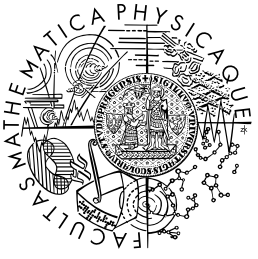

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MODERN STATISTICAL METHODS

NMST 434

Course notes

Last updated: February 20, 2021

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

Contents

1 Clippings from the asymptotic theory **3**

1.1 The convergence of random vectors . . . 3
1.2 Δ-theorem . . . 7
1.3 Moment estimators . . . 10
1.4 Confidence intervals and asymptotic variance-stabilising transformation . . . 13

2 Maximum likelihood methods **16**

2.1 Asymptotic normality of maximum likelihood estimator . . . 16
2.2 Asymptotic efficiency of maximum likelihood estimators . . . 20
2.3 Estimation of the asymptotic variance matrix . . . 21
2.4 Asymptotic tests (without nuisance parameters) . . . 23
2.5 Asymptotic confidence sets . . . 25
2.6 Asymptotic tests with nuisance parameters . . . 26
2.7 Profile likelihood . . . 32
2.8 Some notes on maximum likelihood in case of not i.i.d. random vectors . . . 37
2.9 Conditional and marginal likelihood . . . 40

3 M- and Z-estimators **44**

3.1 Identifiability of parameters via M- and/or Z-estimators . . . 45
3.2 Asymptotic distribution of Z-estimators . . . 46
3.3 Likelihood under model misspecification . . . 51
3.4 Asymptotic normality of M-estimators defined by convex minimization . . . 53
3.4.1 Sample median . . . 54

4 M-estimators and Z-estimators in robust statistics **55**

4.1 Robust estimation of location . . . 56
4.2 Studentized M/Z-estimators . . . 59
4.3 Robust estimation in linear models . . . 60
4.3.1 The least squares method . . . 60
4.3.2 Method of the least absolute deviation . . . 61
4.3.3 Huber estimator of regression . . . 62
4.3.4 Studentized Huber estimator of regression . . . 64

5 Quantile regression **64**

5.1 Introduction . . . 65
5.2 Regression quantiles . . . 66

5.3 Interpretation of the regression quantiles . . . 70
5.4 Inference for regression quantiles . . . 71
5.5 Asymptotic normality of sample quantiles . . . 72

6 EM-algorithm 73
6.1 General description of the EM-algorithm . . . 75
6.2 Convergence of the EM-algorithm . . . 77
6.3 Rate of convergence of EM-algorithm . . . 77
6.4 The EM algorithm in exponential families . . . 78
6.5 Some further examples of the usage of the EM algorithm . . . 80

7 Missing data 80
7.1 Basic concepts for the mechanism of missing . . . 81
7.2 Methods for dealing with missing data . . . 83

8 Bootstrap and other resampling methods 87
8.1 Monte Carlo principle . . . 87
8.2 Standard nonparametric bootstrap . . . 90
8.2.1 Comparison of nonparametric bootstrap and normal approximation . . . 96
8.2.2 Smooth transformations of sample means . . . 97
8.2.3 Limits of the standard nonparametric bootstrap . . . 98
8.3 Confidence intervals . . . 99
8.3.1 Basic bootstrap confidence interval . . . 99
8.3.2 Studentized bootstrap confidence interval . . . 101
8.4 Parametric bootstrap . . . 102
8.5 Testing hypotheses and bootstrap . . . 104
8.6 Permutation tests . . . 106
8.7 Bootstrap in linear models . . . 108
8.8 Variance estimation and bootstrap . . . 109
8.9 Bias reduction and bootstrap . . . 109
8.10 Jackknife . . . 110

9 Kernel density estimation 112
9.1 Consistency and asymptotic normality . . . 113
9.2 Bandwidth choice . . . 119
9.2.1 Normal reference rule . . . 121
9.2.2 Least-squares cross-validation . . . 122

9.2.3 Biased cross-validation 124
9.3 Higher order kernels 124
9.4 Mirror-reflection 125

10 Kernel regression **125**
10.1 Local polynomial regression 125
10.2 Nadaraya-Watson estimator 127
10.3 Local linear estimator 130
10.4 Locally polynomial regression (general p) 133
10.5 Bandwidth selection 133
10.5.1 Asymptotically optimal bandwidths 133
10.5.2 Rule of thumb for bandwidth selection 134
10.5.3 Cross-validation 135
10.5.4 Nearest-neighbour bandwidth choice 136
10.6 Robust locally weighted regression (LOWESS) 137
10.7 Conditional variance estimation 137

Appendix **139**

Last update: February 20, 2021

1 Clippings from the asymptotic theory

1.1 The convergence of random vectors

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

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

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

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

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

Definition. We say that

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

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

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

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

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

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

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

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

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

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

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

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

Proof. (i) *Almost sure convergence.*

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

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

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

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

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

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

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

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

□

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

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

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

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

Proof. Note that it is sufficient to prove

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

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

To prove (1) note that

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

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

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

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

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

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

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

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

Remark 2. Note that

(i) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ implies $\boldsymbol{X}_n = O_P(1)$ (Prohorov's theorem);

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

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

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

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

Note that $\boldsymbol{X}_n = O_P\big(\frac{1}{r_n}\big)$ implies there exists a finite constant K such that

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

* *omezená v pravděpodobnosti*

The statement now follows from the fact that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Δ-theorem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which implies the first statement of the theorem.

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

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

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

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

$$\sqrt{n}\left(\mathbf{T}_n - \boldsymbol{\mu}\right) = \sqrt{n}\left(\overline{\boldsymbol{X}}_n - \mathsf{E}\ \boldsymbol{X}_1\right).$$

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 $\boldsymbol{\mu}$. But the proof would be more complicated.

The end of lecture 1 (17. 2. 2020)

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

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

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

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

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

Example 2. 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 3. 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 4. Suppose you have two independent random samples from Bernoulli distribution. Derive the asymptotic distribution of the logarithm of odds-ratio.

Example 5. Suppose we observe independent identically distributed random vectors

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

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

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

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

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

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

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

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

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

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

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

Example 6. 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 7. Suppose that we observe $X_1, \ldots, X_n$ of a moving average sequence of order 1 given by

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

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

Using the fact that the autocorrelation function at lag 1 satisfies

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

derive the estimator of θ and find its asymptotic distribution.

Hint. Note that by Bartlett's formula

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

where

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

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

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

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 8. Moment estimation in Beta distribution.

Put

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

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

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

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

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

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

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

* *skutečná hodnota* † *Momentový odhad*

Proof. By the assumptions of the theorem and the inverse function theorem (Theorem A1) there exists an open neighbourhood U containing $\boldsymbol{\theta}_X$ and an open neighbourhood V containing $\boldsymbol{\tau}(\boldsymbol{\theta}_X)$ such that $\boldsymbol{\tau} : U \to V$ is a differentiable bijection with a differentiable inverse $\boldsymbol{\tau}^{-1} : V \to U$. Further note that $\mathbf{T}_n$ defined in (5) satisfies $\mathsf{P}\left(\mathbf{T}_n \in V\right) \xrightarrow[n\to\infty]{} 1$. Thus one can use (6) and apply the Δ-theorem (Theorem 3) with $\mathbf{g} = \boldsymbol{\tau}^{-1}$, $\boldsymbol{\mu} = \boldsymbol{\tau}(\boldsymbol{\theta}_X)$ and $A = V$ to get

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

The statement of the theorem now follows from the identity

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

□

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

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

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

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

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

Confidence intervals for θ_{Xj}

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

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

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

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

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

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

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

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

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

Applications of moment estimators

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

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

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

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

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

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

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

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

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

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

Example 9. Let $X_1, \dots, X_n$ be independent identically distributed random variables from the discrete distribution given as

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

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

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

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

has two roots

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

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

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

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

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

1.4 Confidence intervals and asymptotic variance-stabilising transformation

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

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

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

Standard asymptotic confidence interval of 'Wald' type

This interval is based on the fact that

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

and thus

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

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

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

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

Implicit (asymptotic) confidence interval of 'Wilson' type

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

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

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

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

with the critical region

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

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

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

Confidence interval based on the transformation stabilizing the asymptotic variance

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

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

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

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

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

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

Example 10. 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 11. Fisher's Z-transformation and various confidence intervals for the correlation coefficient.

Example 12. 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 an unknown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \Theta$. Let $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$ be the true value of the parameter.

Define the *likelihood function* as

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

and the *log-likelihood function* as

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

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

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

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

2.1 Asymptotic normality of maximum likelihood estimator

The end of lecture 2 (18. 2. 2020)

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 everywhere if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. (*Identifiability*)

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

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

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

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

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

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

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

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

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

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

Example 13. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $\mathsf{N}(\mu_1 + \mu_2, 1)$. Then the identifiability assumption [R0] is not satisfied for the vector parameter $\boldsymbol{\theta} = (\mu_1, \mu_2)^\mathsf{T}$.

Example 14. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution $\mathsf{U}(0, \theta)$. Note that assumption [R2] is not satisfied.

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

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

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

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

The random vector

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

is called *the score statistic*.

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

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

Further define *the observed information matrix* as

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

where

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

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

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

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

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

where

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

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

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

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

which implies the statement of the lemma. □

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

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

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

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

The corollary now follows from Lemma 1. □

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

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

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

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

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

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

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

where I_n^* is a matrix with the elements

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

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

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

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

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

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

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

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

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

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

□

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

The end of lecture 3 (24. 2. 2020)

Remark 7. Note that strictly speaking Theorem 5 does not guarantee the asymptotic normality of the maximum likelihood estimator but of an appropriately chosen root of the likelihood equations (13). As illustrated in Example 17 it can happen that the maximum likelihood estimator defined by (12) is not a consistent estimator of $\boldsymbol{\theta}_X$ even if all the regularity assumptions **[R0]**-**[R6]** are satisfied. It can also happen that the maximum likelihood estimator does not exist (see the example on page 22). That is why some authors define the maximum likelihood estimator in regular families as a (an appropriately chosen) root of the likelihood equations.

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

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

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

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

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

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

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

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

2.3 Estimation of the asymptotic variance matrix

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

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

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

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

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

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

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

for μ-almost all $\mathbf{x}$ and

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

Example 15. Let $X_1,\dots,X_n$ be a random sample from the Pareto distribution with the density

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

where both parameters are unknown.

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

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

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

Example 16. Let $X_1,\dots,X_n$ 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 17. Let $X_1,\dots,X_n$ be a random sample from the mixture of distributions $\mathsf{N}(0,1)$ and $\mathsf{N}(\theta,\exp\{-2/\theta^2\})$ with equal weights and the parameter space given by $\Theta=(0,\infty)$. Define the estimator of the parameter θ as $\widehat{\theta}_n^{(ML)} = \arg\max_{\theta\in\Theta}\ell_n(\theta)$. Then it can be shown that $\widehat{\theta}_n^{(ML)} \xrightarrow[n\to\infty]{P} 0$, thus $\widehat{\theta}_n^{(ML)}$ is not consistent estimator.

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

Example 18. Let $X_1,\dots,X_n$ be a random sample from the mixture of distributions $\mathsf{N}(0,1)$ and $\mathsf{N}(\mu,\sigma^2)$ with equal weights and the parameter space for the parameter $\boldsymbol{\theta}=(\mu,\sigma)^\mathsf{T}$ is given by $\Theta=\mathbb{R}\times(0,\infty)$. Show that

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

and that the maximum likelihood estimator does not exist. But similarly as in Example 17 Theorem 5 still holds.

Literature: Anděl (2007) Chapter 7.6.5, Lehmann and Casella (1998) Chapter 6.5, Kulich (2014).

2.4 Asymptotic tests (without nuisance parameters)

Suppose we are interested in testing the null hypothesis

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

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

Likelihood ratio test is based on the test statistic

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

Wald test is based on the test statistic

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

Rao score test is based on the test statistic

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

Note that the advantage of the likelihood ratio test (LR_n) is that one does not need to estimate the Fisher information matrix. On the other hand the advantage of Rao score test (R_n) is that you do not need to calculate the maximal likelihood estimator $\widehat{\boldsymbol{\theta}}_n$. That is why in Rao score statistic (R_n) one uses usually either $I(\theta_0)$ or $I_n(\theta_0)$ as $\widehat{I}_n$. On the other hand usually (for historical reasons) $I(\widehat{\boldsymbol{\theta}}_n)$ or $I_n(\widehat{\boldsymbol{\theta}}_n)$ is used for Wald statistic (W_n).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

Example 19. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of K-variate random vectors from the multinomial distribution $\mathsf{Mult}_K(1, \mathbf{p})$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})^\mathsf{T}$ and $\mathbf{p} = (p_1, \ldots, p_K)^\mathsf{T}$. Suppose we are interested in testing the null hypothesis

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

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

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

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

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

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

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

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

2.5 Asymptotic confidence sets

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

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

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

Confidence intervals for θ_{Xj}

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

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

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

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 $\widehat{I}_n^{-1}$. Thus the two-sided (asymptotic) confidence interval for θ_{Xj} is given by

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

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

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

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

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

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

In what follows all the vectors and matrices appearing in the notation of maximum likelihood estimation theory are decomposed into the first q (part 1) and the remaining $p - q$ components (part 2), i.e.

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

and

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

* *rušivé*

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

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

Denote

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

Then

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

where

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

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

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

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

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

The end of lecture 4 (25. 2. 2020)

Let $\widehat{I}_n^{11}$ be an estimate of the corresponding block $I^{11}(\boldsymbol{\theta}_X)$ in the inverse of Fisher information matrix $I^{-1}(\boldsymbol{\theta}_X)$.

The three asymptotic tests of the null hypothesis $H_0 : \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$ are as follows.

Likelihood ratio test is based on the test statistic

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

Wald test is based on the test statistic

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

Rao score test is based on the test statistic

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

Remark 11. As $\mathbf{U}_{2n}\big(\widetilde{\boldsymbol{\theta}}_n\big) = \mathbf{0}_{p-q}$, 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 (18) of the Rao score test in case of no nuisance parameters.

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

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

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

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

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

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

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

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

which further with the CMT (Theorem 1) implies

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

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

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

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

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

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

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

As the null hypothesis holds, using Theorem 5 one gets

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

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

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

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

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

one gets

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

where

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

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

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

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

Further by Theorem 5

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

which together with (25) implies

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

where

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

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

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

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

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

Note that matrix $\mathbb{D}$ is idempotent, thus also $\mathbb{I}_p - \mathbb{D}$ and $\mathbb{B}\mathbb{V} = (\mathbb{I}_p - \mathbb{D})(\mathbb{I}_p - \mathbb{D})$ are idempotent. Now using (27), (28), CS (Theorem 2), Lemma 3 and CMT (Theorem 1) one gets

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

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

The end of lecture 5 (2. 3. 2020)

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

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

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

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

Example 20. The following data gives the number of male children among the first 12 children of family size 13 in 6115 families taken from hospital records in the 19th century Saxony. The 13th child is ignored to assuage the effect of families non-randomly stopping when a desired gender is reached. Test the null hypothesis that the gender of the babies can be viewed as realisations of independent random variables having the same probability of a baby boy for each family.

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

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

and the table can be viewed as a realisation of a random vector $(N_0, N_1, \dots, N_{12})^\mathsf{T}$ that follows multinomial distribution $\mathsf{Mult}_{13}(n, \boldsymbol{\pi})$.

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

Example 21. Breusch-Pagan test of heteroscedasticity.

Example 22. 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

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

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 set for the parameter $\boldsymbol{\beta}$.

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.

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

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

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

In this subsection we will assume that there exists $\widehat{\boldsymbol{\theta}}_n$ which is a unique maximiser of $\ell_n(\boldsymbol{\theta})$ and also a consistent estimator of $\boldsymbol{\theta}_X$.

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

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

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

First of all put

$$\widehat{\boldsymbol{\tau}}_n^{(p)} = \arg\max_{\boldsymbol{\tau}\in\Theta_{\boldsymbol{\tau}}} \ell_n^{(p)}(\boldsymbol{\tau}).$$

Note that

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

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

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

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

Further denote

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

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

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

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

Lemma 4. *Suppose that assumptions* [R0]-[R6] *are satisfied. Then (with probability tending to one) on a neighbourbood of $\boldsymbol{\tau}_X$*

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

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

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

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

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

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

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

Further by differentiating both sides of the identity

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

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

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

which implies that

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

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

Tests based on profile likelihood

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

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

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

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

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

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

and further

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

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

$\underline{W_n^{(p)}}$: Follows from Theorem 7 and the fact that by Lemmas 1, 2 and 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 $\widehat{I}_n^{11} = \big[\widehat{I}_n^{(p)}\big]^{-1}$, where R_n^* is Rao score test statistic in the presence of nuisance parameters defined in (22). The statement of the theorem now follows by (34) and Theorem 7. □

Confidence interval for θ_{Xj}

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

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

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

Example 23. 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 24. Box-Cox transformation. See Zvára (2008) pp. 149–151.

