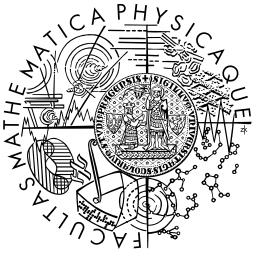

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MODERN STATISTICAL METHODS

NMST 434

Course notes

Contents

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 1. 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

$$\lim_{n\to\infty} F_{\boldsymbol{X}_n}(\mathbf{x}) = F_{\boldsymbol{X}}(\mathbf{x}) \tag{1}$$

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

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

Definition 2. 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

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

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

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

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

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

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

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

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) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\textit{alm. surely}} \boldsymbol{X} \Rightarrow \mathbf{g}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\textit{alm. surely}} \mathbf{g}(\boldsymbol{X})$;

(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.*

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

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

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

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

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

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

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 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

$$(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow[n\to\infty]{d} (\boldsymbol{X}, \mathbf{c}). \tag{2}$$

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

To prove (2) 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 3. 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) $\boldsymbol{X}_n = o_P(r_n)$, if $\frac{\boldsymbol{X}_n}{r_n} \xrightarrow[n\to\infty]{P} \mathbf{0}_k$, where $\mathbf{0}_k = (0, \ldots, 0)^\mathsf{T}$ is a zero point in $\mathbb{R}^k$;

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

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

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

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);

(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 $\boldsymbol{X}_n = O_P\big(\frac{1}{r_n}\big)$.

Remark 3. Further note that the calculus with (the possible random) quantities $o_p(1)$ and $O_P(1)$ is analogous to the calculus in 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) $o_P(1) + o(1) = o_P(1)$;

(v) $O_P(1) + O(1) = O_P(1)$.

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 $\boldsymbol{X}_n\, \boldsymbol{Y}_n$ makes sense. Let $\varepsilon > 0$ be given. Then one can find $K < \infty$ such that $\sup_{n\in\mathbb{N}} \mathsf{P}\left(\|\boldsymbol{X}_n\| > K\right) \le \frac{\varepsilon}{2}$. Thus for all sufficiently large $n \in \mathbb{N}$

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

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

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 an estimator of a p-dimensional parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\mathsf{T}$ and $\mathbf{g} = (g_1, \ldots, g_m)$ be a function from $\mathbb{R}^p \to \mathbb{R}^m$. Denote the Jacobi matrix of the function $\mathbf{g}$ at the point $\mathbf{x}$ as $\mathbb{D}_{\mathbf{g}}(\mathbf{x})$, i.e.

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

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 $\mathbf{g}$ has continuous first-order partial derivatives 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 $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \boldsymbol{\Sigma}\big)$, *then*

$$\sqrt{n}\,\big(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\mu})\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_m\big(\mathbf{0}_m, \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\boldsymbol{\Sigma}\,\mathbb{D}_{\mathbf{g}}^\mathsf{T}(\boldsymbol{\mu})\big). \tag{3}$$

Proof. Statement (i):. For $j = 1, \ldots, m$ 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 point $\boldsymbol{\mu}$ such that the function g_j has continuous partial derivatives in this neighbourhood and $\mathsf{P}\left(\mathbf{T}_n \in \mathcal{U}_\delta(\boldsymbol{\mu})\right) \xrightarrow[n\to\infty]{} 1$. Thus without loss of generality we 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

$$\begin{aligned} \sqrt{n}\,\big(g_j(\mathbf{T}_n) - g_j(\boldsymbol{\mu})\big) &= \nabla g_j(\boldsymbol{\mu}_n^{j*})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) \\ &= \nabla g_j(\boldsymbol{\mu})\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) + \big[\nabla g_j(\boldsymbol{\mu}_n^{j*}) - \nabla g_j(\boldsymbol{\mu})\big]\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}). \end{aligned} \tag{4}$$

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

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

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

$$\big[\nabla g_j(\boldsymbol{\mu}_n^{j*}) - \nabla g_j(\boldsymbol{\mu})\big]\,\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\mu}) = o_P(1). \tag{5}$$

Now combining (4) and (5) yields that for each $j = 1, \ldots, m$

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

which implies the first statement of the theorem.

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

$$\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\mu})\right) = \mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\sqrt{n}\left(\mathbf{T}_n - \boldsymbol{\mu}\right) + o_P(1).$$

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

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 $\boldsymbol{\mu}$.

Sometimes instead of (3) we write shortly $\mathbf{g}(\mathbf{T}_n) \overset{\text{as}}{\approx} \mathsf{N}_p\big(\mathbf{g}(\boldsymbol{\mu}), \frac{1}{n}\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\boldsymbol{\Sigma}\,\mathbb{D}_{\mathbf{g}}^{\mathsf{T}}(\boldsymbol{\mu})\big)$. The quantity $\frac{1}{n}\mathbb{D}_{\mathbf{g}}(\boldsymbol{\mu})\,\boldsymbol{\Sigma}\,\mathbb{D}_{\mathbf{g}}^{\mathsf{T}}(\boldsymbol{\mu})$ is then called the **asymptotic variance** matrix of $\mathbf{g}(\mathbf{T}_n)$ and it is denoted as $\mathsf{avar}\left(\mathbf{g}(\mathbf{T}_n)\right)$. Do not confuse the asymptotic variance and the variance of $\mathbf{g}(\mathbf{T}_n)$. As the following example shows these quantities can be substantially different.

Example 1. 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 2. 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 3. Suppose you have two independent random samples from Bernoulli distribution. Derive the asymptotic distribution of the logarithm of odds-ratio.

Example 4. Derive the asymptotic distribution of the standard (Pearson's) correlation coefficient.

Example 5. 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 6. Suppose that we observe $X_1, \dots, X_n$ 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 $\mathsf{var}(Y_t) = \sigma^2$.

Derive the asymptotic distribution of the estimator of θ given by

$$\widehat{\theta}_n = \frac{1 - \sqrt{1 - 4\,\widehat{r}_n^2(1)}}{2\,\widehat{r}_n(1)}\,,$$

where $\widehat{r}_n(1)$ is the sample autocorrelation function at lag 1.

Hint. Note that by Bartlett's formula

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

where

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

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 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^{\mathsf{T}} \in \Theta$. Let $\boldsymbol{\theta}_X$ be the true value of this unknown parameter. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be a random sample from this distribution and $t_1, \dots, t_p$ be given real functions. For instance if the observations are one-dimensional one can take $t_j(x) = x^j$, $j = 1, \dots, p$. For $j = 1, \dots, p$ define the function $\tau_j : \Theta \to \mathbb{R}$ as

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

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

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

Example 7. Moment estimation in Beta distribution

Put

$$\mathbf{T}_n = \Big(\frac{1}{n}\sum_{i=1}^{n} t_1(\boldsymbol{X}_i), \dots, \frac{1}{n}\sum_{i=1}^{n} t_p(\boldsymbol{X}_i)\Big)^{\mathsf{T}} \tag{6}$$

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

$$\sqrt{n}\,\Big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\Big) = \sqrt{n}\,\Big(\boldsymbol{\tau}^{-1}(\mathbf{T}_n) - \boldsymbol{\tau}^{-1}\big(\boldsymbol{\tau}(\boldsymbol{\theta}_X)\big)\Big). \tag{7}$$

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 $\max_{j=1,\dots,p} \mathsf{var}_{\boldsymbol{\theta}}(t_j(\boldsymbol{X}_1)) < \infty$. Further let the function $\boldsymbol{\tau}$ have continuous first-order partial derivatives in a neighbourhood of $\boldsymbol{\theta}_X$ and the Jacobi matrix $\mathbb{D}_{\boldsymbol{\tau}}(\boldsymbol{\theta}_X)$ is regular. Then*

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

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

Proof. By the assumptions of the theorem and the implicit function theorem there exists an open neighbourbood U containing $\boldsymbol{\theta}_X$ and an open neighbourbood V containing $\boldsymbol{\tau}(\boldsymbol{\theta}_X)$ such that $\boldsymbol{\tau} : U \mapsto V$ is a differentiable bijection with a differentiable inverse $\boldsymbol{\tau}^{-1} : V \mapsto U$. Further note that $\mathbf{T}_n$ defined in (6) satisfies $\mathsf{P}\left(\mathbf{T}_n \in V\right) \xrightarrow[n\to\infty]{} 1$. Thus one can use (7) and apply the $\boldsymbol{\Delta}$-theorem (Theorem 3) with $\mathbf{g} = \boldsymbol{\tau}^{-1}$, $\boldsymbol{\mu} = \boldsymbol{\tau}(\boldsymbol{\theta}_X)$ and $A = V$ to get

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

The statement of the theorem now follows from the identity

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

□

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

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

Alternatively the matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_X)$ can be also estimated as

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

where $\mathbf{Z}_i = \big(t_1(\boldsymbol{X}_i), \dots, t_p(\boldsymbol{X}_i)\big)^{\mathsf{T}}$.

Example 8. Let $X_1, \dots, X_n$ be independent identically distributed random variables from the discrete distribution given as

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

where $p \in (0, \frac{1}{2})$. Consider the moment estimator of the parameter p and derive its asymptotic distribution. Based on these results derive the confidence interval for the parameter p.

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

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

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

Asymptotic confidence interval of 'Wald' type

This interval is based on the fact that

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

and thus

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

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

Asymptotic confidence interval of 'Wilson' type

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

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

Asymptotic variance stabilising transformation

Let the function g be such that $[g'(\theta)]^2\,\sigma^2(\theta)$ does not depend on θ. Put $v^2 := [g'(\theta)]^2\,\sigma^2(\theta)$. Then with the help (8) and Δ-theorem it holds $\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta_X)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, v^2\big)$. Thus

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

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

Example 9. 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 10. Fisher Z-transformation and various confidence intervals for the correlation coefficient.

Example 11. 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.

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 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

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}).$$

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 $I(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}} \left[\frac{\partial \log f(\boldsymbol{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, \frac{\partial \log f(\boldsymbol{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} \right]$ 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 surely if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

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

[R2] The support set $S = \{\mathbf{x} \in \mathbb{R}^k : f(\mathbf{x}; \boldsymbol{\theta}) > 0\}$ 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 $\boldsymbol{\theta}_X$. Further for each j, k, l in $\{1, \ldots, p\}$ there exists s function $M_{jkl}(\mathbf{x})$ such that

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

for μ-almost all $\mathbf{x}$ and

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

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

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

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

where $h(\mathbf{x};\boldsymbol{\theta})$ is either $f(\mathbf{x};\boldsymbol{\theta})$ or $\partial f(\mathbf{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

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

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

see for instance Theorem 7.27 of Anděl (2007).

Remark 5. Note that in particular assumption **[R4]** is rather strict. There are ways how to arrive at 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 likelihood equations

$$\mathbf{U}_n\big(\widehat{\boldsymbol{\theta}}_n\big) \stackrel{!}{=} \mathbf{0}. \tag{12}$$

Further define *the observed information matrix* as

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

where

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

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 the matrix* I_n^* *be a matrix with the dimension* $p \times p$ *and with the elements*

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

where $\widehat{\mathbf{t}}_n^{(jk)} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ *for each* $j,k \in \{1,\ldots,p\}$. *Then*

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

Proof. Fix $j,k \in \{1,\ldots,p\}$. Put

$$i_{n,jk}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{-\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k},$$

and let $i_{jk}(\boldsymbol{\theta}_X)$ be the (j,k) element of the Fisher information matrix $I(\boldsymbol{\theta}_X)$. Note that one can bound

$$\left|i_{n,jk}^* - i_{jk}(\boldsymbol{\theta}_X)\right| \le \left|i_{n,jk}^* - i_{n,jk}(\boldsymbol{\theta}_X)\right| + \left|i_{n,jk}(\boldsymbol{\theta}_X) - i_{jk}(\boldsymbol{\theta}_X)\right| \tag{13}$$

The second term on the right-hand side of (13) converges in probability to zero by the law of large numbers. Now with the help of assumption [R4] the first term on the right-hand side of (13) can be bounded by

$$\begin{aligned}
\left|i_{n,jk}^* - i_{n,jk}(\boldsymbol{\theta}_X)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\Big|_{\boldsymbol{\theta}=\widehat{\mathbf{t}}_n^{(jk)}} - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_X}\right| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{p} M_{jkl}(\boldsymbol{X}_i)\left|\widehat{t}_{nl}^{(jk)} - \theta_{Xl}\right| \\
&= o_P(1)\,\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{p} M_{jkl}(\boldsymbol{X}_i) \;= o_P(1)\,O_p(1) = o_P(1),
\end{aligned}$$

where $\widehat{t}_{nl}^{(jk)}$ is the l-th element of $\widehat{\mathbf{t}}_n^{(jk)}$. □

Theorem 5. *Suppose that assumptions* [R0]-[R6] *hold. Then with probability tending to one as* $n \to \infty$ *there exists a consistent solution* $\widehat{\boldsymbol{\theta}}_n$ *of the likelihood equations* (12) *such that*

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

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. First, we need to prove the consistency, that is $\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$. This can be found in the proof of Theorem 5.1 of Lehmann and Casella (1998, Chapter 6).

Once the consistency of $\widehat{\boldsymbol{\theta}}_n$ is proved then by the mean value theorem (applied to each component of $\mathbf{U}_n(\boldsymbol{\theta})$) one gets that

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

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 Lemma 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

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

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

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

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

□

Remark 6. While the proof of consistency is for $p = 1$ relatively simple, 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.

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 $\mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$, and $\mathbf{T}_n$ be an *unbiased* estimator of $\boldsymbol{\theta}_X$ (based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$). Then

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

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

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

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

Remark 7. 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. 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.

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

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

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

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

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

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

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

for μ-almost all $\mathbf{x}$ and

$$\mathsf{E}_{\boldsymbol{\theta}_X} M_{jk}^2(\boldsymbol{X}_1) < \infty.$$

Example 12. Random sample from a uniform distribution.

Example 13. 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 $\widehat{\beta}_n$.

Example 14. 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.

Example 15. 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. Then the maximum likelihood estimator of θ is not consistent.

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

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

Wald test is based on the test statistic

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

Rao score test is based on the test statistic

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

Theorem 6. *Suppose that the null hypothesis holds, assumptions* [R0]-[R6] *are satisfied and* $\widehat{I}_n \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_0)$. *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

$$R_n = \Big(\tfrac{1}{\sqrt{n}}\,\hat{I}_n^{-\frac{1}{2}}\,\mathbf{U}_n(\boldsymbol{\theta}_0)\Big)^{\mathsf{T}}\Big(\tfrac{1}{\sqrt{n}}\,\hat{I}_n^{-\frac{1}{2}}\,\mathbf{U}_n(\boldsymbol{\theta}_0)\Big). \tag{19}$$

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

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

where $\mathbb{I}_p$ is an identity matrix of dimension $p\times p$. Now the statement follows by using CMT (Theorem 1).

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

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

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

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

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

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

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

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

□

Remark 8. 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.

2.5 Asymptotic confidence sets

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

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

where $\hat{I}_n$ 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 $\hat{I}_n$. 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

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

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

$$\left(\widehat{\theta}_{jn} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{i_n^{jj}}{n}},\ \widehat{\theta}_{jn} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{i_n^{jj}}{n}}\right). \tag{20}$$

Remark 9. 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 $\boldsymbol{\theta}$ as

$$\left\{\boldsymbol{\theta} :\ 2\left(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n(\boldsymbol{\theta})\right) \leq \chi^2_p(1-\alpha)\right\}.$$

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 $(1 \leq q < p)$ 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 $H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$ against $H_1 : \boldsymbol{\tau}_X \neq \boldsymbol{\tau}_0$.

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 $p-q$ 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

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

Lemma 2. *Let $\mathbb{J}$ be a symmetric regular 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 $\mathbb{J}\,\mathbb{J}^{-1}$ and use the fact that by the symmetry of the matrix $\mathbb{J}$ it holds that $\mathbb{J}_{12} = \mathbb{J}_{21}^{\mathsf{T}}$. □

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.

$$\widetilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \boldsymbol{\tau}_0 \\ \widetilde{\boldsymbol{\psi}}_n \end{pmatrix}, \quad \text{where} \quad \widetilde{\boldsymbol{\psi}}_n = \underset{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}}{\arg\max}\, L_n(\boldsymbol{\tau}_0, \boldsymbol{\psi}).$$

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

$$LR_n^* = 2\left(\ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) - \ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big)\right). \tag{22}$$

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

$$R_n^* = \frac{1}{n}\, \mathbf{U}_{1n}^{\mathsf{T}}\big(\widetilde{\boldsymbol{\theta}}_n\big)\, \widehat{I}_n^{11}\, \mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big). \tag{23}$$

Remark 10. As $\mathbf{U}_{2n}\big(\widetilde{\boldsymbol{\theta}}_n\big) = \mathbf{0}_{p-q}$, the test statistic of the Rao score test can be also written in a form

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

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

Theorem 7. *Suppose that the null hypothesis holds, assumptions* **[R0]**-**[R6]** *are satisfied and* $\widehat{I}_n^{11} \xrightarrow[n\to\infty]{P} I^{11}(\boldsymbol{\theta}_X)$. *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 $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(0, I^{-1}(\boldsymbol{\theta}_X)\big)$, which yields

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

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

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

which further with the CMT (Theorem 1) implies

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

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

$$\frac{1}{\sqrt{n}}\, \mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) = \frac{1}{\sqrt{n}}\, \mathbf{U}_{1n}(\boldsymbol{\theta}_X) - I_{12n}^*\, \sqrt{n}\left(\widetilde{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_X\right), \tag{24}$$

where I_{12n}^* is 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 $\boldsymbol{\theta}_n^{j*} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$, Lemma 1 implies that

$$I_{12n}^* \xrightarrow[n\to\infty]{P} I_{12}(\boldsymbol{\theta}_X). \tag{25}$$

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

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

As the null hypothesis holds, using Theorem 5 one gets

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

Combining (24), (25) and (26) yields

$$\frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) = \frac{1}{\sqrt{n}}\,\mathbf{U}_{1n}(\boldsymbol{\theta}_X) - I_{12}(\boldsymbol{\theta}_X) I_{22}^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\,\mathbf{U}_{2n}(\boldsymbol{\theta}_X) + o_P(1). \tag{27}$$

Now using (27) 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 $\sqrt{n}\,\mathbf{U}_{1n}\big(\widetilde{\boldsymbol{\theta}}_n\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_q\left(0, \big[I^{11}(\boldsymbol{\theta}_X)\big]^{-1}\right)$, 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

$$\ell_n\big(\widetilde{\boldsymbol{\theta}}_n\big) = \ell_n\big(\widehat{\boldsymbol{\theta}}_n\big) + \underbrace{\mathbf{U}_n^{\mathsf{T}}\big(\widehat{\boldsymbol{\theta}}_n\big)}_{=\mathbf{0}_p}\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big) - \frac{n}{2}\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big)^{\mathsf{T}} I_n(\boldsymbol{\theta}_n^*)\big(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\big), \tag{28}$$

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 Lemma 1 implies $I_n(\boldsymbol{\theta}_n^*) \xrightarrow[n\to\infty]{P} I(\boldsymbol{\theta}_X)$.

Further by Theorem 5

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big) = I^{-1}(\boldsymbol{\theta}_X)\,\frac{1}{\sqrt{n}}\,\mathbf{U}_n(\boldsymbol{\theta}_X) + o_P(1), \tag{29}$$

which together with (26) implies

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

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 $\mathbb{A}(\boldsymbol{\theta}_X)$

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(0, \mathbb{A}(\boldsymbol{\theta}_X)\,I(\boldsymbol{\theta}_X)\,\mathbb{A}(\boldsymbol{\theta}_X)\big). \tag{30}$$

Now we will use the following lemma (Anděl, 2007, Theorem 4.16).

Lemma 3. *Let $\mathbf{Z} \sim \mathsf{N}_p(\mathbf{0}_p, \mathbb{V})$, where $\mathbb{V}$ is $p \times p$ matrix. Let $\mathbb{B}\mathbb{V}$ be an idempotent (nonzero) matrix. Then $\mathbf{Z}^\mathsf{T}\mathbb{B}\mathbf{Z} \sim \chi^2_{\mathrm{tr}(\mathbb{B}\mathbb{V})}$.*

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

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

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 (28), (30), CS (Theorem 2), Lemma 3 and CMT (Theorem 1) one gets

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

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

Example 16. Breusch-Pagan test of heteroscedasticity

Example 17. Suppose that you observe independent identically distributed random vectors $(\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$ such that

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

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

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

(ii) Find the confidence interval for the parameter β.

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

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, ...

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

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

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

Further denote

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

and define the profile score function 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 $I_n(\boldsymbol{\theta})$.

Lemma 4. *Suppose that assumptions* [R0]-[R6] *are satisfied. Then (with probability tending to one)*

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

where $I_{jkn}(\boldsymbol{\theta})$ *(for* $j,k \in \{1,2\}$*) were introduced in* (21).

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

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

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

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

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 (22). 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 4

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

$\underline{R_n^{(p)}}$: By Lemma 4 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 $\hat{I}_n^{(11)} = \big[\hat{I}_n^{(p)}\big]$, where R_n^* is Rao score test statistic in the presence of nuisance parameters defined in (23). 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 (20), 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 18. 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 19. Box-Cox transformation. See Zvára (2008) pp. 149–151.

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

Let observations $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ 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

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

the *maximum likelihood estimator* (of parameter $\boldsymbol{\theta}_X$) as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}),$$

the score function as

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

and *the 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

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

In 'nice (regular) models' it holds that

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

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

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

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.

Example 20. Suppose we have K independent samples, that is for each $i = 1, \ldots, K$ the random variables $\boldsymbol{X}_{ij}, j = 1, \ldots, n_i$ are independent and identically distributed with density $f_i(\mathbf{x}; \boldsymbol{\theta})$ (with respect to some σ-finite measure μ). Further let all the random variables be

independent and let $\lim_{n\to\infty} \frac{n_i}{n} = w_i$, where $n = n_1 + \ldots + n_K$. Then

$$\begin{aligned}
L_n(\boldsymbol{\theta}) &= \prod_{i=1}^{K}\prod_{j=1}^{n_i} f_i(\boldsymbol{X}_{ij};\boldsymbol{\theta}),\\
\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^{K}\sum_{j=1}^{n_i} \log f_i(\boldsymbol{X}_{ij};\boldsymbol{\theta}),\\
\mathbf{U}_n(\boldsymbol{\theta}) &= \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{\partial \log f_i(\boldsymbol{X}_{ij};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},\\
I_n(\boldsymbol{\theta}) &= -\frac{1}{n}\frac{\partial \mathbf{U}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = -\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{\partial^2 \log f_i(\boldsymbol{X}_{ij};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\mathsf{T}}\\
\overline{I}(\boldsymbol{\theta}) &= \lim_{n\to\infty} \mathsf{E}\, I_n(\boldsymbol{\theta}) = \lim_{n\to\infty} \sum_{i=1}^{K} \underbrace{\frac{n_i}{n}}_{\to w_i} I^{(i)}(\boldsymbol{\theta}) = \sum_{i=1}^{K} w_i\, I^{(i)}(\boldsymbol{\theta}),
\end{aligned}$$

where $I^{(i)}(\boldsymbol{\theta})$ is Fisher information matrix of $\boldsymbol{X}_{i1}$ (i.e. for the density $f_i(\mathbf{x};\boldsymbol{\theta})$).

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 variables. Thus (in the most simple situation) we assume that we observe independent and identically distributed random vectors

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

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

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

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 $f(y_1|\mathbf{x}_1;\boldsymbol{\theta})$, ...,

$f(y_n|\mathbf{x}_n;\boldsymbol{\theta})$. 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 21. *(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 $\lambda(\boldsymbol{X}_i) = \exp\{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}_i\}$ and $\boldsymbol{X}_i = (X_{i1},\ldots,X_{ip})^{\mathsf{T}}$. Then (provided assumptions **[R0]**-**[R6]** are satisfied)

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

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

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

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

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

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 22. Maximum likelihood estimation in AR(1) process.

Example 23. A comparison of K independent binomial distributions.

Literature: Hoadley (1971).

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 (where q is fixed) and $\boldsymbol{\psi}^{(n)}$ containing the remaining $p_n - q$ components.

Example 24. *Strongly stratified sample* Let Y_{ij}, $i = 1, \ldots, N$, $j = 1, 2$ 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?

Note that in the previous example 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 $\boldsymbol{\psi} = (\mu_1, \ldots, \mu_N)^\mathsf{T}$ is increasing to infinity (too quickly). Marginal and conditional likelihoods are two attempts to modify the likelihood so that it yields consistent (and hopefully also asymptotically normal) estimators of the parameters of interest $\boldsymbol{\tau}$.

Suppose that the data $\mathbb{X}$ can be transformed (or simply decomposed) to $\mathbb{V}$ and $\mathbb{W}$.

Let the distribution of $\mathbb{V}$ depends only on parameter $\boldsymbol{\tau}$ (and not on $\boldsymbol{\psi}^{(n)}$). Then *the marginal (log-)likelihood* of parameter $\boldsymbol{\tau}$ is defined as

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

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

Let the conditional distribution of $\mathbb{V}$ given $\mathbb{W}$ depends only on parameter $\boldsymbol{\tau}$ (and not on $\boldsymbol{\psi}^{(n)}$). Then *the conditional (log-)likelihood* of parameter $\boldsymbol{\tau}$ is defined as

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

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

Remark 11. (i) If $\mathbb{V}$ is independent of $\mathbb{W}$, then $p(\mathbb{V}|\mathbb{W}; \boldsymbol{\tau}) = p(\mathbb{V}; \boldsymbol{\tau})$ and thus $L_n^{(M)}(\boldsymbol{\tau}) = L_n^{(C)}(\boldsymbol{\tau})$.

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

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

where $\ell_n(\boldsymbol{\tau}, \boldsymbol{\psi}^{(n)})$ is the log-likelihod of $(\mathbb{V}, \mathbb{W})$ and $\ell_{n,\mathbb{W}}(\boldsymbol{\tau}, \boldsymbol{\psi}^{(n)})$ is the log-likelihod of $\mathbb{W}$.

(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.

Exponential family

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

$$p(\mathbf{x}; \boldsymbol{\tau}, \boldsymbol{\psi}^{(n)}) = \exp\left\{\sum_{j=1}^{q} Q(\boldsymbol{\tau})\, T_j(\mathbf{x}) + \sum_{j=1}^{p_n - q} R(\boldsymbol{\tau}, \boldsymbol{\psi}^{(n)})\, S_j(\mathbf{x})\right\} a\big(\boldsymbol{\tau}, \boldsymbol{\psi}^{(n)}\big)\, h(\mathbf{x}),$$

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 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 $\mathbb{W}$ and $\mathbb{X}$ as $\mathbb{V}$.

Example 25. *Strongly stratified sample (cont.).* Using marginal and conditional likelihood.

