+ If the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable then this method is not biased and does not lose any information provided MAR holds;

− The observed log-likelihood $\ell_{obs}(\boldsymbol{\theta})$ might be difficult to calculate. Nevertheless, sometimes the EM algorithm can be helpful.

Imputation

In this method the missing observations are estimated ('imputed') and then one works with the data as if there were no missing values.
Advantages and disadvantages:

+ If the missing values are estimated appropriately, it can give 'reasonable' estimates of the unknown parameters;

+ One can use the *completed* dataset also for other analyses;

− The standard estimates of the (asymptotic) variances of the estimates of the parameters computed from the completed dataset are too optimistic (too low). The reason is that an appropriate estimate of variance should reflect that part of the data has been imputed.

Example 53. Suppose that $X_1, \dots, X_n$ is a random sample. Further suppose that we observe only $X_1, \dots, X_{n_0}$ for some $n_0 < n$ and the remaining observations $X_{n_0+1}, \dots, X_n$ are missing. For $i = n_0 + 1, \dots, n$ let the missing observations be estimated as $\widehat{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_j$. Then the standard estimate of $\mu = \mathsf{E}\, X_1$ is given by

$$\widehat{\mu}_n = \frac{1}{n} \left(\sum_{i=1}^{n_0} X_i + \sum_{i=n_0+1}^{n} \widehat{X}_i \right) = \frac{1}{n_0} \sum_{j=1}^{n_0} X_j$$

and seems to be reasonable.

But the standard estimate of the variance of $\widehat{\mu}_n$ computed from the completed dataset

$$\widehat{\mathsf{var}(\widehat{\mu}_n)} = \frac{S_n^2}{n}, \quad \text{where} \quad S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n_0} (X_i - \widehat{\mu}_n)^2 + \sum_{i=n_0+1}^{n} (\widehat{X}_i - \widehat{\mu}_n)^2 \right)$$

is too small. The first reason is that S_n^2 as the estimate of $\mathsf{var}(X_1)$ is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n_0} (X_i - \widehat{\mu}_n)^2 = \frac{n_0 - 1}{n-1}\, S_{n_0}^2 < S_{n_0}^2.$$

The second reason is that the factor $\frac{1}{n}$ assumes that there are n independent observations, but in fact there are only n_0 independent observations.

The expected end of class 24 (17. 5. 2024)

Multiple imputation

In this method the missing observations are imputed several times. Formally, for $j = 1, \dots, M$ let $\widehat{\mathbb{Y}}_{mis}^{(j)}$ be the imputed values in the j-th round. Further let $\widehat{\boldsymbol{\theta}}_j$ be the estimate of the parameter $\boldsymbol{\theta}$ from the completed data $\left(\mathbb{Y}_{obs}, \widehat{\mathbb{Y}}_{mis}^{(j)}\right)$. Then the final estimate of the parameter $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}}_{MI} = \frac{1}{M} \sum_{j=1}^{M} \widehat{\boldsymbol{\theta}}_j.$$

The advantage of this method is that one can also estimate the (asymptotic) variance of this estimator by

$$\widehat{\mathsf{var}(\widehat{\boldsymbol{\theta}}_{MI})} = \overline{\mathbb{V}}_M + \left(1 + \tfrac{1}{M}\right) \mathbb{B}_M, \tag{89}$$

where

$$\overline{\mathbb{V}}_M = \frac{1}{M} \sum_{j=1}^{M} \widehat{\mathbb{V}}_j \quad \text{and} \quad \mathbb{B}_M = \frac{1}{M-1} \sum_{j=1}^{M} \left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_{MI}\right) \left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_{MI}\right)^{\mathsf{T}},$$

with $\widehat{\mathbb{V}}_j$ being a standard estimate of the asymptotic variance calculated from the completed data $\widehat{\mathbb{Y}}^{(j)} = (\mathbb{Y}_{obs}, \widehat{\mathbb{Y}}_{mis}^{(j)})$.