Remark 12. Although we have shown that one can work with the profile likelihood as with the standard likelihood not all the properties are shared. For instance for standard score statistic one has $\mathsf{E}\ \mathbf{U}_n(\boldsymbol{\theta}_X) = \mathbf{0}_p$. But this is not guaranteed for profile score statistic as by Lemma 4

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

and the expectation on the right-hand side of the previous equation is typically not zero due to the random argument $\widetilde{\boldsymbol{\psi}}_n(\boldsymbol{\tau}_X)$ (for illustration think of $\mathsf{E}\ U_n^{(p)}(\beta_X)$ in Example 23). From the proof of Theorem 7 we only know that $\frac{1}{\sqrt{n}}\mathbf{U}_n^{(p)}(\boldsymbol{\tau}_X)$ converges in distribution to a zero-mean Gaussian distribution.

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

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

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

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

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

the *log-likelihood function* as

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

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

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

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 25. 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 a 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 $f_i(\mathbf{x}; \boldsymbol{\theta})$).

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

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

See also Example 29.

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

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

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

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

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

Example 26. *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 $\lambda(\boldsymbol{X}_i) = \exp\{\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}\}$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathsf{T}}$. Then (provided assumptions **[R0]**-**[R6]** are satisfied)

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$)

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

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 27. Note that alternatively one can view the K-sample problem described in Example 25 also within i.i.d framework. Consider the data as a realization of the random sample $(\mathbf{Z}_1^\mathsf{T}, J_1)^\mathsf{T}, \ldots, (\mathbf{Z}_n^\mathsf{T}, J_n)^\mathsf{T}$, where J_i takes values in $\{1, \ldots, K\}$ and the conditional distribution of $\mathbf{Z}_i$ given $J_i = j$ is given by the density $f_j(\mathbf{x}; \boldsymbol{\theta})$.

Example 28. Maximum likelihood estimation in AR(1) process.

Example 29. Suppose that X_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ be independent random variables such that X_{ij} follows Bernoulli distribution with parameter p_i. We are interested in testing the hypothesis

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

Note that one can easily construct a likelihood ratio test.

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

Literature: Hoadley (1971).

The end of lecture 6 (3. 3. 2020)

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 30. *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 $\boldsymbol{\psi}^{(N)} = (\mu_1, \ldots, \mu_N)^\mathsf{T}$ 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 $\mathbf{V}$ and $\mathbf{W}$.

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

$$L_n^{(M)}(\boldsymbol{\tau}) = p(\mathbf{V}; \boldsymbol{\tau}), \qquad \ell_n^{(M)}(\boldsymbol{\tau}) = \log\left(L_n^{(M)}(\boldsymbol{\tau})\right),$$

* *Podmíněná a marginální věrohodnost.*

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

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

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

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

Remark 13. (i) If $\mathbf{V}$ is independent of $\mathbf{W}$, then $p(\mathbf{V}|\mathbf{W};\boldsymbol{\tau}) = p(\mathbf{V};\boldsymbol{\tau})$ and thus $L_n^{(M)}(\boldsymbol{\tau}) = L_n^{(C)}(\boldsymbol{\tau})$.

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

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

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

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

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

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

Exponential family

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

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

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 one can take $\mathbf{S}_n(\mathbb{X})$ as $\mathbf{W}$ and $\mathbb{X}$ as $\mathbf{V}$.

Example 31. *Strongly stratified sample (cont.).* Using marginal and conditional likelihood.

Example 32. Let Y_{ij}, $i = 1, \dots, N$, $j = 1, 2$ be independent random variables such that $Y_{i1} \sim \mathsf{Exp}(\psi_i)$ and $Y_{i2} \sim \mathsf{Exp}(\tau\,\psi_i)$ where $\tau > 0$ and ψ_i are unknown parameters. Show that the distribution of $V_i = \frac{Y_{i2}}{Y_{i1}}$ depends only on the parameter τ (and not on ψ_i). Derive the marginal likelihood of τ that is based on $\mathbf{V} = (V_1, \dots, V_N)^\mathsf{T}$.

Example 33. Let Y_{ij}, $i = 1, \dots, I$, $j = 0, 1$ 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 Chapter 2.6 require that I is fixed and all the sample sizes n_{ij} tend to infinity. This implies that using conditional likelihood is reasonable in situations when (some) n_{ij} are small.

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

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

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 34. Consider in Example 33 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 33 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 $\mathcal{K} = \left\{\max\{0, y_+ - n_0\}, \dots, \min\{y_+, n_1\}\right\}$.

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

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

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

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

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

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

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

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 35. Consider in Example 33 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.

Example 36. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from the Poisson distributions. Let λ_X be the true value of the parameter for the first sample and λ_Y for the second sample. Note that $\mathbf{S} = (S_1, S_2)^\mathsf{T} = \left(\sum_{i=1}^{n_1} X_i, \sum_{i=1}^{n_2} Y_i\right)^\mathsf{T}$ is a sufficient statistic for the parameter $\boldsymbol{\theta} = (\lambda_X, \lambda_Y)^\mathsf{T}$. Derive the conditional distribution of S_1 given $S_1 + S_2$. Use this result to find an exact test of

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

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

Literature: Pawitan (2001) Chapters 10.1–10.5.

The end of lecture 7 (9. 3. 2020)

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 (ML) 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 squares (LS) estimator of regression parameters as an M-estimator with

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

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

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

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

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

* *M-odhad*

Note that

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

Generally let $\boldsymbol{\psi}$ be a p-dimensional vector function (not necessarily a derivative of 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 squares (LS) estimators can be also viewed as Z-estimators with

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

Literature: van der Vaart (2000) – Chapter 5.1.

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

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

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

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

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

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

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

Example 37. Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be i.i.d. observations from a distribution with a density $f(\mathbf{x})$ (with respect to a σ-finite measure μ). By assuming that f belongs to a parametric family of densities $\mathcal{F} = \big\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\big\}$ we are estimating (identifying) $\boldsymbol{\theta}_X$ such that

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

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

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

---

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

Now by Jensen's inequality

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

Suppose that our (parametric) assumption **is right** and there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $f(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_0)$. Then $\mathsf{E}\,\log\left[\tfrac{f(\boldsymbol{X}_1;\boldsymbol{\theta})}{f(\boldsymbol{X}_1;\boldsymbol{\theta}_0)}\right]$ 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

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

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

3.2 Asymptotic distribution of Z-estimators

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

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

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

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

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

[Z3] 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 $\alpha > 0$ and an open neighbourhood U of $\boldsymbol{\theta}_X$ so that for each $j, k \in \{1, \ldots, p\}$ there exists a function $M_{jk}(\mathbf{x})$ such that for each $\boldsymbol{\theta} \in U$

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

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

[Z5] The matrix

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

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

[Z6] The *variance matrix*

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

is finite.

Introduce

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

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

Lemma 5. *Suppose that assumptions* [Z1]-[Z5] *are satisfied. Let* ε_n *be a sequence of positive numbers going to zero. Then*

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

where

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

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

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

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

which implies the statement of the lemma. □

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

(i) Then with probability going to one there exists a consistent solution $\widehat{\boldsymbol{\theta}}_n$ *to the estimating equations* (41).

(ii) Further, if $\widehat{\boldsymbol{\theta}}_n$ is a consistent root of the estimating equations (41), then

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

which further implies that

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

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

Proof. Consistency: Introduce the vector function

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

where

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

In what follows we will show that with probability going to one (as $n\to\infty$) the mapping h_n is a contraction on $U_{\varepsilon_n} = \left\{\boldsymbol{\theta}\in\Theta : \|\boldsymbol{\theta}-\boldsymbol{\theta}_X\| \le \varepsilon_n\right\}$, where ε_n is a sequence of positive numbers going to zero such that $\varepsilon_n\sqrt{n} \xrightarrow[n\to\infty]{} \infty$. Having proved that then by the Banach fixed point theorem (Theorem A2) there exists a unique fixed point $\widehat{\boldsymbol{\theta}}_n \in U_{\varepsilon_n}$ such that $h_n(\widehat{\boldsymbol{\theta}}_n) = \widehat{\boldsymbol{\theta}}_n$ and thus also $\mathbf{Z}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}_p$. This implies the existence of a consistent root of the estimating equations (41).

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

$$\begin{aligned}\left\|h_n(\boldsymbol{\theta}_1) - h_n(\boldsymbol{\theta}_2)\right\| &= \left\|\left(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right) - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\left(\mathbf{Z}_n(\boldsymbol{\theta}_1) - \mathbf{Z}_n(\boldsymbol{\theta}_2)\right)\right\| \\ &= \left\|\left(\mathbb{I}_p - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbb{\Gamma}_n^*\right)\left(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right)\right\|,\end{aligned} \tag{46}$$

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

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

evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\boldsymbol{\theta}_1$ a $\boldsymbol{\theta}_2$. Note that $\boldsymbol{\theta}_n^{j*} \in U_{\varepsilon_n}$. Now by Lemma 5 and assumption **[Z5]**

$$a_n = \max_{j,k\in\{1,\ldots,p\}}\sup_{\boldsymbol{\theta}\in U_{\varepsilon_n}}\left|\left(\mathbb{I}_p - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbb{\Gamma}_n(\boldsymbol{\theta})\right)_{jk}\right| = o_P(1). \tag{47}$$

So with the help of (46) and (47) it holds that

$$\left\|h_n(\boldsymbol{\theta}_1) - h_n(\boldsymbol{\theta}_2)\right\| \le p^2\,a_n\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|, \tag{48}$$

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

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

Thus to show that h_n is a contraction on U_{ε_n} it remains to prove that (with probability going to one) $h_n : U_{\varepsilon_n} \to U_{\varepsilon_n}$. Note that for each $\boldsymbol{\theta} \in U_{\varepsilon_n}$ the inequality (48) implies

$$h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}_X) = o_P(1)\,\varepsilon_n, \tag{49}$$

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

$$h_n(\boldsymbol{\theta}_X) = \boldsymbol{\theta}_X - \left[\mathbb{\Gamma}(\boldsymbol{\theta}_X)\right]^{-1}\mathbf{Z}_n(\boldsymbol{\theta}_X) = \boldsymbol{\theta}_X + O_P\big(\tfrac{1}{\sqrt{n}}\big), \tag{50}$$

where we have used that by the central limit theorem $\mathbf{Z}_n(\boldsymbol{\theta}_X) = O_P\big(\frac{1}{\sqrt{n}}\big)$. Now combining (49) and (50) yields that

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

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

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

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

where similarly as in the proof of consistency $\mathbb{\Gamma}_n^*$ is $(p\times p)$-matrix whose j-th row is the j-th row of the matrix $\mathbb{\Gamma}_n(\boldsymbol{\theta})$ evaluated at some $\boldsymbol{\theta}_n^{j*}$ that is between $\widehat{\boldsymbol{\theta}}_n$ a $\boldsymbol{\theta}_X$. Thus $\boldsymbol{\theta}_n^{j*} \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}_X$ as $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_X$. So one can use Lemma 5 to conclude that $\mathbb{\Gamma}_n^* \xrightarrow[n\to\infty]{P} \mathbb{\Gamma}(\boldsymbol{\theta}_X)$. Now with the help of CS (Theorem 2) one can write

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

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

Remark 14. If there exists a real function $\rho(\mathbf{x};\boldsymbol{\theta})$ such that $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial\rho(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, then the matrix $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$ is symmetric and one can simply write $\mathbb{\Gamma}(\boldsymbol{\theta}_X)^{-1}$ instead of $[\mathbb{\Gamma}(\boldsymbol{\theta}_X)^{-1}]^{\mathsf{T}}$ in (45).

The end of lecture 8 (10. 3. 2020)

Asymptotic variance estimations

Note that by Theorem 9 one has

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

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

$$\widehat{\mathsf{avar}\,(\widehat{\boldsymbol{\theta}}_n)} = \frac{1}{n}\,\widehat{\mathbb{\Gamma}}_n^{-1}\,\widehat{\mathbb{\Sigma}}_n\,\big[\widehat{\mathbb{\Gamma}}_n^{-1}\big]^{\mathsf{T}}, \tag{51}$$

where

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

Note that Lemma 5 together with the consistency of $\widehat{\boldsymbol{\theta}}_n$ implies that

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

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

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

To derive such assumptions rewrite

$$\begin{aligned}
\widehat{\mathbb{\Sigma}}_n = {} & \frac{1}{n}\sum_{i=1}^{n}\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]\,\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]^{\mathsf{T}} \\
& + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\,\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]^{\mathsf{T}} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[\boldsymbol{\psi}\big(\boldsymbol{X}_i;\widehat{\boldsymbol{\theta}}_n\big) - \boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\big]\,\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X) \\
& + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{X}_i;\boldsymbol{\theta}_X)\,\boldsymbol{\psi}^{\mathsf{T}}(\boldsymbol{X}_i;\boldsymbol{\theta}_X).
\end{aligned} \tag{52}$$

Now by the law of large numbers the last summand in (52) converges in probability to $\mathbb{\Sigma}(\boldsymbol{\theta}_X)$, thus it is sufficient to show that the remaining terms are of order $o_P(1)$. With the help of assumption **[Z4]** this can be done for instance by assuming that for each $j,k \in \{1,\dots,p\}$

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

Confidence sets and confidence intervals

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

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

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

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

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

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

3.3 Likelihood under model misspecification

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample with a density f (with respect to a σ-finite measure μ). Then the maximum likelihood estimator can be viewed as the M-estimator with $\rho(\mathbf{x}; \boldsymbol{\theta}) = -\log f(\mathbf{x}; \boldsymbol{\theta})$ or Z-estimator with $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. From Example 37 we know that when assuming $f \in \mathcal{F} = \{f(\mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$, the method of the maximum likelihood identifies the parameter

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

Further by Theorem 9 we also know that (with probability going to one there exists a consistent solution $\widehat{\boldsymbol{\theta}}_n$ of (41) which satisfies)

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

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

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

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

Often in practice we are not completely sure that $f \in \mathcal{F}$. If we are not sure about the parametric assumption then it is safer to view the estimator $\widehat{\boldsymbol{\theta}}_n$ as an Z-estimator with $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = -\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. The asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ can now be estimated with the help of 'sandwich estimator' (51) where

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

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

Example 38. *Misspecified normal linear model.* Let $\binom{\boldsymbol{X}_1}{Y_1}, \dots, \binom{\boldsymbol{X}_n}{Y_n}$ be independent and identically distributed random vectors, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\mathsf{T}$. Note that if one assumes that $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \sim \mathsf{N}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}, \sigma^2)$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$, then the maximum likelihood estimation of $\boldsymbol{\beta}$ corresponds to the method of least squares given by $\rho(\mathbf{x}, y; \boldsymbol{\beta})$.

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

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

and it holds that

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(0, \mathbb{V}), \quad \text{where} \quad \mathbb{V} = \big[\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]^{-1}\big[\mathsf{E}\ \sigma^2(\boldsymbol{X}_1)\,\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]\big[\mathsf{E}\ \boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T}\big]^{-1},$$

with $\sigma^2(\boldsymbol{X}_1) = \mathsf{var}(Y_1|\boldsymbol{X}_1)$.

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

Example 39. *Misspecified Poisson regression.* Let $\binom{\boldsymbol{X}_1}{Y_1}, \dots, \binom{\boldsymbol{X}_n}{Y_n}$ be independent and identically distributed random vectors, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\mathsf{T}$. Assume that the conditional distribution of Y_i given $\boldsymbol{X}_i$ is Poisson, i.e. $\mathcal{L}\big(Y_i|\boldsymbol{X}_i\big) \sim \mathsf{Po}\big(\lambda(\boldsymbol{X}_i)\big)$, where $\lambda(\mathbf{x}) = \mathrm{e}^{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}}$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\mathsf{T}$. The score function for the maximum likelihood estimation is given by

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

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

$$\boldsymbol{\psi}(\mathbf{x}, y; \boldsymbol{\beta}) = \mathbf{x}\big(y - \mathrm{e}^{\mathbf{x}^\mathsf{T}\boldsymbol{\beta}}\big) \tag{53}$$

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

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

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

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

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

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

then we can still use the score function (53) which identifies the parameter $\boldsymbol{\beta}_0$. By Theorem 9 we know that the estimator $\widehat{\boldsymbol{\beta}}_n$ is asymptotically normal with the matrices $\mathbb{\Gamma}(\boldsymbol{\beta}_X)$ and $\mathbb{\Sigma}(\boldsymbol{\beta}_X)$ given by

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

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

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

where

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

Literature: White (1980), White (1982).

3.4 Asymptotic normality of M-estimators defined by convex minimization

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

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

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

A natural estimate of the parameter $\boldsymbol{\theta}_X$ is given by

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

The function $\rho(\mathbf{x}; \boldsymbol{\theta})$ does not have to be smooth in $\boldsymbol{\theta}$, but the convexity guarantees that it is differentiable almost everywhere. Thus for each $\boldsymbol{\theta}$ let $\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \rho(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ for almost all $\mathbf{x}$. Further suppose that

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

For formulating the main result it is useful to introduce the 'reminder function'

$$R(\mathbf{x}; \mathbf{t}) = \rho(\mathbf{x}; \boldsymbol{\theta}_X + \mathbf{t}) - \rho(\mathbf{x}; \boldsymbol{\theta}_X) - \mathbf{t}^\mathsf{T} \boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta}_X) \tag{54}$$

and the asymptotic objective function

$$M(\boldsymbol{\theta}) = \mathsf{E}\ \rho(\boldsymbol{X}_1; \boldsymbol{\theta}).$$

Theorem 10. *Suppose that (54) holds and that*

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

$$M(\boldsymbol{\theta}_X + \mathbf{t}) = M(\boldsymbol{\theta}_X) + \tfrac{1}{2}\,\mathbf{t}^{\mathsf{T}}\mathbb{\Gamma}(\boldsymbol{\theta}_X)\,\mathbf{t} + o(\|\mathbf{t}\|^2), \;\; \textit{as } \mathbf{t} \to \mathbf{0}_p;$$

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

Then the asymptotic representation (44) *holds for* $\widehat{\boldsymbol{\theta}}_n$*, which further gives the asymptotic normality result* (45).