Example 26. Y_{ij}, $i = 1, \ldots, I$, $j = 0, 1$ 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 Section 2.6 requires 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

$$R_n^{(c)} = \frac{\Big(\sum_{i=1}^I Y_{i1} - \mathsf{E}_{H_0}[Y_{i1} \,|\, Y_{i+}]\Big)^2}{\sum_{i=1}^I \mathsf{var}_{H_0}[Y_{i1} \,|\, Y_{i+}]} = \frac{\Big(\sum_{i=1}^I Y_{i1} - Y_{i+}\,\frac{n_{i1}}{n_{i+}}\Big)^2}{\sum_{i=1}^I Y_{i+}\,\frac{n_{i1}n_{i0}}{n_{i+}^2}\,\frac{n_{i+}-Y_{i+}}{n_{i+}-1}}, \tag{38}$$

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 27. Consider in Example 26 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 26 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

$$\mathsf{P}_\tau\left(Y_1 = k \,|\, Y_+ = y_+\right) = \frac{\binom{n_1}{k}\binom{n_0}{y_+-k}\mathrm{e}^{\tau k}}{\sum_{l\in\mathcal{K}}\binom{n_1}{l}\binom{n_0}{y_+-l}\mathrm{e}^{\tau l}}, \quad k \in \mathcal{K}, \tag{39}$$

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

$$\mathrm{p}(\tau_0) = 2 \min \big\{ \mathsf{P}_{\tau_0}(Y_1 \leq y_1 \,|\, Y_+ = y_+), \mathsf{P}_{\tau_0}(Y_1 \geq y_1 \,|\, Y_+ = y_+) \big\}, \tag{40}$$

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 for τ as the set of those values for which we do not reject the null hypothesis, i.e.

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

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

The special case presents testing the null hypothesis $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. Then (39) 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 (40) 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 (40).

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

y_0	y_1
$n_0 - y_0$	$n_1 - y_1$

,

which is an asymptotic test.

Example 28. Consider in Example 26 the special case $n_{i0} = n_{i1} = 1$ for each $i = 1, \ldots, I$. Introduce

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

Then the test statistic (38) simplifies to

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

which is known as McNemar's test.

Literature: Pawitan (2001) Chapters 10.1–10.5.

3 M- and Z-estimators

M-estimator

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from a distribution F and one is interested in estimating some quantity (p-dimensional parameter) of this distribution, say $\boldsymbol{\theta}_X = \boldsymbol{\theta}(F)$. Let ρ be a function defined on $S_{\boldsymbol{X}} \times \Theta$, where $S_{\boldsymbol{X}}$ is the support of F and Θ is a set of possible values of $\boldsymbol{\theta}(F)$ for different distributions F (parameter space). The **M-estimator** is defined as

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

Note that the maximum likelihood estimator can be viewed as an M-estimator with

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

For regression problems when one observes $\mathbf{Z}_1 = (\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \ldots, \mathbf{Z}_n = (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$, one can view the least square (LS) estimator of regression parameters as an M-estimator with

$$\rho(\mathbf{z}; \boldsymbol{\beta}) = \rho(\mathbf{x}, y; \boldsymbol{\beta}) = \left(y - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}\right)^2.$$

Also the least absolute deviation (LAD) estimator can be viewed as an M-estimator with

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

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

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}\left(\boldsymbol{X}_i; \widehat{\boldsymbol{\theta}}_n\right) = \mathbf{0}_p, \qquad \text{where} \qquad \boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \rho(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{41}$$

Note that

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

Generally let $\boldsymbol{\psi}$ be a p-dimensional vector function (not necessarily a derivative of some function ρ) defined on $S_{\boldsymbol{X}} \times \Theta$. Then we define the Z-estimator as the solution of the system of equations (41).

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

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

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 29. 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 $\mathcal{F} = \{f(\mathbf{x};\boldsymbol{\theta}),\, \boldsymbol{\theta}\in\Theta\}$ we are estimating (identifying) $\boldsymbol{\theta}_X$ such that

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

Now by Jensen's inequality

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

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. 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

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

Thus $\boldsymbol{\theta}_X$ is the point of parameter space Θ for which the Kullback–Leibler divergence of $f(\mathbf{x})$ from $\mathcal{F}$ is minimised.

3.2 Asymptotic distribution of M/Z-estimators

Put $M(\boldsymbol{\theta}) = \mathsf{E}\,\rho(\boldsymbol{X}_1;\boldsymbol{\theta})$, $\mathbf{Z}(\boldsymbol{\theta}) = \mathsf{E}\,\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta})$ and $\mathbb{D}_{\boldsymbol{\psi}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\intercal}$ (the Jacobi matrix of $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$). Further let $\boldsymbol{X}_1$ has density f with respect to a σ-finite measure μ.

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.* For M-estimators $\boldsymbol{\theta}_X$ is a unique maximizer of the function $M(\boldsymbol{\theta})$. For Z-estimators for any $\delta > 0$ there exists $\varepsilon > 0$ such that $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_X\|\geq\delta} \left\|\mathbf{Z}(\boldsymbol{\theta})\right\| \geq \varepsilon$.

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

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

$$\left|\frac{\partial\psi_j(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_k} - \frac{\partial\psi_j(\mathbf{x};\boldsymbol{\theta}_X)}{\partial\theta_k}\right| \leq M_{jk}(\mathbf{x}) \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_X\right\|^\alpha$$

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

[Z5] The matrix

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathsf{E}\, \mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_1;\boldsymbol{\theta}) \tag{42}$$

is finite and *positive definite* in a neighbourhood of $\boldsymbol{\theta}_X$.

[Z6] The *variance matrix*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_X) = \mathsf{var}\left(\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\right) = \mathsf{E}\left[\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\, \boldsymbol{\psi}^\mathsf{T}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\right] \tag{43}$$

is finite.

Theorem 9. *Suppose that assumptions* [Z0]-[Z6] *are satisfied. Then with probability going to oner there exists a solution* $\widehat{\boldsymbol{\theta}}_n$ *to the estimating equations* (41) *such that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X)\, \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) + o_P(1), \tag{44}$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\Big(\mathbf{0}, \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X)\, \boldsymbol{\Sigma}(\boldsymbol{\theta}_X)\, \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X)\Big), \tag{45}$$

where the matrices $\boldsymbol{\Gamma}(\boldsymbol{\theta}_X)$ *and* $\boldsymbol{\Sigma}(\boldsymbol{\theta}_X)$ *are defined in* (42) *and* (43) *respectively.*

Proof. Consistency: For an M-estimator this can be proved analogously as the existence of a consistent solution of the maximum likelihood equations, see the proof of Theorem 5.1 of Lehmann and Casella (1998, Chapter 6). For a Z-estimator one can adapt the proof of Theorem 5.1 of Jurečková et al. (2012).

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

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

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

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

evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\widehat{\boldsymbol{\theta}}_n$ a $\boldsymbol{\theta}_X$. Further, similarly as in the proof of Lemma 1 using the law of large numbers (for i.i.d. vectors) one gets

$$\boldsymbol{\Gamma}_n^* - \boldsymbol{\Gamma}(\boldsymbol{\theta}_X = \Big(\boldsymbol{\Gamma}_n^* - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_X)\Big) + \Big(\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_X) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_X)\Big) = \Big(\boldsymbol{\Gamma}_n^* - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_X)\Big) + o_P(1). \tag{46}$$

Now using assumption [Z4] and the law of large numbers one can bound the (j,k)-element of the matrix on the right hand side of (46) as

$$\Big|\Big(\boldsymbol{\Gamma}_n^* - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_X)\Big)_{jk}\Big| \le \frac{1}{n}\sum_{i=1}^{n} M_{jk}(\boldsymbol{X}_i)\,\|\boldsymbol{\theta}_n^{j*} - \boldsymbol{\theta}_X\|^{\alpha} = O_P(1)\,o_P(1) = o_P(1).$$

This combined with (46) implies $\boldsymbol{\Gamma}_n^* \xrightarrow[n\to\infty]{P} \boldsymbol{\Gamma}(\boldsymbol{\theta}_X)$. Now with the help CS (Theorem 2) one can write

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

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

Asymptotic variance estimations

Note that by Theorem 9 one has

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

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

$$\widehat{\mathsf{avar}\,(\widehat{\boldsymbol{\theta}}_n)} = \frac{1}{n}\,\boldsymbol{\Gamma}_n^{-1}\,\boldsymbol{\Sigma}_n\,\boldsymbol{\Gamma}_n^{-1}, \tag{47}$$

where

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

Note that in the same way as in the proof of Theorem 9 one can show that by [Z4] it follows that

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

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

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

To arrive at such assumptions rewrite

$$\mathbf{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \left[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\right]\left[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\right]^{\mathsf{T}} \tag{48}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\left[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\right]^{\mathsf{T}} \tag{49}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \left[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\right]\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) \tag{50}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\,\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X). \tag{51}$$

Now by the law of large numbers the quanity in (51) converges in probability to $\mathbf{\Sigma}(\boldsymbol{\theta}_X)$, 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 $\mathbf{V}_n = \mathbf{\Gamma}_n^{-1}\,\mathbf{\Sigma}_n\,\mathbf{\Gamma}_n^{-1}$ is a consistent estimator of $\mathbf{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\mathbf{\Sigma}(\boldsymbol{\theta}_X)\,\mathbf{\Gamma}^{-1}(\boldsymbol{\theta}_X)$.

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

$$\Big\{\boldsymbol{\theta} :\, n\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big)^{\mathsf{T}}\mathbf{V}_n^{-1}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \le \chi_p^2(1-\alpha)\Big\}.$$

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

$$\left[\widehat{\theta}_{nk} - \tfrac{u_{1-\alpha/2}\sqrt{v_n^{kk}}}{\sqrt{n}},\ \widehat{\theta}_{nk} + \tfrac{u_{1-\alpha/2}\sqrt{v_n^{kk}}}{\sqrt{n}}\right],$$

where $\widehat{\theta}_{nk}$ is the k-th coordinate of $\widehat{\boldsymbol{\theta}}_n$ and v_n^{kk} is the k-th diagonal element of the matrix $\mathbf{V}_n$.

Literature: Sen et al. (2010) Chapter 8.2., White (1980)

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). From Example 29 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

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

Further by Theorem 9 we also know that

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

Suppose that our parametric assumption is right and $f \in \mathcal{F}$, i.e. there exists $\boldsymbol{\theta}_X \in \Theta$ such that $f(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_X)$. Then $\boldsymbol{\Gamma}(\boldsymbol{\theta}_X) = \boldsymbol{\Sigma}(\boldsymbol{\theta}_X) = I(\boldsymbol{\theta}_X)$ and thus $\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X)\boldsymbol{\Sigma}(\boldsymbol{\theta}_X)\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X) = I^{-1}(\boldsymbol{\theta}_X)$. Thus 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 M-estimator with $\rho(\mathbf{x};\boldsymbol{\theta}) = -\log f(\mathbf{x};\boldsymbol{\theta})$. The asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ can now be estimated with the help of 'sandwich estimator' (47) where

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

Example 30. *Misspecified Poisson regression* Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathsf{T}}$ and

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

be independent and identically distributed random vectors. Assume that $Y_i|\boldsymbol{X}_i \sim \mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, where $\lambda(\mathbf{x}) = \mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathsf{T}}$. The score function for the maximum likelihood estimation is given by

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

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

$$\boldsymbol{\psi}(\mathbf{x}, y; \boldsymbol{\beta}) = \mathbf{x}\left(y - \mathrm{e}^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}}\right) \tag{52}$$

and $\boldsymbol{\beta}_X$ solves the equation

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

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

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

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 $Y_i|\boldsymbol{X}_i$ is $\mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, but we are willing to assume that $\mathsf{E}\,[Y_i|\boldsymbol{X}_i] = \mathrm{e}^{\boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X}_i}$, we can still use the score function (52) 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 $\boldsymbol{\Gamma}(\boldsymbol{\beta}_X)$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}_X)$ given by

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

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

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

where

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

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

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

where for each fixed $\mathbf{x}$ the functions $\rho(\mathbf{x};\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$.

Let estimate the parameter $\boldsymbol{\theta}_X$ as

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

The function $\rho(\mathbf{x};\boldsymbol{\theta})$ does not have to be smooth in $\boldsymbol{\theta}$, but it needs to be differentiable at least almost everywhere. Thus we suppose that there exists a function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ such that $\mathsf{E}\,\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X) = \mathbf{0}_p$ and one can write

$$\rho(\mathbf{x};\boldsymbol{\theta}_X + \mathbf{t}) - \rho(\mathbf{x};\boldsymbol{\theta}_X) = \mathbf{t}^\mathsf{T}\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}_X) + R(\mathbf{x};\mathbf{t}). \tag{53}$$

Theorem 10. *Suppose that (53) holds and that*

(i) there exists a positive definitive matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta}_X)$ such that

$$\mathsf{E}\left[\rho(\boldsymbol{X}_1;\boldsymbol{\theta}_X + \mathbf{t}) - \rho(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\right] = \tfrac{1}{2}\,\mathbf{t}^\mathsf{T}\boldsymbol{\Gamma}(\boldsymbol{\theta}_X)\,\mathbf{t} + o(\|\mathbf{t}\|^2), \;\; \textit{as}\;\; \mathbf{t}\to\mathbf{0}_p;$$

(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 $\boldsymbol{\Sigma}(\boldsymbol{\theta}_X) = \mathsf{var}\left(\boldsymbol{\psi}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)\right)$.

Then the statements of Theorem 9 holds

Proof. See proof of Theorem 2.1 of Hjort and Pollard (2011). □

Note that if assumptions **[Z3]** and **[Z4]** hold then also assumptions (i) and (ii) of Theorem 10 are satisfied. The nice thing about Theorem 10 is that the matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta}_X)$ does not have to be computed as $\boldsymbol{\Gamma}(\boldsymbol{\theta}_X) = \mathsf{E}\,\mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_1;\boldsymbol{\theta}_X)$ but one can compute it 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 smooth assumptions about $\boldsymbol{\psi}$ can be replaced with the smooth assumptions on the distribution of $\boldsymbol{X}_1$ so that the function $M(\boldsymbol{\theta})$ is sufficiently smooth.

Application to LAD regression

Suppose independent and identically distributed random vectors $\mathbf{Z}_1 = (\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \ldots, \mathbf{Z}_n = (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$ are observed. The LAD estimator of parameter $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \mathbf{b}^\mathsf{T}\boldsymbol{X}_i\right|.$$

To formulate the result we will assume the following (strict linear) model.

(M) The observations satisfy

$$Y_i = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_i + \varepsilon_i, \qquad i = 1,\ldots,n, \tag{54}$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are independent identically distributed random variables that are independent of $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$.

Theorem 11. *Let the model* **(M)** *hold. Further, let* $\mathsf{E}\,\|\boldsymbol{X}_1\|^3 < \infty$*, the matrix* $\mathsf{E}\,\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}$ *be positive definite,* ε_1 *have a zero median and the density* $f_\varepsilon(x)$ *of* ε_1 *be positive and continuous in a neighbourhood of* 0*. Then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathsf{E}\,\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\right]^{-1}\frac{\boldsymbol{X}_i\,\operatorname{sign}(\varepsilon_i)}{2\,f_\varepsilon(0)} + o_P(1),$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\Big(\mathbf{0}, \left[\mathsf{E}\,\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\right]^{-1}\tfrac{1}{4\,f_\varepsilon^2(0)}\Big).$$

Proof. We show that the assumptions of Theorem 10 are satisfied. First

Identification of $\boldsymbol{\beta}$: Note that $\boldsymbol{\beta}_X = \arg\min_{\mathbf{b}} g(\mathbf{b})$, where $g(\mathbf{b}) = \mathsf{E}\,|Y_1 - \mathbf{b}^\mathsf{T}\boldsymbol{X}_1|$. Now with the help of independence of ε_1 and $\boldsymbol{X}_1$ (we will write shortly $\boldsymbol{X}_1 \perp \varepsilon_1$)

$$\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathsf{E}\left[\operatorname{sign}(Y_1 - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_1)\,(-\boldsymbol{X}_1)\right] = -\mathsf{E}\left[\operatorname{sign}(\varepsilon_1)\,\boldsymbol{X}_1\right] \overset{\varepsilon_1 \perp \boldsymbol{X}_1}{=} -\mathsf{E}\operatorname{sign}(\varepsilon_1)\,\mathsf{E}\,\boldsymbol{X}_1 = \mathbf{0}_p, \tag{55}$$

as $\operatorname{med}(\varepsilon_1) = 0$ and $\mathsf{P}(\varepsilon_1 = 0) = 0$. Thus $\boldsymbol{\beta}_X = \boldsymbol{\beta}$.

Introducing $\boldsymbol{\psi}$ and R as in (53): Note that the function

$$\rho(\mathbf{z}; \mathbf{b} + \mathbf{t}) = \left|y - \mathbf{b}^\mathsf{T}\mathbf{x}\right|$$

is almost everywhere differentiable with respect to $\mathbf{b}$ with the derivative given by

$$\boldsymbol{\psi}(\mathbf{z}; \boldsymbol{\theta}_X) = -\mathbf{x}\,\operatorname{sign}\left(y - \mathbf{b}^\mathsf{T}\mathbf{x}\right).$$

Further by the same calculation as in (55) one gets

$$\mathsf{E}\,\boldsymbol{\psi}(\mathbf{Z}_1; \boldsymbol{\beta}_X) = -\mathsf{E}\left[\boldsymbol{X}_1\,\operatorname{sign}(Y_1 - \boldsymbol{\beta}_X^\mathsf{T}\boldsymbol{X}_1)\right] = \mathbf{0}_p.$$

Thus the function $\boldsymbol{\psi}(\mathbf{z}; \boldsymbol{\theta}_X)$ seems to be a reasonable candidate for the function $\boldsymbol{\psi}$ from (53) and the 'remainder' function R is defined as

$$\begin{aligned} R(\mathbf{z}; \mathbf{t}) &= \rho(\mathbf{z}; \boldsymbol{\beta}_X + \mathbf{t}) - \rho(\mathbf{z}; \boldsymbol{\beta}_X) - \mathbf{t}^\mathsf{T}\boldsymbol{\psi}(\mathbf{z}; \boldsymbol{\beta}_X) \\ &= \left|y - (\boldsymbol{\beta}_X + \mathbf{t})^\mathsf{T}\mathbf{x}\right| - \left|y - \boldsymbol{\beta}_X^\mathsf{T}\mathbf{x}\right| + \mathbf{t}^\mathsf{T}\mathbf{x}\,\operatorname{sign}\left(y - \boldsymbol{\beta}_X^\mathsf{T}\mathbf{x}\right). \end{aligned}$$

Showing (i): In what follows let $\mathsf{E}_{\boldsymbol{X}_1}$ and $\mathsf{E}_{\varepsilon_1}$ stand for the expected values with respect to $\boldsymbol{X}_1$ and ε_1 respectively. With the help of this convention one can calculate

$$\begin{aligned} \mathsf{E}\left[\rho(\mathbf{Z}_1; \boldsymbol{\beta}_X + \mathbf{t}) - \rho(\mathbf{Z}_1; \boldsymbol{\beta}_X)\right] &= \mathsf{E}\left[\left|\underbrace{Y_1 - (\boldsymbol{\beta}_X + \mathbf{t})^\mathsf{T}\boldsymbol{X}_1}_{=\varepsilon_1 - \mathbf{t}^\mathsf{T}\boldsymbol{X}_1}\right| - \left|\underbrace{Y_1 - \boldsymbol{\beta}_X^\mathsf{T}\boldsymbol{X}_1}_{=\varepsilon_1}\right|\right] \\ &= -\mathsf{E}\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} \operatorname{sign}(\varepsilon_1 - s)\,\mathrm{d}s \\ &\overset{\boldsymbol{X}_1 \perp \varepsilon_1}{=} -\mathsf{E}_{\boldsymbol{X}_1}\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} \underbrace{\mathsf{E}_{\varepsilon_1}\operatorname{sign}(\varepsilon_1 - s)}_{=1\cdot\mathsf{P}(\varepsilon_1 > s) + (-1)\cdot\mathsf{P}(\varepsilon_1 < s) = 1 - 2F_\varepsilon(s)}\,\mathrm{d}s \\ &= -\mathsf{E}_{\boldsymbol{X}_1}\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} [\underbrace{1}_{=2F_\varepsilon(0)} - 2F_\varepsilon(s)]\,\mathrm{d}s = 2\,\mathsf{E}\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} \underbrace{[F_\varepsilon(s) - F_\varepsilon(0)]}_{s\cdot f_\varepsilon(0) + s\cdot o(1)}\,\mathrm{d}s \\ &= 2\,\mathsf{E}_{\boldsymbol{X}_1}\left[\frac{s^2}{2}f_\varepsilon(0) + \frac{s^2}{2}o(1)\right]_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} = \mathbf{t}^\mathsf{T}\mathsf{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\right)\mathbf{t}\cdot f_\varepsilon(0) + o\left(\|\mathbf{t}\|^2\right). \end{aligned}$$

Thus (ii) of Theorem 10 is satisfied by putting $\Gamma(\boldsymbol{\theta}_X) = 2\, f_\varepsilon(0)\, \mathsf{E}\left(\boldsymbol{X}_1 \boldsymbol{X}_1^\mathsf{T}\right)$.

Showing (ii) Using Cauchy–Schwarz inequality (C-S ineq)

$$\begin{aligned}
\mathsf{var}\left(R(\mathbf{Z}_1;\mathbf{t})\right) \le \mathsf{E}\left[R(\mathbf{Z}_1;\mathbf{t})\right]^2 &= \mathsf{E}\left[\underbrace{\left|\varepsilon_1 - \mathbf{t}^\mathsf{T}\boldsymbol{X}_1\right| - \left|\varepsilon_1\right|}_{=\int -\operatorname{sign}(\varepsilon_1 - s)\ldots} + \mathbf{t}^\mathsf{T}\boldsymbol{X}_1 \operatorname{sign}(\varepsilon_1)\right]^2 \\
&= \mathsf{E}\left[\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} -\operatorname{sign}(\varepsilon_1 - s)\,\mathrm{d}s + \mathbf{t}^\mathsf{T}\boldsymbol{X}_1 \operatorname{sign}(\varepsilon_1)\right]^2 \\
&= \mathsf{E}\left[\int_0^{\mathbf{t}^\mathsf{T}\boldsymbol{X}_1} \operatorname{sign}(\varepsilon_1) - \operatorname{sign}(\varepsilon_1 - s)\,\mathrm{d}s\right]^2 \\
&\le \mathsf{E}\left[\int_{-|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|} 1 \cdot \left|\operatorname{sign}(\varepsilon_1) - \operatorname{sign}(\varepsilon_1 - s)\right| \mathrm{d}s\right]^2 \\
&\overset{\text{C-S ineq}}{\le} \mathsf{E}\left\{\left[\int_{-|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|} 1^2\,\mathrm{d}s\right] \cdot \left[\int_{-|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|} [\operatorname{sign}(\varepsilon_1) - \operatorname{sign}(\varepsilon_1 - s)]^2\,\mathrm{d}s\right]\right\} =: (\Delta),.
\end{aligned}$$

Note that $\operatorname{sign}(\varepsilon_1) - \operatorname{sign}(\varepsilon_1 - s)$ can be different from zero only when ε_1 and $\varepsilon_1 - s$ are of different signs, which is a subset of the event $\varepsilon_1 \in [-|s|, |s|]$. Further using independence of $\boldsymbol{X}_1$ and ε_1 one gets

$$\begin{aligned}
(\Delta) &\le \mathsf{E}_{\boldsymbol{X}_1}\left\{2\left|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1\right| \int_{-|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|} 4\ \mathsf{P}\left(|\varepsilon_1| \le s\right) \mathrm{d}s\right\} \\
&= 8\,\mathsf{E}_{\boldsymbol{X}_1}\left\{\left|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1\right| \int_{-|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|} \underbrace{\left|F_\varepsilon(s) - F_\varepsilon(-s)\right|}_{\le C|s|} \mathrm{d}s\right\} \le 2\,\mathsf{E}_{\boldsymbol{X}_1}\left\{\left|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1\right| 2\,C\left[\tfrac{s^2}{2}\right]_0^{|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1|}\right\} \\
&\le 2\,C\,\mathsf{E}\left|\mathbf{t}^\mathsf{T}\boldsymbol{X}_1\right|^3 \le 2\,C\,\|\mathbf{t}\|^3 \underbrace{\mathsf{E}\,\|\boldsymbol{X}_1\|^3}_{<\infty} = O\left(\|\mathbf{t}\|^3\right) = o\left(\|\mathbf{t}\|^2\right), \text{for}\, \mathbf{t} \to \mathbf{0}_p,
\end{aligned}$$

where we have used that by the assumptions of the theorem there exists a finite constant C such that for $|F_\varepsilon(s) - F_\varepsilon(-s)| \le C\,|s|$ for each $s \in \mathbb{R}$.

Showing (iii) This follows from

$$\begin{aligned}
\mathsf{var}\left(\operatorname{sign}(\varepsilon_1)\boldsymbol{X}_1\right) &= \mathsf{var}\left(\mathsf{E}\left[\boldsymbol{X}_1\ \operatorname{sign}(\varepsilon_1)\,|\,\boldsymbol{X}_1\right]\right) + \mathsf{E}\left(\mathsf{var}\left[\boldsymbol{X}_1\ \operatorname{sign}(\varepsilon_1)\,|\,\boldsymbol{X}_1\right]\right) \\
&= \mathbf{0}_{p\times p} + \mathsf{E}\left(\boldsymbol{X}_1\, \mathsf{var}[\operatorname{sign}(\varepsilon_1)]\boldsymbol{X}_1^\mathsf{T}\right) = \mathsf{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\right).
\end{aligned}$$

□

Note that Theorem 11 covers an asymptotic behaviour of a sample median as a special case. This is explicitly formulated in the following corollary.

Corollary 1. *Let $Y_1, \ldots, Y_n$ be independent identically random variables with density $f(y)$ that is positive and continuous in a neighbourhood of median. Then*

$$\sqrt{n}\left(F_n^{-1}(0.5) - F^{-1}(0.5)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\Big(0, \frac{1}{4\,f^2\big(F^{-1}(0.5)\big)}\Big).$$

Proof. The proof follows from Theorem 11 by taking $\boldsymbol{X}_i = 1$, $\varepsilon_i = Y_i - F^{-1}(0.5)$ and noting that $f_\varepsilon(0) = f\big(F^{-1}(0.5)\big)$ and $F_n^{-1}(0.5) = \arg\min_{b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |Y_i - b|$. □

Literature: Hjort and Pollard (2011) Section 2A

3.5 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) valid even when (some) of the assumptions (under which the procedure) was 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.

(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).

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 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 $IF(\mathbf{x}) = -\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_X)\,\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}_X)$ is called *the influence function*. Thus if one can ignore the reminder term $o_P\big(\frac{1}{\sqrt{n}}\big)$, then changing $\boldsymbol{X}_i$ to $\boldsymbol{X}_i + \boldsymbol{\Delta}$ results that $\widehat{\boldsymbol{\theta}}_n$ changes 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.

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) that if the function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ is bounded then the breakdown point of the associated $M(Z)$-estimator is $\frac{1}{2}$.

3.5.1 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 $\widetilde{X}_n = F_n^{-1}(0.5)$ 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 $\widetilde{X}_n$ estimate different quantities. But when one can hope that the distribution F is symmetric, then both $\overline{X}_n$ and $\widetilde{X}_n$ 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 $\widetilde{X}_n$ is efficient if F is doubly exponential.

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

$$F(x) = (1-\varepsilon)\,\Phi\big(\tfrac{x-\mu}{\sigma}\big) + \varepsilon\, G(x),$$

where ε is usually interpreted as probability of having an outlying observations and G is a (hopefully symmetric) distribution of outlying observations. It was found that if ε is 'small' then using sample median is too pessimistic (and inefficient). We will mention here several alternative options.