The rationale of the formula (89) is as follows. Note that

$$\mathsf{var}\left(\widehat{\boldsymbol{\theta}}_{MI}\right) = \mathsf{E}\,\left(\mathsf{var}(\widehat{\boldsymbol{\theta}}_{MI}\,|\,\widehat{\mathbb{Y}}^{(j)})\right) + \mathsf{var}\left(\mathsf{E}\,(\widehat{\boldsymbol{\theta}}_{MI}\,|\,\widehat{\mathbb{Y}}^{(j)})\right).$$

Now the first term on right-hand side of the above equation is estimated by $\overline{\mathbb{V}}_M$ and the second term is estimated by $\mathbb{B}_M$.

Example 54. In Example 53 one can for instance impute the values $X_{n_0+1}, \ldots, X_n$ by a random sample from $\mathsf{N}(\widehat{\mu}, \widehat{\sigma}^2)$, where $\widehat{\mu} = \overline{X}_{n_0}$ and $\widehat{\sigma}^2 = S^2_{n_0}$ are the sample mean and variance calculated from the observed data. Put $\widehat{V}_j = \frac{S^{2(j)}_n}{n}$, where $S^{2(j)}_n$ is the sample variance calculated from the j-th completed sample. Then one can show that

$$\overline{V}_M \xrightarrow[M\to\infty]{\text{a.s.}} \frac{S^2_{n_0}}{n}. \tag{90}$$

Further let $\widehat{\theta}_j = \overline{Y}^{(j)}_n$ be the sample mean calculated from the j-th completed sample. Then it can be shown that

$$B_M \xrightarrow[M\to\infty]{\text{a.s.}} \frac{S^2_{n_0}(n-n_0)}{n^2}. \tag{91}$$

Now combining (90) and (91) yields that

$$\overline{V}_M + B_M \xrightarrow[M\to\infty]{\text{a.s.}} S^2_{n_0}\left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right).$$

Further it is straightforward to prove that for $n_0 < n$

$$S^2_{n_0}\left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right) < \frac{S^2_{n_0}}{n_0},$$

where the right-hand side of the above inequality represents the standard estimate of the variance of $\overline{X}_{n_0}$ (that assumes MCAR). This indicates that when doing multiple imputation, one needs to take into consideration also the variability that comes from the fact that one uses the estimates $\widehat{\mu}$, $\widehat{\sigma}^2$ instead of the true values of μ and σ. This can be done very naturally within the framework of Bayesian statistics.

Advantages and disadvantages:

+ If the missing values are estimated appropriately, it can give 'reasonable' estimate of the unknown parameter as well as of the variance of this estimate.

− To be done properly it requires the knowledge of Bayesian approach to statistics.

Re-weighting

Roughly speaking in this method each observation is given a weight (w_i) that is proportional to the inverse probability of being observed (π_i), i.e.

$$w_i = \frac{\frac{1}{\pi_i}}{\sum_{j:R_j=1} \frac{1}{\pi_j}}, \quad i \in \{j : R_j = 1\}.$$

All the procedures are now weighted with respect to these weights, e.g. the M-estimator of a parameter $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i:R_i=1} w_i \, \rho(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

Example 55. Suppose we have a study where for a large number of patients some basic and cheap measurements have been done resulting in $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$. Now a random subsample $\mathbb{S}$ of size n from these patients has been done for some more expensive measurements. Note that then $\mathbb{S} = \big\{j \in \{1, \ldots, N\} : R_j = 1\big\}$.

This method can be also used where one has some auxiliary variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ that can be used to estimate the probabilities π_i with the help of for instance a logistic regression.

The expected end of class 25 (21. 5. 2024)

Appendix

Inverse function theorem

The following theorem is sometimes also called the theorem about the local diffeomorphism. It follows easily from the implicit function theorem applied to the function $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{f}(\mathbf{y})$.

Theorem A13. *Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ have continuous first order partial derivatives in a neighbourhood of the point $\mathbf{a} \in \mathbb{R}^n$ and the Jacobi matrix $\mathbb{D}_{\mathbf{f}}(\mathbf{a})$ is a non-singular matrix. Then there exist open neighbourhoods U of the point $\mathbf{a}$ and V of the point $\mathbf{f}(\mathbf{a})$ such that $\mathbf{f}$ is a bijection of U on V. Further there exists an inverse function $\mathbf{f}^{-1}$ on V with the continuous first order partial derivatives.*

Lemma about the distribution of a quadratic form

The following lemma can be found as Theorem 4.16 in Anděl [2007].