Proof. See the 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. On the other hand it is worth noting that Theorem 10 allows for $\rho(\mathbf{x}; \boldsymbol{\theta})$ that does not meet assumptions **[Z3]** and **[Z4]**. Note that the matrix $\mathbb{\Gamma}(\boldsymbol{\theta}_X)$ does not have to be computed as $\mathbb{\Gamma}(\boldsymbol{\theta}_X) = \mathsf{E}\,\mathbb{D}_{\boldsymbol{\psi}}(\boldsymbol{X}_1; \boldsymbol{\theta}_X)$ (as in Theorem 9) 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 assumption about the smoothness of $\boldsymbol{\psi}$ (i.e. **[Z3]** and **[Z4]**) can be replaced with the assumptions on the distribution of $\boldsymbol{X}_1$ so that the function $M(\boldsymbol{\theta})$ is sufficiently smooth.

Another important difference in comparison to Theorem 9 is that Theorem 10 guarantees the asymptotic normality for the minimizer of $\frac{1}{n}\sum_{i=1}^{n}\rho(\boldsymbol{X}_i; \boldsymbol{\theta})$ and not only for a consistent root of the estimating equations which might be difficult to find in case that there are more roots to the estimating equations.

3.4.1 Sample median

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with density $f(y)$ that is positive and continuous in a neighbourhood of median $F^{-1}(0.5)$.

It is well known (see Lemma 6 and Remark 16) that the sample median $\widetilde{m}_n$ can be written as

$$\widetilde{m}_n = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |X_i - \theta|.$$

Thus one can view $\widetilde{m}_n$ as an M-estimator with $\rho(x; \theta) = |x - \theta|$. For theoretical reasons it is advantageous to consider

$$\rho(x; \theta) = |x - \theta| - |x|,$$

which does not require that $\mathsf{E}\,|X_1| < \infty$ in order to define $M(\theta) = \mathsf{E}\,\rho(X; \theta)$. Note that then

$$F^{-1}(0.5) = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\,\rho(X; \theta).$$

Now one uses Theorem 10 to derive the asymptotic distribution of $\widetilde{m}_n$. Note that it is natural to take

$$\psi(x;\theta) = -\operatorname{sign}(x-\theta).$$

Then (the matrix) $\mathbb{\Sigma}(\theta_X)$ reduces to

$$\sigma_\psi^2 = \mathsf{var}\left(\psi(X_1; F^{-1}(0.5))\right) = 1.$$

Further

$$\frac{\partial M(\theta)}{\partial \theta} = -\,\mathsf{E}\left[\operatorname{sign}(X_1-\theta)\right] = -\,\mathsf{P}(X_1 > \theta) + \mathsf{P}(X_1 < \theta) = 2\,F(\theta) - 1,$$

which implies that (the matrix) $\mathbb{\Gamma}(\theta_X)$ reduces to

$$\gamma = \frac{\partial^2 M(\theta)}{\partial \theta^2}\Big|_{\theta = F^{-1}(0.5)} = 2\,f\big(F^{-1}(0.5)\big).$$

One can show that in our situation the assumptions of Theorem 10 are satisfied, so one gets

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

Literature: Hjort and Pollard (2011) Section 2A.

The end of the self study for the week (16.-20. 3. 2020)

4 M-estimators and Z-estimators in robust statistics*

In statistics the word 'robust' has basically two meanings.

(i) We say that a procedure is robust, if it stays (approximately/asymptotically) valid even when some of the assumptions (under which the procedure is derived) are not satisfied. For instance the standard ANOVA F-statistic is robust against the violation of the normality of the observations provided that the variances of all the observations are the same (and finite).

(ii) People interested in robust statistics say that a procedure is robust, if it is not 'too much' influenced by the outlying observations. In what follows we will concentrate on this meaning of the robustness.

One of the standard measures of robustness is the **breakdown point**. Vaguely speaking the breakdown point of an estimator is the smallest percentage of observations that one has to change so that the estimator produces a nonsense value (e.g. $\pm\infty$ for location or regression estimator; 0 or $+\infty$ when estimating the scale).

* *Robustní statistika*

Let $\widehat{\boldsymbol{\theta}}_n$ be an M- or Z-estimator of a parameter $\boldsymbol{\theta}_X$. Note that thanks to Theorems 9 or 10 (under appropriate assumptions) one has the following representation

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X = \frac{1}{n}\sum_{i=1}^{n} IF(\boldsymbol{X}_i) + o_P\big(\tfrac{1}{\sqrt{n}}\big),$$

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

$$\frac{1}{n}\big[IF(\boldsymbol{X}_i + \boldsymbol{\Delta}) - IF(\boldsymbol{X}_i)\big].$$

Thus provided that $IF(\mathbf{x})$ is bounded then also this change is bounded (and of order $O(\frac{1}{n})$).

Note that the above reasoning was not completely correct as the term $o_P\big(\frac{1}{\sqrt{n}}\big)$ was ignored. Nevertheless it can be proved that (under some mild assumptions excluding 'singular' cases) if the function $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})$ is bounded then the breakdown point of the associated $M(Z)$-estimator is $\frac{1}{2}$.

4.1 Robust estimation of location*

Suppose that we observe a random sample $X_1, \ldots, X_n$ from a distribution F and we are interested in characterising the location.

Note that for the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ it is sufficient to change only one observation to get an arbitrary value of $\overline{X}_n$.

On the other hand when considering the sample median $\widetilde{m}_n = \widehat{F}_n^{-1}(0.5)$ then one needs to change at least half of the observations so that one can for instance change the estimator to $\pm\infty$.

When deciding between a sample mean and a sample median one has to take into consideration that if the distribution F is not symmetric then $\overline{X}_n$ and $\widetilde{m}_n$ estimate different quantities. But when one can hope that the distribution F is symmetric, then both $\overline{X}_n$ and $\widetilde{m}_n$ estimate the centre of the symmetry and one can be interested which of the estimators is more appropriate. By the maximum likelihood theory we know that $\overline{X}_n$ is efficient if F is normal while $\widetilde{m}_n$ is efficient if F is doubly exponential (i.e. it has a density $f(x) = \frac{1}{2\sigma}\exp\{-\frac{|x-\theta|}{\sigma}\}$).

In robust statistics it is usually assumed that most of our observations follow normal distributions but there are some outlying values. This can be formalised by assuming that the distribution function F of each of the observations satisfies

$$F(x) = (1-\eta)\,\Phi\big(\tfrac{x-\mu}{\sigma}\big) + \eta\,G(x), \tag{55}$$

where η is usually interpreted as probability of having an outlying observation and G is a distribution (hopefully symmetric around μ) of outlying observations. It was found that if η

* *Robustní odhad polohy*

is 'small' then using sample median is too pessimistic (and inefficient). We will mention here several alternative options.

Before we proceed note that both the sample mean $\overline{X}_n$ and the sample median $\widetilde{m}_n$ can be viewed as M-estimators as

$$\overline{X}_n = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} (X_i - \theta)^2 \quad \text{and} \quad \widetilde{m}_n = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} |X_i - \theta|. \tag{56}$$

Huber estimator

This estimator is defined as

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

where

$$\rho_H(x) = \begin{cases} \frac{x^2}{2}, & |x| \le k, \\ k \cdot \left(|x| - \frac{k}{2}\right), & |x| > k. \end{cases} \tag{57}$$

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

$$\psi_H(x) = \rho'_H(x) = \begin{cases} x, & |x| \le k, \\ k \cdot \operatorname{sgn}(x), & |x| > k. \end{cases} \tag{58}$$

Thus one can see that for $x \in (-k, k)$ the function ψ_H corresponds to a score function of a sample mean (which is $\psi(x) = x$) while for $x \in (-\infty, k) \cup (k, \infty)$ it corresponds to a score function of a sample median (which is $\psi(x) = \operatorname{sgn}(x)$). Thus Huber estimator presents a compromise between a sample mean and a sample median. So it is not surprising that $\widehat{\theta}_n^{(H)}$ is usually a value between the sample median and the sample mean.

When using Huber estimator one has to keep in mind that the identified parameter is

$$\theta_H = \arg\min_{\theta\in\mathbb{R}} \mathsf{E}\ \rho_H(X_1 - \theta).$$

Thus if the distribution F **is not symmetric** then $\mathsf{E}\ X_1$ generally does not coincide with $F^{-1}(0.5)$ and θ_H lies between $\mathsf{E}\ X_1$ and $F^{-1}(0.5)$.

On the other hand if the distribution F **is symmetric**, then θ_H coincides with the centre of symmetry, i.e. with $F^{-1}(0.5)$ (the median of F) and also with $\mathsf{E}\ X_1$, if the expectation exists. It was observed that for the contamination model (55) with G symmetric, Huber estimator usually performs better than the sample mean as well as the sample median. This can be proved analytically by showing that for $\eta > 0$ and G heavy tailed, then usually

$$\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right) < \min\left\{\mathsf{var}(\overline{X}_n), \mathsf{avar}(\widetilde{m}_n)\right\},$$

where the asymptotic variance $\mathsf{avar}\left(\widehat{\theta}_n^{(H)}\right)$ is derived in Example 40.

The nice thing about Huber estimator is that its loss function $\rho(x;\theta) = \rho_H(x-\theta)$ is convex (in θ) thus $\widehat{\theta}_n^{(H)}$ is not too difficult to calculate and with the help of Theorem 10 one can derive its asymptotic distribution (see also Example 40).

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 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 $k = 1.03$.

Example 40. With the help of Theorem 10 one can show that (under appropriate regularity assumptions)

$$\sqrt{n}\left(\widehat{\theta}_n^{(H)} - \theta_H\right) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \tfrac{\sigma_\psi^2}{\gamma^2}\big),$$

where

$$\gamma = \frac{\partial^2 \,\mathsf{E}\, \rho_H(X_1 - \theta)}{\partial \theta^2}\bigg|_{\theta=\theta_H} = F(\theta_H + k) - F(\theta_H - k)$$

and

$$\sigma_\psi^2 = \mathsf{var}\left(\psi_H(X_1 - \theta_H)\right) = \int_{\theta_H - k}^{\theta_H + k} (x - \theta_H)^2 \,\mathrm{d}F(x) + k^2\big(1 - F(\theta_H + k) + F(\theta_H - k)\big).$$

Thus $\mathsf{avar}(\widehat{\theta}_n^{(H)}) = \frac{\sigma_\psi^2}{n\gamma^2}$.

Other robust M/Z-estimators of location

The other most common M/Z-estimators are the following.

(i) **Cauchy-pseudolikelihood**: $\rho(x;\theta) = \log(1 + (x-\theta)^2)$. The problem is that this function is not convex in θ and the estimating equation

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

has usually more roots.

(ii) **Tukey's biweight:**

$$\psi(x) = \begin{cases} x\left(1 - \frac{x^2}{k^2}\right)^2, & |x| \leq k, \\ 0, & |x| > k. \end{cases}$$

But also here the corresponding loss function ρ ($\psi = \rho'$) is not convex.

4.2 Studentized M/Z-estimators

The problem is that the M/Z-estimators presented above (except for the sample mean and the sample median) are not scale equivariant (i.e. $\widehat{\theta}_n(c\,\boldsymbol{X}) \neq c\,\widehat{\theta}_n(\boldsymbol{X})$ for each $c \in \mathbb{R}$). That is why in practice M/Z-estimators are usually defined as

$$\widehat{\theta}_n = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \rho\big(\tfrac{X_i-\theta}{S_n}\big), \text{ or as } \sum_{i=1}^{n} \psi\big(\tfrac{X_i-\widehat{\theta}_n}{S_n}\big) \stackrel{!}{=} 0,$$

where S_n is an appropriate estimator of scale*, which satisfies $S_n(c\,\boldsymbol{X}) = |c|\,S_n(\boldsymbol{X})$ for each $c \in \mathbb{R}$. The most common estimators of scale are as follows.

Sample standard deviation

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

Note that in robust statistics S_n is rather rarely used as it is not robust (i.e. it is sensitive to outlying observations).

Interquartile range†

$$S_n = IQR = \widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25),$$

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

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

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

Note that the breakdown point of interquartile range is 0.25.

Median absolute deviation‡

This measure is given as the median absolute deviation from the median, i.e.

$$MAD = \mathrm{med}_{1\leq i\leq n}|X_i - \widehat{F}_n^{-1}(0.5)|,$$

or its modification

$$\widetilde{MAD} = \frac{MAD}{\Phi^{-1}(0.75)},$$

* *odhad měřítka* † *mezikvartilové rozpětí* ‡ *mediánová absolutní odchylka*

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 15. Note that due to the studentization the functions $\rho(x;\theta) = \rho\big(\frac{x-\theta}{S_n}\big)$ and $\psi(x;\theta) = \psi\big(\frac{x-\theta}{S_n}\big)$ (when viewed as functions of x and θ) are random. Thus one can use neither Theorem 9 nor Theorem 10 to derive the asymptotic distribution of studentized M/Z-estimators.

Nevertheless, if $S_n \xrightarrow[n\to\infty]{P} S(F)$ and the distribution F is symmetric, then (under some regularity assumptions) it can be shown that the asymptotic distribution of studentized Z/M-estimators is the same as the the asymptotic distribution of M/Z-estimators with $\rho(x;\theta) = \rho\big(\frac{x-\theta}{S(F)}\big)$ and $\psi(x;\theta) = \psi\big(\frac{x-\theta}{S(F)}\big)$ for which one can (usually) use either Theorem 9 or Theorem 10.

4.3 Robust estimation in linear models

Suppose we observe independent random vectors $\binom{\boldsymbol{X}_1}{Y_1}, \dots, \binom{\boldsymbol{X}_n}{Y_n}$ each of them having the same distribution as the generic random vector $\binom{\boldsymbol{X}}{Y}$.

4.3.1 The least squares method

This method results in the estimator

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

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

From Example 38 we know that the method of the least squares identifies the parameter

$$\boldsymbol{\beta}_X^{(LS)} = \big[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\big]^{-1}\ \mathsf{E}\ \boldsymbol{X}Y$$

and it holds that

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n^{(LS)} - \boldsymbol{\beta}_X^{(LS)}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(0,\mathbb{V}),\ \text{where}\ \mathbb{V} = \big[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\big]^{-1}\big[\mathsf{E}\ \sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\big]\big[\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\big]^{-1},$$

with $\sigma^2(\boldsymbol{X}) = \mathsf{var}(Y|\boldsymbol{X})$. Further provided $\mathsf{E}\,\big[Y|\boldsymbol{X}\big] = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X^{(LS)} = \boldsymbol{\beta}_0$.

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 model holds

$$Y = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + \varepsilon,\ \text{where}\quad \varepsilon \perp \widetilde{\boldsymbol{X}}. \tag{59}$$

Then $\mathsf{E}\,[Y \mid \boldsymbol{X}] = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\widetilde{\boldsymbol{X}} + \mathsf{E}\ \varepsilon$ and the method of the least squares identifies the parameter

$$\boldsymbol{\beta}_X^{(LS)} = \begin{pmatrix} \beta_0 + \mathsf{E}\ \varepsilon \\ \boldsymbol{\beta} \end{pmatrix}. \tag{60}$$

Further the asymptotic variance matrix $\mathbb{V}$ simplifies to

$$\mathbb{V} = \sigma^2 \left(\mathsf{E}\, \boldsymbol{X}\boldsymbol{X}^\mathsf{T} \right)^{-1}, \quad \text{where} \quad \sigma^2 = \mathsf{var}(\varepsilon). \tag{61}$$

4.3.2 Method of the least absolute deviation*

This method is usually considered as a robust alternative to the least squares methods. The estimate of the regression parameter is given by

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

As we will see later (see Chapter 5) the LAD method models $\mathsf{med}[Y \,|\, \boldsymbol{X}] = F^{-1}_{Y|\boldsymbol{X}}(0.5)$ as $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$. So if indeed $\mathsf{med}[Y \,|\, \boldsymbol{X}] = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X^{(LAD)} = \boldsymbol{\beta}_0$.