Huber's estimator is defined as $\widehat{\theta}_n^{(H)} = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \rho_H(X_i-\theta)$, where

$$\rho_H(x) = \begin{cases} \frac{x^2}{2}, & |x| \le k \\ k\cdot\big(|x| - \frac{k}{2}\big), & |x| > k \end{cases}, \tag{56}$$

and k is a given constant. Note that the 'score function' $\psi_H(x) = \rho_H'(x)$ of the estimator is

$$\psi_H(x) = \rho_H'(x) = \begin{cases} x, & |x| \le k \\ k\cdot\operatorname{sgn}(x), & |x| > k \end{cases}. \tag{57}$$

Thus one can see that for $x \in (-k, k)$ the function ψ_H corresponds to a score function of a sample mean while for $x \in (-\infty, k) \cup (k, \infty)$ it corresponds to a score function of a sample median. Thus Huber's estimator presents a compromise between a sample mean and a sample median.

The nice thing about Huber's 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.

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

$$\frac{\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right)}{\mathsf{var}\left(\overline{X}_n\right)} \leq 1+\delta,$$

where δ stands for the efficiency loss of Huber's 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 $k = 1.03$.

When using Huber's estimator one has to remember that if the population is not symmetric then $\widehat{\theta}_n^{(H)} \xrightarrow[n\to\infty]{P} \theta^{(H)}(F)$, which lies between $\mathsf{E}\, X_1$ and $F^{-1}(0.5)$.

Among other common M-estimators of location let us mention:

(i) **Cauchy-pseudolikelihood**: $\rho(x;\theta) = -\log(1+(x-\theta)^2)$. The problem with this estimator is that the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{2\,(X_i-\widehat{\theta}_n)}{1+(X_i-\widehat{\theta}_n)^2}}_{\psi(x,\widehat{\theta}_n)} \stackrel{!}{=} 0$$

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.

3.5.2 Studentized $M(Z)$-estimators

The problem of the $M(Z)$-estimators presented above is that the estimators are not scale equivariant (i.e. $\widehat{\theta}_n(c\,\boldsymbol{X}) \neq c\,\widehat{\theta}_n(\boldsymbol{X})$). That is why in practice M-estimators are usually defined as

$$\widehat{\theta}_n = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \rho\big(\tfrac{X_i-\theta}{S_n}\big),$$

where S_n is an appropriate estimator of scale. The most common estimators of scale are as follows.

Sample standard deviation $S_n = \sqrt{\frac{1}{n-1}\sum_i (X_i-\overline{X}_n)^2}$, but this is used rather rarely as it is not robust.

Interquartile range:

$$S_n = IQR = F_n^{-1}(0.75) - F_n^{-1}(0.25),$$

where F_n is the empirical distribution function $\big($i.e. $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \le x\}\big)$. Some people prefer to use

$$\widetilde{S}_n = \frac{F_n^{-1}(0.75) - F_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

as it is desired that $\widetilde{S}_n$ estimates σ, when $X_1, \dots, X_n$ is a random sample from $\mathsf{N}(\mu, \sigma^2)$.

Note that the breakdown point of this estimator is 0.25.

Median absolute deviation:

$$MAD = \mathrm{med}_{1\le i\le n}|X_i - F_n^{-1}(0.5)|,$$

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 12. The asymptotic distribution of studentized M-estimators is difficult to derive and rather complex.

3.5.3 Robust estimation in linear models

Suppose we observe independent random vectors

$$\begin{pmatrix} \boldsymbol{X}_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} \boldsymbol{X}_n \\ Y_n \end{pmatrix}.$$

The least square method results in an estimator

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^n \big(Y_i - \mathbf{b}^\mathsf{T}\boldsymbol{X}_i\big)^2 = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i Y_i\right).$$

Generally, the LS method models $\mathsf{E}\,[Y_1 \,|\, \boldsymbol{X}_1]$ as $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_1$.

Suppose now that 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 $\boldsymbol{X}_i = \big(1, \widetilde{\boldsymbol{X}}_i^\mathsf{T}\big)^\mathsf{T}$. Further suppose that the following models holds

$$Y_i = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\widetilde{\boldsymbol{X}}_i + \varepsilon_i, \text{ where } \varepsilon_1, \dots, \varepsilon_n \quad \text{are i.i.d. and} \quad \varepsilon_i \perp \widetilde{\boldsymbol{X}}_i. \tag{58}$$

Then $\mathsf{E}\,[Y_1 \,|\, \boldsymbol{X}_1] = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\widetilde{\boldsymbol{X}}_1 + \mathsf{E}\,\varepsilon_1$ and $\widehat{\boldsymbol{\beta}}_n$ estimates (identifies)

$$\boldsymbol{\beta}_X = \begin{pmatrix} \beta_0 + \mathsf{E}\,\varepsilon_1 \\ \boldsymbol{\beta} \end{pmatrix}.$$

Note that if $\boldsymbol{X}_{ik} \neq 0$ then by changing Y_k one can arrive at any arbitrary value of $\widehat{\beta}_{nk}$.

Method of the least absolute deviation (LAD), i.e.

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} |Y_i - \mathbf{b}^{\mathsf{T}}\boldsymbol{X}_i|,$$

is usually considered as a robust alternative to the least square method. Generally, the LAD method models $\mathsf{med}[Y_1 \,|\, \boldsymbol{X}_1]$ as $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}_1$. But if moreover model (58) holds, then $\mathsf{med}(Y_1 \,|\, \boldsymbol{X}_1) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{X}}_1 + F_\varepsilon^{-1}(0.5)$, where F_ε^{-1} is the quantile function of ε_1 and thus

$$\boldsymbol{\beta}_X = \begin{pmatrix} \beta_0 + F_\varepsilon^{-1}(0.5) \\ \boldsymbol{\beta} \end{pmatrix}.$$

By Theorem 11

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X = \frac{1}{n}\sum_{i=1}^{n} \Big(\mathsf{E}\,\boldsymbol{X}_1\boldsymbol{X}_1^{\mathsf{T}}\Big)^{-1}\boldsymbol{X}_i \frac{\operatorname{sign}\big(\varepsilon_i - F_\varepsilon^{-1}(0.5)\big)}{2f_\varepsilon(F_\varepsilon^{-1}(0.5))} + o_P\big(\tfrac{1}{\sqrt{n}}\big).$$

Thus one can expect that the change of Y_i (or equivalently the change of ε_i) has only a bounded effect on $\widehat{\boldsymbol{\beta}}_n$. On the other hand note that the change of $\boldsymbol{X}_i$ has an unbounded effect on $\widehat{\boldsymbol{\beta}}_n$. Thus LAD method is robust with respect to the response but not with respect to the covariates.

Analogously as the Huber's estimator of location is a compromise between a sample mean and a sample median, Huber's estimator of regression is a compromise between LS and LAD. Put

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_H\big(Y_i - \mathbf{b}^{\mathsf{T}}\boldsymbol{X}_i\big),$$

where ρ_H is defined in (56). Generally, it is difficult to interpret what is being model with Huber's estimator of regression $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}_1$ (it is something between $\mathsf{E}\,(Y_1 \,|\, \boldsymbol{X}_1)$ a $\mathsf{med}(Y_1 \,|\, \boldsymbol{X}_1)$). Note that it identifies

$$\boldsymbol{\beta}_X = \operatorname*{arg\,min}_{\mathbf{b}\in\mathbb{R}^p} \mathsf{E}\,\rho_H(Y_1 - \mathbf{b}^{\mathsf{T}}\boldsymbol{X}_1).$$

Equivalently $\boldsymbol{\beta}_X$ solves

$$\mathsf{E}\left[\psi_H\big(Y_1 - \boldsymbol{\beta}_X^{\mathsf{T}}\boldsymbol{X}_1\big)\,\boldsymbol{X}_1\right] \stackrel{!}{=} \mathbf{0}_p,$$

where ψ_H is defined in (57). Thus if model (58) holds then one needs to solve

$$\mathsf{E}\left[\psi_H\big(\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{X}}_1 + \varepsilon_1 - \beta_{X0} - \widetilde{\boldsymbol{\beta}}_X^{\mathsf{T}}\widetilde{\boldsymbol{X}}_1\big)\,\boldsymbol{X}_1\right] \stackrel{!}{=} \mathbf{0}_p, \tag{59}$$

where we put $\boldsymbol{\beta}_X = \big(\beta_{X0}, \widetilde{\boldsymbol{\beta}}_X^{\mathsf{T}}\big)^{\mathsf{T}}$. Thus $\boldsymbol{\beta}_X$ identifies the following parameter

$$\boldsymbol{\beta}_X = \begin{pmatrix} \beta_0 + \theta_H \\ \boldsymbol{\beta} \end{pmatrix},$$

where θ_H solves $\mathsf{E}\,\psi_H(\varepsilon_1 - \theta_H) \stackrel{!}{=} 0$.

Thus if model (58) holds then the interpretation of the regression slope coefficient ($\boldsymbol{\beta}$) is the same for each of the methods described above (LS, LAD, Huber's regression).

Analogously as in Section 3.5.2 in practice *the studentized Huber's estimator* is usually used. This estimator is defined as

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_H\left(\tfrac{Y_i - \mathbf{b}^\mathsf{T}\boldsymbol{X}_i}{S_n}\right),$$

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 $\widehat{\varepsilon_i} = Y_i - \widehat{\boldsymbol{\beta}}_{n,LAD}^\mathsf{T}\boldsymbol{X}_i$.

Inference:

- With the help of Theorem 10 one can show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ of the (non-Studentized) Huber's estimator.
- If model (58) holds, then it can be shown, that the estimate of the scale influences only the asymptotic distribution of the estimate of intercept and not of the slope.

Literature: Maronna et al. (2006) Chapters 2.1-2.2 and Chapters 4.1-4.4.

4 Bootstrap and other resampling methods

Suppose we observe independent and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution F and we are interested in some characteristic of F, say $\boldsymbol{\theta}_X$. Let $\boldsymbol{\theta}(F)$ be the quantity of interest and $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(F_n)$ its estimator. Now let $\mathbf{R}_n$ be either an appropriately standardised version of $\widehat{\boldsymbol{\theta}}_n$ or a test statistics,

$$\text{i.e} \qquad \mathbf{R}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \quad \text{or} \quad R_n = \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)^\mathsf{T}\left[\widehat{\mathsf{avar}(\widehat{\boldsymbol{\theta}}_n)}\right]^{-1}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$$

For doing inference about parameter $\boldsymbol{\theta}$, one needs to know the distribution of $\mathbf{R}_n$. Usually we are not able to derive the exact distribution of $\mathbf{R}_n$ analytically. For instance consider the distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)$ where $\widehat{\boldsymbol{\theta}}_n$ is a maximum likelihood estimator whose formula cannot be explicitly given. In such situations the inference is often based on an asymptotic distribution of $\mathbf{R}_n$. For example for a MLE estimator in regular models one has $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} N\big(0, I^{-1}(\boldsymbol{\theta}_X)\big)$. Bootstrap presents an alternative to using the asymptotic normality. As we will see later, bootstrap combines the 'Monte-Carlo principle' and substitution (plug-in) principle.

4.1 Monte Carlo principle

Sometimes one knows the distribution of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and thus he/she is (at least theoretically) able to derive the distribution of $\mathbf{R}_n = (R_{n1}, \ldots, R_{nk})^\mathsf{T}$. But the derivations are too complicated and/or the resulting distribution is too complex to work with. For instance consider the standard maximum likelihood tests without nuisance parameters as in Section 2.4.

Another way how to utilize the knowledge of data-generating mechanism of $\mathbb{X}$ is to use Monte-Carlo principle, which runs as follows. Choose B sufficiently large and for each $b \in \{1, \ldots, B\}$ independently generate the samples $\mathbb{X}_b^* = (\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*)^\mathsf{T}$ such that the distribution of $\mathbb{X}_b^*$ is the same as the distribution of $\mathbb{X}$. Thus we get B independent samples $\mathbb{X}_1^*, \ldots, \mathbb{X}_B^*$. Let $\mathbf{R}_{n,b}^*$ be the quantity $\mathbf{R}_n$ calculated from the b-th sample $\mathbb{X}_b^*$. The unknown distribution function $H_n(\mathbf{x})$ of $\mathbf{R}_n$ can now be estimated as

$$H_{n,B}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\{\mathbf{R}_{n,b}^* \leq \mathbf{x}\}.$$

As $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$ are independent and identically distributed random variables and each variable has the same distribution as $\mathbf{R}_n$, by the Glivenko-Cantelli Theorem (Theorem A1) we know that

$$\sup_{\mathbf{x} \in \mathbb{R}^k} \left| H_{n,B}(\mathbf{x}) - H_n(\mathbf{x}) \right| \xrightarrow[B \to \infty]{\text{alm. surely}} 0, \tag{60}$$

thus for a sufficiently large B one can use $H_{n,B}(\mathbf{x})$ as an approximation of $H_n(\mathbf{x})$.

Note that if R_n is one dimensional then also for each fixed $u \in (0, 1)$:

$$H_{n,B}^{-1}(u) \xrightarrow[B \to \infty]{\text{alm. surely}} H_n^{-1}(u),$$

provided that $H_n^{-1}(u)$ is a unique solution of $H_n(x_-) \leq u \leq H_n(x)$ (see e.g. Theorem of Section 2.1.3 in Serfling, 1980).

Further if R_n is a (one-dimensional) test statistic whose large values are in favour of the alternative hypothesis, then with the help of the Monte-Carlo principle the p-value of the test can be estimated as

$$\widehat{\mathrm{p}}_B = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\{R_{n,b}^* \geq R_n\}}{B + 1},$$

as $\widehat{\mathrm{p}}_B \xrightarrow[B \to \infty]{\text{alm. surely}} 1 - H_n(R_{n-})$.

Example 31. Suppose we observe a random variable with the multinomial distribution $\mathsf{M}(n; p_1, \ldots, p_k)$. Denote $\mathbf{p} = \big(p_1, \ldots, p_k\big)^\mathsf{T}$ and $\mathbf{p}_X$ the true value of the parameter $\mathbf{p}$. In some applications we are interested in testing

$$H_0 : \mathbf{p}_X = \mathbf{p}^{(0)} \quad \text{vs.} \quad H_0 : \mathbf{p}_X \neq \mathbf{p}^{(0)},$$

where $\mathbf{p}^{(0)} = \left(p_1^{(0)}, \ldots, p_k^{(0)}\right)^\mathsf{T}$ is a given vector. Explain how the Monte Carlo principle can be used to estimate the critical value and the p-value of the χ^2-test of goodness of fit.

Example 32. Note that the significance of all the test statistics introduced in Section 2.4 for testing the null hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ can be assessed with the help of Monte Carlo principle.

In the following examples we will utilize that in fact it is not necessary to know the data-generating mechanism of $\mathbb{X}$ exactly, provided we are able to generate independent copies of $\mathbf{R}_n$.

Example 33. Suppose that we observe independent and identically distributed random vectors $(Y_1, X_1)^\mathsf{T}, \ldots, (Y_n, X_n)^\mathsf{T}$ from the bivariate normal distribution. Then the distribution of the sample correlation coefficient $\widehat{\rho}_n$ depends only on the true value of the parameter ρ which we denote by ρ_X. Thus when testing the null hypothesis

$$H_0 : \rho_X = \rho_0, \qquad \text{vs.} \qquad H_1 : \rho_X \neq \rho_0$$

one should be able (at least theoretically) calculate the distribution of the test statistic $R_n = \sqrt{n}\,(\widehat{\rho}_n - \rho_0)$ when the null hypothesis holds. Now let $r_n(\alpha)$ be an α-quantile of the distribution R_n when the null distribution holds. Then one rejects the null hypothesis if

$$R_n \leq r_n(\alpha/2), \qquad \text{or} \qquad R_n \geq r_n(1-\alpha/2).$$

Although the quantiles $r_n(\alpha/2)$ and $r_n(1-\alpha/2)$ are difficult to calculate analytically, it is straightforward to estimate them by Monte-Carlo simulations.

Example 34. Let $X_1, \ldots, X_n$ be a random sample from the logistic distribution with the density

$$f(x) = \frac{\exp\{-(x-\theta)\}}{\left(1+\exp\{-(x-\theta)\}\right)^2}, \quad x \in \mathbb{R},$$

where $\theta \in \mathbb{R}$.

Let θ_X be the true value of the parameter θ and $\widehat{\theta}_n$ be for instance its maximum likelihood estimator Then the distribution of $R_n = \sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$ does not depend on θ. Thus the distribution of R_n can be approximated by simulating from the logistic distribution with $\theta = 0$ and calculating $R^*_{n,b} = \sqrt{n}\,\widehat{\theta}^*_n$.

Usually in practice we do not know the data generating process completely. But very often we are able to estimate the distribution of $\mathbb{X}$. Depending on whether this distribution is estimated parametrically or nonparametrically we distinguish parametric or nonparametric bootstrap.

4.2 Standard nonparametric bootstrap

Suppose we observe independent and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution F. Let $\boldsymbol{\theta}(F)$ be the quantity of interest and $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(F_n)$ its estimator with F_n being the empirical distribution

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \leq \mathbf{x}\}.$$

Suppose we are interested in the distribution of

$$\mathbf{R}_n = \mathbf{g}_n\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}\big) = \mathbf{g}_n\big(\boldsymbol{\theta}(F_n), \boldsymbol{\theta}(F)\big) \qquad \Big(\text{e.g. } \mathbf{R}_n = \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big)\Big).$$

In nonparametric bootstrap the unknown F is estimated by the empirical distribution function F_n. Now generating independent random vectors $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ from the distribution F_n is equivalent to drawing a simple random sample with replacement of size n from the observed values $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The bootstrap algorithm now runs as follows.

Choose B sufficiently large and for each $b \in \{1, \ldots, B\}$ independently generate the datasets $\mathbb{X}_b^* = (\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*)^\mathsf{T}$ (i.e. the datasets $\mathbb{X}_1^*, \ldots, \mathbb{X}_B^*$ are independent). Let

$$\mathbf{R}_{n,b}^* = \mathbf{g}_n\big(\widehat{\boldsymbol{\theta}}_{n,b}^*, \widehat{\boldsymbol{\theta}}_n\big) = \mathbf{g}_n\big(\boldsymbol{\theta}(F_{n,b}^*), \boldsymbol{\theta}(F_n)\big) \qquad \Big(\text{e.g. } \mathbf{R}_{n,b}^* = \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_{n,b}^* - \widehat{\boldsymbol{\theta}}_n\big)\Big),$$

where $\widehat{\boldsymbol{\theta}}_{n,b}^*$ is an estimator of $\boldsymbol{\theta}$ based on $\mathbb{X}_b^*$ and analogously $F_{n,b}^*$ is an empirical distribution function based on $\mathbb{X}_b^*$. The unknown distribution function $H_n(\mathbf{x})$ of $\mathbf{R}_n$ is now estimated by

$$H_{n,B}^*(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\big\{\mathbf{R}_{n,b}^* \leq \mathbf{x}\big\}.$$

Note that the random variables/vectors $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$ are identically distributed and put $\mathbf{R}_n^*$ for any of the random variables. As $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$ are also independent then by the Glivenko-Cantelli Theorem

$$\sup_{\mathbf{x}} \big|H_{n,B}^*(\mathbf{x}) - H_n^*(\mathbf{x})\big| \xrightarrow[B\to\infty]{\text{alm. surely}} 0,$$

where

$$H_n^*(\mathbf{x}) = \mathsf{P}\left(\mathbf{R}_n^* \leq \mathbf{x} \,|\, \mathbb{X}\right) = \mathsf{P}\left(\mathbf{g}_n\big(\boldsymbol{\theta}(F_n^*), \boldsymbol{\theta}(F_n)\big) \leq \mathbf{x} \,|\, \mathbb{X}\right) = \mathsf{P}\left(\mathbf{g}_n\big(\widehat{\boldsymbol{\theta}}_n^*, \widehat{\boldsymbol{\theta}}_n\big) \leq \mathbf{x} \,|\, \mathbb{X}\right).$$

The crucial question for the success of the nonparametric bootstrap is whether H_n^* is 'close' (at least asymptotically) to H_n. To answer this question it is to useful to introduce the supremum metric on the space of distribution functions (of random vectors on $\mathbb{R}^k$) as

$$\rho_\infty(H_1, H_2) = \sup_{\mathbf{x}\in\mathbb{R}^k} \big|H_1(\mathbf{x}) - H_2(\mathbf{x})\big|.$$

The following lemma states that if the distribution function of the limiting distribution is continuous, then ρ_∞ can be used for metrizing the convergence in distribution.

Lemma 5. *Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots$ and $\boldsymbol{Y}$ be random vectors (with values in $\mathbb{R}^k$) with the corresponding distribution functions $G_1, G_2, \dots$ and G. Further let the distribution function G be continuous. Then $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{Y}$ if and only if $\rho_\infty(G_n, G) \to 0$ as $n \to \infty$.*

Proof. We would like to show that

$$\rho_\infty(G_n, G) \xrightarrow[n\to\infty]{} 0 \quad \Longleftrightarrow \quad G_n \xrightarrow[n\to\infty]{w} G.$$

The implication $\Rightarrow$ is straightforward as $\sup_{\mathbf{y}\in\mathbb{R}^k} |G_n(\mathbf{y}) - G(\mathbf{y})| \to 0$ implies that $G_n(\mathbf{y}) \to G(\mathbf{y})$ for each $\mathbf{y} \in \mathbb{R}^k$.

The implication $\Leftarrow$ is more difficult. By the continuity of G for each $\varepsilon > 0$ there exists a finite set of points $B_\varepsilon = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ such that for each $\mathbf{y} \in \mathbb{R}^k$ one can find $\mathbf{y}_L, \mathbf{y}_U \in B_\varepsilon$ that

$$\mathbf{y}_L \le \mathbf{y} \le \mathbf{y}_U, \qquad \text{and} \qquad G(\mathbf{y}_U) - G(\mathbf{y}_L) \le \tfrac{\varepsilon}{2}.$$

Thus for each $\mathbf{y} \in \mathbb{R}^k$ one can bound

$$G_n(\mathbf{y}) - G(\mathbf{y}) \le G_n(\mathbf{y}_U) - G(\mathbf{y}) \le G_n(\mathbf{y}_U) - G(\mathbf{y}_U) + \tfrac{\varepsilon}{2} \tag{61}$$

and analogously also

$$G_n(\mathbf{y}) - G(\mathbf{y}) \ge G_n(\mathbf{y}_L) - G(\mathbf{y}) \ge G_n(\mathbf{y}_L) - G(\mathbf{y}_L) - \tfrac{\varepsilon}{2}. \tag{62}$$

Now combining (61) and (62) together with $G_n \xrightarrow[n\to\infty]{w} G$ one gets that for all sufficiently large n

$$\sup_{\mathbf{y}} \big|G_n(\mathbf{y}) - G(\mathbf{y})\big| \le \max_{\mathbf{y}\in B_\varepsilon} \big|G_n(\mathbf{y}) - G(\mathbf{y})\big| + \tfrac{\varepsilon}{2} \le \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon,$$

which implies the statement of the lemma. □

Suppose that a metric ρ can be used for metrizing weak convergence. Let $\mathbf{R}$ be a random vector with the distribution function H. Then we say that conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ the random variable $\mathbf{R}_n^*$ converges in distribution to $\mathbf{R}$ in probability if

$$\rho(H_n^*, H) \xrightarrow[n\to\infty]{P} 0 \quad \Big(\text{i.e. for each } \varepsilon > 0 \lim_{n\to\infty} \mathsf{P}\left[\omega : \rho\big(H_n^*(\omega), H_n\big) \ge \varepsilon\right] = 0\Big).$$

Analogously we say that conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ the random variable $\mathbf{R}_n^*$ converges in distribution to $\mathbf{R}$ almost surely if

$$\rho(H_n^*, H) \xrightarrow[n\to\infty]{\text{alm. surely}} 0 \quad \Big(\text{i.e. } \mathsf{P}\left[\omega : \lim_{n\to\infty} \rho\big(H_n^*(\omega), H_n\big) = 0\right] = 1\Big).$$

Theorem 12. *Suppose that $\mathbf{R}_n \xrightarrow[n\to\infty]{d} \mathbf{R}$, where $\mathbf{R}$ is a random vector with a continuous distribution function. Further suppose that*

$$\rho_\infty(H_n^*, H_n) \xrightarrow[n\to\infty]{P} 0 \quad (\textit{or} \xrightarrow[n\to\infty]{\textit{alm. surely}} 0), \tag{63}$$

then conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ one gets $\mathbf{R}_n^* \xrightarrow[n\to\infty]{d} \mathbf{R}$ in probability (or almost surely).

Proof. By the triangular inequality, (60) and Lemma 5

$$\rho_\infty(H_n^*, H) \le \rho_\infty(H_n^*, H_n) + \rho_\infty(H_n, H) \xrightarrow[n\to\infty]{P} 0 \quad (\text{or } \xrightarrow[n\to\infty]{\text{alm. surely}} 0).$$

□

Typically we know that $\mathbf{R}_n$ converges to a multivariate normal distribution. Thus in view of Theorem 12 the crucial question to answer is if convergence (63) holds. The next theorem states that (63) holds for a sample mean (for the proof see e.g. Theorem 23.4 of van der Vaart (2000), pp. 330–331).

Theorem 13. *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent identically distributed random vectors such that $\mathsf{E}\,\|\boldsymbol{X}_1\|^2 < \infty$ and consider $\mathbf{R}_n = \sqrt{n}\,(\overline{\boldsymbol{X}}_n - \mathsf{E}\,\boldsymbol{X}_1)$ and $\mathbf{R}_n^* = \sqrt{n}\,(\overline{\boldsymbol{X}}_n^* - \overline{\boldsymbol{X}}_n)$. Then*

$$\rho_\infty(H_n^*, H_n) \xrightarrow[n\to\infty]{\textit{alm. surely}} 0. \tag{64}$$

Note that for $\boldsymbol{X}_1$ being a p-variate random vector, then the central limit theorem implies that the distribution function H_n converges weakly to the distribution function of $\mathsf{N}_p\big(\mathbf{0}_p, \mathsf{var}(\boldsymbol{X}_1)\big)$. Now Theorems 12 and 13 imply that conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$

$$\mathbf{R}_n^* \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \mathsf{var}(\boldsymbol{X}_1)\big), \qquad \text{almost surely.}$$

Thus one can say that H_n^* estimates also the distribution function of $\mathsf{N}_p\big(\mathbf{0}_p, \mathsf{var}(\boldsymbol{X}_1)\big)$.

4.2.1 Comparison of nonparametric bootstrap and normal approximation

Note that Theorem 12 implies only the asymptotic validity of bootstrap provided that (63) holds. The question is whether bootstrap estimate H_n^* is a better estimate of H_n than the asymptotic distribution of H where one estimates the unknown parameters.