Lemma A6. *Let $\mathbf{Z} \sim \mathsf{N}_p(\mathbf{0}_p, \mathbb{V})$, where $\mathbb{V}$ is $p \times p$ matrix. Let $\mathbb{B}$ be a positively semidefinite matrix such that $\mathbb{B}\mathbb{V}$ is an idempotent (nonzero) matrix. Then $\mathbf{Z}^\mathsf{T}\mathbb{B}\mathbf{Z} \sim \chi^2_{\mathrm{tr}(\mathbb{B}\mathbb{V})}$.*

Banach fixed point theorem

Definition. Let (P, ρ) be a metric space. Then a map $T : P \mapsto P$ is called a *contraction mapping* on P if there exists $q \in [0, 1)$ such that for all $x, y \in P$

$$\rho\big(T(x), T(y)\big) \leq q\rho(x, y).$$

Theorem A14. *Let (P, ρ) be a non-empty complete metric space with a contraction mapping $T : P \mapsto P$. Then T admits a unique fixed-point $x^* \in P$ (i.e. $T(x^*) = x^*$).*

Bayes theorem for densities

Theorem A15. *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_k)^\mathsf{T}$ and $\mathbf{Z} = (Z_1, \ldots, Z_G)^\mathsf{T}$ be random vectors defined on the same probability space. Let $f_{\boldsymbol{X}}$ and $f_{\mathbf{Z}}$ be the densities of $\boldsymbol{X}$ and $\mathbf{Z}$ respectively and $f_{\boldsymbol{X}|\mathbf{Z}}$ be the conditional density of $\boldsymbol{X}$ given $\mathbf{Z}$. Then the conditional density of $\mathbf{Z}$ given $\boldsymbol{X}$ equals*

$$f_{\mathbf{Z}|\boldsymbol{X}}(\mathbf{z}|\mathbf{x}) = \begin{cases} \frac{f_{\boldsymbol{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})}{f_{\boldsymbol{X}}(\mathbf{x})}, & \text{for } f_{\boldsymbol{X}}(\mathbf{x}) > 0, \\ 0, & \text{for } f_{\boldsymbol{X}}(\mathbf{x}) = 0. \end{cases}$$

Proof. The proof follows from the fact that $f_{\boldsymbol{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z}) = f_{\boldsymbol{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$ is the joint density of $\binom{\boldsymbol{X}}{\mathbf{Z}}$ and then by the definition of the conditional density. For details see e.g. Chapter 3.5 of Anděl [2007]. □

References

J. Anděl. *Základy matematické statistiky*. Matfyzpress, Praha, 2007.

N. L. Hjort and D. Pollard. Asymptotics for minimisers of convex processes. *arXiv preprint, arXiv:1107.3806*, 2011.

B. Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Annals of Mathematical Statistics*, 42:1977–1991, 1971.

J. Jiang. *Large sample techniques for statistics*. Springer Texts in Statistics. Springer, New York, 2010. ISBN 978-1-4419-6826-5. doi: 10.1007/978-1-4419-6827-2. URL http://dx.doi.org/10.1007/978-1-4419-6827-2.

R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005.

M. Kulich. Maximum likelihood estimation theory, 2014. URL https://www.karlin.mff.cuni.cz/~kulich/vyuka/glm/index.html.

P. Lachout. *Teorie pravděpodobnosti*. Karolinum, 2004. Skripta.

H. Leeb and B. M. Pötscher. Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer, New York, 1998.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics*. Wiley, Chichester, 2006.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 2008. Second Edition.

S. Nagy. Mathematical statistics 2. Course notes for NMSA332. URL https://www.karlin.mff.cuni.cz/~nagy/NMSA332/NMSA332.pdf.

Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

P. K. Sen, J. M. Singer, and A. C. P. de Lima. *From finite sample to asymptotic methods in statistics*. Cambridge University Press, 2010.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, 2000. ISBN 0521784506.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

C. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11: 95–103, 1983.

K. Zvára. *Regrese*. MATFYZPRESS, 2008.