The *asymptotic distribution* of $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$ can be heuristically derived by Theorem 10 as follows. The score function is given by

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

Now put $M(\mathbf{b}) = \mathsf{E}\left[|Y - \boldsymbol{X}^\mathsf{T}\mathbf{b}| - |Y|\right]$, where the random vector $\binom{\boldsymbol{X}}{Y}$ has the same distribution as $\binom{\boldsymbol{X}_i}{Y_i}$. Then

$$\begin{aligned}\frac{\partial M(\mathbf{b})}{\partial \mathbf{b}} &= \mathsf{E}\left[\operatorname{sign}(Y - \boldsymbol{X}^\mathsf{T}\mathbf{b})\,(-\boldsymbol{X})\right] = -\,\mathsf{E}\,\boldsymbol{X}\left[\mathbb{I}\{Y > \boldsymbol{X}^\mathsf{T}\mathbf{b}\} - \mathbb{I}\{Y < \boldsymbol{X}^\mathsf{T}\mathbf{b}\}\right] \\ &= -\,\mathsf{E}\,\boldsymbol{X}\left[1 - 2F_{Y|\boldsymbol{X}}\left(\boldsymbol{X}^\mathsf{T}\mathbf{b}\right)\right].\end{aligned}$$

Thus

$$\frac{\partial^2 M(\mathbf{b})}{\partial \mathbf{b}\, \partial \mathbf{b}^\mathsf{T}} = 2\,\mathsf{E}\,\boldsymbol{X} f_{Y|\boldsymbol{X}}\left(\boldsymbol{X}^\mathsf{T}\mathbf{b}\right)\boldsymbol{X}^\mathsf{T},$$

which finally implies that

$$\mathbb{\Gamma}(\boldsymbol{\beta}_X^{(LAD)}) = \frac{\partial^2 M(\mathbf{b})}{\partial \mathbf{b}\, \partial \mathbf{b}^\mathsf{T}}\Big|_{\mathbf{b}=\boldsymbol{\beta}_X^{(LAD)}} = 2\,\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\left([\boldsymbol{\beta}_X^{(LAD)}]^\mathsf{T}\boldsymbol{X}\right)\right] = 2\,\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\left(F^{-1}_{Y|\boldsymbol{X}}(0.5)\right)\right].$$

Further as

$$\mathbb{\Sigma}(\boldsymbol{\beta}_X^{(LAD)}) = \mathsf{var}\left(\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}_X^{(LAD)})\right) = \mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T},$$

one gets that under appropriate regularity assumptions

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{(LAD)} - \boldsymbol{\beta}_X^{(LAD)}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}),$$

where

$$\mathbb{V} = \left(\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\left(F^{-1}_{Y|\boldsymbol{X}}(0.5)\right)\right]\right)^{-1} \mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T} \left(\mathsf{E}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\left(F^{-1}_{Y|\boldsymbol{X}}(0.5)\right)\right]\right)^{-1}.$$

* *Metoda nejmenších absolutních odchylek, mediánová regrese*

Note that if model (59) holds, then $\mathsf{med}(Y_1 \,|\, \boldsymbol{X}_1) = \beta_0 + \widetilde{\boldsymbol{X}}_1^{\mathsf{T}}\boldsymbol{\beta} + F_\varepsilon^{-1}(0.5)$, where F_ε^{-1} is the quantile function of ε_1 and thus

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

Thus when compared with the method of the least squares (60) one can see, that if model (59) holds then both methods identify the same slope parameter $\boldsymbol{\beta}$. The only difference is in intercept.

Further if model (59) holds then

$$f_{Y|\boldsymbol{X}}\big(F_{Y|\boldsymbol{X}}^{-1}(0.5)\big) = f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big),$$

which implies that

$$\mathbb{\Gamma}(\boldsymbol{\beta}_X^{(LAD)}) = 2\, f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\ \mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$$

and

$$\mathbb{V} = \frac{1}{4\big[f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\big]^2}\,\big(\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\big)^{-1}. \tag{62}$$

Now when one compares (61) with (62), one can see that the least absolute deviation method is favourable if

$$\frac{1}{\big[4f_\varepsilon\big(F_\varepsilon^{-1}(0.5)\big)\big]^2} < \mathsf{var}(\varepsilon).$$

Regarding the robustness of the least absolute deviation estimator note that in this special situation (i.e. if model (59) holds) $Y_i - \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}_X^{(LAD)} = \varepsilon_i - F_\varepsilon^{-1}(0.5)$ and the asymptotic representation (44) of $\widehat{\boldsymbol{\beta}}_n$ implies

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

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

4.3.3 Huber estimator of regression

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

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

where ρ_H is defined in (57). Generally, it is difficult to interpret what is being modelled with Huber estimator of regression (it is something between $\mathsf{E}\,(Y \mid \boldsymbol{X})$ and $\mathsf{med}(Y \mid \boldsymbol{X})$). Note that it identifies

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

Equivalently $\boldsymbol{\beta}_X^{(H)}$ solves

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

where ψ_H is defined in (58).

Analogously as in Example 40 one can derive that under appropriate assumptions

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n^{(H)} - \boldsymbol{\beta}_X^{(H)}\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}), \quad \text{with} \quad \mathbb{V} = \mathbb{\Gamma}^{-1}(\boldsymbol{\beta}_X^{(H)})\,\mathbb{\Sigma}(\boldsymbol{\beta}_X^{(H)})\,\mathbb{\Gamma}^{-1}(\boldsymbol{\beta}_X^{(H)}),$$

where

$$\mathbb{\Gamma}\big(\boldsymbol{\beta}_X^{(H)}\big) = \mathsf{E}_{\boldsymbol{X}}\left[F_{Y|\boldsymbol{X}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)} + k) - F_{Y|\boldsymbol{X}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)} - k)\right]$$

and

$$\mathbb{\Sigma}\big(\boldsymbol{\beta}_X^{(H)}\big) = \mathsf{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\,\mathsf{var}\left(\psi(Y - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_X^{(H)})|\boldsymbol{X}\right)\right].$$

If model (59) holds then $\boldsymbol{\beta}_X^{(H)} = \binom{\beta_{X0}^{(H)}}{\widetilde{\boldsymbol{\beta}}_X^{(H)}}$ solves

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

Thus $\boldsymbol{\beta}_X$ identifies the following parameter

$$\boldsymbol{\beta}_X^{(H)} = \begin{pmatrix} \beta_0 + \theta_H \\ \boldsymbol{\beta} \end{pmatrix},$$

where θ_H solves $\mathsf{E}\ \psi_H(\varepsilon - \theta_H) \stackrel{!}{=} 0$. So if model (59) holds then the interpretation of the regression slope coefficient ($\boldsymbol{\beta}$) is the same for each of the methods described above (LS, LAD, Huber regression).

Further the asymptotic variance matrix simplifies to

$$\mathbb{V} = \frac{\sigma_\psi^2}{\gamma^2}\left(\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}, \tag{63}$$

with

$$\gamma = F_\varepsilon(\theta_H + k) - F_\varepsilon(\theta_H - k)$$

and

$$\sigma_\psi^2 = \int_{\theta_H - k}^{\theta_H + k} (x - \theta_H)^2\,\mathrm{d}F_\varepsilon(x) + k^2\big(1 - F_\varepsilon(\theta_H + k) + F_\varepsilon(\theta_H - k)\big).$$

Using (61), (62) and (63) one sees that to compare the efficiency of the estimators $\widehat{\boldsymbol{\beta}}_n^{(LS)}$, $\widehat{\boldsymbol{\beta}}_n^{(LAD)}$ and $\widehat{\boldsymbol{\beta}}_n^{(H)}$ it is sufficient to compare $\mathsf{var}(\varepsilon)$, $\frac{1}{4f_\varepsilon^2(F_\varepsilon^{-1}(0.5))}$ and $\frac{\sigma_\psi^2}{\gamma^2}$.

Regarding the robustness properties the influence function is given by

$$IF(\mathbf{x}, y) = \left(\mathsf{E}\ \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1} \frac{1}{\gamma}\, \psi_H(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}_X^{(H)})\, \mathbf{x},$$

thus the estimator is robust in response but not in the covariate.

4.3.4 Studentized Huber estimator of regression

Analogously as in Chapter 4.2 in practice *the studentized Huber estimator* is usually used. This estimator is defined as

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

where S_n is an estimator of scale of ε_i. For instance one can take MAD or IQR calculated from the residuals from LAD regression $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n^{(LAD)}$.

Inference:

- With the help of Theorem 10 one can show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ of the (non-Studentized) Huber estimator.
- If model (59) holds, then it can be shown, that the estimate of the scale influences only the asymptotic distribution of the estimate of the intercept and not of the slope.

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

The end of the self study for the week (23.-27. 3. 2020)

5 Quantile regression*

Generally speaking, while the least squares method aims at estimating (modelling) a conditional expectation, quantile regression aims at estimating (modelling) a conditional quantile. This is of interest if the covariate may have different effect on different quantiles of the response.

Applications of the quantile regression can be found in medicine (e.g. constructing reference charts), finance (e.g. estimating value at risk), economics (e.g. wage and income studies, modelling household electricity demand) and environment modelling (e.g. modelling flood height).

* *Kvantilová regrese.*

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. Then*

$$F^{-1}(\tau) = \arg\min_{\theta \in \mathbb{R}} \mathsf{E}\,\big[\rho_\tau(X-\theta) - \rho_\tau(X)\big]. \tag{64}$$

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

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

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

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

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

Analogously for $\theta > F^{-1}(\tau)$

$$M'(\theta_-) = -\tau + F(\theta_-) \ge -\tau + F\big(F^{-1}(\tau)\big) = 0 \text{ and } M'(\theta_+) \ge 0,$$

thus the function $M(\theta)$ is non-decreasing on $\big(F^{-1}(\tau), +\infty\big)$. This further implies that $F^{-1}(\tau)$ is the point of the global minimum of the function $M(\theta)$. □

Remark 16. Suppose we observe a random sample $X_1, \ldots, X_n$. Let $\widehat{F}_n$ be the corresponding empirical distribution function. Then by

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(X_i - \theta) = \mathsf{E}_{\widehat{F}_n}\, \rho_\tau(X-\theta),$$

where the random variable X has the distribution given by the the empirical distribution function $\widehat{F}_n$ and $\mathsf{E}_{\widehat{F}_n}$ stands for the expectation with respect to this distribution.

Thus by Lemma 6

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

Note that for $\tau = 0.5$ one gets the characterization of the sample median as in (56).

Further note that from the proof of Lemma 6 it follows that the $\arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(X_i-\theta)$ is not unique if there exists a root of the function $-\tau + \widehat{F}_n(\theta)$. This happens if $n\,\tau = i_0 \in \mathbb{N}$ and $X_{(i_0)} < X_{(i_0+1)}$. Then $M(\theta)$ is minimised by any value from the interval $\left[X_{(i_0)}, X_{(i_0+1)}\right]$. In this situation $\widehat{F}_n^{-1}(\tau) = X_{(i_0)}$ is the left point of this interval.

5.2 Regression quantiles*

Suppose that one observes independent and identically distributed random vectors

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

being distributed as the generic vector $\binom{\boldsymbol{X}}{Y}$.

The τ-th regression quantile is defined as

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

At the population level the regression quantile identifies the parameter

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

Note that thanks to (64)

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

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 $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}_0$, then $\boldsymbol{\beta}_X(\tau) = \boldsymbol{\beta}_0$.

Often in applications we assume that $\boldsymbol{X}_i = \big(1, \widetilde{\boldsymbol{X}}_i^\mathsf{T}\big)^\mathsf{T}$ and that

$$Y = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + \varepsilon, \text{ where } \quad \varepsilon \perp \widetilde{\boldsymbol{X}}. \tag{65}$$

Then $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + F^{-1}_{\varepsilon}(\tau)$, where $F^{-1}_{\varepsilon}(\tau)$ is the τ-th quantile of the random error ε. Thus provided model (65) holds

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

* *Regresní kvantily*

Thus if model (59) holds, then for $\tau_1 \neq \tau_2$ the regression quantiles $\boldsymbol{\beta}_X(\tau_1)$ and $\boldsymbol{\beta}_X(\tau_2)$ differ only in the intercepts. That is the effect of the covariate is the same for all quantiles of the response. But this is not true in general. In fact the regression quantiles are interesting in situations where the effect of the covariate can be different for different quantiles of the response.

As also illustrated by the following simple examples, the regression quantiles gives us a more detailed idea about the effect of the covariate on the response. This can be interest on its own or as a check that we do not simplify the situation too much by considering only the effect of the covariate on the conditional expectation.

Example 41. To illustrate consider one-dimensional covariate X_i which is generated from the uniform distribution on the interval $(0, 1)$ and the error term ε_i which has an exponential distribution with mean 1 and which is independent of X_i. Further consider the following two models

- *The homoscedastic model* given by

$$Y_i = 1 + 2\,X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

- *The heteroscedastic model* given by

$$Y_i = 1 + 2\,X_i + 2\,X_i\,\varepsilon_i, \quad i = 1, \ldots, n.$$

On Figure 41 one can find a random sample of size 1 000 from these models. The solid lines represent the fitted regression quantiles for $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ assuming that the conditional quantile is in the simple linear form

$$F^{-1}_{Y|X}(\tau) = \beta_1(\tau) + \beta_2(\tau)\,X.$$

The standard least square estimator is included for the reason of comparison.

Note that in the homoscedastic model all the fitted lines are approximately parallel. This is in agreement with the above finding that in the 'strict linear model' (59) the slope of the (theoretical) regression quantiles is the same (up to the random variations that decreases as the sample size increases).

On the other hand in the heteroscedastic model the slopes differ and in this simple example we see that the effect of the covariate is stronger on larger conditional quantiles.

Homework exercise. In the homoscedastic as well as heteroscedastic model find the theoretical conditional quantile $F^{-1}_{Y|X}(\tau)$ for different values of τ and compare it with the conditional expectation $\mathsf{E}\,[Y|X]$. Compare the results with the fitted lines on Figure 41.

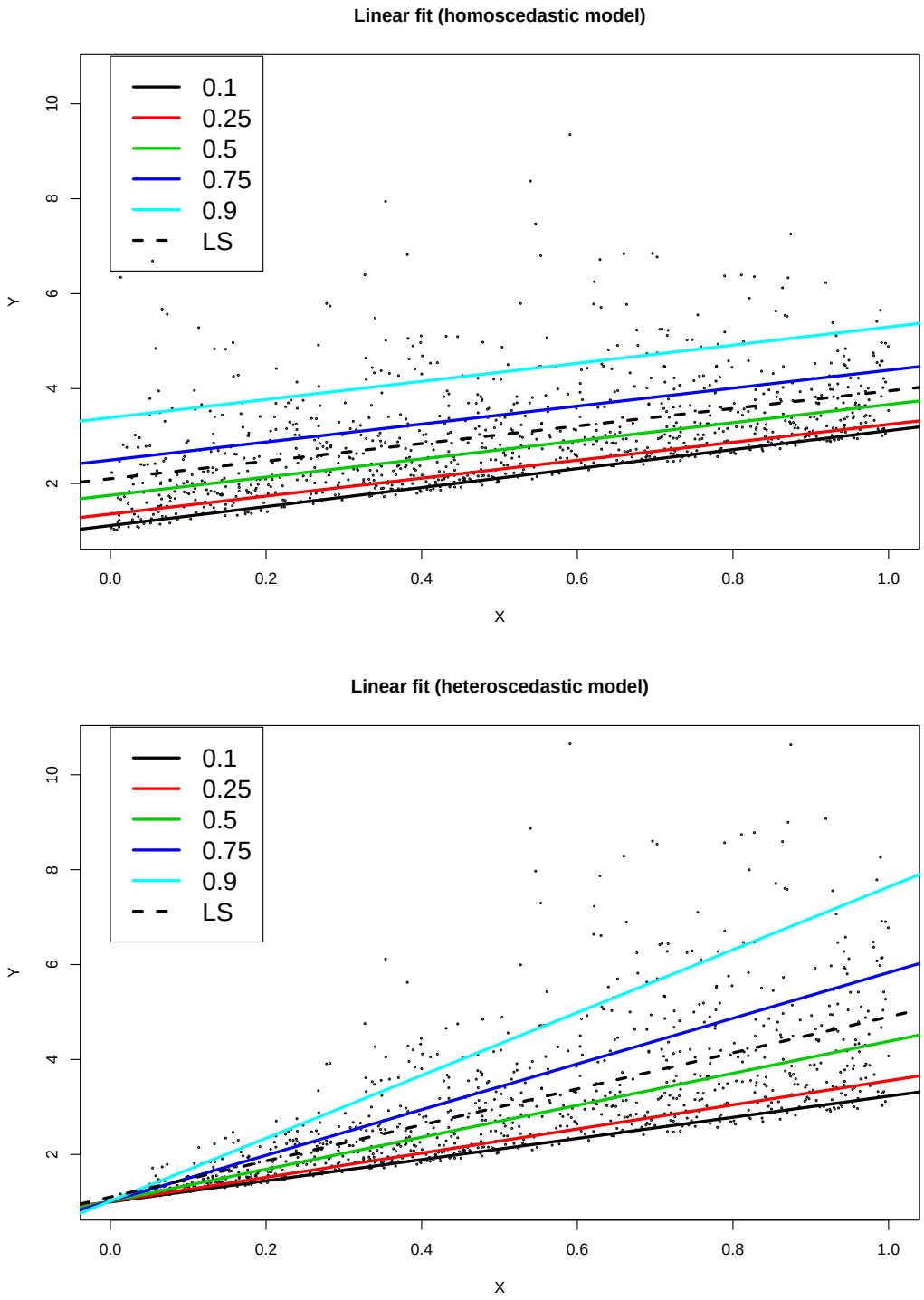

Figure 1: Fitted regression quantiles for $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ (solid lines with different colours) for homoscedastic model (the upper figure) and heteroscedastic model (the lower figure). The least squares fit is included for the reason of comparison (dashed line).