To answer the above question, consider $p = 1$. Further let X_1 have a continuous distribution and put $\gamma_1 = \mathsf{E}\left(\frac{X_1-\mu}{\sigma}\right)^3$, where $\mu = \mathsf{E}\,X_1, \sigma^2 = \mathsf{var}(X_1)$. Further let $\mathsf{E}\,X_1^4 < \infty$. Then it can be proved that

$$H_n(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}_n-\mu)}{S_n} \le x\right) = \Phi(x) + \frac{\gamma_1}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O\big(\tfrac{1}{n}\big), \tag{65}$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, S_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Further it can be shown that an analogous approximation also holds for $H_n^*(x)$, i.e.

$$H_n^*(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}_n^*-\overline{X}_n)}{S_n^*} \le x \,\middle|\, \mathbb{X}\right) = \Phi(x) + \frac{\gamma_{1n}}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O_P\big(\tfrac{1}{n}\big), \tag{66}$$

where $\overline{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$, $S_n^{2*} = \frac{1}{n-1}\sum_{i=1}^n (X_i^* - \overline{X}_n^*)^2$ and $\gamma_{1n} = \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i-\overline{X}_n}{S_n}\right)^3$. Thus comparing (65) and (66) one gets

$$H_n^*(x) - H_n(x) = O_p\big(\tfrac{1}{n}\big).$$

On the other if $\gamma_1 \neq 0$, then by the normal approximation one gets only

$$\Phi(x) - H_n(x) = O\big(\tfrac{1}{\sqrt{n}}\big).$$

Thus if $\gamma_1 \neq 0$ then one can expect that in comparison with Φ the bootstrap estimator H_n^* is closer to H_n.

4.2.2 Smooth transformations of sample means

The standard nonparametric bootstrap also works for 'smooth' transformations of sample means.

Theorem 14. *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent identically distributed random (p-variate) vectors such that $\mathsf{E}\,\|\boldsymbol{X}_1\|^2 < \infty$. Further suppose that there exists a neighbourhood U of $\boldsymbol{\mu}$ such the function $\mathbf{g} : U \to \mathbb{R}^m$ have continuous partial derivatives in this neighbourhood. Consider $\mathbf{R}_n = \sqrt{n}\,\big(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\big)$ and $\mathbf{R}_n^* = \sqrt{n}\,\big(\mathbf{g}(\overline{\boldsymbol{X}}_n^*) - \mathbf{g}(\overline{\boldsymbol{X}}_n)\big)$. Then (64) holds.*

Remark 13. Suppose for simplicity that $g : \mathbb{R}^p \to \mathbb{R}$. Note that if $\nabla g^{\mathsf{T}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\,\nabla g(\boldsymbol{\mu}) = 0$, then although (64) holds, the bootstrap might be not useful as the limiting distribution of $\mathbf{R}_n$ is degenerate.

To illustrate this consider $p = 1$. Let $X_1, \ldots, X_n$ be a random sample from the distribution with $\mathsf{E}\,X_1 = \mu_X$. Further let g be a twice continuously differentiable function in μ_X such that $g'(\mu_X) = 0$ and $g''(\mu_X) \neq 0$. Then by Theorem 3 one gets $R_n = \sqrt{n}\,\big(g(\overline{X}_n) - g(\mu_X)\big) \xrightarrow[n\to\infty]{P} 0$. Thus although by Theorem 14 convergence (64) holds, one cannot say if bootstrap works as the limiting distribution is not continuous.

Nevertheless a finer analysis shows that (see Theorem B of Section 3.1 in Serfling, 1980)

$$\widetilde{R}_n = 2\,n\,\big(g(\overline{X}_n) - g(\mu_X)\big) \xrightarrow[n\to\infty]{d} \big[g''(\mu_X)\big]\,\sigma^2\,\chi_1^2.$$

So the bootstrap would work if the convergence (63) holds also for $\widetilde{R}_n^* = 2\,n\,\big(g(\overline{X}_n^*) - g(\overline{X}_n)\big)$. But this is not true as it is shown in Example 3.6 of Shao and Tu (1996).

Roughly speaking one can say that (64) holds provided that $\widehat{\boldsymbol{\theta}}_n$ can be approximated as a mean of independent identically distributed random vectors plus a reminder term of order $o_P\big(\frac{1}{\sqrt{n}}\big)$. This can be formalised through the concept of Hadamard-differentiability of the functional $\boldsymbol{\theta}(F)$, but this is out of the scope of this course.

4.2.3 Limits of the standard nonparametric bootstrap

Although the standard nonparametric bootstrap often presents an interesting alternative to the inference based on the asymptotic normality, it often fails in situations when the asymptotic normality does not hold. These include for instance extremal statistics and non-smooth

transformations of sample means. Note also that the standard nonparametric bootstrap assume that the observations are realisations of **independent** and **identically distributed** random vectors. Thus among others the standard nonparametric bootstrap is not appropriate in regression problems with fixed design or in time series problems.

Example 35. Let $X_1, \dots, X_n$ be a random sample from the uniform distribution on $R(0, \theta)$. Then the maximum likelihood estimator of θ is given by $\widehat{\theta}_n = \max_{1 \le i \le n} X_i =: X_{(n)}$. Note that for $x < 0$

$$\mathsf{P}\left(n(X_{(n)} - \theta) \le x\right) = \mathsf{P}\left(X_{(n)} \le \theta + \tfrac{x}{n}\right) = F_{X_1}^n\left(\theta + \tfrac{x}{n}\right) = \left[\tfrac{\theta + \frac{x}{n}}{\theta}\right]^n = \left[1 + \tfrac{x}{n\theta}\right]^n \xrightarrow[n\to\infty]{} \mathrm{e}^{\frac{x}{\theta}}.$$

Thus $n\left(X_{(n)} - \theta\right) \xrightarrow[n\to\infty]{d} Y$, where Y has a cumulative distribution function

$$\mathsf{P}(Y \le x) = \begin{cases} \mathrm{e}^{\frac{x}{\theta}}, & x < 0 \\ 1, & x \ge 0 \end{cases}$$

On the other side

$$\mathsf{P}(X_{(n)}^* = X_{(n)} \,|\, \mathbb{X}) = 1 - \mathsf{P}\left(X_{(n)} \notin \{X_1^*, \dots, X_n^*\} \,|\, \mathbb{X}\right) = 1 - \left(\tfrac{n-1}{n}\right)^n \xrightarrow[n\to\infty]{} 1 - \mathrm{e}^{-1}$$

and thus (63) cannot hold for $R_n^* = n\left(X_{(n)}^* - X_{(n)}\right)$.

Literature: Prášková (2004), Shao and Tu (1996) Chapter 3.2.2, Chapter 3.6, A.10

4.3 Confidence intervals

Suppose for simplicity that $\theta(F)$ is one-dimensional.

4.3.1 Basic bootstrap confidence interval

Consider $R_n = \sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$. Then the quantiles $r_n^*(u)$ of the bootstrap distribution $R_n^* = \sqrt{n}\left(\widehat{\theta}_n^* - \widehat{\theta}_n\right)$ estimate the unknown quantiles of the distribution R_n. Thus if 'bootstrap works' (i.e. Theorem 12 holds for R_n^*) then

$$\lim_{n\to\infty} \mathsf{P}\left[r_n^*(\alpha/2) \le \sqrt{n}\left(\widehat{\theta}_n - \theta\right) \le r_n^*(1 - \alpha/2)\right] = 1 - \alpha. \tag{67}$$

Now with the help of (67) one can construct an asymptotic confidence interval for θ_X as

$$\left(\widehat{\theta}_n - \tfrac{r_{n,B}^*(1-\alpha/2)}{\sqrt{n}}, \widehat{\theta}_n - \tfrac{r_{n,B}^*(\alpha/2)}{\sqrt{n}}\right), \tag{68}$$

where $r_{n,B}^*(\alpha)$ is a Monte-Carlo approximation (estimate) of $r_n^*(\alpha)$. The confidence interval in (68) is usually called *basic bootstrap confidence interval*.

Remark 14. Note that as $r^*_{n,B}(\alpha)$ is a sample α-quantile of $R^*_{n,1}, \dots, R^*_{n,B}$, where $R^*_{n,b} = \sqrt{n}\left(\widehat{\theta}^*_{n,b} - \widehat{\theta}_n\right)$. Thus the confidence interval (68) can be also rewritten as

$$\left(2\,\widehat{\theta}_n - q^*_{n,B}(1-\alpha/2), 2\,\widehat{\theta}_n - q^*_{n,B}(\alpha/2)\right), \tag{69}$$

where $q^*_{n,B}(\alpha)$ is a sample α-quantile calculated from the values $\widehat{\theta}^*_{n,1}, \dots, \widehat{\theta}^*_{n,B}$.

Sometimes in literature you can find the bootstrap confidence interval of the form

$$\left(q^*_{n,B}(\alpha/2), q^*_{n,B}(1-\alpha/2)\right),$$

which is usually called the *percentile confidence interval*.

4.3.2 Studentized bootstrap confidence interval

Usually it is recommended to 'bootstrap' a variable whose limit distribution does not depend on the unknown parameters (such a variable is called pivot). Thus consider $R_n = \frac{\sqrt{n}\,(\widehat{\theta}_n - \theta_X)}{\widehat{\sigma}_n}$, where $\widehat{\sigma}^2_n$ is an estimate of the asymptotic variance of $\sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$. Let $\tilde{r}^*_n(u)$ be the u-th quantile of the distribution $\widetilde{R}^*_n = \frac{\sqrt{n}\,(\widehat{\theta}^*_n - \widehat{\theta}_n)}{\widehat{\sigma}^*_n}$, where $\widehat{\sigma}^*_n$ is an estimate of the asymptotic variance of $\sqrt{n}\left(\widehat{\theta}^*_n - \widehat{\theta}_n\right)$ calculated from the bootstrap sample. Thus if 'bootstrap works' (i.e. Theorem 12 holds), then

$$\lim_{n\to\infty} \mathsf{P}\left[\tilde{r}^*_n(\alpha/2) \le \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}_n} \le \tilde{r}^*_n(1-\alpha/2)\right] = 1-\alpha,$$

which yields an asymptotic confidence interval

$$\left(\widehat{\theta}_n - \frac{\tilde{r}^*_{n,B}(1-\alpha/2)\,\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n - \frac{\tilde{r}^*_{n,B}(\alpha/2)\,\widehat{\sigma}_n}{\sqrt{n}}\right), \tag{70}$$

where $\tilde{r}^*_{n,B}(\alpha)$ is a Monte-Carlo approximation of $\tilde{r}^*_n(\alpha)$. The confidence interval in (70) is usually called *studentized bootstrap confidence interval*.

Literature: Efron and Tibshirani (1993) Chapters 15 and 16

4.4 Parametric bootstrap

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be random vectors having the joint distribution $F(\cdot;\boldsymbol{\theta})$ that is known only up to an unknown parameter $\boldsymbol{\theta}$. In parametric bootstrap we generate the bootstrap vectors $\boldsymbol{X}^*_{1,b}, \dots, \boldsymbol{X}^*_{n,b}$ from $F(\cdot;\widehat{\boldsymbol{\theta}}_n)$, where $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$.

Example 36. Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be two independent random samples from the the exponential distributions with the density $f(x,\lambda) = \lambda \mathrm{e}^{-\lambda x}\mathbb{I}[x > 0]$. Let λ_X be the true value of the parameter for the first sample and λ_Y for the second sample. Find the confidence interval for $\frac{\lambda_X}{\lambda_Y}$.

Solution. The maximum likelihood estimators are given by $\widehat{\lambda}_X = \frac{1}{\overline{X}_{n_1}}, \widehat{\lambda}_Y = \frac{1}{\overline{Y}_{n_2}}$. Now generate $X_1^*, \ldots, X_{n_1}^*$ and $Y_1^*, \ldots, Y_{n_2}^*$ as two independent random samples from the exponential distributions with the parameters $\widehat{\lambda}_X$ and $\widehat{\lambda}_Y$ respectively. Put

$$R_n = \left(\frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y} - \frac{\lambda_X}{\lambda_Y}\right) \quad \text{and} \quad R_n^* = \left(\frac{\widehat{\lambda}_X^*}{\widehat{\lambda}_Y^*} - \frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y}\right),$$

where $\widehat{\lambda}_X^* = \frac{1}{\overline{X}_{n_1}^*}$ and $\widehat{\lambda}_Y^* = \frac{1}{\overline{Y}_{n_2}^*}$. The confidence interval for the ratio $\frac{\lambda_X}{\lambda_Y}$ can now be calculated as

$$\left(\frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y} - r_{n,B}^*\left(1-\tfrac{\alpha}{2}\right), \frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y} - r_{n,B}^*\left(\tfrac{\alpha}{2}\right)\right),$$

where $r_{n,B}^*(\alpha)$ is the estimate of the α-kvantil R_n^*.

Alternatively instead of bootstrap one can use Δ-theorem (Theorem 3), which implies that

$$\left(\frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y} - \frac{\lambda_X}{\lambda_Y}\right) \overset{as}{\approx} \mathsf{N}\left(0, \frac{\lambda_X^2}{\lambda_Y^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

By combining Δ-theorem and bootstrap one can also use

$$\widetilde{R}_n = \frac{\frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y} - \frac{\lambda_X}{\lambda_Y}}{\frac{\widehat{\lambda}_Y}{\widehat{\lambda}_X}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \quad \text{and} \quad \widetilde{R}_n^* = \frac{\frac{\widehat{\lambda}_X^*}{\widehat{\lambda}_Y^*} - \frac{\widehat{\lambda}_X}{\widehat{\lambda}_Y}}{\frac{\widehat{\lambda}_Y^*}{\widehat{\lambda}_X^*}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

Goodness of fit testing

Parametric bootstrap is often used in **goodness of fit testing**. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of k-variate random vectors with the distribution function F. Suppose we are interested in testing that F belongs to a given parametric family, i.e.

$$H_0 : F \in \mathcal{F} = \left\{F(\mathbf{x}; \boldsymbol{\theta}), \theta \in \Theta\right\}$$

against the alternative

$$H_1 : F \notin \mathcal{F}.$$

As a test statistic one can use for instance

$$KS_n = \sup_{\mathbf{x}\in\mathbb{R}^k} \left|F_n(\mathbf{x}) - F(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)\right|,$$

where F_n is an empirical distribution function and $\widehat{\boldsymbol{\theta}}_n$ is an estimate of $\boldsymbol{\theta}$ under the null hypothesis. As the asymptotic distribution of the test statistic under the null hypothesis is rather difficult, the significance of the test statistic is derived as follows.

1. For $b = 1, \ldots, B$ generate a random sample $\mathbb{X}_b^* = (\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*)$ (of size n), where each random vector $\boldsymbol{X}_{i,b}^*$ has the distribution function $F(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)$.

2. Calculate

$$KS^*_{n,b} = \sup_{\mathbf{x}\in\mathbb{R}^k} \left|F^*_{n,b}(\mathbf{x}) - F(\mathbf{x};\widehat{\boldsymbol{\theta}}^*_{n,b})\right|,$$

where $F^*_{n,b}(\mathbf{x})$ is the empirical distribution function calculated from $\mathbb{X}^*_b$ and $\widehat{\boldsymbol{\theta}}^*_{n,b}$ is the estimate of $\boldsymbol{\theta}$ (under H_0) calculated from $\mathbb{X}^*_b$.

3. Estimate the p-value as

$$\frac{1+\sum_{b=1}^{B}\mathbb{I}\{KS^*_{n,b}\geq KS_n\}}{B+1},$$

where B is usually chosen as 999 or 9 999.

Remark 15. Instead of the test statistic KS_n it is usually recommended to use one of the following statistics.

Cramér-von-Mises:

$$CM = \int \big(F_n(\mathbf{x}) - F(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\big)^2 f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\,\mathrm{d}\mathbf{x}, \quad \text{or} \quad CM = \frac{1}{n}\sum_{i=1}^{n}\big(F_n(\boldsymbol{X}_i) - F(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n)\big)^2.$$

Anderson-Darling:

$$AD = \int \frac{\big(F_n(\mathbf{x}) - F(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\big)^2}{F(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\big(1-F(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\big)} f(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)\,\mathrm{d}\mathbf{x}, \quad \text{or} \quad AD = \frac{1}{n}\sum_{i=1}^{n}\frac{\big(F_n(\boldsymbol{X}_i) - F(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n)\big)^2}{F(\boldsymbol{X}_i)\big(1-F(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n)\big)}.$$

Example 37. Bootstrap estimation of the distribution of estimators of parameters in AR(p) process.

4.5 Testing and bootstrap

Suppose that we have a test statistic $T_n = T(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ and that large values of T_n speaks against the null hypothesis. Let $\mathbb{X}^*_1 = (\boldsymbol{X}^*_{1,1},\ldots,\boldsymbol{X}^*_{n,1})^\mathsf{T}$, …, $\mathbb{X}^*_B = (\boldsymbol{X}^*_{1,B},\ldots,\boldsymbol{X}^*_{n,B})^\mathsf{T}$ be independently resampled datasets by a procedure that mimics generating data under the null hypothesis. Let $T^*_{n,b} = T_n(\mathbb{X}^*_b)$ be the test statistic calculated from the b-th generated sample $\mathbb{X}^*_b$ $(b=1,\ldots,B)$. Then the p-value of the test is estimated as

$$\widehat{\text{pvalue}} = \frac{1+\sum_{b=1}^{B}\mathbb{I}\{T^*_{n,b}\geq T_n\}}{B+1}. \tag{71}$$

Comparison of expectations in two-sample problems

Let $X_1,\ldots,X_{n_1}$ and $Y_1,\ldots,Y_{n_2}$ be two independent random samples from the distributions F and G respectively. Suppose we are interested in testing the null hypothesis

$$H_0 : \mathsf{E}\,X_1 = \mathsf{E}\,Y_1.$$

In what follows we will mention several options how to test for the above null hypothesis.
1. Standard t-test is based on the test statistics

$$T_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$S^{*2} = \frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1) S_X^2 + (n_2 - 1) S_Y^2 \right], \; S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X}_{n_1})^2, \; S_Y^2 = \dots .$$

The crucial assumption of this test is the homoscedasticity, i.e. $\mathsf{var}\, X_1 = \mathsf{var}\, Y_1 \in (0, \infty)$ or that $\frac{n_1}{n_1 + n_2} \to \frac{1}{2}$. Then under the null hypothesis $T_n \xrightarrow[n \to \infty]{d} \mathsf{N}(0, 1)$.
2. Welch t-test is based on the test statistics

$$\widetilde{T}_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}.$$

The advantage of this test is that it does not require $\mathsf{var}\, X_1 = \mathsf{var}\, Y_1$ in order to have that under the null hypothesis $\widetilde{T}_n \xrightarrow[n \to \infty]{d} \mathsf{N}(0, 1)$.
3. Parametric bootstrap. Suppose that $F = \mathsf{N}(\mu_1, \sigma_1^2)$ and $G = \mathsf{N}(\mu_2, \sigma_2^2)$. Thus the null hypothesis can be written as $H_0 : \mu_1 = \mu_2$. Let us generate $X_{1,b}^*, \dots, X_{n_1,b}^*$ and $Y_{1,b}^*, \dots, Y_{n_2,b}^*$ as independent random samples from the distributions $\mathsf{N}(0, S_X^2)$ and $\mathsf{N}(0, S_Y^2)$ respectively. Based on these bootstrap samples calculate $|\widetilde{T}_{n,1}^*|, \dots, |\widetilde{T}_{n,B^*}|$. Alternatively one could use also a test statistic $T_{n0} = \left| \overline{X}_{n_1} - \overline{Y}_{n_2} \right|$, but it is recommended to use a test statistic whose asymptotic distribution under the null hypothesis does not depend on unknown parameters.
4. Standard nonparametric bootstrap. Suppose that $\mathsf{var}\, X_1, \mathsf{var}\, Y_1 \in (0, \infty)$. Let us generate $X_{1,b}^*, \dots, X_{n_1,b}^*$ and $Y_{1,b}^*, \dots, Y_{n_2,b}^*$ as independent random samples with replacement from $X_1 - \overline{X}_{n_1}, \dots, X_{n_1} - \overline{X}_{n_1}$ and $Y_1 - \overline{Y}_{n_2}, \dots, Y_{n_2} - \overline{Y}_{n_2}$ respectively.

4.6 Permutation tests

Permutation tests are interesting in particular in two (or more generally k) sample problems and when testing for independence.

Two-sample problems

Let $X_1, \dots, X_{n1}$ and $Y_1, \dots, Y_{n2}$ be two independent random samples with the distribution functions F and G respectively. Let the null hypothesis states that the distributions functions F and G coincide, i.e. $H_0 : F(x) = G(x)$ for all $x \in \mathbb{R}$.

Put $n = n_1 + n_2$ and denote $\mathbb{Z} = (Z_1, \ldots, Z_n)$ the joint sample, that is $Z_i = X_i$ for $i = 1, \ldots, n_1$ and $Z_i = Y_{i-n_1}$ for $i = n_1 + 1, \ldots, n$. Let $\mathbb{Z}_{(\cdot)} = (Z_{(1)}, \ldots, Z_{(n)})$ be the ordered sample, that is $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$. Note that under the null hypothesis the random variables $Z_1, \ldots, Z_n$ are independent and identically distributed. Thus the conditional distribution of $\mathbb{Z}$ given $\mathbb{Z}_{(\cdot)}$ is a discrete uniform distribution on the set of all permutations of $\mathbb{Z}_{(\cdot)}$. More formally,

$$\mathsf{P}\left(\mathbb{Z} = (z_1, \ldots, z_n) \,|\, \mathbb{Z}_{(\cdot)} = (z_{(1)}, \ldots, z_{(n)})\right) = \frac{1}{n!}\,\mathbb{I}\big\{(z_1, \ldots, z_n) \text{ is a permutation of } (z_{(1)}, \ldots, z_{(n)})\big\},$$

where $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$.

The samples $\mathbb{Z}_1^*, \ldots, \mathbb{Z}_B^*$ are now generated by randomly permuting the joint sample $\mathbb{Z}$. Now for each $b \in \{1, \ldots, B\}$ the test statistic $T_{n,b}^*$ is recalculated from

$$\big(X_{1,b}^*, \ldots, X_{n_1,b}^*\big) = \big(Z_{1,b}^*, \ldots, Z_{n_1,b}^*\big), \qquad \big(Y_{1,b}^*, \ldots, Y_{n_2,b}^*\big) = \big(Z_{n_1+1,b}^*, \ldots, Z_{n,b}^*\big)$$

and the p-value is estimated by (71).

Note that the test assumes that under the null hypothesis the distribution functions F and G coincide. Then the permutation test is called *exact*. In practice it is of interest to know whether the permutation test is useful also to test for instance the null hypothesis that $\mathsf{E}\, X_1 = \mathsf{E}\, Y_1$ without assuming that $F \equiv G$. Usually it can be proved that if the test statistic T_n under null hypothesis has a limiting distribution that does not depend on the unknown parameters, then the permutation test holds the prescribed level asymptotically. In this situation the permutation test is called *approximate*. It was shown by simulations in many different setting that the level of the approximate permutation test is usually closer to the prescribed value α than the level of the test that directly uses the asymptotic distribution of T_n.

Testing independence

Suppose we observe independent and identically distributed random vectors

$$\mathbf{Z}_1 = (X_1, Y_1)^\mathsf{T}, \ldots, \mathbf{Z}_n = (X_n, Y_n)^\mathsf{T}$$

and we are interested in testing the null hypothesis that X_1 is independent with Y_1. Then under the null hypothesis

$$\mathsf{P}\left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} \,\Big|\, \begin{pmatrix} X_1 \\ Y_{(1)} \end{pmatrix} = \begin{pmatrix} x_1 \\ y_{(1)} \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_{(n)} \end{pmatrix} = \begin{pmatrix} x_n \\ y_{(n)} \end{pmatrix}\right) = \frac{1}{n!}\,\mathbb{I}\big\{(y_1, \ldots, y_n) \text{ is a permutation of } (y_{(1)}, \ldots, y_{(n)})\big\}.$$

Thus one can generate $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$ by permuting $Y_1, \ldots, Y_n$ while keeping $X_1, \ldots, X_n$ fixed.

The permutation scheme as described above can be used for instance for assessing the significance of the test statistic based on a correlation coefficient or of the χ^2-test of independence.

Example 38. Let $X_1, \ldots, X_n$ be a random sample such that $\mathsf{var}\, X_1 \in (0, \infty)$ and $H_0 : \mathsf{E}\, X_1 = \mu_0$. In this situation no permutation test is available. But one can use nonparametric bootstrap and generate $X_{b,1}^*, \ldots, X_{b,n}^*$ as a simple random sample with replacement from $X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n$. A possible test statistic is then

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_0)}{S_n},$$

and $T_{n,b}^* = \frac{\sqrt{n}\,(\overline{X}_{n,b}^* - 0)}{S_{n,b}^*}$, where $\overline{X}_{n,b}^*$ and $S_{X,b}^*$ are the sample mean and sample deviation calculated from the bootstrap sample.

Example 39. Permutation test and χ^2-test of independence.

Literature: Davison and Hinkley (1997) Chapters 4.1–4.4, Efron and Tibshirani (1993) Chapters 15 and 16

4.7 Bootstrap in linear models

Suppose we observe $(\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$ a random sample, where, $\boldsymbol{X}_i$ is a p-dimensional random vector. The standard nonparametric bootstrap generate $(\boldsymbol{X}_1^{*T}, Y_1^*)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^{*T}, Y_n^*)^\mathsf{T}$ as a simple random sample with replacement from the vectors $(\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$. Note that this bootstrap method works as long as there exists an asymptotic distribution of the estimator $\widehat{\boldsymbol{\beta}}_n$.

In linear models we usually assume a more specific structure

$$Y_i = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{72}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed zero-mean random variables independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Then the **model-based boostrap** runs as follows. Let $\widehat{\boldsymbol{\beta}}_n$ be the estimate of $\boldsymbol{\beta}$. Calculate the standardized residuals as

$$\hat{\varepsilon}_i = \frac{Y_i - \widehat{\boldsymbol{\beta}}_n^\mathsf{T} X_i}{\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n,$$

where h_{ii} is the i-th diagonal element of the projection matrix $\mathbb{H} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$. Then one can generate the response in the bootstrap sample as

$$Y_i^* = \widehat{\boldsymbol{\beta}}_n^\mathsf{T} \boldsymbol{X}_i + \varepsilon_i^*, \quad i = 1, \ldots, n,$$

where $\varepsilon_1^*, \ldots, \varepsilon_n^*$ is a simple random sample with replacement from the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. As the covariate values are fixed the bootstrap sample is given by $(\boldsymbol{X}_1^\mathsf{T}, Y_1^*)^\mathsf{T}, \ldots, (\boldsymbol{X}_n^\mathsf{T}, Y_n^*)^\mathsf{T}$.

The advantage of the nonparametric bootstrap is that it does not require model (72) to hold. On the other hand if model (72) holds then the distribution of $\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)$ from the model based bootstrap is closer to the conditional distribution of $\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ given the values of the covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ than the corresponding distribution from the nonparametric bootstrap. Further, the model based bootstrap can be also used in the case of a fixed design. On the other hand this method is not appropriate for instance in the presence of heteroscedasticity.

Literature: Davison and Hinkley (1997) Chapter 6.3

4.8 Variance estimation and boostrap

Often one knows that

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, \boldsymbol{\Sigma}\big),$$

but matrix $\boldsymbol{\Sigma}$ typically depends on unknown parameters (or it might be 'too difficult' to derive the analytic form of $\boldsymbol{\Sigma}$). In such a situation a straightforward bootstrap estimation of the asymptotic variance matrix $\boldsymbol{\Sigma}_n = \frac{1}{n}\,\boldsymbol{\Sigma}$ is given by

$$\widehat{\Sigma}_{n,B}^* = \frac{1}{B-1}\sum_{b=1}^{B}\big(\widehat{\boldsymbol{\theta}}_{n,b}^* - \overline{\boldsymbol{\theta}}_{n,B}^*\big)\big(\widehat{\boldsymbol{\theta}}_{n,b}^* - \overline{\boldsymbol{\theta}}_{n,B}^*\big)^\mathsf{T}, \qquad \text{where } \overline{\boldsymbol{\theta}}_{n,B}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\theta}}_{n,b}^*. \tag{73}$$

Note that

$$\widehat{\Sigma}_{n,B}^* \xrightarrow[B\to\infty]{\text{alm. surely}} \mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}_n^* \,|\, \mathbb{X}\big).$$

Thus for a valid inference we need that

$$n\;\mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}_n^* \,|\, \mathbb{X}\big) \xrightarrow[n\to\infty]{P} \boldsymbol{\Sigma}. \tag{74}$$

Note that $\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{d} \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ generally **does not imply** that (74) holds. The reason is that $\mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}_n^* \,|\, \mathbb{X}\big)$ estimates $\mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}_n\big)$ rather than $\frac{1}{n}\boldsymbol{\Sigma}$.

Example 40. Let $X_1, \ldots, X_n$ be a random sample from the distribution with the density $f(x) = \frac{3}{x^4}\mathbb{I}[x \geq 1]$. Then by the central limit theorem

$$\sqrt{n}\,\big(\overline{X}_n - \tfrac{3}{2}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \tfrac{3}{4}\big).$$

Further consider the transformation $g(x) = \mathrm{e}^{x^4}$. Then with the help of Δ-theorem (Theorem 3) one gets

$$\sqrt{n}\,\big[g(\overline{X}_n) - g\big(\tfrac{3}{2}\big)\big] \xrightarrow[n\to\infty]{d} \mathsf{N}\Big(0, \big[g'\big(\tfrac{3}{2}\big)\big]^2 \cdot \tfrac{3}{4}\Big).$$

But it is straightforward to calculate that $\mathsf{E}\,\big(g(\overline{X}_n)\big) = \infty$ and thus $\mathsf{var}\,\big(g(\overline{X}_n)\big)$ does not exist. Further it can be proved that $\mathsf{var}\,\big(g(\overline{X}_n^*)|\mathbb{X}\big) \xrightarrow[n\to\infty]{\text{alm. surely}} \infty$.

Literature: Efron and Tibshirani (1993) Chapters 6 and 7, Shao and Tu (1996) Chapter 3.2.2

4.9 Bias reduction and bootstrap

In practice one can get unbiased estimators for only very simple models. Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of $\boldsymbol{\theta}_X$ and put $\mathbf{b}_n = \mathsf{E}\,\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X$ for the bias of $\widehat{\boldsymbol{\theta}}_n$. The bias $\mathbf{b}_n$ can be estimated by $\mathbf{b}_n^* = \mathsf{E}\,[\widehat{\boldsymbol{\theta}}_n^*|\boldsymbol{X}] - \widehat{\boldsymbol{\theta}}_n$. The bias corrected estimator of $\boldsymbol{\theta}$ is then defined as $\widehat{\boldsymbol{\theta}}_n^{(bc)} := \widehat{\boldsymbol{\theta}}_n - \mathbf{b}_n^*$.

Example 41. Let $X_1, \ldots, X_n$ be a random sample, $\mathsf{E}\,X_1^4 < \infty$ and $g : \mathbb{R} \to \mathbb{R}$ be such that g''' is bounded and continuous in a neighbourbood of $\mu = \mathsf{E}\,X_1$. Then $\overline{X}_n$ is an unbiased estimator of μ. But if g is not linear then $g(\overline{X}_n)$ is not an unbiased estimator of $g(\mu)$. Put $\sigma^2 = \mathsf{var}(X_1)$. Then the bias of $g(\overline{X}_n)$ can be approximated by

$$\begin{aligned} \mathsf{E}\,g(\overline{X}_n) - g(\mu) &= \mathsf{E}\left\{g'(\mu)(\overline{X}_n - \mu) + \tfrac{g''(\mu)}{2}(\overline{X}_n - \mu)^2\right\} + \frac{R_n}{3!} \\ &= \frac{g''(\mu)\,\sigma^2}{2\,n} + O\Big(\tfrac{1}{n^{3/2}}\Big), \end{aligned} \tag{75}$$

where we have used that

$$|R_n| \le \sup_x \left|g'''(x)\right| \mathsf{E}\left|\overline{X}_n - \mu\right|^3 \le \sup_x \left|g'''(x)\right| \left[\mathsf{E}\left|\overline{X}_n - \mu\right|^4\right]^{3/4} = \left[O\big(\tfrac{1}{n^2}\big)\right]^{3/4} = O\Big(\tfrac{1}{n^{3/2}}\Big).$$

Analogously one can calculate that

$$\begin{aligned} b_n^* = \mathsf{E}\,[g(\overline{X}_n^*)\,|\,\mathbb{X}] - g(\overline{X}_n) &= \frac{g''(\overline{X}_n)}{2n}\ \mathsf{var}[X_1^*\,|\,\mathbb{X}] + O_P\Big(\tfrac{1}{n^{3/2}}\Big) \\ &= \frac{g''(\overline{X}_n)\,\widehat{\sigma}_n^2}{2n} + O_P\Big(\tfrac{1}{n^{3/2}}\Big). \end{aligned} \tag{76}$$

where $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Now by comparing (75) and (76) one gets that the bias of the estimator $\widehat{\theta}_n^{(bc)}$ is given by

$$b_n - b_n^* = \frac{1}{2n}\Big(g''(\mu)\sigma^2 - g''(\overline{X}_n)\widehat{\sigma}_n^2\Big) + O_P\Big(\tfrac{1}{n^{3/2}}\Big) = O_P\Big(\tfrac{1}{n^{3/2}}\Big),$$

where we used that by the delta-theorem

$$g''(\overline{X}_n) = g''(\mu) + O_P\big(\tfrac{1}{\sqrt{n}}\big), \qquad \widehat{\sigma}_n^2 = \sigma^2 + O_P\big(\tfrac{1}{\sqrt{n}}\big).$$

Literature: Efron and Tibshirani (1993) Chapter 10.

4.10 Jackknife

Jackknife can be considered as an ancestor of bootstrap. It was originally suggested to reduce bias of an estimator. Later it was found out that it can be often also used to estimate the variance of an estimator.

Bias reduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample and denote $\mathbf{T}_n = \mathbf{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ the estimator of the parameter of interest $\boldsymbol{\theta}_X$. Put

$$\mathbf{T}_{n-1,i} = \mathbf{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \ldots \boldsymbol{X}_n)$$

for the estimate when the i-th observation is left out. Further put $\overline{\mathbf{T}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{T}_{n-1,i}$. Then the bias of the estimator $\mathbf{T}_n$ is estimated by

$$\widehat{\mathbf{b}}_n = (n-1)\left(\overline{\mathbf{T}}_n - \mathbf{T}_n\right) \tag{77}$$

and the 'bias-corrected' estimator is defined as

$$\mathbf{T}_n^{(bc)} = \mathbf{T}_n - \widehat{\mathbf{b}}_n. \tag{78}$$

Remark 16. The rational of the estimator (78) is as follows. For simplicity let θ be a one-dimensional parameter and suppose that the bias of estimator T_n is given by

$$\mathsf{E}\, T_n - \theta_X = \frac{a}{n} + \frac{b}{n^{3/2}} + \frac{c}{n^2} + o\Big(\tfrac{1}{n^{5/2}}\Big). \tag{79}$$

Then also analogously

$$\mathsf{E}\, T_{n-1,i} - \theta_X = \frac{a}{n-1} + \frac{b}{(n-1)^{3/2}} + \frac{c}{(n-1)^2} + o\Big(\tfrac{1}{(n-1)^{5/2}}\Big),$$

and the same holds true also for $\overline{T}_n = \frac{1}{n}\sum_{i=1}^n T_{n-1,i}$. This further implies that

$$\begin{aligned}
\mathsf{E}\,\widehat{b}_n &= (n-1)\Big(\frac{a}{n-1} + \frac{b}{(n-1)^{3/2}} + \frac{c}{(n-1)^2} + o\Big(\tfrac{1}{(n-1)^{5/2}}\Big) \\
&\qquad - \frac{a}{n} - \frac{b}{n^{3/2}} - \frac{c}{n^2} - o\Big(\tfrac{1}{n^{5/2}}\Big)\Big), \\
&= (n-1)\left(\frac{a}{n(n-1)} + \frac{b(1-\frac{1}{n})^{3/2}}{(n-1)^{3/2}} + \frac{c(1-\frac{1}{n})^2}{(n-1)^2}\right) + O\Big(\tfrac{1}{n^{3/2}}\Big) \\
&= \frac{a}{n} + O\Big(\tfrac{1}{n^{3/2}}\Big).
\end{aligned} \tag{80}$$

Now combining (79) and (80) gives that

$$\mathsf{E}\, T_n^{(bc)} - \theta_X = O\big(\tfrac{1}{n^{3/2}}\big) \quad \text{while} \quad \mathsf{E}\, T_n - \theta_X = O\big(\tfrac{1}{n}\big).$$

Variance estimation

To estimate the variance, let us define *jackknife pseudovalues* as

$$\widetilde{\mathbf{T}}_{n,i} = n\,\mathbf{T}_n - (n-1)\,\mathbf{T}_{n-1,i}, \qquad i = 1, \ldots, n.$$

Then (under some regularity assumptions) the variance of $\mathbf{T}_n$ can be estimated as if $\mathbf{T}_n$ was a mean of jackknife pseudovalues $\widetilde{\mathbf{T}}_{n,1}, \ldots, \widetilde{\mathbf{T}}_{n,n}$ that are independent, i.e.

$$\widehat{\mathsf{var}\,(\mathbf{T}_n)} = \frac{1}{n}\,S^2_{\widetilde{\mathbf{T}}_n}, \quad \text{where} \quad S^2_{\widetilde{\mathbf{T}}_n} = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\widetilde{\mathbf{T}}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\widetilde{\mathbf{T}}_{n,j}\Big)\Big(\widetilde{\mathbf{T}}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\widetilde{\mathbf{T}}_{n,j}\Big)^{\mathsf{T}}.$$

Literature: Shao and Tu (1996) Chapter 1.3.

5 Quantile regression

Generally speaking, while the least square method aims at estimating (modelling) a conditional expectation, quantile regression aims at estimating (modelling) a conditional quantile.

5.1 Introduction

For a given $\tau \in (0,1)$ consider the following loss function

$$\rho_\tau(x) = \tau\, x\, \mathbb{I}\{x > 0\} + (1-\tau)\,(-x)\,\mathbb{I}\{x \le 0\}.$$

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 6. *Let the random variable X have a cumulative distribution function F and $\mathsf{E}\,|X| < \infty$. Then*

$$F^{-1}(\tau) = \arg\min_{\theta\in\mathbb{R}} \mathsf{E}\,\rho_\tau(X-\theta). \tag{81}$$

Proof. Put $M(\theta) = \mathsf{E}\,\rho_\tau(X-\theta) - \mathsf{E}\,\rho_\tau(X)$. One can calculate

$$\begin{aligned} M(\theta) &= -\mathsf{E}\int_0^\theta \psi_\tau(X-t)\,dt = -\int_0^\theta \mathsf{E}\,\psi_\tau(X-t)\,dt \\ &= -\int_0^\theta \tau\;\mathsf{P}(X > t) - (1-\tau)\,\mathsf{P}(X < t)\,dt. \\ &= -\int_0^\theta \tau - \tau\,F(t) - (1-\tau)F(t)\,dt. \\ &= -\tau\,\theta + \int_0^\theta F(t)\,dt. \end{aligned}$$

Now for each $\theta < F^{-1}(\tau)$

$$M'(\theta_-) = -\tau + F(\theta_-) \le -\tau + F(\theta) < 0 \text{ and } M'(\theta_+) = -\tau + F(\theta_+) = -\tau + F(\theta) < 0.$$

As the function $M(\theta)$ is continuous, this implies that $M(\theta)$ is decreasing on $\big(-\infty, F^{-1}(\tau)\big)$.

Analogously one can show that 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 $X_1, \dots, X_n$ from a continuous distribution. Then by Lemma 6

$$F_n^{-1}(\tau) = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(X_i - \theta).$$

Let $X_{(1)}, \dots, X_{(n)}$ be the ordered sample. Note that from the proof of Lemma 6 it also follows that if $i_0 = n\,\tau$ is an integer then the function $M(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(X_i - \theta)$ satisfies $M'(\theta_-) = 0$ for each $\theta \in \big(X_{(i_0)}, X_{(i_0+1)}\big]$ and $M'(\theta_+) = 0$ for each $\theta \in \big[X_{(i_0)}, X_{(i_0+1)}\big)$. Thus $M(\theta)$ is minimised by any value from the interval $\big[X_{(i_0)}, X_{(i_0+1)}\big]$. In this situation $F_n^{-1}(\tau) = X_{(i_0)}$ is the left point of this interval.

5.2 Regression quantiles

Suppose that one observes independent and identically distributed random vectors

$$(\boldsymbol{X}_1^\mathsf{T}, Y_1)^\mathsf{T}, \dots, (\boldsymbol{X}_n^\mathsf{T}, Y_n)^\mathsf{T}$$

being distributed as the generic vector $(\boldsymbol{X}^\mathsf{T}, Y)^\mathsf{T}$,

The τ-th regression quantile is defined as

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

At the population level the regression quantile identifies the parameter

$$\boldsymbol{\beta}_X(\tau) = \arg\min_{\mathbf{b} \in \mathbb{R}^p} \mathsf{E}\, \rho_\tau(Y - \mathbf{b}^\mathsf{T} \boldsymbol{X}).$$

Note that thanks to (81)

$$\begin{aligned} \mathsf{E}\, \rho_\tau(Y - \mathbf{b}^\mathsf{T} \boldsymbol{X}) &= \mathsf{E}\Big\{ \mathsf{E}\left[\rho_\tau(Y - \mathbf{b}^\mathsf{T} \boldsymbol{X}) \,|\, \boldsymbol{X} \right] \Big\} \\ &\ge \mathsf{E}\Big\{ \mathsf{E}\left[\rho_\tau\big(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau)\big) \,|\, \boldsymbol{X} \right] \Big\} = \mathsf{E}\, \rho_\tau\big(Y - F^{-1}_{Y|\boldsymbol{X}}(\tau)\big), \end{aligned}$$

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 $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}$, then $\boldsymbol{\beta}_X(\tau) = \boldsymbol{\beta}$.

Often in applications we assume that $\boldsymbol{X}_i = \left(1, \widetilde{\boldsymbol{X}}_i^{\mathsf{T}}\right)^{\mathsf{T}}$ and that model (58) holds. Then $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X} + F^{-1}_{\varepsilon}(\tau)$, where $F^{-1}_{\varepsilon}(\tau)$ is the τ-th quantile of the random error ε. Thus provided model (58) holds

$$\boldsymbol{\beta}_X(\tau) = \begin{pmatrix} \beta_0 + F^{-1}_{\varepsilon}(\tau) \\ \boldsymbol{\beta} \end{pmatrix}. \tag{82}$$

Thus if model (58) 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.

Example 42. 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

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{83}$$

where

$$x_i = \begin{cases} 0, & i = 1, \ldots, n_1 \\ 1, & i = n_1 + 1, \ldots, n_1 + n_2 \end{cases}$$

and ε_i has a cumulative distribution function F. Usually we are interested in estimating β_1. By the LS method one gets

$$\hat{\beta}_1 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} Y_i - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i \xrightarrow[n_1,n_2\to\infty]{P} \underbrace{\mu_G - \mu_F}_{=:\mu} =: \beta_1^{LS},$$

where μ_F and μ_G stand for the expectation of an observation from the first and second sample respectively.

On the other hand the quantile regression yields

$$\begin{aligned} \widehat{\beta}(\tau) &= \arg\min_{b_0,b_1} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - b_0 - b_1 x_i) \\ &= \arg\min_{b_0,b_1} \frac{1}{n}\left(\sum_{i=1}^{n_1} \rho_\tau(Y_i - b_0) + \sum_{i=n_1+1}^{n_1+n_2} \rho_\tau(Y_i - b_0 - b_1)\right), n = n_1 + n_2. \end{aligned}$$

The first sum is minimised by

$$\widehat{\beta_0}(\tau) = F^{-1}_{n_1}(\tau)$$

and the second sum by

$$\widehat{\beta_0(\tau) + \beta_1}(\tau) = G^{-1}_{n_2}(\tau)$$

Thus we get

$$\widehat{\beta}_1(\tau) = G_{n_2}^{-1}(\tau) - F_{n_1}^{-1}(\tau) \underset{n_1,n_2\to\infty}{\xrightarrow{P}} G^{-1}(\tau) - F^{-1}(\tau) := \beta_1(\tau).$$

Further if model (83) 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

$$\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{b}^\mathsf{T} \boldsymbol{X}_i)$$

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

$$\begin{aligned}
&\sum_{j=1}^{p} X_{ij}\, b_j + r_i^+ - r_i^- = Y_i, \qquad i = 1,\dots,n,\\
&r_i^+ \geq 0,\ r_i^- \geq 0, \qquad i = 1,\dots,n,\\
&b_j \in \mathbb{R}, \qquad j = 1,\dots,p.
\end{aligned}$$

Note that one can think of r_i^+ and r_i^- as the positive or negative part of the residuals, i.e.

$$r_i^+ = \left(Y_i - \mathbf{b}^\mathsf{T} \boldsymbol{X}_i\right)_+, \quad r_i^- = \left(Y_i - \mathbf{b}^\mathsf{T} \boldsymbol{X}_i\right)_- .$$

This can be solved for instance with the help of *the simplex algorithm*.

5.3 Inference for regression quantiles

The following theorem could be proved completely analogously as Theorem 11.

Theorem 15. *Let model (58) holds and $\boldsymbol{\beta}_X$ be given by (82). Further, let $\mathsf{E}\,\|\boldsymbol{X}\|^3 < \infty$, the matrix $\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ is positive definite and the density $f_\varepsilon(x)$ of ε is positive and continuous in a neighbourhood of $F_\varepsilon^{-1}(\tau)$. Then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}_X(\tau)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]^{-1} \frac{\boldsymbol{X}_i\, \psi_\tau\left(\varepsilon_i - F_\varepsilon^{-1}(\tau)\right)}{f_\varepsilon(F_\varepsilon^{-1}(\tau))} + o_P(1), \tag{84}$$

which further implies that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}_X(\tau)\right) \underset{n\to\infty}{\xrightarrow{d}} \mathsf{N}_p\Big(\mathbf{0}, \left[\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right]^{-1} \tfrac{\tau(1-\tau)}{f_\varepsilon^2(F_\varepsilon^{-1}(\tau))}\Big). \tag{85}$$

Note that (85) follows from (84) and the central limit theorem (for i.i.d. random vectors) where we have used that

$$\begin{aligned}\mathsf{var}\left(\psi_\tau(\varepsilon_1^\tau)\boldsymbol{X}_1\right) &= \mathsf{var}\left(\mathsf{E}\left[\boldsymbol{X}\,\psi_\tau(\varepsilon_1^\tau)\,|\,\boldsymbol{X}\right]\right) + \mathsf{E}\left(\mathsf{var}\left[\boldsymbol{X}\,\psi_\tau(\varepsilon_1^\tau)\,|\,\boldsymbol{X}\right]\right)\\ &= \mathbf{0}_{p\times p} + \mathsf{E}\left(\boldsymbol{X}\,\mathsf{var}[\psi_\tau(\varepsilon_1^\tau)]\boldsymbol{X}^\mathsf{T}\right) = \mathsf{E}\left(\boldsymbol{X}\tau(1-\tau)\boldsymbol{X}^\mathsf{T}\right),\end{aligned}$$

with $\varepsilon_1^\tau = \varepsilon_1 - F_\varepsilon^{-1}(\tau)$.

Estimation of asymptotic variance of $\widehat{\boldsymbol{\beta}}_n(\tau)$

Note that by (85) one gets

$$\widehat{\mathsf{avar}\left(\widehat{\boldsymbol{\beta}}_n(\tau)\right)} = \tfrac{1}{n}\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}\tfrac{\tau(1-\tau)}{f_\varepsilon^2\left(F_\varepsilon^{-1}(\tau)\right)}.$$

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) the author suggests that one can use the following estimate

$$\widehat{s}_n(\tau) = \frac{\widehat{F}_{n\varepsilon}^{-1}(\tau+h_n) - \widehat{F}_{n\varepsilon}^{-1}(\tau-h_n)}{2\,h_n},$$

where $\widehat{F}_{n\varepsilon}(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i - \boldsymbol{X}_i^\mathsf{T}\,\widehat{\boldsymbol{\beta}}_n(\tau) \le y\}$ is the empirical distribution function of the residuals and (the bandwidth) h_n is sequence going to zero as $n\to\infty$. A possible choice of h_n (derived when assuming normal errors $\varepsilon,\ldots,\varepsilon_n$) is given by

$$h_n = n^{-1/3}\,u_{1-\alpha/2}^{2/3}\left[\tfrac{1.5\,\varphi^2(u_\tau)}{2\,u_\tau^2+1}\right]^{1/3},$$

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.

An alternative option for doing the inference would be to use **bootstrap**.

5.4 Interpretation of the regression quantiles

Provided $F_{Y|\boldsymbol{X}}^{-1}(\tau) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$ and the model is correctly specified then one can interpret $\widehat{\beta}_{nk}$ $\left(\text{the } k\text{-th element of } \widehat{\boldsymbol{\beta}}_n\right)$ as the estimated change of the conditional quantile of the response when the k-th element of the explanation variable increases by 1.

It is worth noting that if one models the conditional quantile of the transformed response, that is one assumes that $F_{h(Y)|\boldsymbol{X}}^{-1}(\tau) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$ for a given increasing transformation h, then

$$\tau = \mathsf{P}\left(h(Y) \le \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\,|\,\boldsymbol{X}\right) = \mathsf{P}\left(Y \le h^{-1}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})\,|\,\boldsymbol{X}\right),$$

which implies that $F^{-1}_{Y|\boldsymbol{X}}(\tau) = h^{-1}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})$. Analogously $F^{-1}_{Y|\boldsymbol{X}}(1-\tau) = h^{-1}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})$ for h decreasing. That is unlike for modelling of conditional expectation (through the least square method), here we still have a link between $\boldsymbol{\beta}$ and and the quantile of the original (not transformed) response $F^{-1}_{Y|\boldsymbol{X}}(\tau)$.

A very common and popular transformation is log-transformation, i.e. $h(y) = \log y$. This results in $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}}$ and e^{β_k} 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.

5.5 Asymptotic normality of sample quantiles

Suppose that you have a random sample $Y_1, \ldots, Y_n$, where Y_1 has a cumulative distribution function F. For a fixed $\tau \in (0,1)$ put

$$F_n^{-1}(\tau) = \inf\big\{x : F_n(x) \geq \tau\big\},$$

where F_n is the empirical distribution function.

The following theorem is a consequence of Theorem 15.

Theorem 16. *Let $f(y)$ (the density of Y_1) be positive and continuous in a neighbourhood of $F^{-1}(\tau)$. Then*

$$\sqrt{n}\left(F_n^{-1}(\tau) - F^{-1}(\tau)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\psi_\tau\big(Y_i - F^{-1}(\tau)\big)}{f(F^{-1}(\tau))} + o_P(1),$$

which further implies that

$$\sqrt{n}\left(F_n^{-1}(\tau) - F^{-1}(\tau)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\Big(0, \tfrac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}\Big).$$

Proof. The proof follows from Theorem 11 by taking $\boldsymbol{X}_i = 1$, $\varepsilon_i = Y_i$ and noting that by Lemma 6 and Remark 17 one has $F_n^{-1}(\tau) = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - \theta)$. □

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.

Example 43. Let $X_1, \ldots, X_n$ be a random sample from the distribution with the density

$$f(x) = \sum_{j=1}^{G} \pi_j \, f_j(x),$$

where $f_1, \ldots, f_G$ are known densities and $\pi_1, \ldots, \pi_G$ are 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 Θ stands for the parametric space given by the possible values of $\boldsymbol{\pi}$.

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 $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^\mathsf{T}$, 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 $(X_1, \mathbf{Z}_1^\mathsf{T})^\mathsf{T}, \ldots, (X_n, \mathbf{Z}_n^\mathsf{T})^\mathsf{T}$, where $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are missing.

Put $\mathbb{X} = (X_1, \ldots, X_n)^\mathsf{T}$. The joint density of $(X_1, \mathbf{Z}_1^\mathsf{T})^\mathsf{T}$ is given by

$$f_{X,\mathbf{Z}}(x, \mathbf{z}; \boldsymbol{\pi}) = f_{X|\mathbf{Z}}(x|\mathbf{z}; \boldsymbol{\pi}) \, f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\pi}) = \left(\sum_{j=1}^{G} z_j f_j(x) \right) \cdot \left(\prod_{j=1}^{G} \pi_j^{z_j} \right).$$

The complete log-loglikehood is given by

$$\begin{aligned} \ell_n^C(\boldsymbol{\pi}) &= \log \left\{ \prod_{i=1}^{n} \left[\left(\sum_{j=1}^{G} Z_{ij} f_j(X_i) \right) \left(\prod_{j=1}^{G} \pi_j^{Z_{ij}} \right) \right] \right\} \\ &= \sum_{i=1}^{n} \left[\log \left(\sum_{j=1}^{G} Z_{ij} f_j(X_i) \right) \right] + \sum_{i=1}^{n} \left[\sum_{j=1}^{G} Z_{ij} \log \pi_j \right]. \end{aligned}$$

If we knew $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, then we would estimate simply $\widehat{\pi}_j = \frac{1}{n}\sum_{i=1}^n Z_{ij}, j = 1, \ldots, G$. The EM algorithm runs in two steps:

(i) E-step (Expectation step): Let $\widehat{\boldsymbol{\pi}}^{(k)}$ be the current estimate of $\boldsymbol{\pi}$. In this step we calculate

$$Q\left(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\right) = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}[\ell_n^C(\boldsymbol{\pi}) \,|\, \mathbb{X}],$$

where the expectation is taken with respect to the unobserved random variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ provided that $\mathbf{Z}_i$ follows a multinomial distribution $\mathsf{Mult}(1, \hat{\boldsymbol{\pi}}^{(k)})$.