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

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{66}$$

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

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

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

On the other hand let $n = n_1 + n_2$. Then the quantile regression yields

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

The first sum is minimised by

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

and the second sum by

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

Thus we get

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

Further if model (66) really holds, then $G^{-1}(\tau) = F^{-1}(\tau) + \mu$ and one gets $\beta_1(\tau) = \mu = \beta_1^{LS}$ for each $\tau \in (0, 1)$.

Computing regression quantiles*

The optimisation task

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

* Not done at the lecture.

can be rewritten with the help of linear programming as minimisation of the objective function

$$\tau \sum_{i=1}^{n} r_i^+ + (1-\tau) \sum_{i=1}^{n} r_i^-,$$

subject to the following constrains

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

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

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

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

5.3 Interpretation of the regression quantiles

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

Intersection of the fitted regression quantiles

Note that it might happen that for a given value of the covariate $\mathbf{x}$ and given quantiles $0 < \tau_1 < \tau_2 < 1$

$$\widehat{F}^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_1) = \mathbf{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau_1) > \mathbf{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau_2) = \widehat{F}^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_2). \tag{67}$$

which is rather strange as we know that the theoretical quantiles for $\tau_1 < \tau_2$ must satisfy

$$F^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_1) \leq F^{-1}_{Y|\boldsymbol{X}=\mathbf{x}}(\tau_2).$$

Thus if one gets the inequality (67) (we also say that the regression quantiles cross) for $\mathbf{x}$ from the support of the covariate, it might indicate the the assumed linear model for the conditional quantile is not correct.

Transformed response

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

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

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

Thus from the practical point of view even if $\widehat{\boldsymbol{\beta}}_n(\tau)$ is estimated from the response-transformed data $\binom{\boldsymbol{X}_1}{h(Y_1)}, \ldots, \binom{\boldsymbol{X}_n}{h(Y_n)}$, one can still estimate the conditional quantile of the original (not transformed) data $\widehat{F}^{-1}_{Y|\boldsymbol{X}}(\tau) = h^{-1}\big(\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n(\tau)\big)$ (for h increasing). On the other hand if we estimate the conditional expectation of $\mathsf{E}\,[h(Y)|\boldsymbol{X}]$ as $\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n$, there is no general way how to use $\widehat{\boldsymbol{\beta}}_n$ to get an estimate of $\mathsf{E}\,[Y|\boldsymbol{X}]$.

A very common and popular transformation is log-transformation, i.e. $h(y) = \log y$. This results in $F^{-1}_{Y|\boldsymbol{X}}(\tau) = \mathrm{e}^{\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}(\tau)}$ and $\mathrm{e}^{\beta_k(\tau)}$ measures how many times the conditional quantile $F^{-1}_{Y|\boldsymbol{X}}(\tau)$ changes when the k-th coordinate of the covariate is increased by adding one.

5.4 Inference for regression quantiles

Analogously as in Chapter 4.3.2 one can heuristically derive that under appropriate regularity assumption for fixed $\tau \in (0,1)$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}_X(\tau)\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}),$$

where

$$\mathbb{V} = \Big(\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\big]\Big)^{-1} \tau(1-\tau)\,\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big(\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\big]\Big)^{-1}. \quad (68)$$

Note that if model (59) holds, then $F_{Y|\boldsymbol{X}} = \beta_0 + \widetilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{\beta} + F_\varepsilon^{-1}(\tau)$, where F_ε^{-1} is the quantile function of ε_1 and thus

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

Further if model (59) holds then

$$f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big) = f_\varepsilon\big(F_\varepsilon^{-1}(\tau)\big),$$

which implies that

$$\mathsf{E}\,\big[\boldsymbol{X}\boldsymbol{X}^\mathsf{T} f_{Y|\boldsymbol{X}}\big(F^{-1}_{Y|\boldsymbol{X}}(\tau)\big)\big] = f_\varepsilon\big(F_\varepsilon^{-1}(\tau)\big)\,\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$$

and

$$\mathbb{V} = \frac{\tau(1-\tau)}{\big[f_\varepsilon\big(F_\varepsilon^{-1}(\tau)\big)\big]^2}\left(\mathsf{E}\,\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}. \quad (69)$$

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

Note that in general the asymptotic variance matrix (68) of $\widehat{\boldsymbol{\beta}}_n(\tau)$ is rather complicated and it is not clear how to estimate it. That is why nonparametric bootstrap is of interest.

If model (59) holds, then the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_n(\tau)$ simplifies considerably and one gets

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

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) it is suggested 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 a sequence going to zero as $n \to \infty$. A possible choice of h_n (derived when assuming normal errors $\varepsilon_1, \ldots, \varepsilon_n$) is given by

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

where φ is the density of $\mathsf{N}(0,1)$. For details and other possible choices of h_n see Chapter 4.10.1 in Koenker (2005) and the references therein.

As estimating $\frac{1}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)}$ is rather delicate, also in this situation the nonparametric bootstrap (see Chapter 8.2 below) is of interest.

5.5 Asymptotic normality of sample quantiles*

Suppose that we have a random sample $X_1, \ldots, X_n$, where X_1 has a cumulative distribution function F. Note that for a given $\tau \in (0,1)$ thanks to Remark 16 one can view the sample quantile $\widehat{F}_n^{-1}(\tau)$ as the argument of minimum of a convex function. Thus analogously as in Chapter 3.4.1 one can derive that if $f(x)$ (the density of X_1) is positive and continuous in a neighbourhood of $F^{-1}(\tau)$, then

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

Literature: Koenker (2005), Sections 2.1, 2.4, 4.2. 4.10.

The end of the self study for the week (30. 3. - 3. 4. 2020)

* Not done at the lecture. It is assumed that it is known from the bachelor degree.

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

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

where $f_1, \ldots, f_G$ are known densities and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^{\mathsf{T}}$ is a vector of unknown non-negative *mixing proportions* such that $\sum_{j=1}^{G} \pi_j = 1$. Find the maximum likelihood estimator of the parameter $\boldsymbol{\pi}$, i.e.

$$\widehat{\boldsymbol{\pi}}_n = \arg\max_{\boldsymbol{\pi} \in \Theta} \left(\prod_{i=1}^{n} f(X_i; \boldsymbol{\pi}) \right),$$

where $\Theta = \left\{ (\pi_1, \ldots, \pi_G)^{\mathsf{T}} \, : \, \pi_j \in [0,1], \ \sum_{j=1}^{G} \pi_j = 1 \right\}$.

Solution. A straightforward approach would be to maximize the log-likelihood

$$\ell_n(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{G} \pi_j f_j(X_i) \right).$$

Using for instance the parametrization $\pi_G = 1 - \sum_{j=1}^{G-1} \pi_j$, the system of score equations is given by

$$U_{jn}(\boldsymbol{\pi}) = \frac{\partial \ell_n(\boldsymbol{\pi})}{\partial \pi_j} = \sum_{i=1}^{n} \left[\frac{f_j(X_i)}{\sum_{l=1}^{G} \pi_l f_l(X_i)} - \frac{f_G(X_i)}{\sum_{l=1}^{G} \pi_l \, f_l(X_i)} \right] \stackrel{!}{=} 0, \quad j = 1, \ldots, G-1,$$

which requires some numerical routines.

Alternatively one can use the EM-algorithm, which runs as follows. Introduce $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^{\mathsf{T}} \sim \mathsf{Mult}_G(1; \boldsymbol{\pi})$, where

$$Z_{ij} = \begin{cases} 1, & X_i \text{ is generated from } f_j(x), \\ 0, & \text{otherwise.} \end{cases}$$

Note that one can think of our data as the realizations of the independent and identically distributed random vectors $\binom{X_1}{\mathbf{Z}_1}^{\mathsf{T}}, \ldots, \binom{X_n}{\mathbf{Z}_n}^{\mathsf{T}}$, where $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are missing.

Put $\mathbb{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$. The joint density of a random vector $\binom{X_i}{\mathbf{Z}_i}$ is given by

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

In the context of EM algorithm the random sample $\binom{X_1}{\mathbf{Z}_1}^{\mathsf{T}}, \ldots, \binom{X_n}{\mathbf{Z}_n}^{\mathsf{T}}$ is called *complete data*. The corresponding log-likehood is called *complete log-likehood* and it 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 the following two steps:

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

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

where the expectation is taken with respect to the unobserved random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. More precisely one has to take the expectation with respect to the conditional distribution of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ given $X_1, \ldots, X_n$. As this distribution depends on the unknown parameter $\boldsymbol{\pi}$, this parameter is replaced with the current version of the estimate $\widehat{\boldsymbol{\pi}}^{(k)}$. This is indicated by $\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}$. Note that in this step one gets rid of the unobserved random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$.

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

$$\widehat{\boldsymbol{\pi}}^{(k+1)} = \operatorname*{arg\,max}_{\boldsymbol{\pi}\in\Theta} Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big).$$

E-step in a detail:

$$Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^{n}\log\left(\sum_{j=1}^{G} Z_{ij} f_j(X_i)\right)\,\middle|\,\mathbb{X}\right] + \mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\log\pi_j\,\middle|\,\mathbb{X}\right]. \quad (70)$$

Note that the first term on the right-hand side of the above equation does not depend on $\boldsymbol{\pi}$. Thus we do not need to calculate this term for M-step. To calculate the second term it is sufficient to calculate $\mathsf{E}_{\widehat{\boldsymbol{\pi}}^{(k)}}\left[Z_{ij} \,|\, \mathbb{X}\right]$. To do that denote $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathsf{T}}$ for the j-th canonical vector. Now with the help of Bayes theorem for densities (Theorem A4) one can calculate

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

M-step in a detail: Note that with the help of the previous step and (70)

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

Analogously as when calculating the maximum likelihood estimator in a multinomial distribution one can show that the updated value of the estimate of $\boldsymbol{\pi}$ is given by

$$\widehat{\boldsymbol{\pi}}^{(k+1)} = \arg\max_{\boldsymbol{\pi}\in\Theta} Q\big(\boldsymbol{\pi}, \widehat{\boldsymbol{\pi}}^{(k)}\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^{(k)},$$

where $\mathbf{z}_i^{(k)} = \big(z_{i1}^{(k)}, \ldots, z_{iG}^{(k)}\big)^{\mathsf{T}}$ and so $\widehat{\pi}_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(k)}$ for $j \in \{1, \ldots, G\}$.

6.1 General description of the EM-algorithm

Denote the observed random variables as $\mathbb{Y}_{obs}$ and the unobserved (missing) random variables $\mathbb{Y}_{mis}$. Let $f(\mathbf{y}; \boldsymbol{\theta})$ be the joint density (with respect to a σ-finite measure μ) of $\mathbb{Y} = (\mathbb{Y}_{obs}, \mathbb{Y}_{mis})$ and denote $\ell_n^C(\boldsymbol{\theta})$ the *complete log-likelihood* of $\mathbb{Y}$. Our task is to maximize the *observed log-likelihood* $\ell_{obs}(\boldsymbol{\theta}) = \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta})$, where $f(\mathbf{y}_{obs}; \boldsymbol{\theta})$ is the density of $\mathbb{Y}_{obs}$. Note that

$$\begin{aligned}\ell_n^C(\boldsymbol{\theta}) &= \log f(\mathbb{Y}_{obs}, \mathbb{Y}_{mis}; \boldsymbol{\theta}) = \log\big(f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta})\, f(\mathbb{Y}_{obs}; \boldsymbol{\theta})\big)\\ &= \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta}) = \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \ell_{obs}(\boldsymbol{\theta}),\end{aligned}$$

where $f(\mathbf{y}_{mis}|\mathbf{y}_{obs}; \boldsymbol{\theta})$ stands for the conditional density of $\mathbb{Y}_{mis}$ given $\mathbb{Y}_{obs} = \mathbf{y}_{obs}$. Thus one can express observed log-likelihood with the help of complete log-likelihood as

$$\ell_{obs}(\boldsymbol{\theta}) = \ell_n^C(\boldsymbol{\theta}) - \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs}; \boldsymbol{\theta}). \tag{71}$$

Finally denote

$$Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \mathsf{E}_{\widetilde{\boldsymbol{\theta}}}\left[\ell_n^C(\boldsymbol{\theta}) \,|\, \mathbb{Y}_{obs}\right]. \tag{72}$$

EM-algorithm runs as follows:

Let $\widehat{\boldsymbol{\theta}}^{(k)}$ be the result of the k-th iteration of the EM-algorithm. The next iteration $\widehat{\boldsymbol{\theta}}^{(k+1)}$ is computed in two steps:

E-step: Calculate $Q\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\big)$.

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

Note that at this moment it is not at all clear, if the EM-algorithm is a good idea. Remember that our task is to maximize the observed likelihood. The following theorem is the first answer in this aspect.

Theorem 11. *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*

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

Proof. Note that the left-hand side of (71) does not depend on $\mathbb{Y}_{mis}$. Thus applying $\mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}[\,\cdot\,|\,\mathbb{Y}_{obs}]$ on both sides of (71) yields that

$$\begin{aligned}\ell_{obs}(\boldsymbol{\theta}) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}[\ell_n^C(\boldsymbol{\theta})\,|\,\mathbb{Y}_{obs}] - \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})\,\middle|\,\mathbb{Y}_{obs}\right] \\ &=: Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) - H\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right). \end{aligned} \tag{73}$$

Now note that

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

Thus to verify $\ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right) \geq \ell_{obs}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$ it is sufficient to show that

$$Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \geq Q\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \quad \text{and also} \quad H\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \leq H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right). \tag{74}$$

Showing *the first inequality* in (74) is easy as from the M-step

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

which implies that $Q\left(\widehat{\boldsymbol{\theta}}^{(k+1)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \geq Q\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right)$ for each $\boldsymbol{\theta}\in\Theta$.

To show *the second inequality* in (74) one gets with the help of Jensen's inequality that for each $\boldsymbol{\theta}\in\Theta$:

$$\begin{aligned} H\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})\,\middle|\,\mathbb{Y}_{obs}\right] \\ &= \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log\left(\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\right)\middle|\,\mathbb{Y}_{obs}\right] + \mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\log f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)|\mathbb{Y}_{obs}\right] \\ &\overset{\text{Jensen}}{\leq} \log\left(\mathsf{E}_{\widehat{\boldsymbol{\theta}}^{(k)}}\left[\frac{f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\middle|\,\mathbb{Y}_{obs}\right]\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \\ &= \log\left(\int \frac{f(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)}\cdot f\left(\mathbf{y}_{mis}|\mathbb{Y}_{obs};\widehat{\boldsymbol{\theta}}^{(k)}\right)\mathrm{d}\mu(\mathbf{y}_{mis})\right) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) \\ &= \log(1) + H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right) = H\left(\widehat{\boldsymbol{\theta}}^{(k)},\widehat{\boldsymbol{\theta}}^{(k)}\right). \end{aligned} \tag{75}$$

□

6.2 Convergence of the EM-algorithm

Although from Theorem 11 we know that EM algorithm increases (more precisely does not decrease) the observed log-likelihood, it is still not clear whether the sequence $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{\infty}$ converges. And if it converges what is the limit.

To answer this question we need to introduce the following regularity assumptions.

- The parameter space Θ is a subset of $\mathbb{R}^p$.
- The set $\Theta_0 = \left\{\boldsymbol{\theta} \in \Theta : \ell_{obs}(\boldsymbol{\theta}) \geq \ell_{obs}(\boldsymbol{\theta}_0)\right\}$ 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 12. *Let the function $Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ defined in (72) be continuous both in $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$. Then all the limit 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 12 the EM-algorithm finds a solution of the system of log-likelihood equations but in generally there is no guarantee that this is a global maximum of $\ell_{obs}(\boldsymbol{\theta})$.

Corollary 2. *Let the assumptions of Theorem 12 be satisfied. Further suppose that the function $\ell_{obs}(\boldsymbol{\theta})$ has a unique maximum $\widehat{\boldsymbol{\theta}}_n$ that is the only stationary point. Then $\widehat{\boldsymbol{\theta}}^{(k)} \to \widehat{\boldsymbol{\theta}}_n$ as $k \to \infty$.*

6.3 Rate of convergence of EM-algorithm

Note that in the M-step of the algorithm there might not be a unique value that maximizes $Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)$. Thus denote the set of maximizing points as $\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$, i.e.

$$\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) = \left\{\widetilde{\boldsymbol{\theta}} : Q\left(\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^{(k)}\right) = \max_{\boldsymbol{\theta}\in\Theta} Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)\right\}.$$

Then one needs to choose $\widehat{\boldsymbol{\theta}}^{(k+1)}$ as an element of the set $\mathcal{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)$. Thus let $\mathbf{M} : \Theta \to \Theta$ be a mapping such that

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \mathbf{M}\left(\widehat{\boldsymbol{\theta}}^{(k)}\right).$$

Let $\widehat{\boldsymbol{\theta}}^{(k)} \to \boldsymbol{\theta}^*$ as $k \to \infty$. Note that then $\boldsymbol{\theta}^* = \mathbf{M}(\boldsymbol{\theta}^*)$. Assuming that $\mathbf{M}$ is sufficiently smooth one gets by the one term Taylor expansion around the point $\boldsymbol{\theta}^*$ the following approximation

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

Thus

$$\widehat{\boldsymbol{\theta}}^{(k+1)} - \boldsymbol{\theta}^* = \left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \left(\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right) + o\left(\left\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\|\right) \tag{76}$$

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

$$\left.\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = [I_n^C(\boldsymbol{\theta}^*)]^{-1} I_n^{mis}(\boldsymbol{\theta}^*), \tag{77}$$

where

$$I_n^C(\boldsymbol{\theta}) = -\,\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell_n^C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}} \,\middle|\, \mathbb{Y}_{obs}\right]$$

can be considered as the empirical Fisher information matrix from the complete data and

$$I_n^{mis}(\boldsymbol{\theta}) = -\,\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f(\mathbb{Y}_{mis}|\mathbb{Y}_{obs};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}}\middle|\mathbb{Y}_{obs}\right],$$

can be considered as the empirical Fisher information matrix of the contribution of the missing data (that is not explained by the observed data).

Note that by (76) and (77) in the presence of missing data the convergence is only linear. Further the bigger proportion of missing data the 'bigger' $I_n^{mis}(\boldsymbol{\theta})$ and the slower is the convergence.

6.4 The EM algorithm in exponential families

Let the complete data $\mathbb{Y}$ have a density with respect to a σ-finite measure μ given by

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbf{y})\right\} b(\boldsymbol{\theta})\, c(\mathbf{y}) \tag{78}$$

and the standard choice of the parametric space is

$$\Theta = \left\{\boldsymbol{\theta} : \int \exp\left\{\sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbf{y})\right\} c(\mathbf{y})\, \mathrm{d}\mu(\mathbf{y}) < \infty\right\}.$$

Note that $\mathbf{T}(\mathbb{Y}) = \big(T_1(\mathbb{Y}), \ldots, T_p(\mathbb{Y})\big)^{\mathsf{T}}$ is a sufficient statistic for $\boldsymbol{\theta}$.

The log-likelihood of the complete data is now given by

$$\ell_n^C(\boldsymbol{\theta}) = \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, T_j(\mathbb{Y}) + \log b(\boldsymbol{\theta}) + const.,$$

which yields that the function Q from the EM-algorithm is given by

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

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

The nice thing about exponential families is that in the E-step of the algorithm we do not need to calculate $Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(k)}\right)$ for each $\boldsymbol{\theta}$ separately but it is sufficient to calculate

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

and in the M-step we maximize

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{j=1}^{p} a_j(\boldsymbol{\theta})\, \widehat{T}_j^{(k)} + \log b(\boldsymbol{\theta}) \right\}. \tag{79}$$

Interval censoring

Let $-\infty = d_0 < d_1 < \ldots < d_M = \infty$ be a division of $\mathbb{R}$. Further let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables whose exact values are not observed. Instead of each Y_i we only know that $Y_i \in (d_{q_i - 1}, d_{q_i}]$, for some $q_i \in \{1, \ldots, M\}$. Thus we observed independent and identically distributed random variables $X_1, \ldots, X_n$ such that $X_i = q_i$ if $Y_i \in (d_{q_i - 1}, d_{q_i}]$.

Suppose now that Y_i has a density $f(y; \boldsymbol{\theta})$ of the form

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

Thus the joint density of the random sample $Y_1, \ldots, Y_n$ is of the form (78) 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

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

and the M-step is given by (79) 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 (79) one gets that

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

6.5 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.

The end of the self study for the week (6.4.-10.4.)

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 available data are given by

$$(\mathbb{Y}_{obs}, \mathbf{R}_1, \ldots, \mathbf{R}_I) = (\mathbb{Y}_{obs}, \boldsymbol{R}),$$

* *Chybějící data*

where $\boldsymbol{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_I)$. Note that the complete data can be represented as

$$(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_I, \boldsymbol{R}) = (\mathbb{Y}_{obs}, \mathbb{Y}_{mis}\, \boldsymbol{R}) =: (\mathbb{Y}, \boldsymbol{R}).$$

Suppose that the distribution of $\mathbb{Y}$ depends on a parameter $\boldsymbol{\theta}$ (which we are interested in) and the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ depends on $\boldsymbol{\psi}$. Then the joint density of the complete data can be written as

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

Now integrating the above density with respect to $\mathbf{y}_{mis}$ yields the density of the available data

$$f(\mathbf{y}_{obs}, \mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\theta})\, f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\psi})\, \mathrm{d}\mu(\mathbf{y}_{mis}). \tag{80}$$

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

7.1 Basic concepts for the mechanism of missing

Depending on what can be assumed about the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ we distinguish three situations.

Missing completely at random (MCAR). Suppose that $\boldsymbol{R}$ is independent of $\mathbb{Y}$, thus one can write $f(\mathbf{r}|\mathbf{y}; \boldsymbol{\psi}) = f(\mathbf{r}; \boldsymbol{\psi})$ and with the help of (80) 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. $p_i = \frac{1}{1+i}$).

Missing at random (MAR). Suppose that the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ is the same as the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}_{obs}$. Thus one can write $f(\mathbf{r}|\mathbf{y}; \boldsymbol{\psi}) = f(\mathbf{r}|\mathbf{y}_{obs}; \boldsymbol{\psi})$ and with the help of (80)

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

which further implies that the observed log-likelihood is of the form

$$\ell_{obs}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \log f(\mathbb{Y}_{obs}; \boldsymbol{\theta}) + \log f(\boldsymbol{R}|\mathbb{Y}_{obs}; \boldsymbol{\psi}).$$

Note that although MAR is not so strict in assumptions as MCAR, also here the second term on the right-hand side of the above equation does not depend on $\boldsymbol{\theta}$ provided that $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable.

Example 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 stand for the indicator of missing of Y_i and

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

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

Missing not at random (MNAR). In this concept neither the distribution of $\boldsymbol{R}$ is independent of $\mathbb{Y}$ nor the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}_{obs}$ is independent of $\mathbb{Y}_{mis}$. Thus the density of the observed data is generally given by (80). To proceed one has to make some other assumptions about the conditional distribution of $\boldsymbol{R}$ given $\mathbb{Y}$ (i.e. about the density $f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}; \boldsymbol{\psi})$).

Example 48. *Maximum likelihood estimator for the right-censored data from an exponential distribution.* Suppose that $Y_1, \ldots, Y_n$ is a random sample from the exponential distribution with the density $f(x; \lambda) = \lambda \mathrm{e}^{-\lambda x} \mathbb{I}\{x > 0\}$. Nevertheless we observe Y_i only if $Y_i \leq C$, where C is a known constant (e.g. duration of the study). If $Y_i > C$ then we do not know observe the value of Y_i (we only now that Y_i is greater than C).

Note that

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}; \lambda) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda y_i}$$

and

$$f(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{mis}) = \prod_{i=1}^{n} \big[\mathbb{I}\{y_i \leq C\}\big]^{r_i} \big[\mathbb{I}\{y_i > C\}\big]^{1-r_i}.$$

Although this conditionally density depends on $\mathbf{y}_{mis}$ (thus the we are in a situation of MNAR), we can proceed because this conditionally density is completely known.

Let n_0 be the number of fully observed Y_i (i.e. $n_0 = \sum_{i=1}^{n} \mathbb{I}\{Y_i \leq C\}$). For simplicity of notation assume that $Y_1, \ldots, Y_n$ are ordered in such a way that $Y_1, \ldots, Y_{n_0}$ are fully observed and $Y_{n_0+1}, \ldots, Y_n$ are censored (i.e. $Y_i > C$ for $i \in \{n_0 + 1, \ldots, n\}$). Thus the corresponding components of $\boldsymbol{R}$ are given by $R_i = 1$ for $i \in \{1, \ldots, n_0\}$ and zero otherwise.

Now with the help of (80) one can calculate

$$f(\boldsymbol{Y}_{obs}, \mathbf{R}; \lambda) = \prod_{i=1}^{n_0} \lambda \mathrm{e}^{-\lambda Y_i} \int_C^\infty \cdots \int_C^\infty \prod_{i=n_0+1}^{n} \lambda \, \mathrm{e}^{-\lambda y_i} \mathrm{d}y_{n_0+1}, \ldots, \mathrm{d}y_n$$
$$= \lambda^{n_0} \mathrm{e}^{-\lambda \sum_{i=1}^{n_0} Y_i} \left[\mathrm{e}^{-\lambda C}\right]^{n-n_0}.$$

The corresponding log-likelihood of the observed data is

$$\ell_{obs}(\lambda) = n_0 \log \lambda - \lambda \sum_{i=1}^{n_0} Y_i - (n - n_0) C \lambda,$$

which is maximised at

$$\widehat{\lambda}_n = \frac{1}{\frac{1}{n_0} \sum_{i=1}^{n_0} Y_i + \frac{(n-n_0)C}{n_0}}.$$

Note that the above example is in fact rather exceptional as the missing mechanism is given by the design of the study and thus known.

The general problem of all the concepts is that if missing is not a part of the design of the study then no assumptions about the relationship of $\mathbb{Y}_{mis}$ and $\boldsymbol{R}$ can be verified as we do not observe $\mathbb{Y}_{mis}$.

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 59 % ($0.9^5 \doteq 0.59$) of the records will be complete.

Available case analysis (ACA)

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

Example 50. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from $\mathsf{N}((\mu_1, \mu_2, \mu_3)^\mathsf{T}, \mathbb{\Sigma}_{3\times 3})$. Then the covariance $\sigma_{ij} = \mathsf{cov}(X_{1i}, X_{1j})$ is estimated from all the vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ for which both the i-th and the j-th coordinate is observed.

Advantages and disadvantages:

+ simplicity;
+ more data can be used than with CCA;
− the inference about $\boldsymbol{\theta}$ is biased, if MCAR does not hold;
− it can result in estimates with strange features (e.g. there is no guarantee that the estimate of the variance matrix $\widehat{\mathbb{\Sigma}}$ in Example 50 is positive semidefinite).

Direct (ignorable) observed likelihood

The inference is based on $\log f(\mathbb{Y}_{obs}; \boldsymbol{\theta})$, that is the distribution of $\boldsymbol{R}$ is 'ignored'.
Advantages and disadvantages:

+ If the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable then this method is not biased and does not lose any information provided MAR holds;
− The observed log-likelihood $\ell_{obs}(\boldsymbol{\theta})$ might be difficult to calculate. Nevertheless, sometimes the EM algorithm can be helpful.

Imputation

In this method the missing observations are estimated ('imputed') and then one works with the data as if there were no missing values.
Advantages and disadvantages:

+ If the missing values are estimated appropriately, it can give 'reasonable' estimates of the unknown parameters;
+ One can use the *completed* dataset also for other analyses;
− The standard estimates of the (asymptotic) variances of the estimates of the parameters computed from the completed dataset are too optimistic (too low). The reason is that an appropriate estimate of variance should reflect that part of the data has been imputed.

Example 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 $i = n_0 + 1, \dots, n$ let the missing observations be estimated as $\widehat{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_j$. Then the standard estimate of $\mu = \mathsf{E}\, X_1$ is given by

$$\widehat{\mu}_n = \frac{1}{n} \left(\sum_{i=1}^{n_0} X_i + \sum_{i=n_0+1}^{n} \widehat{X}_i \right) = \frac{1}{n_0} \sum_{j=1}^{n_0} X_j$$

and seems to be reasonable.

But the standard estimate of the variance of $\widehat{\mu}_n$ computed from the completed dataset

$$\widehat{\mathsf{var}(\widehat{\mu}_n)} = \frac{S_n^2}{n}, \quad \text{where} \quad S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n_0} (X_i - \widehat{\mu}_n)^2 + \sum_{i=n_0+1}^{n} (X_i - \widehat{\mu}_n)^2 \right)$$

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

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

The second reason is that the factor $\frac{1}{n}$ assumes that there are n independent observations, but in fact there are only n_0 independent observations.

Multiple imputation

In this method the missing observations are imputed several times. Formally, for $j = 1, \dots, M$ let $\widehat{\mathbb{Y}}_{mis}^{(j)}$ be the imputed values in the j-th round. Further let $\widehat{\boldsymbol{\theta}}_j$ be the estimate of the parameter $\boldsymbol{\theta}$ from the completed data $\left(\mathbb{Y}_{obs}, \widehat{\mathbb{Y}}_{mis}^{(j)}\right)$. Then the final estimate of the parameter $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}}_{MI} = \frac{1}{M} \sum_{j=1}^{M} \widehat{\boldsymbol{\theta}}_j.$$

The advantage of this method is that one can also estimate the (asymptotic) variance of this estimator by

$$\widehat{\mathsf{var}(\widehat{\boldsymbol{\theta}}_{MI})} = \overline{\mathbb{V}}_M + \left(1 + \tfrac{1}{M}\right) \mathbb{B}_M, \tag{81}$$

where

$$\overline{\mathbb{V}}_M = \frac{1}{M} \sum_{j=1}^{M} \widehat{\mathbb{V}}_j \quad \text{and} \quad \mathbb{B}_M = \frac{1}{M-1} \sum_{j=1}^{M} \left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_{MI}\right) \left(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{\theta}}_{MI}\right)^{\mathsf{T}},$$

with $\widehat{\mathbb{V}}_j$ being a standard estimate of the asymptotic variance calculated from the completed data $\widehat{\mathbb{Y}}^{(j)} = (\mathbb{Y}_{obs}, \widehat{\mathbb{Y}}_{mis}^{(j)})$.

The rationale of the formula (81) is as follows. Note that

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

Now the first term on right-hand side of the above equation is estimated by $\overline{\mathbb{V}}_M$ and the second term is estimated by $\mathbb{B}_M$.

Example 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 $\widehat{\sigma}^2 = S^2_{n_0}$ 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

$$\lim_{M\to\infty} \overline{V}_M = \frac{S^2_{n_0}}{n} \quad \text{a.s..} \tag{82}$$

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

$$\lim_{M\to\infty} B_M = \frac{S^2_{n_0}(n-n_0)}{n^2} \quad \text{a.s..} \tag{83}$$

Now combining (82) and (83) yields that

$$\lim_{M\to\infty} \overline{V}_M + B_M = S^2_{n_0}\left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right) \quad \text{a.s.}$$

Further it is straightforward to prove that for $n_0 < n$

$$S^2_{n_0}\left(\tfrac{2}{n} - \tfrac{n_0}{n^2}\right) < \frac{S^2_{n_0}}{n_0},$$

where the right-hand side of the above inequality represents the standard estimate of the variance of $\overline{X}_{n_0}$ (that assumes MCAR). This indicates that when doing multiple imputation, one needs to take into consideration also the variability that comes from the fact that one uses the estimates $\widehat{\mu}$, $\widehat{\sigma}^2$ instead of the true values of μ and σ. This can be done very naturally within the framework of Bayesian statistics.

Advantages and disadvantages:

+ If the missing values are estimated appropriately, it can give 'reasonable' estimate of the unknown parameter as well as of the variance of this estimate.

– To be done properly it requires the knowledge of Bayesian approach to statistics.

Re-weighting

Roughly speaking in this method each observation is given a weight (w_i) that is proportional to the inverse probability of being observed (π_i), i.e.

$$w_i = \frac{\frac{1}{\pi_i}}{\sum_{j:R_j=1} \frac{1}{\pi_j}}, \quad i \in \{j : R_j = 1\}.$$

All the procedures are now weighted with respect to these weights, e.g. the M-estimator of a parameter $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i:R_i=1} w_i\, \rho(\boldsymbol{X}_i;\boldsymbol{\theta}).$$

Example 53. Suppose we have a study where for a large number of patients some basic and cheap measurements have been done resulting in $\mathbf{Z}_1,\dots,\mathbf{Z}_N$. Now a random subsample $\mathbb{S}$ of size n from these patients has been done for some more expensive measurements. Note that then $\mathbb{S} = \big\{j \in \{1,\dots,N\} : R_j = 1\big\}$.

This method can be also used where one has some auxiliary variables $\mathbf{Z}_1,\dots,\mathbf{Z}_n$ that can be used to estimate the probabilities π_i with the help of for instance a logistic regression.

The end of the self study for the week (13. 4. - 17. 4. 2020)

8 Bootstrap and other resampling methods

Suppose we observe independent and identically distributed k-dimensional random vectors $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ from the distribution F_X and let $\boldsymbol{\theta}_X = \boldsymbol{\theta}(F_X)$ be the quantity of interest. Let $\mathbf{R}_n = \mathbf{g}(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n;\boldsymbol{\theta}_X)$ be a p-dimensional random vector that we want to use for doing inference about $\boldsymbol{\theta}_X$, e.g.

$$\mathbf{R}_n = \sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big) \quad \text{or} \quad R_n = \big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)^{\mathsf{T}} \Big[\widehat{\mathsf{avar}(\widehat{\boldsymbol{\theta}}_n)}\Big]^{-1} \big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big),$$

where $\widehat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}_X$.