(ii) M-step (Maximization): The updated value of the estimate of $\boldsymbol{\pi}$ is calculated as

$$\hat{\boldsymbol{\pi}}^{(k+1)} = \arg\max_{\boldsymbol{\pi}\in\Theta} Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big).$$

E-step in a detail:

$$Q\left(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\right) = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^n \log\left(\sum_{j=1}^G Z_{ij} f_j(X_i)\right) \,\middle|\, \mathbb{X}\right] + \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^n \sum_{j=1}^G Z_{ij} \log \pi_j \,\middle|\, \mathbb{X}\right]. \quad (86)$$

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 $\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\big[Z_{ij} \,|\, \mathbb{X}\big]$. To do that denote $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\mathsf{T}$ for the j-th canonical vector. Now with the help of Bayes theorem for densities (see e.g. Theorem 3.21 of Anděl, 2007) one can calculate

$$\begin{aligned}\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\big[Z_{ij} \,|\, \mathbb{X}\big] &= \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\big[Z_{ij} \,|\, X_i\big] = \mathsf{P}_{\widehat{\boldsymbol{\pi}}^{(k)}}(Z_{ij} = 1 \,|\, X_i) = f_{\mathbf{Z}|X}(\mathbf{e}_j | X_i; \widehat{\boldsymbol{\pi}}^{(k)}) \\ &= \frac{f_{X|\mathbf{Z}}(X_i|\mathbf{e}_j; \widehat{\boldsymbol{\pi}}^{(k)}) f_{\mathbf{Z}}(\mathbf{e}_j; \widehat{\boldsymbol{\pi}}^{(k)})}{f_X(X_i; \boldsymbol{\pi}^{(k)})} = \frac{f_j(X_i)\,\widehat{\pi}_j^{(k)}}{\sum_{l=1}^G f_l(X_i)\,\widehat{\pi}_l^{(k)}} =: z_{ij}^{(k)}.\end{aligned}$$

M-step in a detail: Note that with the help of the previous step and (86)

$$Q\left(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\right) = const + \sum_{i=1}^n \sum_{j=1}^G z_{ij}^{(k)} \log \pi_j.$$

Analogously as when calculating the maximum likelihood estimator in 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 $\widehat{\pi}_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^n z_{ij}^{(k)}$.

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 $f(\mathbf{y};\boldsymbol{\theta})$ is 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 $\boldsymbol{Y}$. 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 with the help of the complete log-likelihood one can write

$$\ell_{obs}(\boldsymbol{\theta}) = \ell_n^C(\boldsymbol{\theta}) - \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta}), \tag{87}$$

where $f(\mathbf{y}_{mis}|\mathbf{y}_{obs};\boldsymbol{\theta})$ stands for the conditional density of $\mathbb{Y}_{mis}$ given $\mathbb{Y}_{obs} = \mathbf{y}_{obs}$. Finally denote

$$Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}) = \mathsf{E}_{\widetilde{\boldsymbol{\theta}}}\left[\ell_n^C(\boldsymbol{\theta}) \,|\, \mathbb{Y}_{obs}\right]. \tag{88}$$

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 $\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q\big(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\big)$.

Theorem 17. *Let $\ell_{obs}(\boldsymbol{\theta})$ be the observed likelihood and $\widehat{\boldsymbol{\theta}}^{(k)}$ be a result of the k-th iteration of the EM-algorithm. Then*

$$\ell_{obs}\Big(\widehat{\boldsymbol{\theta}}^{(k+1)}\Big) \geq \ell_{obs}\Big(\widehat{\boldsymbol{\theta}}^{(k)}\Big).$$

Proof. Note that the left-hand side of (87) does not depend on $\mathbb{Y}_{mis}$ and thus the same holds true also for the right-hand side. So one can write

$$\begin{aligned} \ell_{obs}(\boldsymbol{\theta}) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}[\ell_n^C(\boldsymbol{\theta}) \,|\, \mathbb{Y}_{obs}] - \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f_{\boldsymbol{\theta}}(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}) \,\middle|\, \mathbb{Y}_{obs}\right] \\ &=: Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right). \end{aligned} \tag{89}$$

From the M-step one knows that

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) \Rightarrow Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \geq Q\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right). \tag{90}$$

Further with the help of Jensen's inequality one gets

$$
\begin{aligned}
H\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log\left(\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\right)\middle|\,\mathbb{Y}_{obs}\right] + \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)|\mathbb{Y}_{obs}\right] \\
&\overset{\text{Jensen}}{\leq} \log\left(\mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log\left(\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\right)\middle|\,\mathbb{Y}_{obs}\right]\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) \\
&= \log\left(\int \frac{f(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)} \cdot f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right) \mathrm{d}\mu(\mathbf{y}_{mis})\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) \\
&= \log(1) + H\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) = H\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right). \qquad (91)
\end{aligned}
$$

Now with the help of (89), (90) and (91) one gets

$$
\begin{aligned}
\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right) - \ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) &= Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) - Q\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) \\
&\quad - \left[H\left(\widehat{\boldsymbol{\theta}}^{(k+1)}, \widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)}\right)\right] \geq 0,
\end{aligned}
$$

which completes the proof. □

In what follows we make the following regularity assumptions.

- The parameter space Θ is a subset of $\mathbb{R}^p$.
- The set $\Theta_0 = \left\{\boldsymbol{\theta} \in \Theta : \ell_{obs}(\boldsymbol{\theta}) \geq \ell_{obs}(\boldsymbol{\theta}_0)\right\}$ 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 18. *Let the function $Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ defined in (88) be continuous both in $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$. Then all the limits points of any instance $\left\{\widehat{\boldsymbol{\theta}}^{(k)}\right\}$ are stationary points of $\ell_{obs}(\boldsymbol{\theta})$. Further $\left\{\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)\right\}$ 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 18 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 18 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 $k \to \infty$.*

6.2 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 $\widehat{\boldsymbol{\theta}}^{(k)}$ 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 $k \to \infty$. Assuming that $\mathbf{M}$ is sufficiently smooth one gets the 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

$$\widehat{\boldsymbol{\theta}}^{(k+1)} - \boldsymbol{\theta}^* = \left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \left(\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right) + o\left(\left\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\|\right) \tag{92}$$

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

$$\left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = [I_n^C(\boldsymbol{\theta}^*)]^{-1} I_n^{mis}(\boldsymbol{\theta}^*), \tag{93}$$

where

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

is the empirical Fisher information matrix from the complete data and

$$I_n^{mis}(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial^2 \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}},$$

is the empirical Fisher information matrix of the contribution of the missing data.

Note that by (92) and (93) 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 speed of convergence.

6.3 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

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbf{y})\right\} b(\boldsymbol{\theta})\, c(\mathbf{y}) \tag{94}$$

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

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\big[\ell_n^C(\boldsymbol{\theta}) | \mathbb{Y}_{obs}\big] = \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\big[T_j(\mathbb{Y}) \,\big|\, \mathbb{Y}_{obs}\big] + \log b(\boldsymbol{\theta}) + const. \\ &= \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \widehat{T}_j^{(k)} + \log b(\boldsymbol{\theta}) + const., \end{aligned}$$

where we put $\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\big[T_j(\mathbb{Y}) \,\big|\, \mathbb{Y}_{obs}\big]$.

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

$$\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\big[T_j(\mathbb{Y}) \,\big|\, \mathbb{Y}_{obs}\big], \quad j = 1, \ldots, p,$$

and in the M-step we maximize

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \widehat{T}_j^{(k)} + \log b(\boldsymbol{\theta}) \right\}. \tag{95}$$

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

$$f(y; \boldsymbol{\theta}) = \exp\Big\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, t_j(y) \Big\}\, b_1(\boldsymbol{\theta})\, c_1(y).$$

Thus the joint density of the random sample $Y_1, \ldots, Y_n$ is of the form (94) where

$$T_j(\mathbb{Y}) = \sum_{i=1}^{n} t_j(Y_i), \qquad j = 1, \ldots, p.$$

Thus in the E-step of the EM-algorithm it is sufficient to calculate

$$\widehat{T}_j^{(k)} = \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[T_j(\mathbb{Y}) \,\middle|\, X_1, \ldots, X_n\right] = \sum_{i=1}^{n} \mathsf{E}\left[t_j(Y_i) \,|\, X_i\right], \quad j = 1, \ldots, p,$$

and the M-step is given by (95) where $b(\boldsymbol{\theta}) = b_1^n(\boldsymbol{\theta})$.

Example 44. 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 (95) one gets that

$$\hat{\lambda}^{(k+1)} = \operatorname*{arg\,max}_{\lambda > 0} \left\{ Q\left(\lambda, \hat{\lambda}^{(k)}\right) \right\} = \operatorname*{arg\,max}_{\lambda > 0} \left\{ -\lambda \sum_{i=1}^{n} \widehat{Y}_i^{(k)} + n \log \lambda \right\} = \frac{1}{\frac{1}{n} \sum \widehat{Y}_i^{(k)}}.$$

6.4 Some further examples of the usage of the EM algorithm

Example 45. 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.

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 $i = 1, \ldots, I$ let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}$ represent the data of the i-th subject that could be ideally observed. Let $\mathbf{R}_i = (R_{i1}, \ldots, R_{in_i})^\mathsf{T}$, where

$$R_{ij} = \begin{cases} 1, & \text{if } Y_{ij} \text{ is observed }, \\ 0, & \text{otherwise.} \end{cases}$$

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 observed 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 $\boldsymbol{\theta}$ 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 observed data

$$f(\mathbf{y}_{obs}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\theta})\, f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\psi})\, \mathrm{d}\mu(\mathbf{y}_{mis}). \tag{96}$$

In what follows we will say that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are *separable* if $\boldsymbol{\theta} \in \Omega_1$, $\boldsymbol{\psi} \in \Omega_2$ and $(\boldsymbol{\theta}, \boldsymbol{\psi})^\mathsf{T} \in \Omega_1 \times \Omega_2$.

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 (96) 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 46. 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. $p_i = \frac{1}{i}$).

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 $f(\mathbf{r}|\mathbf{y};\boldsymbol{\psi}) = f_{\boldsymbol{\psi}}(\mathbf{r}|\mathbf{y}_{obs};\boldsymbol{\psi})$ and with the help of (96)

$$f(\mathbf{y}_{obs}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_{obs}; \boldsymbol{\theta}) f(\mathbf{r}|\mathbf{y}_{obs}; \boldsymbol{\psi}),$$

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 $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable.

Example 47. 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 stands for the indicator of missing of Y_i and

$$\mathsf{P}(R_i = 1 \,|\, \boldsymbol{X}_i, Y_i) = f(\boldsymbol{X}_i),$$

where $f(\mathbf{x})$ is a given (but possibly unknown) function.

Missing not at random (MNAR). In this concept neither the distribution of $\boldsymbol{R}$ is not 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 (96). To proceed one has to make some other assumptions about the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$.

Example 48. Maximum likelihood estimator for the right-censored data from an exponential distribution.

The general problem of all the concepts is that if missing is not a part of the design of the study then any assumptions about the relationship of $\mathbb{Y}_{mis}$ and $\boldsymbol{R}$ cannot be verified as we do not observe $\mathbb{Y}_{mis}$.

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 49. 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 $0.9^5 \doteq 0.59$ per cent of the records will be complete.

Available case analysis (ACA)

In each of the analysis one uses all the data that are available for this particular analysis.

Example 50. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from $\mathsf{N}((\mu_1, \mu_2, \mu_3)^\mathsf{T}, \boldsymbol{\Sigma}_{3\times 3})$. 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 feauters (e.g. there is no guarantee that the estimate of the variance matrix $\widehat{\boldsymbol{\Sigma}}$ in Example 50 is positive semidefinite).

Direct (ignorable) observed likelihood

The inference is based on $\ell_{obs}(\boldsymbol{\theta}) = \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta})$, 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 loose any information provided MAR holds;
– The observed log-likelihood $\ell_{obs}(\boldsymbol{\theta})$ might be difficult to calculate.

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 51. 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 $i = n_0, \dots, n$ 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

$$\widehat{\mathsf{var}(\widehat{\mu}_n)} = \frac{S_n^2}{n}, \quad \text{where} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu}_n)^2$$

is too low. The first reason is that S_n^2 as the estimate of $\mathsf{var}(X_1)$ is too low as

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n_0} (X_i - \widehat{\mu}_n)^2 = \frac{n_0 - 1}{n-1}\, S_{n_0}^2 < S_{n_0}^2.$$

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 observations.

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 $\widehat{\boldsymbol{\theta}}_1, \dots, \widehat{\boldsymbol{\theta}}_M$ be the estimates 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

$$\widehat{\mathsf{var}}(\widehat{\boldsymbol{\theta}}_{MI}) = \overline{\mathbb{V}}_M + \left(1 + \tfrac{1}{M}\right) \mathbb{B}_M, \tag{97}$$

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 $(\mathbb{Y}_{obs}, \widehat{\mathbb{Y}}_{mis}^{(j)})$.

The rational of the formula (97) is as follows. Note that one can think of the imputed values $\widehat{\mathbb{Y}}_{mis}$ as a random vector and write

$$\mathsf{var}\left(\widehat{\boldsymbol{\theta}}_{MI}\right) = \mathsf{E}\left(\mathsf{var}(\widehat{\boldsymbol{\theta}}_{MI} \,|\, \widehat{\mathbb{Y}}_{mis})\right) + \mathsf{var}\left(\mathsf{E}\left(\widehat{\boldsymbol{\theta}}_{MI} \,|\, \widehat{\mathbb{Y}}_{mis}\right)\right).$$

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 52. In Example 51 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 $\widehat{\sigma}^2 = S_{n_0}^2$ are the sample mean and variance calculated from the observed data. Put $\widehat{V}_j = \frac{S_n^{2(j)}}{n}$, where $S_n^{2(j)}$ is the sample variance calculated from the j-th completed sample. Then one can show that

$$\lim_{M\to\infty} \overline{V}_M = \frac{S_{n_0}^2}{n} \quad \text{a.s.}, \tag{98}$$

Further let $\widehat{\theta}_j = \overline{Y}_n^{(j)}$ be the sample mean calculated from the j-th completed sample. Then it can be shown that

$$\lim_{M\to\infty} B_M = \frac{S_{n_0}^2 (n - n_0)}{n^2} \quad \text{a.s.} \tag{99}$$

Now combining (98) and (99) yields that

$$\lim_{M\to\infty} \overline{V}_M + B_M = S_{n_0}^2 \left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right) \quad \text{a.s.}$$

Further it is straightforward to prove that for $n_0 < n$

$$S_{n_0}^2 \left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right) < \frac{S_{n_0}^2}{n_0},$$

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 of the variance of this estimate.

− Requires the knowledge of Bayesian approach to statistics to be done properly.

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 53. 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 subsample $\mathbb{S}$ of size n from these patients has been done for some more expensive measurements resulting in $\{\boldsymbol{X}_i : i \in \mathbb{S}\}$, where $\mathbb{S} = \{j : R_j = 1\}$.

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 wit the help of for instance a logistic regression.

Literature: Little and Rubin (2014) Chapters 1.6, 3, 5.3

8 Kernel density estimation

Suppose we have independent identically distributed random variables $X_1, \ldots, X_n$ drawn from a distribution with the density $f(x)$ with respect to a Lebesgue measure and we are interested in estimating this density nonparametrically.

As

$$f(x) = \lim_{h\to 0_+} \frac{F(x+h) - F(x-h)}{2\,h},$$

a *naive estimator* of $f(x)$ would be

$$\widetilde{f}_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2\,h_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}\{X_i \in (x-h_n, x+h_n]\}}{2h_n}, \tag{100}$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\}$ is the empirical distribution function and (the bandwidth) h_n is a sequence of positive constants going to zero.

It is straightforward to show

$$\mathsf{E}\,\widetilde{f}_n(x) \xrightarrow[n\to\infty]{} f(x) \qquad \text{and} \qquad \mathsf{var}\left(\widetilde{f}_n(x)\right) \xrightarrow[n\to\infty]{} 0,$$

provided that $h_n \to 0$ and at the same time $(n\,h_n) \to \infty$.

Note that the estimator (100) can be rewritten as

$$\widetilde{f}_n(x) = \frac{1}{2\,n\,h_n}\sum_{i=1}^{n} \mathbb{I}\big\{-1 \le \tfrac{x-X_i}{h_n} < +1\big\} = \frac{1}{n\,h_n}\sum_{i=1}^{n} w\big(\tfrac{x-X_i}{h_n}\big), \tag{101}$$

where $w(y) = \frac{1}{2}\,\mathbb{I}\{y \in [-1,1)\}$ can be viewed as the density of the uniform distribution on $[-1,1)$. Generalising (101) we define the kernel estimator of a density as

$$\widehat{f}_n(x) = \frac{1}{n\,h_n}\sum_{i=1}^{n} K\big(\tfrac{x-X_i}{h_n}\big), \tag{102}$$

where the function K is called a kernel function. Usually the function K is taken as a symmetric density of a probability distribution. The common choices of K are summarised in Table 1.

Epanechnikov kernel:	$K(x) = \frac{3}{4}(1-x^2)\,\mathbb{I}\{\lvert x\rvert \le 1\}$
Triangular kernel:	$K(x) = (1-\lvert x\rvert)\,\mathbb{I}\{\lvert x\rvert \le 1\}$
Uniform kernel:	$K(x) = \frac{1}{2}\,\mathbb{I}\{\lvert x\rvert \le 1\}$
Biweight kernel:	$K(x) = \frac{15}{16}(1-x^2)^2\,\mathbb{I}\{\lvert x\rvert \le 1\}$
Tricube kernel:	$K(x) = \frac{70}{81}(1-\lvert x\rvert^3)^3\,\mathbb{I}\{\lvert x\rvert \le 1\}$
Gaussian kernel:	$K(x) = \frac{1}{\sqrt{2\pi}}\,\exp\{-x^2/2\}$

Table 1: Commonly used kernel functions.

Remark 18. Note that:

(i) When compared to a histogram both estimators $\widetilde{f}_n(x)$ and $\widehat{f}_n(x)$ do not require to specify the starting point to calculate the intervals.

(ii) As we usually assume that the density f is continuous, the estimator $\widehat{f}_n(x)$ with a continuous function K is preferred.

(iii) If K is a density of a probability distribution, then $\int \widehat{f}_n(x)\,\mathrm{d}x = 1$.

8.1 Consistency and asymptotic normality

Theorem 19 (Bochner's theorem)**.** *Let the function K satisfy*

$$\text{(B1)}\ \int_{-\infty}^{+\infty} |K(y)|\,dy < \infty, \qquad \text{(B2)}\ \lim_{|y|\to\infty} |y\,K(y)| = 0. \tag{103}$$

Further let the function g satisfy $\int_{-\infty}^{+\infty} |g(y)|\,dy < \infty$. Put

$$g_n(x) = \frac{1}{h_n}\int_{-\infty}^{+\infty} g(z)\,K\big(\tfrac{x-z}{h_n}\big)\,dz,$$

where $h_n \searrow 0$ as $n \to \infty$. Then in each point x of the continuity of g it holds that

$$\lim_{n\to\infty} g_n(x) = g(x)\int_{-\infty}^{+\infty} K(y)\,dy. \tag{104}$$

Proof. Let x be the point of continuity g. We need to show that

$$\lim_{n\to\infty}\left|g_n(x) - g(x)\int K(y)\,\mathrm{d}y\right| \to 0.$$

Using the substitutions $y = x - z$ and $z = \frac{y}{h_n}$ one can write

$$\begin{aligned} g_n(x) - g(x)\int K(z)\,\mathrm{d}z &= \frac{1}{h_n}\int g(x-y)K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y - \frac{g(x)}{h_n}\int K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y \\ &= \frac{1}{h_n}\int [g(x-y) - g(x)]K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y. \end{aligned}$$

Before we proceed note that for each fixed $\delta > 0$:

$$\frac{\delta}{h_n} \to \infty \qquad \text{and} \qquad \frac{1}{\delta}\sup_{t:|t|\ge\frac{\delta}{h_n}} \big|t\,K(t)\big| \to 0, \qquad \text{as} \quad n \to \infty.$$

Thus there exists a sequence of positive constants $\{\delta_n\}$ such that

$$\delta_n \to 0, \quad \frac{\delta_n}{h_n} \to \infty \quad \text{and} \quad \frac{1}{\delta_n}\sup_{t:|t|\ge\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \to 0, \quad \text{as} \quad n \to \infty. \tag{105}$$

Taking δ_n satisfying (105) one can bound

$$\begin{aligned} \left|g_n(x) - g(x)\int K(y)\,\mathrm{d}y\right| \le{}& \underbrace{\frac{1}{h_n}\int_{-\delta_n}^{\delta_n} \big|g(x-y) - g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:A_n} \\ &+ \underbrace{\frac{1}{h_n}\int_{y:|y|\ge\delta_n} \big|g(x-y) - g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_n}. \end{aligned} \tag{106}$$

Dealing with A_n. As g is continuous in the point x

$$A_n \leq \sup_{y:|y|\leq\delta_n} \big|g(x-y)-g(x)\big| \int_{-\delta_n}^{\delta_n} \tfrac{1}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y = o(1) \underbrace{\int_{\mathbb{R}} |K(t)|\,\mathrm{d}t}_{<\infty;\,\text{(B1)}} = o(1), \tag{107}$$

as $n \to \infty$.

Dealing with B_n. Further one can bound B_n with

$$B_n \leq \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x-y)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{1n}} + \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{2n}}. \tag{108}$$

Using the substitution $t = \frac{y}{h_n}$ and (105) one gets

$$B_{2n} = \big|g(x)\big| \int_{y:|y|\geq\delta_n} \tfrac{1}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y = \big|g(x)\big| \int_{t:|t|\geq\frac{\delta_n}{h_n}} |K(t)|\,\mathrm{d}t \xrightarrow[n\to\infty]{} 0. \tag{109}$$

Finally using (105)

$$\begin{aligned} B_{1n} &= \int_{y:|y|\geq\delta_n} \underbrace{\tfrac{|y|}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|}_{\leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}}|t\,K(t)|} \frac{\big|g(x-y)\big|}{|y|}\,\mathrm{d}y \leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \int_{y:|y|\geq\delta_n} \frac{\big|g(x-y)\big|}{|y|}\,\mathrm{d}y \\ &\leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \frac{1}{\delta_n} \underbrace{\int \big|g(x-y)\big|\,\mathrm{d}y}_{=\int|g(y)|\,\mathrm{d}y<\infty} \xrightarrow[n\to\infty]{} 0. \end{aligned} \tag{110}$$

Now combining (106), (107), (108), (109) and (110) yields the statement of the theorem. □

Remark 19. Note that:

(i) If K is density, then $\int |K(y)|\,\mathrm{d}y = 1$ and assumption (B1) holds.

(ii) Assumption (B2) holds true if K has a bounded support. Further from the last part of the proof of Theorem 19 (dealing with B_{1n}) it follows that for K with a bounded support one can drop assumption $\int_{-\infty}^{+\infty} |g(y)|\,dy < \infty$ from Theorem 19.

(iii) If K is a density but with an unbounded support, then assumption (B2) is satisfied for instance when $\int |y|K(y)\,\mathrm{d}y < \infty$, that is there exists the first moment of the distribution given by the density K.

(iv) If g is uniformly continuous then one can show that also the convergence in (104) is uniform.

Theorem 20 (Variance and consistency of $\widehat{f}_n(x)$)**.** *Let the estimator $\widehat{f}_n(x)$ be given by (102) and the function K satisfies* (B1) *and* (B2) *introduced in (103). Further, let $\int K(y)\,dy = 1$, $\sup_{y\in\mathbb{R}} |K(y)| < \infty$, $h_n \searrow 0$ as $n \to \infty$ and $(n\,h_n) \to \infty$ as $n \to \infty$. Then at each point of continuity of f*

(i) $\lim_{n\to\infty} n\,h_n\ \mathsf{var}\left(\widehat{f}_n(x)\right) = f(x)\int K^2(y)\,dy$;

(ii) $\widehat{f}_n(x) \xrightarrow[n\to\infty]{P} f(x)$.

Proof. Let x be the point of continuity of f.

Showing (i). Let us calculate

$$\begin{aligned}\mathsf{var}\left(\widehat{f}_n(x)\right) &= \mathsf{var}\left[\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\tfrac{x-X_i}{h_n}\right)\right] = \frac{1}{n\,h_n^2}\,\mathsf{var}\left[K\left(\tfrac{x-X_1}{h_n}\right)\right]\\ &= \frac{1}{n\,h_n^2}\left[\mathsf{E}\,K^2\left(\tfrac{x-X_1}{h_n}\right) - \left(\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right)\right)^2\right]. \qquad (111)\end{aligned}$$

Now using Theorem 19

$$\frac{1}{h_n}\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right) = \int \frac{1}{h_n}K\left(\tfrac{x-y}{h_n}\right)f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} (x)\int K(y)\,\mathrm{d}y = f(x). \qquad (112)$$

Analogously

$$\frac{1}{h_n}\mathsf{E}\,K^2\left(\tfrac{x-X_1}{h_n}\right) = \frac{1}{h_n}\int K^2\left(\tfrac{x-y}{h_n}\right)f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} (x)\int K^2(y)\,\mathrm{d}y. \qquad (113)$$

where we have used again Theorem 19 with K replaced by K^2. Note that assumptions (B1) and (B2) are satisfied as

$$\text{ad (B1)}: \int |K^2(y)|\,\mathrm{d}y \le \underbrace{\sup_{y\in\mathbb{R}}|K(y)|}_{<\infty}\underbrace{\int |K(y)|\,\mathrm{d}y}_{<\infty} < \infty$$

and

$$\text{ad (B2)}: \lim_{|y|\to\infty}|yK^2(y)| \le \underbrace{\sup_{y\in\mathbb{R}}|K(y)|}_{<\infty}\underbrace{\lim_{|y|\to\infty}|yK(y)|}_{=0} = 0.$$

Now combining (111), (112) and (113) yields

$$n\,h_n\ \mathsf{var}\left(\widehat{f}_n(x)\right) = \underbrace{\frac{1}{h_n}\mathsf{E}\,K^2\left(\tfrac{x-X_1}{h_n}\right)}_{\to f(x)\int K^2(y)\,\mathrm{d}y} - \underbrace{\left[\frac{1}{h_n}\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right)\right]^2}_{\to f(x)} h_n \xrightarrow[n\to\infty]{} f(x)\int K^2(y)\,\mathrm{d}y.$$

Showing (ii). Note that with the help of (112)

$$\mathsf{E}\,\widehat{f}_n(x) = \frac{1}{h_n}\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right) \xrightarrow[n\to\infty]{} f(x). \qquad (114)$$

Now with the help of (i) and (114)

$$\mathsf{E}\left[\widehat{f}_n(x) - f(x)\right]^2 = \mathsf{var}\left[\widehat{f}_n(x)\right] + \left[\mathsf{E}\,\widehat{f}_n(x) - f(x)\right]^2 \xrightarrow[n\to\infty]{} 0,$$

which implies the consistency of $\widehat{f}_n(x)$. □

Remark 20. Note that Theorem 20 implies only pointwise consistency. It would be much more difficult to show that $\sup_{x\in\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \xrightarrow[n\to\infty]{P} 0$.

Theorem 21 (Asymptotic normality of $\widehat{f}_n(x)$)**.** *Let the assumptions of Theorem 20 be satisfied and further that $f(x) > 0$. Then*

$$\frac{\widehat{f}_n(x) - \mathsf{E}\,\widehat{f}_n(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1).$$

Proof. From Theorem 20 we know that

$$\frac{\mathsf{var}\left(\widehat{f}_n(x)\right)}{\frac{f(x)R(K)}{n\,h_n}} \xrightarrow[n\to\infty]{} 1, \tag{115}$$

where $R(K) = \int K^2(y)\,\mathrm{d}y$. Thus thanks to CS (Theorem 2) it is sufficient to consider

$$\frac{\widehat{f}_n(x) - \mathsf{E}\,\widehat{f}_n(x)}{\sqrt{\frac{f(x)R(K)}{n\,h_n}}} = \frac{\frac{1}{\sqrt{n\,h_n}}\sum_{i=1}^n \left[K\left(\frac{x-X_i}{h_n}\right) - \mathsf{E}\,K\left(\frac{x-X_i}{h_n}\right)\right]}{\sqrt{f(x)R(K)}} = \sum_{i=1}^n X_{n,i},$$

where

$$X_{n,i} = \frac{1}{\sqrt{n\,h_n}} \frac{K\left(\frac{x-X_i}{h_n}\right) - \mathsf{E}\,K\left(\frac{x-X_i}{h_n}\right)}{\sqrt{f(x)R(K)}}, \quad i = 1,\ldots,n,$$

are independent identically random variables (with the distribution depending on n). Thus the statement would follow from the Lindeberg-Feller central limit theorem (see e.g. Proposition 2.27 in van der Vaart, 2000), provided its assumptions are satisfied. It is straightforward to verify the assumptions as

$$\mathsf{E}\,X_{n,1} = \cdots = \mathsf{E}\,X_{n,n} = 0 \qquad \text{and} \qquad \sum_{i=1}^n \mathsf{var}(X_{n,i}) \xrightarrow[n\to\infty]{} 1.$$

Further for each $\varepsilon > 0$ for all sufficiently large n it holds that uniformly in $i = 1,\ldots,n$:

$$\begin{aligned}\mathbb{I}\{|X_{n,i}| \geq \varepsilon\} &= \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \left|\frac{K\left(\frac{x-X_i}{h_n}\right) - \mathsf{E}\,K\left(\frac{x-X_i}{h_n}\right)}{\sqrt{f(x)R(K)}}\right| \geq \varepsilon\right\} \\ &\leq \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \frac{2\sup_y |K(y)|}{\sqrt{f(x)R(K)}} \geq \varepsilon\right\} = 0,\end{aligned}$$

which implies that the 'Feller-Lindeberg condition'

$$\lim_{n\to\infty}\sum_{i=1}^{n}\mathsf{E}\left[X_{n,i}^2\,\mathbb{I}\{|X_{n,i}|\geq\varepsilon\}\right]=0$$

is satisfied. □

Remark 21. Note that Theorem 21 implies

$$\frac{\widehat{f}_n(x)-f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}}\xrightarrow[n\to\infty]{d}\mathsf{N}(0,1), \tag{116}$$

only if

$$\frac{\mathsf{E}\,\widehat{f}_n(x)-f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}}=\frac{\mathrm{bias}(\widehat{f}_n(x))}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}}\xrightarrow[n\to\infty]{}0,$$

which depends on the rate of h_n. As we will see later, typically we have

$$\frac{\mathsf{E}\,\widehat{f}_n(x)-f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}}=\frac{O(h_n^2)}{\sqrt{O\left(\frac{1}{n\,h_n}\right)}}=O\left(\sqrt{n\,h_n^5}\right)$$

and thus $\lim_{n\to\infty} n\,h_n^5=0$ is needed to show (116). But this would require that $h_n=o\big(n^{-1/5}\big)$ which would exclude the optimal bandwidth choice, see the next section.

8.2 Bandwidth choice

Basically we distinguish two situations:

(i) h_n depends on x (on the point where we estimate the density f), then we speak about the *local bandwidth*;

(ii) the same h_n is used for all x, then we speak about the *global bandwidth*.

The standard methods of choosing the bandwidth are based on **the mean squared error**

$$\mathrm{MSE}(\widehat{f}_n(x))=\mathsf{var}\left(\widehat{f}_n(x)\right)+\left[\mathrm{bias}\big(\widehat{f}_n(x)\big)\right]^2.$$

Note that by Theorem 20

$$\mathsf{var}\left(\widehat{f}_n(x)\right)=\frac{f(x)R(K)}{n\,h_n}+o\big(\tfrac{1}{n\,h_n}\big), \tag{117}$$

where $R(K)=\int K^2(y)\,dy$.

To approximate the bias suppose that f is twice differentiable in x that is an interior point of the support of f. Further let the kernel K be a bounded symmetric function with

a bounded support such that $\int K(t)\,\mathrm{d}t = 1, \int t\,K(t)\,\mathrm{d}t = 0$ and $\int |t^2 K(t)|\,\mathrm{d}t < \infty$. Then for all sufficiently large n

$$\begin{aligned}\mathsf{E}\,\widehat{f}_n(x) &= \tfrac{1}{h_n}\,\mathsf{E}\,K\big(\tfrac{x-X_1}{h_n}\big) = \int \frac{1}{h_n} K\big(\tfrac{x-y}{h_n}\big) f(y)\,\mathrm{d}y \\ &= \int K(t) f(x - t h_n)\,\mathrm{d}t = \int K(t)\left[f(x) - t h_n f'(x) + \tfrac{1}{2}\,t^2\,h_n^2\,f''(x) + o(h_n^2)\right]\,\mathrm{d}t \\ &= f(x) + \tfrac{1}{2}\,h_n^2\,f''(x)\,\mu_{2K} + o(h_n^2),\end{aligned}$$

where $\mu_{2K} = \int y^2\,K(y)\,dy$. Thus one gets

$$\mathrm{bias}\big(\widehat{f}_n(x)\big) = \mathsf{E}\,\widehat{f}_n(x) - f(x) = \tfrac{1}{2}\,h_n^2\,f''(x)\,\mu_{2K} + o(h_n^2),$$

which together with (117) implies

$$\mathrm{MSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\,f(x)\,R(K) + \tfrac{1}{4}\,h_n^4\,[f''(x)]^2\mu_{2K}^2 + o\big(\tfrac{1}{n\,h_n}\big) + o\big(h_n^4\big). \tag{118}$$

Ignoring the remainder $o(\cdot)$ terms in (118), AMSE (asymptotic mean squared error) of $\widehat{f}_n(x)$ is given by

$$\mathrm{AMSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\,f(x)\,R(K) + \tfrac{1}{4}\,h_n^4\,[f''(x)]^2\mu_{2K}^2 \tag{119}$$

Minimising (119) one gets *asymptotically optimal local bandwidth* (i.e. bandwidth that minimises the AMSE)

$$h_n^{(opt)}(x) = n^{-1/5}\left[\frac{f(x)\,R(K)}{[f''(x)]^2\,\mu_{2K}^2}\right]^{1/5}. \tag{120}$$

To get a global bandwidth it is useful to define **(A)MISE - (asymptotic) mean integrated squared error**. Introduce

$$\mathrm{MISE}\big(\widehat{f}_n\big) = \int \mathrm{MSE}\big(\widehat{f}_n(x)\big)\,\mathrm{d}x = \int \mathsf{E}\left[\widehat{f}_n(x) - f(x)\right]^2\,\mathrm{d}x,$$

and its asymptotic approximation

$$\begin{aligned}\mathrm{AMISE}\big(\widehat{f}_n\big) &= \int \mathrm{AMSE}\big(\widehat{f}_n(x)\big)\,\mathrm{d}x = \int \tfrac{1}{n\,h_n} f(x) R(K) + \tfrac{[f''(x)]^2\mu_{2K}^2}{4}\,h_n^4\,\mathrm{d}x. \\ &= \frac{R(K)}{n\,h_n} + h_n^4\,\frac{R(f'')\,\mu_{2K}^2}{4},\end{aligned} \tag{121}$$

where $R(f'') = \int \left[f''(x)\right]^2\,\mathrm{d}x$.

Minimising (121) one gets *asymptotically optimal global bandwidth* (i.e. bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5}\left[\frac{R(K)}{R(f'')\,\mu_{2K}^2}\right]^{1/5}. \tag{122}$$

Remark 22. Note that after substitution of the optimal bandwidth (122) into (121) one gets that the optimal AMISE is given by

$$\frac{5\,\big[R(f'')\big]^{1/5}}{4\,n^{4/5}}\,\big[R(K)\big]^{4/5}\,\mu_{2K}^{2/5}.$$

It can be shown that if we consider kernels that are densities of probability distributions then $\big[R(K)\big]^{4/5}\,\mu_{2K}^{2/5}$ is minimised for K being an Epanechnikov kernel. Further note that for $\widetilde{K}(x) = \sqrt{\mu_{2K}}\,K\big(\sqrt{\mu_{2K}}\,x\big)$ one has

$$\mu_{2\widetilde{K}} = 1 \quad \text{and} \quad \big[R(\widetilde{K})\big]^{4/5} = \big[R(K)\big]^{4/5}\,\mu_{2K}^{2/5}$$

and the optimal AMISE is the same for $\widetilde{K}$ and K. That is why some authors prefer to use the kernels in a standardised form so that $\mu_{2K} = 1$. Some of the most common kernels having this property are summarised in Table 2.

Epanechnikov kernel:	$K(x) = \frac{3}{4\sqrt{5}}\big(1 - \frac{x^2}{5}\big)\,\mathbb{I}\big\{\lvert x\rvert \le \sqrt{5}\big\}$
Triangular kernel:	$K(x) = \frac{1}{\sqrt{6}}\big(1 - \lvert x\rvert\big)\,\mathbb{I}\big\{\lvert x\rvert \le \sqrt{6}\big\}$
Uniform kernel:	$K(x) = \frac{1}{2\sqrt{3}}\,\mathbb{I}\big\{\lvert x\rvert \le \sqrt{3}\big\}$
Biweight kernel:	$K(x) = \frac{15}{16\sqrt{7}}(1 - x^2)^2\,\mathbb{I}\big\{\lvert x\rvert \le \sqrt{7}\big\}$
Tricube kernel:	$K(x) = \frac{70\sqrt{243}}{81\sqrt{35}}(1 - \lvert x\rvert^3)^3\,\mathbb{I}\Big\{\lvert x\rvert \le \sqrt{\frac{35}{243}}\Big\}$
Gaussian kernel:	$K(x) = \frac{1}{\sqrt{2\pi}}\,\exp\{-x^2/2\}$

Table 2: Some kernel functions standardised so that $\mu_{2K} = 1$.

8.2.1 Normal reference rule

The problem of asymptotically optimal bandwidths given in (120) and (122) is that the quantities $f(x)$, $f''(x)$ and $R(f'')$ are unknown. Normal reference rule assumes that $f(x) = \frac{1}{\sigma}\,\varphi(\frac{x-\mu}{\sigma})$, where $\varphi(x)$ is density of a standard normal distribution.

Then

$$f'(x) = \tfrac{1}{\sigma^2}\varphi'\big(\tfrac{x-\mu}{\sigma}\big), \qquad f''(x) = \tfrac{1}{\sigma^3}\varphi''\big(\tfrac{x-\mu}{\sigma}\big),$$

where

$$\begin{aligned}
\varphi'(x) &= \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}(-x) = \tfrac{-x}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\\
\varphi''(x) &= \tfrac{-1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} + \tfrac{x^2}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} = (x^2 - 1)\,\varphi(x).
\end{aligned}$$

Thus with the help of (120) one gets

$$\widehat{h}_n(x) = n^{-\frac{1}{5}}\,\widehat{\sigma}\,\left[\frac{R(K)}{\mu_{2K}^2}\,\frac{1}{\big[\frac{(x-\widehat{\mu})^2}{\widehat{\sigma}^2} - 1\big]^2\,\varphi\big(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\big)}\right]^{\frac{1}{5}},$$

where $\widehat{\mu}$ a $\widehat{\sigma}^2$ are some estimates of the unknown parameters μ and σ^2, for instance $\widehat{\mu} = \overline{X}_n, \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

For the global bandwidth choice we need to calculate

$$\begin{aligned}
R(f'') &= \int \left[f''(x)\right]^2 \mathrm{d}x = \int \left\{ \frac{1}{\sigma^3}\left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]\varphi\left(\tfrac{x-\mu}{\sigma}\right)\right\}^2 \mathrm{d}x \\
&= \frac{1}{\sigma^6}\int \left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]^2 \varphi^2\left(\tfrac{x-\mu}{\sigma}\right)\mathrm{d}x \\
&= \left| \begin{array}{rcl} t &=& \frac{x-\mu}{\sigma} \\ \mathrm{d}t &=& \frac{\mathrm{d}x}{\sigma} \end{array} \right| = \frac{1}{\sigma^5}\int (t^2-1)^2\varphi^2(t)\,\mathrm{d}t \\
&= \frac{1}{\sigma^5}\int (t^4 - 2t^2 + 1)\frac{1}{2\pi}\mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{1}{\sigma^5 2\sqrt{\pi}}\int (t^4-2t^2+1)\underbrace{\frac{1}{\sqrt{\pi}}\mathrm{e}^{-t^2}}_{\sim \mathsf{N}\left(0,\frac{1}{2}\right)}\,\mathrm{d}t \\
&= \frac{1}{2\sigma^5\sqrt{\pi}}\,\mathsf{E}\left(Y^4 - 2Y^2 + 1\right) = \frac{1}{2\sigma^5\sqrt{\pi}}\left[3\cdot\left(\tfrac{1}{2}\right)^2 - 2\cdot\tfrac{1}{2} + 1\right] = \frac{3}{8\sigma^5\sqrt{\pi}},
\end{aligned}$$

where $Y \sim \mathsf{N}\left(0, \frac{1}{2}\right)$. Thus the asymptotically optimal global bandwidth would be

$$h_n^{(opt)} = \sigma\, n^{-1/5}\left[\frac{8\sqrt{\pi}\,R(K)}{3\,\mu_{2K}^2}\right]^{1/5}.$$

Further if one uses a Gaussian kernel $K(y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}$, one gets

$$\begin{aligned}
\mu_{2K} &= \int y^2 K(y)\,\mathrm{d}y = 1, \\
R(K) &= \int K^2(y)\,\mathrm{d}y = \tfrac{1}{2\sqrt{\pi}}\int \tfrac{1}{\sqrt{\pi}}\mathrm{e}^{-y^2}\,\mathrm{d}y = \tfrac{1}{2\sqrt{\pi}},
\end{aligned}$$

which results in

$$h_n^{(opt)} = \sigma\, n^{-1/5}\left[\frac{4}{3}\right]^{1/5} \doteq 1.06\,\sigma\, n^{-1/5}$$

The standard normal reference rule is now given by

$$h_n = 1.06\, n^{-1/5}\,\min\left\{S_n, \widetilde{IQR}_n\right\}, \tag{123}$$

where

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \qquad \text{and} \qquad \widetilde{IQR}_n = \frac{F_n^{-1}(0.75) - F_n^{-1}(0.25)}{1.34}.$$

It was found out that the bandwidth selector (123) works well if the true distribution is 'very close' to the normal distribution. But at the same time the bandwidth is usually too large for distributions 'moderately' deviating from normal distribution. That is why some authors prefer to use

$$h_n = 0.9\, n^{-1/5}\,\min\left\{S_n, \widetilde{IQR}_n\right\}. \tag{124}$$

For a more detailed argumentation see Silverman (1986), page 48.

8.2.2 Least-squares cross-validation

By this method we choose the bandwith as

$$h_n^{LSCV} = \arg\min_{h_n>0} \mathcal{L}(h_n),$$

where

$$\mathcal{L}(h_n) = \int \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i),$$

with $\widehat{f}_{-i}(x) = \frac{1}{(n-1)h_n}\sum_{j=1,j\neq i}^{n} K\big(\frac{x-X_j}{h_n}\big)$ being the kernel density estimator based on a sample that leaves out the i-th observation.

The rational behind the above method is as follows. Suppose we are interested in minimizing $\mathrm{MISE}\big(\widehat{f}_n\big)$. Note that $\mathrm{MISE}\big(\widehat{f}_n\big)$ can be rewritten as

$$\begin{aligned}
\mathrm{MISE}\big(\widehat{f}_n\big) &= \int \mathsf{E}\left(\widehat{f}_n(x) - f(x)\right)^2 \mathrm{d}x \overset{\text{Fub.}}{=} \mathsf{E}\int \widehat{f}_n^2(x) - 2\widehat{f}_n(x)f(x) - f^2(x)\,\mathrm{d}x \\
&= \mathsf{E}\int \widehat{f}_n^2(x)\,\mathrm{d}x - 2\,\mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x + \int f^2(x)\,\mathrm{d}x.
\end{aligned}$$

An unbiased estimator for $\mathsf{E}\int \widehat{f}_n^2(x)\,\mathrm{d}x$ is simply given by $\int \widehat{f}_n^2(x)\,\mathrm{d}x$. Further the term $\int f^2(x)\,\mathrm{d}x$ does not depen on h_n. Thus it remains to estimate $\int \widehat{f}_n(x)f(x)\,\mathrm{d}x$. Let us consider the following estimate

$$\widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i),$$

where

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)h_n}\sum_{j=1,j\neq i}^{n} K\big(\tfrac{x-X_j}{h_n}\big)$$

is the estimate of $f(x)$ that is based the sample without the i-th observation X_i. In what follows it is shown that $\widehat{A}_n$ is an unbiased estimator of $\int \widehat{f}_n(x)f(x)\,\mathrm{d}x$. Note that

$$\mathsf{E}\,\widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\,\widehat{f}_{-i}(X_i).$$

Now with the help of (112) and (114)

$$\begin{aligned}
\mathsf{E}\,\hat{f}_{-i}(X_i) &= \mathsf{E}\left[\frac{1}{(n-1)h_n}\sum_{j=1,j\neq i}^{n} K\big(\tfrac{X_i-X_j}{h_n}\big)\right] = \frac{1}{h_n}\,\mathsf{E}\,K\big(\tfrac{X_1-X_2}{h_n}\big) \\
&= \frac{1}{h_n}\int\int K\big(\tfrac{x-y}{h_n}\big)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int \underbrace{\left[\int \frac{1}{h_n}K\big(\tfrac{x-y}{h_n}\big)f(y)\,\mathrm{d}y\right]}_{=\mathsf{E}\,\widehat{f}_n(x)} f(x)\,\mathrm{d}x = \int \mathsf{E}\,\widehat{f}_n(x)f(x)\,\mathrm{d}x \\
&\overset{\text{Fub}}{=} \mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x.
\end{aligned}$$

Thus $\widehat{A}_n$ is an unbiased estimator of $\mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x$ and $\mathcal{L}(h_n)$ is an unbiased estimator of $\mathsf{E}\int \widehat{f}_n^2(x)\,\mathrm{d}x - 2\,\mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x$.

Remark 23. Stone (1984) has proved that

$$\frac{\mathrm{ISE}\left(h_n^{(LSCV)}\right)}{\min_{h_n}\mathrm{ISE}(h_n)} \xrightarrow[n\to\infty]{\text{alm. surely}} 1,$$

where $\mathrm{ISE}(h_n) = \int(\widehat{f}_n(x) - f(x))^2\,\mathrm{d}x$. But the simulations show that the variance of $h_n^{(LSCV)}$ (for not too big sample sizes) is rather large. Thus this method cannot be used blindly.

8.2.3 Biased-cross validation

This method aims at minimizing the AMISE given by (121). Note that to estimate AMISE it is sufficient to estimate $R(f'')$. It was found that the straightforward estimator $R\big(\widehat{f}_n''\big)$ is (positively) biased. To correct for the main term in the bias expansion it is recommended to use $R\big(\widehat{f}_n''\big) - \frac{R(K'')}{n\,h_n^5}$ instead. That is why in this method the bandwidth is chosen as

$$h_n^{(BCV)} = \arg\min_{h_n>0} \mathcal{B}(h_n),$$

where

$$\mathcal{B}(h_n) = \frac{R(K)}{n\,h_n} + \tfrac{1}{4}\,h_n^4\,\mu_{2K}^2\Big[R\big(\widehat{f}_n''\big) - \frac{R(K'')}{n\,h_n^5}\Big].$$

Remark 24. It can be proved that $\frac{\hat{h}_n^{(BCV)}}{h_n^{(opt)}} \xrightarrow[n\to\infty]{P} 1$, where $h_n^{(opt)}$ is given by (120).

8.3 Higher order kernels

By a formal calculation (for sufficiently large n, sufficiently smooth f and x an interior point of the support) one gets

$$\begin{aligned}
\mathsf{E}\,\widehat{f}_n(x) &= \int K(t)f(x-th_n)\,\mathrm{d}t \\
&= f(x)\int K(t)\,\mathrm{d}t - f'(x)h_n\int tK(t)\,\mathrm{d}t \\
&\quad + \frac{f''(x)}{2}h_n^2\int t^2K(t)\,\mathrm{d}t - \frac{f'''(x)}{3!}h_n^3\int t^3K(t)\,\mathrm{d}t + \ldots.
\end{aligned}$$

The kernel of order p is such that $\int K(t)\,dt = 1$ and

$$\int t^j\,K(t)\,dt = 0, \qquad j = 1,\ldots,p-1, \qquad \text{and} \qquad \int t^p\,K(t)\,dt \neq 0.$$

But note that if the above equations holds for $p > 2$, then (among others) $\int t^2K(t)\,\mathrm{d}t = 0$, which implies that K cannot be non-negative. As a consequence it might happen that $\widehat{f}_n(x) < 0$.

One of possible modifications of a Gaussian kernel to get a kernel of order 4 is given by

$$K(y) = \tfrac{1}{2}\,(3 - y^2)\,\tfrac{1}{\sqrt{2\,\pi}}\,e^{-y^2/2}.$$

8.4 Mirror-reflection

The standard kernel density estimator (102) is usually not consistent in the points, where the density f is not continuous. These might be the boundary points of the support. Even if the density is continuous at these points, the bias at these points is usually only of order $O(h_n)$ and not $O(h_n^2)$. There are several ways how to improve the performance of $\widehat{f}_n(x)$ close to the boundary points. The most straightforward is the *mirror-reflection method*.

To illustrate this method suppose we know that the support of the distribution with the density f is $[0, \infty)$. The modified kernel density estimator that uses mirror-reflection is given by

$$\widehat{f}_n^{(MR)}(x) = \begin{cases} \frac{1}{n\,h_n}\sum_{i=1}^n K\big(\frac{x-X_i}{h_n}\big) + \frac{1}{n\,h_n}\sum_{i=1}^n K\big(\frac{x+X_i}{h_n}\big), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{125}$$

Note that the first term on the right-hand side of (125) (for $x \geq 0$) is the standard kernel density estimator $\widehat{f}_n(x)$. The second term on the right-hand side of (125) is in fact also a standard kernel density estimator $\widehat{f}_n(x)$, but based on the 'mirror reflected' observations $-X_1, \ldots, -X_n$. This second term is introduced in order to compensate for the mass of the standard kernel density estimator $\widehat{f}_n(x)$ that falls outside the support $[0, \infty)$.

Literature: Wand and Jones (1995) Chapters 2.5, 3.2, 3.3

9 Kernel regression

Suppose that one observes independent and identically distributed bivariate random vectors $(X_1, Y_1)^\mathsf{T}, \ldots, (X_n, Y_n)^\mathsf{T}$. Our primary interest in this section is to estimate the conditional mean function of Y_1 given $X_1 = x$, i.e.

$$m(x) = \mathsf{E}\,[Y_1 \,|\, X_1 = x]$$

without assuming any parametric form of $m(x)$.

In what follows it will be also useful to denote the conditional variance function as

$$\sigma^2(x) = \mathsf{var}[Y_1 \,|\, X_1 = x].$$

9.1 Local polynomial regression

Suppose that the function m is a p-times differentiable function at the point x then for X_i 'close' to x one can approximate

$$m(X_i) \doteq m(x) + m'(x)\,(X_i - x) + \ldots + \tfrac{m^{(p)}(x)}{p!}\,(X_i - x)^p. \tag{126}$$

Thus 'locally' one can view and estimate the function $m(x)$ as a polynomial. This motivates defininition of the local polynomial estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(x) &= \big(\widehat{\beta}_0(x), \ldots, \widehat{\beta}_p(x)\big)^{\mathsf{T}} \\ &= \operatorname*{arg\,min}_{b_0,\ldots,b_p} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1(X_i - x) - \ldots - b_p(X_i - x)^p\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big),\end{aligned} \tag{127}$$

where K is a given kernel function and h_n is a smoothing parameter (bandwidth) going to zero as $n \to \infty$.

Comparing (126) and (127) one gets that $\widehat{\beta}_j(x)$ estimates $\frac{m^{(j)}(x)}{j!}$. Often we are interested only in $m(x)$ which is estimated by $\widehat{\beta}_0(x)$.

Put

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}, \qquad \mathbb{X}_p(x) = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^p \\ 1 & (X_2 - x) & \ldots & (X_2 - x)^p \\ \ldots & \ldots & \ldots & \ldots \\ 1 & (X_n - x) & \ldots & (X_n - x)^p \end{pmatrix}$$

and $\mathbb{W}(x)$ for the diagonal matrix with the i-th element of the diagonal given by $K\big(\frac{X_i - x}{h_n}\big)$.

Note that the optimisation problem in (127) can be written as the weighted least squares problem

$$\widehat{\boldsymbol{\beta}}(x) = \operatorname*{arg\,min}_{\mathbf{b} \in \mathbb{R}^{p+1}} \Big\{ \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big)^{\mathsf{T}} \mathbb{W}(x) \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big) \Big\}, \tag{128}$$

where $\mathbf{b} = (b_0, b_1, \ldots, b_p)^{\mathsf{T}}$. The solution of (128) can be explicitly written as

$$\widehat{\boldsymbol{\beta}}(x) = \Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{Y},$$

provided that the matrix $\Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p\Big)$ is regular.

The following technical lemma will be useful in deriving the properties of the local polynomial estimator.