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 $\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big)$, 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 the asymptotic distribution of $\mathbf{R}_n$. For example by Theorem 5 for a MLE estimator in regular models one has $\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}_p, I^{-1}(\boldsymbol{\theta}_X)\big)$. 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'*.

8.1 Monte Carlo principle

Sometimes one knows the distribution of $\boldsymbol{X}_i$ and thus also of $\mathbb{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ so one is (at least theoretically) able to derive the distribution of $\mathbf{R}_n = (R_{n1},\dots,R_{np})^{\mathsf{T}}$. 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 Chapter 2.4 when the null hypothesis holds.

Note that if one knows the distribution of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, then one can generate $\mathbb{X}^*$, which has the same distribution as $\mathbb{X}$. Monte-Carlo principle thus runs as follows. Choose B sufficiently large and for each $b \in \{1, \ldots, B\}$ independently generate the random 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

$$H_n(\mathbf{x}) = \mathsf{P}\left(\mathbf{R}_n \leq \mathbf{x}\right),$$

can now be estimated as

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

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 $\mathbf{R}_n$, the Glivenko-Cantelli Theorem (Theorem A3) implies

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left|\widehat{H}_{n,B}(\mathbf{x}) - H_n(\mathbf{x})\right| \xrightarrow[B \to \infty]{\text{a.s.}} 0. \tag{84}$$

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

Note that to achieve (84) it is not necessary to know the distribution of $\mathbb{X}$ exactly nor that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed. The only thing we need is that **we are able to generate independent copies of $\mathbf{R}_n$.**

Application to hypotheses testing

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 approximated (estimated) by

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

as

$$\widehat{p}_B = \frac{1 + B\left(1 - \widehat{H}_{n,B}(R_{n-})\right)}{B+1} \xrightarrow[B \to \infty]{\text{a.s.}} 1 - H_n(R_{n-}),$$

which is the 'true' (precise) p-value. Note that the quality of the approximation of $\widehat{p}_B$ as an estimate of $1 - H_n(R_{n-})$ depends on B which we can take as large as we want (provided that enough computation time is available).

Example 54. Suppose we observe a random variable with the multinomial distribution $\mathsf{M}_K(n; p_1, \ldots, p_K)$. Denote $\mathbf{p} = \left(p_1, \ldots, p_K\right)^\mathsf{T}$ and $\mathbf{p}_X$ be the true value of the parameter $\mathbf{p}$. In some applications we are interested in testing

$$H_0 : \mathbf{p}_X = \mathbf{p}^{(0)} \quad \text{vs.} \quad H_1 : \mathbf{p}_X \neq \mathbf{p}^{(0)},$$

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 p-value of the χ^2-test of goodness of fit.

Example 55. Note that the significance of all the test statistics introduced in Chapter 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 56. Let $(Y_1, X_1)^\mathsf{T}, \ldots, (Y_n, X_n)^\mathsf{T}$ be independent and identically distributed random vectors from the bivariate normal distribution with the true value of the correlation coefficient denoted as ρ_X. Suppose we are interested in testing the null hypothesis

$$H_0 : \rho_X = \rho_0, \qquad \text{vs.} \qquad H_1 : \rho_X \neq \rho_0. \tag{85}$$

It can be showed that the distribution of the sample correlation coefficient $\widehat{\rho}_n$ depends only on ρ_X. Thus 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. But this distribution would be very complicated.

Suggest how one can generate random variables $(Y_1^*, X_1^*)^\mathsf{T}, \ldots, (Y_n^*, X_n^*)^\mathsf{T}$ such that the distribution of $R_n^* = \sqrt{n}\,(\widehat{\rho}_n^* - \rho_0)$ has under the null distribution the same distribution as R_n. Think how this can be used to calculate (estimate) the p-value of the test of the hypothesis (85).

Example 57. Let $X_1, \ldots, X_n$ be a random sample from the distribution F_X. Show how the Monte Carlo principle can be used to test the following hypotheses

$$H_0 : F_X(x) = F_0(x), \forall x \in \mathbb{R}, \qquad H_1 : \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x).$$

Example 58. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ 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. Show how the Monte Carlo principle can be used to to test the following hypotheses

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

Application to confidence intervals*

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

$$\widehat{H}_{n,B}^{-1}(u) \xrightarrow[B\to\infty]{\text{a.s.}} H_n^{-1}(u),$$

* Not done at the lecture.

provided that H_n is continuous and increasing in u.* Thus one can use the quantiles $H_{n,B}^{-1}(u)$ as estimate (approximation) of the quantiles $H_n^{-1}(u)$.

Let $\widehat{\boldsymbol{\theta}}_n = \big(\widehat{\theta}_{n1}, \ldots, \widehat{\theta}_{np}\big)^\mathsf{T}$ be an estimate of $\boldsymbol{\theta}_X = \big(\theta_{X1}, \ldots, \theta_{Xp}\big)^\mathsf{T}$ and $\mathbf{R}_n = \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X$. Suppose that one is able to generate random variable $\mathbf{R}_n^*$ with the same distribution as $\mathbf{R}_n$. Further suppose that we want to find the confidence interval for θ_{Xk} (the k-th component of $\boldsymbol{\theta}_X$). Denote H_n the distribution function of $\widehat{\theta}_{nk} - \theta_{Xk}$ and $H_{n,B}$ the empirical distribution function of the k-th component of $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$. Now provided that the distribution function H_n is continuous and increasing in $H_n^{-1}(\alpha/2)$ and $H_n^{-1}(1-\alpha/2)$, then one gets

$$\lim_{B\to\infty} \mathsf{P}\left(\widehat{H}_{n,B}^{-1}(\alpha/2) < \widehat{\theta}_{nk} - \theta_{Xk} < \widehat{H}_{n,B}^{-1}(1-\alpha/2)\right) = 1-\alpha.$$

Thus the approximate confidence interval for θ_{Xk} can be calculated as

$$\big(\widehat{\theta}_{nk} - \widehat{H}_{n,B}^{-1}(1-\alpha/2), \widehat{\theta}_{nk} - \widehat{H}_{n,B}^{-1}(\alpha/2)\big).$$

Example 59. Let $X_1, \ldots, X_n$ be a random sample from a distribution F_X such that F_X belongs to a location family, i.e.

$$F_X \in \mathcal{F} = \big\{F(\cdot - \theta), \theta \in \mathbb{R}\big\},$$

where F is a known function and θ is an unknown parameter.

Let θ_X be the true value of the parameter θ (i.e. $F_X(x) = F(x - \theta_X)$, for all $x \in \mathbb{R}$) and $\widehat{\theta}_n$ be its estimator that is location equivariant, i.e.

$$\widehat{\theta}_n(X_1 + c, \ldots, X_n + c) = \widehat{\theta}_n(X_1, \ldots, X_n) + c, \quad \forall c \in \mathbb{R}.$$

Then the distribution of $R_n = \widehat{\theta}_n - \theta_X$ depends only on the known function F but it does not depend on θ_X. Thus the distribution of R_n can be approximated by simulating from the distribution with a given θ_0 (i.e. $\theta_0 = 0$) and calculating $R_{n,b}^* = \widehat{\theta}_n^* - \theta_0$.

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.

8.2 Standard nonparametric bootstrap

Suppose we observe **independent and identically** distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution F_X. Let $\boldsymbol{\theta}(F_X)$ be the quantity of interest and $\widehat{\boldsymbol{\theta}}_n$ be its estimator. For

* In fact it is sufficient to assume that $H_n^{-1}(u)$ is a unique solution of $H_n(x_-) \le u \le H_n(x)$, see e.g. Theorem of Section 2.1.3 in Serfling (1980).

presentation purposes it will be instructive to write the estimator as $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\widehat{F}_n)$, with $\widehat{F}_n$ being the empirical distribution

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

Suppose we are interested in the distribution of a p-dimensional random vector

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

In nonparametric bootstrap* the unknown F_X is estimated by the empirical distribution function $\widehat{F}_n$. Now generating independent random vectors $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ from the distribution $\widehat{F}_n$ is equivalent to drawing a simple random sample with replacement† of size n from the observed values $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, i.e. $\mathsf{P}(\boldsymbol{X}_{i,b}^* = \boldsymbol{X}_j | \mathbb{X}) = \frac{1}{n}$ for each $b = 1, \ldots, B$, $i, j = 1, \ldots, n$ and all the random variables $\big\{\boldsymbol{X}_{i,b}^*;\, i = 1, \ldots, n,\, b = 1, \ldots, B\big\}$ are independent.

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

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

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

$$H_n(\mathbf{x}) = \mathsf{P}\big(\mathbf{R}_n \leq \mathbf{x}\big),$$

is now (by the combination of the MC and plug-in principle) estimated by

$$\widehat{H}_{n,B}^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big\{\mathbf{R}_{n,b}^* \leq \mathbf{x}\big\}. \tag{86}$$

Note that the random variables/vectors $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$ are independent and identically distributed as a generic random vector $\mathbf{R}_n^*$. As $\mathbf{R}_{n,1}^*, \ldots, \mathbf{R}_{n,B}^*$ forms a random sample then by the Glivenko-Cantelli Theorem (Theorem A3)

$$\sup_{\mathbf{x}\in\mathbb{R}^p} \big|\widehat{H}_{n,B}^*(\mathbf{x}) - \widehat{H}_n(\mathbf{x})\big| \xrightarrow[B\to\infty]{\text{a.s.}} 0,$$

where

$$\widehat{H}_n(\mathbf{x}) = \mathsf{P}\Big(\mathbf{R}_n^* \leq \mathbf{x} \,|\, \mathbb{X}\Big) = \mathsf{P}\Big(\mathbf{g}_n\big(\boldsymbol{\theta}(\widehat{F}_n^*), \boldsymbol{\theta}(\widehat{F}_n)\big) \leq \mathbf{x} \,|\, \mathbb{X}\Big) = \mathsf{P}\Big(\mathbf{g}_n\big(\widehat{\boldsymbol{\theta}}_n^*, \widehat{\boldsymbol{\theta}}_n\big) \leq \mathbf{x} \,|\, \mathbb{X}\Big). \tag{87}$$

---

* *neparametrický bootstrap* † *prostý náhodný výběr s vracením*

Note that $\widehat{H}_n$ depends on the random sample $\mathbb{X}$ and thus can be viewed as the estimator of the distribution function H_n. The crucial question for the success of the nonparametric bootstrap is whether $\widehat{H}_n$ is 'close' (at least asymptotically) to H_n. To answer this question it is useful to introduce the supremum metric on the space of distribution functions (of random vectors on $\mathbb{R}^p$) as

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

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 7. *Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and $\boldsymbol{Y}$ be random vectors (with values in $\mathbb{R}^p$) with the corresponding distribution functions $G_1, G_2, \ldots$ 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 $\sup_{\mathbf{y}\in\mathbb{R}^p} |G_n(\mathbf{y}) - G(\mathbf{y})| \to 0$ implies that $G_n(\mathbf{y}) \to G(\mathbf{y})$ for each $\mathbf{y} \in \mathbb{R}^p$.

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

$$G_n(\mathbf{y}) - G(\mathbf{y}) \le G_n(\mathbf{y}_U) - G(\mathbf{y}) \le G_n(\mathbf{y}_U) - G(\mathbf{y}_U) + \tfrac{\varepsilon}{2} \tag{88}$$

and analogously also

$$G_n(\mathbf{y}) - G(\mathbf{y}) \ge G_n(\mathbf{y}_L) - G(\mathbf{y}) \ge G_n(\mathbf{y}_L) - G(\mathbf{y}_L) - \tfrac{\varepsilon}{2}. \tag{89}$$

Now combining (88) and (89) together with $G_n \xrightarrow[n\to\infty]{w} G$ one gets that for all sufficiently large n

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

which implies the statement of the lemma. □

Recall the random vector $\mathbf{R}_n^*$ whose distribution function is given by (87). Note that the distribution of $\mathbf{R}_n^*$ depends on (the realizations of our data) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus the

* This implication not shown at the lecture.

distribution $\mathbf{R}_n^*$ is conditionally on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus in what follows we would like to define the convergence of the conditional distributions.

Let ρ be a metric on the space of distribution functions that can be used for metrizing weak convergence (for instance the supremum metric, but in literature other metrics can be found). Let $\mathbf{R}$ be a candidate for the limiting random vector and H be its distribution function. Recall that the distribution $\widehat{H}_n$ given by (87) depends on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus $\rho(\widehat{H}_n, H)$ is in fact a random variable (also depending on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$).

Now we say that **conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ the random variable $\mathbf{R}_n^*$ converges in distribution** to **$\mathbf{R}$ in probability** if

$$\rho(\widehat{H}_n, H) \xrightarrow[n\to\infty]{P} 0 \quad \Big(\text{i.e. for each } \varepsilon > 0 \lim_{n\to\infty} \mathsf{P}\left[\omega \in \Omega : \rho\big(\widehat{H}_n(\omega), H\big) \geq \varepsilon\right] = 0\Big).$$

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

$$\rho(\widehat{H}_n, H) \xrightarrow[n\to\infty]{\text{a.s.}} 0 \quad \Big(\text{i.e. } \mathsf{P}\left[\omega \in \Omega : \lim_{n\to\infty} \rho\big(\widehat{H}_n(\omega), H\big) = 0\right] = 1\Big).$$

Theorem 13. *Suppose that $\mathbf{R}_n \xrightarrow[n\to\infty]{d} \mathbf{R}$, where $\mathbf{R}$ is a random vector with a **continuous** distribution function H. Further suppose that*

$$\rho_\infty(\widehat{H}_n, H_n) \xrightarrow[n\to\infty]{P} 0 \quad (\textit{or } \xrightarrow[n\to\infty]{a.s.} 0), \tag{90}$$

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

Proof. By the triangle inequality, (90) and Lemma 7

$$\rho_\infty(\widehat{H}_n, H) \leq \rho_\infty(\widehat{H}_n, H_n) + \rho_\infty(H_n, H) \xrightarrow[n\to\infty]{P} 0 \quad (\text{or } \xrightarrow[n\to\infty]{\text{a.s.}} 0).$$

□

Although the proof of the above theorem is simply, there are several things worth noting. It is assumed that $\mathbf{R}_n \xrightarrow[n\to\infty]{d} \mathbf{R}$, where $\mathbf{R}$ is a random vector with a continuous distribution function. This requires that we use bootstrap to approximate a distribution that is asymptotically not degenerate. This is analogous to the the use of normal approximation (to which using bootstrap is an alternative), where we also normalize the random vector so that it asymptotically has a non-degenerate distribution.

Typically we know that $\mathbf{R}_n$ converges to a multivariate normal distribution, thus also the continuity of the limit distribution of $\mathbf{R}$ is satisfied. Thus in view of Theorem 13 the crucial question to answer is if the convergence (90) holds. The first answer in this aspect is the next theorem, which states that (90) holds for a sample mean (for the proof see e.g. Theorem 23.4 of van der Vaart, 2000, pp. 330–331). This initial result will be later generalized in Section 8.2.2.

Theorem 14. *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*

$$\rho_\infty(\widehat{H}_n, H_n) \xrightarrow[n\to\infty]{a.s.} 0. \tag{91}$$

Note that for $\boldsymbol{X}_1$ being a p-variate random vector 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 13 and 14 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 $\widehat{H}_n$ estimates also the distribution function of $\mathsf{N}_p\big(\mathbf{0}_p, \mathsf{var}(\boldsymbol{X}_1)\big)$.

Example 60. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables and we are interested in the expectation $\mathsf{E}\,X_i$. The usually approach to find the confidence interval for $\mathsf{E}\,X_i$ is to make use of the convergence

$$\frac{\sqrt{n}\,(\overline{X}_n - \mathsf{E}\,X_i)}{S_n} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1), \tag{92}$$

which holds provided that $\mathsf{var}(X_i) \in (0, \infty)$.

In view of the theory presented above we want to approximate/estimate the distribution function

$$H_n(x) = \mathsf{P}(R_n \le x), \quad \text{where } R_n = \sqrt{n}\,(\overline{X}_n - \mathsf{E}\,X_i).$$

With the help of (92) the estimate of this distribution based on the normal approximation is

$$\widehat{H}_n^{(norm)}(x) = \Phi(x\,S_n).$$

Alternatively one can use the nonparametric bootstrap resulting in an estimator $\widehat{H}_{n,B}^*$, see (86).

Figure 60 illustrates the normal and the bootstrap approximation (with $B = 10\,000$) for the sample sizes $n = 30$ and $n = 1\,000$ when the true distribution of X_i is exponential $\mathsf{Exp}(1)$. In the plots in the first column one can find the densities of the true distribution of $R_n = \sqrt{n}\,(\overline{X}_n - \mathsf{E}\,X_i)$ (black solid), the normal approximation (blue solid) and limit distribution which is $\mathsf{N}(0,1)$ (dotted). The bootstrap approximation is given by the histogram.

In the plots in the second column one can find the difference of the true distribution function H_n of R_n with its estimates. The difference $H_n(x) - \widehat{H}_n^{(norm)}(x)$ is in blue colour, while the difference $H_n(x) - \widehat{H}_{n,B}(x)$ is in red colour. Note that these differences are much smaller for the bigger sample size. It is also worth noting that none of the approximation is evidently preferable in this example.