Lemma 7. *Let the kernel K be bounded, symmetric around zero, positive, with a support $(-1, 1)$ and such that $\int K(x)\,\mathrm{d}x = 1$. For $l \in \mathbb{N} \cup \{0\}$ put*

$$S_{n,l}(x) = \frac{1}{n} \sum_{i=1}^{n} \tfrac{1}{h_n} K\big(\tfrac{X_i - x}{h_n}\big) \big(\tfrac{X_i - x}{h_n}\big)^l.$$

Suppose further that $h_n \to 0$ *and* $(n\,h_n) \to \infty$ *and that the density* f_X *of* X_1 *is positive and twice differentiable in* x*. Then*

$$S_{n,l}(x) = \begin{cases} f_X(x) \int K(t)\, t^l \, dt + \frac{h_n^2}{2} f_X''(x) \int K(t)\, t^{l+2}\, dt + o(h_n^2) + O_P\big(\frac{1}{\sqrt{n\,h_n}}\big), & l \text{ even}, \\ h_n\, f'(x) \int K(t)\, t^{l+1}\, dt + o(h_n^2) + O_P\big(\frac{1}{\sqrt{n\,h_n}}\big), & l \text{ odd}. \end{cases}$$

Proof. Analogously as in the proof of asymptotic normality of $\widehat{f}_n(x)$ (Theorem 21) one can show that

$$\sqrt{nh_n}\,\big(S_{n,l}(x) - \mathsf{E}\, S_{n,l}(x)\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \sigma^2(x)\big), \quad \text{where} \quad \sigma^2(x) = f_X(x) \int t^{2l}\, K^2(t)\, \mathrm{d}t.$$

Thus

$$S_{n,l}(x) = \mathsf{E}\, S_{n,l}(x) + O_P\big(\tfrac{1}{\sqrt{nh_n}}\big)$$

and it remains to calculate $\mathsf{E}\, S_{n,l}(x)$. Using the substitution $t = \frac{y-x}{h_n}$ and the Taylor expansion of the function $f_X(x + t\,h_n)$ around the point x one gets

$$\begin{aligned} \mathsf{E}\, S_{n,l}(x) &= \mathsf{E}\, \tfrac{1}{h_n} K\big(\tfrac{X_1-x}{h_n}\big)\big(\tfrac{X_1-x}{h_n}\big)^l = \int \tfrac{1}{h_n} K\big(\tfrac{y-x}{h_n}\big)\big(\tfrac{y-x}{h_n}\big)^l f_X(y)\, \mathrm{d}y \\ &= \int K(t)\, t^l\, f_X(x + t h_n)\, \mathrm{d}t \\ &= f_X(x) \int K(t) t^l\, \mathrm{d}t + h_n f_X'(x) \int K(t) t^{l+1}\, \mathrm{d}t + \tfrac{h_n^2}{2} f_X''(x) \int K(t) t^{l+2}\, \mathrm{d}t + o(h_n^2). \end{aligned}$$

As K is symmetric, then one gets that $\int K(t) t^{l+1}\, \mathrm{d}t = 0$ for l even and $\int K(t) t^{l+2}\, \mathrm{d}t = 0$ for l odd. □

Remark 25. Note that Lemma 7 implies that

$$S_{n,0}(x) = f_X(x) + \tfrac{h_n^2}{2} f_X''(x)\, \mu_{2K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = f_X(x) + o_P(1), \tag{129}$$

$$S_{n,1}(x) = h_n\, f'(x)\, \mu_{2K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = o_P(1), \tag{130}$$

$$S_{n,2}(x) = f(x)\, \mu_{2K} + o_P(1), \tag{131}$$

$$S_{n,3}(x) = h_n\, f'(x) \int t^4\, K(t)\, \mathrm{d}t + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = o_P(1). \tag{132}$$

9.2 Nadaraya-Watson estimator

For $p = 0$ the local polynomial estimator given by (127) simplifies to

$$\widehat{\beta}_0(x) = \operatorname*{arg\,min}_{b_0 \in \mathbb{R}} \sum_{i=1}^{n} \Big[Y_i - b_0\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big),$$

and solving this optimisation task one gets

$$\widehat{\beta}_0(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i,$$

where

$$w_{ni}(x) = \frac{K\left(\frac{X_i - x}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n}\right)} = \frac{\frac{1}{n\,h_n} K\left(\frac{X_i - x}{h_n}\right)}{S_{n,0}(x)}.$$

This estimator is in the context of the local polynomial regression also called a locally constant estimator.

Put $\mathbb{X} = (X_1, \ldots, X_n)$ and let $\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right)$ and $\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right)$ stand for the conditional bias and variance of the estimator $\widehat{m}_{NW}(x)$ given $\mathbb{X}$.

Theorem 22. *Suppose that the assumptions of Lemma 7 are satisfied and further that* $(n\,h_n^3) \xrightarrow[n\to\infty]{} \infty$*, the density is* $f_X(x)$ *is continuously differentiable and positive at* x*, the function* $m(\cdot)$ *is twice differentiable at the point* x *and the function* $\sigma^2(\cdot)$ *is continuous at the point* x*. Then*

$$\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2K}\left(\frac{m'(x)\,f_X'(x)}{f_X(x)} + \frac{m''(x)}{2}\right) + o_P(h_n^2), \tag{133}$$

$$\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\left(\frac{1}{n\,h_n}\right), \tag{134}$$

where

$$R(K) = \int K^2(x)\,\mathrm{d}x \qquad \textit{and} \qquad \mu_{2K} = \int x^2\,K(x)\,\mathrm{d}x. \tag{135}$$

Proof. Showing (133). Let us calculate

$$\begin{aligned}
\mathsf{E}\left[\widehat{m}_{NW}(x)|\mathbb{X}\right] &= \sum_{i=1}^{n} w_{ni}(x)\,\mathsf{E}\left[Y_i|\mathbb{X}\right] = \sum_{i=1}^{n} w_{ni}(x)\,\mathsf{E}\left[Y_i|X_i\right] = \sum_{i=1}^{n} w_{ni}(x) m(X_i) \\
&= \sum_{i=1}^{n} w_{ni}(x)\left[m(x) + (X_i - x)\,m'(x) + \tfrac{(X_i-x)^2}{2}\,m''(x) + (X_i - x)^2\,\tilde{R}(X_i)\right] \\
&= m(x)\sum_{i=1}^{n} w_{ni}(x) + m'(x)\sum_{i=1}^{n} w_{ni}(x)(X_i - x) + \frac{m''(x)}{2}\sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \\
&\quad + \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2\tilde{R}(X_i), \\
&= m(x) + m'(x)\,A_n + \frac{m''(x)}{2}\,B_n + C_n,
\end{aligned} \tag{136}$$

where $\tilde{R}(z) \to 0$ as $z \to x$ and

$$A_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x),\; B_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2,\; C_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2\tilde{R}(X_i). \tag{137}$$

Now with the help of (129) and (130)

$$A_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x) = \frac{h_n \sum_{i=1}^{n} K\big(\frac{X_i - x}{h_n}\big)(X_i - x)\frac{1}{h_n^2}}{\sum_{j=1}^{n} K\big(\frac{X_j - x}{h_n}\big)\frac{1}{h_n}} = \frac{h_n S_{n,1}(x)}{S_{n,0}(x)}$$
$$= \frac{h_n\left[h_n f'_X(x)\,\mu_{2K} + o(h_n^2) + O_P\left(\frac{1}{\sqrt{nh_n}}\right)\right]}{f_X(x) + o_P(1)} = \frac{h_n^2 f'_X(x)\mu_{2K} + o(h_n^3) + O_P\left(\frac{h_n}{\sqrt{nh_n}}\right)}{f_X(x) + o_P(1)}$$
$$= \frac{h_n^2 f'_X(x)\mu_{2K}}{f_X(x)} + o_P(h_n^2) + O_P\Big(\frac{h_n^2}{\sqrt{nh_n^3}}\Big) = \frac{h_n^2 f'_X(x)\mu_{2K}}{f_X(x)} + o_P(h_n^2), \tag{138}$$

as $(nh_n^3) \to \infty$. Further with the help of (129) and (131)

$$B_n = \sum_{i=1}^{n} w_{ni}(X_i - x)^2 = \ldots = \frac{h_n^2 S_{n,2}(x)}{S_{n,0}(x)}$$
$$= \frac{h_n^2\,[f_X(x)\,\mu_{2K} + o_P(1)]}{f_X(x) + o_P(1)} = h_n^2\mu_{2K} + o_P(h_n^2). \tag{139}$$

Concerning C_n thanks to (139) and the fact that the support of K is $(-1, 1)$ one can bound

$$|C_n| \le \left|\sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2\tilde{R}(X_i)\right| \le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)| \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2$$
$$= o(1)\,O_P(h_n^2) = o_P(h_n^2). \tag{140}$$

Now combining (138), (139) and (140) one gets

$$\mathsf{E}\,[\widehat{m}_{NW}(x)|\mathbb{X}] = m(x) + m'(x)h_n^2\frac{f'_X(x)}{f_X(x)}\mu_{2K} + \frac{m''(x)}{2}h_n^2\mu_{2K} + o_P(h_n^2),$$

which implies (133).

Showing (134). Let us calculate

$$\mathsf{var}[\widehat{m}_{NW}(x)|\mathbb{X}] = \sum_{i=1}^{n} w_{ni}^2(x)\,\mathsf{var}[Y_i|X_i] = \sum_{i=1}^{n} w_{ni}^2(x)\sigma^2(X_i)$$
$$= \frac{\sum_{i=1}^{n} K^2\big(\frac{X_i - x}{h_n}\big)\sigma^2(X_i)}{\Big[\sum_{j=1}^{n} K\big(\frac{X_j - x}{h_n}\big)\Big]^2} = \frac{1}{nh_n}\frac{V_n}{[S_{n,0}(x)]^2},$$

where $V_n = \frac{1}{nh_n}\sum_{i=1}^{n} K^2\big(\frac{X_i - x}{h_n}\big)\sigma^2(X_i)$.

Now completely analogously as in Theorem 20 it is proved that $\widehat{f}_n(x) \xrightarrow[n\to\infty]{P} f(x)$ we will show that

$$V_n \xrightarrow[n\to\infty]{P} f_X(x)\,\sigma^2(x)\,R(K), \tag{141}$$

which combined with (129) implies (134).

Showing (141). First with the help of Bochner's theorem (Theorem 19)

$$\mathsf{E}\, V_n = \frac{1}{h_n}\mathsf{E}\left[K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\right] \tag{142}$$

$$= \int \frac{1}{h_n} K^2\big(\tfrac{z-x}{h_n}\big)\,\sigma^2(z) f_X(z)\,\mathrm{d}z \xrightarrow[n\to\infty]{} \sigma^2(x) f_X(x) \int K^2(t)\,\mathrm{d}t. \tag{143}$$

Now it remains to show that $\mathsf{var}(V_n) \xrightarrow[n\to\infty]{} 0$. Using again Bochner's theorem (Theorem 19)

$$\begin{aligned}
\mathsf{var}(V_n) &= \frac{1}{nh_n^2}\left[\mathsf{E}\, K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1) - \Big(\mathsf{E}\, K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\Big)^2\right] \\
&= \frac{1}{nh_n}\left[\frac{1}{h_n}\mathsf{E}\, K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1)\right] - \frac{1}{n}\left[\frac{1}{h_n}\mathsf{E}\, K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\right]^2 \\
&= \frac{1}{nh_n}\left[\sigma^4(x) f_X(x)\int K^4(t)\,\mathrm{d}t + o(1)\right] - \frac{1}{n}\left[\sigma^2(x) f_X(x)\int K^2(t)\,\mathrm{d}t + o(1)\right]^2 \\
&\xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

□

9.3 Local linear estimator

For $p = 1$ the local polynomial estimator given by (127) simplifies to

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \underset{b_0,b_1}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i-x}{h_n}\big).$$

By solving the above optimisation task one gets

$$\widehat{m}_{LL}(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i, \tag{144}$$

where the weights can be written in the form

$$w_{ni}(x, h_n) = \frac{\frac{1}{n\,h_n} K\big(\frac{X_i-x}{h_n}\big)\big(S_{n,2}(x) - \frac{X_i-x}{h_n} S_{n,1}(x)\big)}{S_{n,0}(x)\, S_{n,2}(x) - S_{n,1}^2(x)}, \qquad i = 1, \ldots, n. \tag{145}$$

Theorem 23. *Suppose that the assumptions of Theorem 22 hold. Then*

$$\mathsf{bias}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2K}\,\frac{m''(x)}{2} + o_P(h_n^2), \tag{146}$$

$$\mathsf{var}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big), \tag{147}$$

where $R(K)$ and μ_{2K} are given in (135).

Proof. Showing (146) Completely analogously as in the proof of Theorem 22 one can arrive at (136) wit the only difference that now the weights $w_{ni}(x)$ are given by (145). Now calculate

$$A_n = \frac{\sum_{i=1}^n \frac{1}{nh_n} K\big(\frac{X_i-x}{h_n}\big)(X_i-x)S_{n,2}(x) - \frac{1}{nh_n^2}\sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)(X_i-x)^2}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}$$
$$= \frac{S_{n,1}(x)\, S_{n,2}(x) - S_{n,2}(x)\, S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} = 0. \tag{148}$$

Further using (129), (130), (131) and (132)

$$B_n = \sum_{i=1}^n w_{ni}(x)\tfrac{(X_i-x)^2}{h_n^2}h_n^2 = h_n^2 \frac{S_{n,2}^2(x) - S_{n,3}(x)S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}$$
$$= h_n^2 \frac{\big[f_X(x)\int t^2K(t)\,\mathrm{d}t + o_P(1)\big]^2 - o_P(1)o_P(1)}{(f_X(x)+o_P(1))\big[f_X(x)\int t^2K(t)\,\mathrm{d}t + o_P(1)\big] - (o_P(1))^2}$$
$$= h_n^2\,\mu_{2K} + o_P(h_n^2). \tag{149}$$

Thus it remains to show that $C_n = o_P(h_n^2)$. Put $D_n = S_{n,0}(x)\, S_{n,2}(x) - S_{n,1}^2(x)$ and note that with the help of (129)–(131) one gets

$$D_n = f_X^2(x)\,\mu_{2K}^2 + o_P(1). \tag{150}$$

Now with the help (150) and Lemma 7 one can bound

$$|C_n| \le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)|\, h_n^2 \sum_{i=1}^n |w_{ni}(x)|\, \tfrac{(X_i-x)^2}{h_n^2}$$
$$\le h_n^2\, o(1)\, \frac{S_{n,2}^2(x) + |S_{n,1}(x)|\,\sum_{i=1}^n \frac{1}{h_n} K\big(\frac{X_i-x}{h_n}\big)\big|\frac{X_i-x}{nh_n}\big|^3}{|D_n(x)|}$$
$$= o(h_n^2)\, \frac{f_X^2(x)\mu_{2K}^2 + o_P(1) + o_P(1)\big[\int K(t)\,|t|^3\,\mathrm{d}t + o_P(1)\big]}{f_X^2(x)\,\mu_{2K} + o_P(1)} = o_P(h_n^2),$$

which together with (137), (148) and (149) yields (146).

Showing (147). With the help of (130), (131), (141) and (150) one can calculate

$$\mathsf{var}[\widehat{m}_{LL}(x)|\mathbb{X}] = \sum_{i=1}^n w_{ni}^2(x)\sigma^2(X_i)$$
$$= \frac{1}{D_n^2(x)}\left[\frac{1}{n^2h_n^2}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(S_{n,2}(x) - \tfrac{X_i-x}{h_n}S_{n,1}(x)\Big)^2\sigma^2(X_i)\right]$$
$$= \frac{1}{nh_n}\frac{1}{D_n^2(x)}\big[S_{n,2}^2(x) + o_P(1)\big]\,\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\sigma^2(X_i)$$
$$= \frac{1}{nh_n}\frac{1}{f_X^4(x)\,\mu_{2K}^2 + o_P(1)}\big[f_X^2(x)\,\mu_{2K}^2 + o_P(1)\big]\,\big[f_X(x)\,\sigma^2(x)\,R(K) + o_P(1)\big],$$

which implies (147). □

9.4 Locally polynomial regression ($p > 1$)

Analogously as for $p \in \{0, 1\}$ one gets the estimator of $m(x)$ in the form

$$\widehat{m}_p(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i,$$

where the weights $w_{ni}(x)$ are given by the first row of the matrix

$$\Big(\mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)\, \mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)$$

and satisfy that $\sum_{i=1}^{n} w_{ni}(x) = 1$ and

$$\sum_{i=1}^{n} w_{ni}(x)(X_i - x)^{\ell} = 0, \quad \ell = 1, \ldots, p-1.$$

Thus analogously as in the proofs of Theorems 22 and 23 one can show that if p is **even** then the (conditional) variances of $\widehat{m}_p(x)$ and $\widehat{m}_{p+1}(x)$ are asymptotically of the order $O_P\big(\frac{1}{n\,h_n}\big)$ and it even holds that

$$\mathsf{var}\left(\widehat{m}_p(x)|\mathbb{X}\right) = \mathsf{var}\left(\widehat{m}_{p+1}(x)|\mathbb{X}\right) + o_P\big(\tfrac{1}{n\,h_n}\big).$$

Further at the same time the biases of $\widehat{m}_p(x)$ and $\widehat{m}_{p+1}(x)$ are of the same order ($O_P(h_n^{p+2})$), but the bias of $\widehat{m}_{p+1}(x)$ has a simpler structure than the bias of $\widehat{m}_p(x)$. That is why in practice usually odd choices of p are preferred.

Literature: Fan and Gijbels (1996) Chapters 3.1 and 3.2.1

9.5 Bandwidth selection

In what follows we will consider $p = 1$.

9.5.1 Asymptotically optimal bandwidths

With the help of Theorem 23 one can approximate the conditional MSE (mean squared error) of $\widehat{m}_{LL}(x)$ as

$$\mathrm{MSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\,h_n}\,\tfrac{\sigma^2(x)\,R(K)}{f_X(x)} + \tfrac{1}{4}\,h_n^4\,[m''(x)]^2\mu_{2K}^2 + o_P\big(\tfrac{1}{n\,h_n}\big) + o_P\big(h_n^4\big), \tag{151}$$

Ignoring the remainder $o_P(.)$ terms in (151), we get that AMSE (asymptotic mean squared error) of $\widehat{m}_{LL}(x)$ is given by

$$\mathrm{AMSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\,h_n}\,\tfrac{\sigma^2(x)\,R(K)}{f_X(x)} + \tfrac{1}{4}\,h_n^4\,[m''(x)]^2\mu_{2K}^2 \tag{152}$$

Minimising (152) one gets asymptotically optimal local bandwidth (i.e. bandwidth that minimises the AMSE).

$$h_n^{(opt)}(x) = n^{-1/5} \left[\frac{\sigma^2(x)\, R(K)}{f_X(x)\, [m''(x)]^2\, \mu_{2K}^2} \right]^{1/5}. \tag{153}$$

The integrated mean squared error (MISE) is usually defined as

$$\mathrm{MISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) = \int \mathrm{MSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \tag{154}$$

where $w_0(x)$ is a given weight function which is introduced in order to guarantee that the integral is finite.

Now with the help of (152) and (154) the asymptotic integrated mean squared error (AMISE) is defined as

$$\begin{aligned} \mathrm{AMISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) &= \int \mathrm{AMSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \\ &= \frac{R(K)}{n\, h_n} \int \sigma^2(x)\, w_0(x)\, \mathrm{d}x + \tfrac{1}{4}\, h_n^4\, \mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x \end{aligned} \tag{155}$$

Minimising (155) one gets asymptotically optimal global bandwidth (i.e. the bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K) \int \sigma^2(x)\, w_0(x)\, \mathrm{d}x}{\mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}. \tag{156}$$

9.5.2 Rule of thumb for bandwidth selection

Suppose that $\sigma(x)$ is constant. Then the asymptotically optimal global bandwidth (156) is given by

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K)\, \sigma^2 \int w_0(x)\, \mathrm{d}x}{\mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}. \tag{157}$$

Now let $\widetilde{m}(x)$ be an estimated mean function fitted by the (global) polynomial regression of order 4 through the standard least squares method.

Now in (156) one replaces the unknown quantity σ^2 by $\widetilde{\sigma}^2 = \frac{1}{n-5} \sum_{i=1}^n \big[Y_i - \widetilde{m}(X_i)\big]^2$ and $m''(x)$ by $\widetilde{m}''(x)$. Finally the integral $\int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x$ is estimated by

$$\frac{1}{n} \sum_{i=1}^n [\widetilde{m}''(X_i)]^2\, w_0(X_i),$$

which results in the bandwidth selector

$$h_n^{(ROT)} = n^{-1/5} \left[\frac{R(K)\, \widetilde{\sigma}^2 \int w_0(x)\, \mathrm{d}x}{\mu_{2K}^2\, \frac{1}{n} \sum_{i=1}^n [\widetilde{m}''(X_i)]^2\, w_0(X_i)} \right]^{1/5}. \tag{158}$$

9.5.3 Cross-validation

$$h_n^{(CV)} = \arg\min_{h_n>0} \mathcal{CV}(h_n),$$

where

$$\mathcal{CV}(h_n) = \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - \widehat{m}_{-i}(X_i)\right]^2 w_0(X_i)$$

with $\widehat{m}_{-i}(x)$ being the estimator based on a sample that leaves out the i-th observation.

The rational of the above procedure is that one aims at minimising the estimated integrated squared error, i.e.

$$\mathrm{ISE}\big(\widehat{m}_{LL}(x)\big) = \int \big(\widehat{m}_{LL}(x) - m(x)\big)^2 f_X(x)\, w_0(x)\, \mathrm{d}x. \tag{159}$$

Now put $\varepsilon_i = Y_i - m(X_i)$ and calculate

$$\begin{aligned} CV(h_n) &= \frac{1}{n}\sum_{i=1}^{n} \left[\varepsilon_i + m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\right]^2 w_0(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, w_0(X_i) + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\right] w_0(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \left[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\right]^2 w_0(X_i). \end{aligned}$$

Now $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, w_0(X_i)$ does not depend on h_n and thus it is not interesting.

Further $\frac{1}{n}\sum_{i=1}^{n} \left[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\right]^2 w_0(X_i)$ can be considered as a reasonable estimate of (159).

Finally $\frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \big[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\big] w_0(X_i)$ does not 'bias' the estimate of (159), as

$$\begin{aligned} \mathsf{E}\left[\varepsilon_i \big[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\big] w_0(X_i)\right] &= \mathsf{E}\left\{\mathsf{E}\left[\varepsilon_i \big[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\big] w_0(X_i) \,|\, \mathbb{X}\right]\right\} \\ &= \mathsf{E}\left\{\mathsf{E}\left[\varepsilon_i | X_i\right] \mathsf{E}\left[\big[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\big] w_0(X_i) \,|\, \mathbb{X}\right]\right\} = 0, \end{aligned}$$

where we have used that $\mathsf{E}\left[\varepsilon_i | X_i\right] = 0$ and that ε_i and $\big[m(X_i) - \widehat{m}_{LL}^{(-i)}(X_i)\big] w_0(X_i)$ are independent conditionally on $\mathbb{X}$.

9.5.4 Nearest-neighbour bandwidth choice

Suppose that the support of the kernel function K is the interval $(-1, 1)$. Note that then $w_{ni}(x) = 0$ if $|X_i - x| > h_n$. The aim of the nearest-neighbour bandwidth choice is to choose such h_n so that for at least k observations $|X_i - x| \le h_n$. This can be technically achieved as follows.

Put

$$d_1(x) = \big|X_1 - x\big|, \ldots, d_n(x) = \big|X_n - x\big|$$

for the distances of the observations $X_1, \ldots, X_n$ from the point of interest x. Let $d_{(1)}(x) \leq \ldots \leq d_{(n)}(x)$ be the ordered sample of $d_1(x), \ldots, d_n(x)$. Then choose h_n as

$$h_n^{(NN)}(x) = d_{(k)}(x). \tag{160}$$

Note that (160) presents a local bandwidth choice.

To get an insight into the bandwidth choice (160) let us approximate

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{|X_i - x| \leq h\} \doteq F_n(x+h) - F_n(x-h) \doteq F_X(x+h) - F_X(x-h) \doteq f_X(x)2h. \tag{161}$$

By plugging $h = d_{(k)}(x) = h_n(x)$ into (161) one gets $\frac{k}{n} \doteq f_X(x)2h_n(x)$ which further implies that

$$h_n^{(NN)}(x) \doteq \frac{k}{2nf_X(x)}.$$

Remark 26. To derive asymptotic properties of $\widehat{m}_{LL}$ when bandwidth h_n is chosen as (160) one needs to consider $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$.

In some textbooks one can also find the following rule for the bandwidth choice

$$h_n(x) = \frac{X_{(l+k)} - X_{(l-k)}}{2},$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ and $X_{(l)}$ is the closest observation to x.

9.6 Robust locally weighted regression (LOWESS)

LOWESS is an algorithm for 'LOcally WEighted Scatterplot Smoothing'. It is used among others in regression diagnostics. It runs as follows.

In the first step the local linear fit $\widehat{m}_{LL}(x)$ with the tricube kernel function, $K(t) = \frac{70}{81}(1 - |t|^3)^3 \mathbb{I}\{|t| \leq 1\}$, is calculated. The bandwidth is chosen by the nearest-neighbour method with $k = \lfloor n\, f \rfloor$, where the default choice of f is $\frac{2}{3}$. Then for a given number of iterations the fit is recalculated as follows.

Let

$$r_i = Y_i - \widehat{m}(X_i), \qquad i = 1, \ldots, n$$

be the residuals of the current fit. Calculate the 'measures of outlyingness'

$$\delta_i = B\Big(\frac{r_i}{6\,\mathsf{med}\,(|r_1|,\ldots,|r_n|)}\Big), \qquad i = 1, \ldots, n,$$

where $B(t) = (1 - t^2)^2\, \mathbb{I}\{|t| \leq 1\}$. With the help of δ_i the outlying observations are down-weighted and the local linear fit is recalculated as $\widehat{m}(x) = \widehat{\beta}_0(x)$, where

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big)\, \delta_i.$$

By default there are 3 iterations.

9.7 Conditional variance estimation

Note that $\sigma^2(x) = \mathsf{E}\left[Y_1^2 | X_1 = x\right] - m^2(x)$, thus the most straightforward estimate is given by

$$\widehat{\sigma}_n^2(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i^2 - \widehat{m}_n^2(x), \tag{162}$$

where $\widehat{m}_n(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i$. This estimator is usually preferred in theoretical papers as its properties can be derived completely analogously as for $\widehat{m}_n(x)$. But in practice it is usually recommended to use the following estimator

$$\widetilde{\sigma}_n^2(x) = \sum_{i=1}^{n} w_{ni}(x) \left(Y_i - \widehat{m}_n(X_i)\right)^2. \tag{163}$$

Note that if the weights $w_{ni}(x)$ are not non-negative, then there is generally no guaranty that either of the estimators (162) or (163) is positive.

Literature: Fan and Gijbels (1996) Chapters 2.4.1, 3.2.3, 4.2, 4.10.1, 4.10.2

Appendix

The following theorem can be found for instance in Section 2.1.4 of Serfling (1980) as Theorem A.

Theorem A1. *(**Glivenko-Cantelli theorem**) Suppose we observe independent and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ (in $\mathbb{R}^k$) from a distribution with the cumulative distribution function F. Let*

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \leq \mathbf{x}\}.$$

be the cumulative empirical distribution function. Then

$$\sup_{\mathbf{x} \in \mathbb{R}^k} \left|F_n(\mathbf{x}) - F(\mathbf{x})\right| \xrightarrow[n\to\infty]{\textit{alm. surely}} 0.$$

References

Anděl, J. (2007). *Základy matematické statistiky.* Matfyzpress, Praha.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, New York.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, London.

Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. *arXiv preprint, arXiv:1107.3806*.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Annals of Mathematical Statistics*, 42:1977–1991.

Jiang, J. (2010). *Large sample techniques for statistics*. Springer Texts in Statistics. Springer, New York.

Jurečková, J., Sen, P. K., and Picek, J. (2012). *Methodology in Robust and Nonparametric Statistics*. Chapman & Hall/CRC.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Kulich, M. (2014). Maximum likelihood estimation theory. Course notes for NMST432. Available at http://www.karlin.mff.cuni.cz/˜kulich/vyuka/pokreg/.

Lachout, P. (2004). *Teorie pravděpodobnosti*. Karolinum. Skripta.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer, New York.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. Wiley, New York.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics*. Wiley, Chichester.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley, New York. Second Edition.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Prášková, Z. (2004). Metoda bootstrap. *Robust 2004*, pages 299–314.

Sen, P. K., Singer, J. M., and de Lima, A. C. P. (2010). *From finite sample to asymptotic methods in statistics*. Cambridge University Press.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shao, J. and Tu, D. (1996). *The jackknife and bootstrap*. Springer, New York.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* CHAPMAN/CRC.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, 12:1285–1297.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, New York.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48:817–838.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103.

Zvára, K. (2008). *Regrese.* MATFYZPRESS.