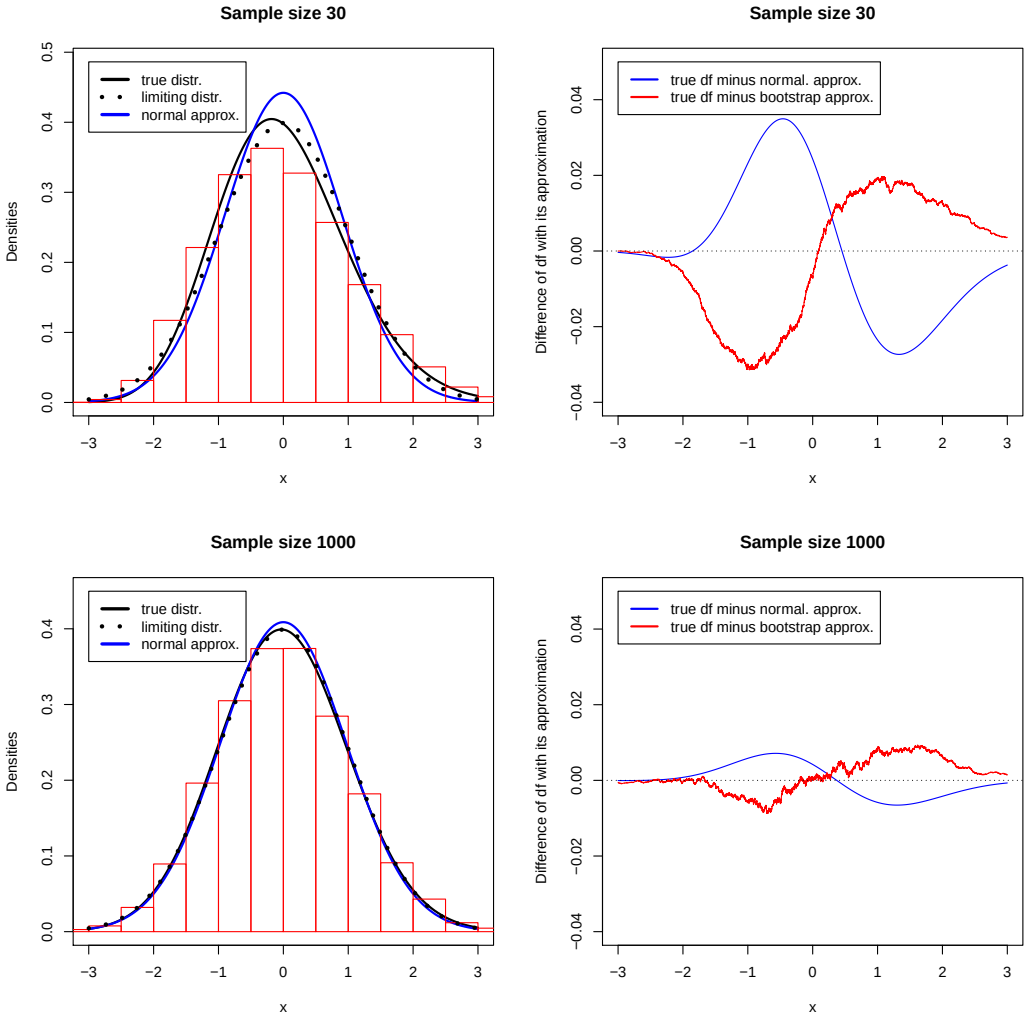

Figure 2: Comparison of the normal and bootstrap approximations of the distribution of the random variable $R_n = \sqrt{n}\,(\overline{X}_n - \mathsf{E}\ X_i)$.

8.2.1 Comparison of nonparametric bootstrap and normal approximation

Note that Theorem 13 implies only the asymptotic validity of bootstrap provided that (90) holds. The question is whether bootstrap estimate $\widehat{H}_n$ 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

$$H_n(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}_n-\mu)}{S_n} \le x\right) = \Phi(x) + \frac{\gamma_1}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O\big(\tfrac{1}{n}\big), \tag{93}$$

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 $\widehat{H}_n(x)$, i.e.

$$\widehat{H}_n(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}^*_n-\overline{X}_n)}{S^*_n} \le x \,\middle|\, \mathbb{X}\right) = \Phi(x) + \frac{\gamma_{1n}}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O_P\big(\tfrac{1}{n}\big), \tag{94}$$

where $\overline{X}^*_n = \frac{1}{n}\sum_{i=1}^n X^*_i$, $S^{2*}_n = \frac{1}{n-1}\sum_{i=1}^n (X^*_i - \overline{X}^*_n)^2$ and $\gamma_{1n} = \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i-\overline{X}_n}{S_n}\right)^3$. Thus comparing (93) and (94) one gets

$$\widehat{H}_n(x) - H_n(x) = O_P\big(\tfrac{1}{n}\big).$$

On the other hand 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 $\widehat{H}_n$ is closer to H_n.

Example 61. We are in the same situation as in Example 60. But instead of approximating/estimating the distribution $\sqrt{n}\,(\overline{X}_n - \mathsf{E}\ X_i)$, we approximate the distribution of its studentized version, i.e.

$$R_n = \frac{\sqrt{n}\,(\overline{X}_n - \mathsf{E}\ X_i)}{S_n}.$$

Note that the normal approximation of the distribution of R_n is simply given by $\widehat{H}_n^{(norm)}(x) = \Phi(x)$. The comparison of the true distribution function with its either normal or bootstrap approximation is found in Figure 61. Similarly as in Example 60 the results are for the random sample from the exponential distribution. Note that in agreement with the theory, the bootstrap approximation is better than the normal approximation.

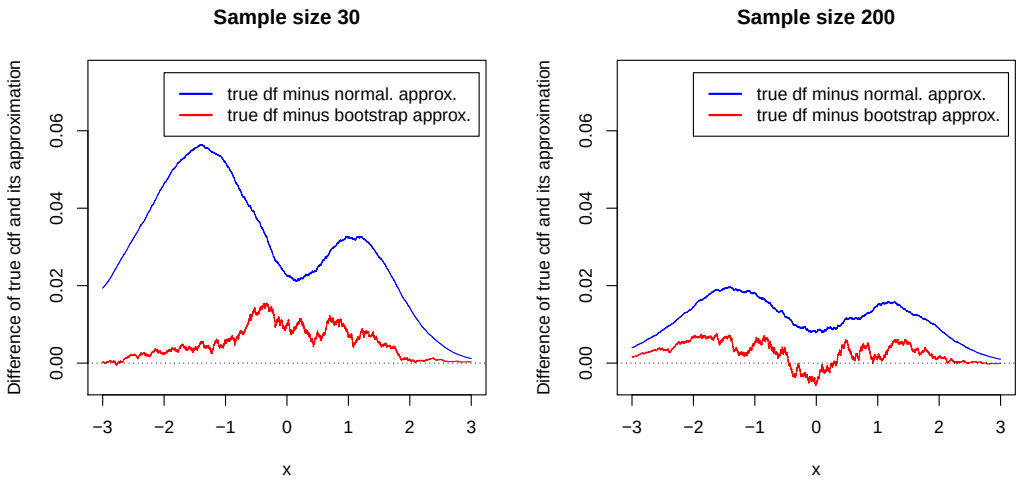

Figure 3: Comparison of the normal and bootstrap approximation of the distribution of the random variable $R_n = \frac{\sqrt{n}\,(\overline{X}_n - \mathsf{E}\, X_i)}{S_n}$.

8.2.2 Smooth transformations of sample means

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

Theorem 15. *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 $\boldsymbol{\mu} = \mathsf{E}\,\boldsymbol{X}_1$ such that the function $\mathbf{g} : U \to \mathbb{R}^m$ has continuous partial derivatives in this neighbourhood. Consider $\mathbf{R}_n = \sqrt{n}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\right)$ and $\mathbf{R}_n^* = \sqrt{n}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n^*) - \mathbf{g}(\overline{\boldsymbol{X}}_n)\right)$. Then (91) holds.*

The above theorem can be of interest for functions of (sample) moments whose asymptotic distribution is difficult to derive (e.g. Pearson's correlation coefficient, skewness, kurtosis,...).

Remark 17. Suppose for simplicity that $g : \mathbb{R}^p \to \mathbb{R}$. Let $\mathbb{\Sigma} = \mathsf{var}(\boldsymbol{X}_1)$. Note that if $\nabla g^{\mathsf{T}}(\boldsymbol{\mu})\mathbb{\Sigma}\,\nabla g(\boldsymbol{\mu}) = 0$, then although (91) 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 $R_n = \sqrt{n}\left(g(\overline{X}_n) - g(\mu_X)\right) \xrightarrow[n\to\infty]{P} 0$. Thus although by Theorem 15 convergence (91) holds, one cannot say if bootstrap works as the limiting distribution is not continuous (i.e. assumptions of Theorem 13 are not satisfied).

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

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

So the bootstrap would work if the convergence (90) holds also for $\widetilde{R}_n^* = 2\,n\,\big(g(\overline{X}_n^*) - g(\overline{X}_n)\big)$, where H_n is now the distribution function of $\widetilde{R}_n$ and $\widehat{H}_n$ is the distribution function of $\widetilde{R}_n^*$. But for this situation (90) does not hold (see Example 3.6 of Shao and Tu, 1996).

Roughly speaking one can say that (91) holds provided that $\widehat{\boldsymbol{\theta}}_n$ satisfies the following asymptotic 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 $IF(\mathbf{x})$ is a given function. This can be formalised through the concept of Hadamard-differentiability of the functional $F \mapsto \boldsymbol{\theta}(F)$ at F_X, but this is out of the scope of this course.

8.2.3 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 assumes 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 62. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta_X)$. Then the maximum likelihood estimator of θ_X is given by $\widehat{\theta}_n = \max_{1\le i\le n} X_i =: X_{(n)}$. Note that for $x < 0$

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

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

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

On the other side

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

and thus (90) 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.

The end of the self study for the week (20. 4. - 24. 4. 2020)

8.3 Confidence intervals

In what follows consider $\mathbf{R}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)$ and suppose that $\mathbf{R}_n \xrightarrow[n\to\infty]{d} \mathbf{R}$, where $\mathbf{R}$ is a random vector with a continuous distribution function. We will be interested in finding the confidence interval for θ_{Xj} (the j-th component of $\boldsymbol{\theta}_X$).

Suppose for a moment the the distribution R_{nj} (the j-th component of $\mathbf{R}_n$) is known and continuous. Then one has

$$\mathsf{P}\left[r_n(\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{nj} - \theta_{Xj}\right) < r_n(1-\alpha/2)\right] = 1-\alpha,$$

where $r_n(\alpha)$ is the α-quantile of R_{nj}. Thus one would get a 'theoretical' confidence interval

$$\left(\widehat{\theta}_{nj} - \tfrac{r_n(1-\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{nj} - \tfrac{r_n(\alpha/2)}{\sqrt{n}}\right). \tag{95}$$

The problem is that the distribution R_{nj} is not known and thus also the quantiles $r_n(\alpha/2)$ and $r_n(1-\alpha/2)$ are not known.

8.3.1 Basic bootstrap confidence interval

Consider $\mathbf{R}^*_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n\right)$ and suppose that the assumptions of Theorem 13 are satisfied. Let $r^*_n(\alpha)$ be the quantile of the bootstrap distribution of $R^*_{nj} = \sqrt{n}\left(\widehat{\theta}^*_{nj} - \widehat{\theta}_{nj}\right)$. Then Theorem 13 implies that $r^*_n(\alpha) \xrightarrow[n\to\infty]{P} r_j(\alpha)$ (or even $r^*_n(\alpha) \xrightarrow[n\to\infty]{\text{a.s.}} r_j(\alpha)$), where $r_j(\alpha)$ is the α-quantile of R_j (the j-th coordinate of the limiting distribution $\mathbf{R}$). Thus one gets

$$\lim_{n\to\infty} \mathsf{P}\left[r^*_n(\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{nj} - \theta_{Xj}\right) < r^*_n(1-\alpha/2)\right] = 1-\alpha. \tag{96}$$

Now with the help of (96) one can construct an asymptotic confidence interval for θ_{Xj} as

$$\left(\widehat{\theta}_{nj} - \tfrac{r^*_{n,B}(1-\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{nj} - \tfrac{r^*_{n,B}(\alpha/2)}{\sqrt{n}}\right), \tag{97}$$

where $r^*_{n,B}(\alpha)$ is a Monte-Carlo approximation (estimate) of $r^*_n(\alpha)$. The confidence interval in (97) is usually called *basic bootstrap confidence interval*.

It is worth noting that the formula for the confidence interval (97) mimics the formula for the theoretical confidence interval (95). The bootstrap idea is to estimate the unknown quantiles $r_n(\alpha)$ with $r^*_n(\alpha)$ that can be calculated only from the observed data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ ('substitution principle'). Further as the quantiles $r^*_n(\alpha)$ are difficult to calculate analytically, one approximates them with $r^*_{n,B}(\alpha)$ ('Monte Carlo principle').

Note that typically

$$\mathbf{R}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}). \tag{98}$$

Then the advantage of the confidence interval given by (96) is that **it does not require to explicitly estimate the asymptotic variance matrix** $\mathbb{V}$. Thus this confidence interval

can be used in situations where deriving or estimating the asymptotic variance of $\mathbf{R}_n$ is rather difficult.

On the other hand the theoretical results stating that the bootstrap confidence interval is more accurate require that the asymptotic distribution of R_{nj} is pivotal (i.e. it does not depend on unknown parameters). If this is not the case, then the simulation studies show that the basic bootstrap confidence interval (97) can be (for finite sample sizes) less accurate than the standard asymptotic confidence interval

$$\left(\widehat{\theta}_{nj} - \frac{u_{1-\alpha/2}\sqrt{v_{n,jj}}}{\sqrt{n}}, \widehat{\theta}_{nj} + \frac{u_{1-\alpha/2}\sqrt{v_{n,jj}}}{\sqrt{n}}\right), \tag{99}$$

where $v_{n,jj}$ is a consistent estimate of the j-th diagonal element of the matrix $\mathbb{V}$. Thus if possible, it is of interest to use R_{nj} which is asymptotically pivotal or at least 'less dependent' on unknown parameters (see Remark 19 below and Chapter 8.3.2).

Example 63. Suppose we observe $\mathbf{Z}_1 = \binom{\boldsymbol{X}_1}{Y_1}, \ldots, \mathbf{Z}_n = \binom{\boldsymbol{X}_n}{Y_n}$ a random sample, where, $\boldsymbol{X}_i$ is a p-dimensional covariate and Y_i is one-dimensional response. In regression models (linear models, generalized linear models, quantile regression models, ...) one aims at estimating $\boldsymbol{\beta}_X$ which specifies how the covariate influences the response. Usually based on theoretical results one can hope that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V})$$

and to find a confidence interval for β_{Xj} (the j-th component of $\boldsymbol{\beta}_X$) one needs to estimate $\mathbb{V}$ (or at least its j-th diagonal element). But this might be rather difficult, see for instance the general asymptotic variance matrix of the least absolute deviation estimator in Section 4.3.2. The bootstrap can thus present an interesting alternative.

Note that in this situation the nonpametric bootstrap corresponds to generating $\mathbf{Z}_1^* = \binom{\boldsymbol{X}_1^*}{Y_1^*}, \ldots, \mathbf{Z}_n^* = \binom{\boldsymbol{X}_n^*}{Y_n^*}$ as a simple random sample with replacement from $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$.

In some textbooks a different formula than (97) can be found. To explain this formula note that $r_{n,B}^*(\alpha)$ is a sample α-quantile of $R_{nj,1}^*, \ldots, R_{nj,B}^*$, where $R_{nj,b}^* = \sqrt{n}\left(\widehat{\theta}_{nj,b}^* - \widehat{\theta}_{nj}\right)$. Further let $q_{n,B}^*(\alpha)$ be a sample α-quantile calculated from the values $\widehat{\theta}_{nj,1}^*, \ldots, \widehat{\theta}_{nj,B}^*$. Then

$$r_{n,B}^*(\alpha) = \sqrt{n}\left(q_{n,B}^*(\alpha) - \widehat{\theta}_{nj}\right) \tag{100}$$

and the confidence interval (97) can be also rewritten as

$$\left(2\,\widehat{\theta}_{nj} - q_{n,B}^*(1-\alpha/2), 2\,\widehat{\theta}_{nj} - q_{n,B}^*(\alpha/2)\right). \tag{101}$$

Thus in practice it is sufficient to calculate $\widehat{\theta}_{nj,b}^*$ instead of $R_{nj,b}^*$ and then use formula (101). On the other hand the approach based on calculating $R_{nj,b}^*$ is more appropriate from the

theoretical point of view. The thing is that to justify the bootstrap one needs (among others) that the limiting distribution R_{nj} has a continuous distribution function (see Theorem 13).

Remark 18. Sometimes in literature one can find a bootstrap confidence interval of the form

$$\Big(q^*_{n,B}(\alpha/2), q^*_{n,B}(1-\alpha/2)\Big), \tag{102}$$

which is usually called the *percentile confidence interval*. Note that with the help of (100) this confidence interval can be rewritten as

$$\Big(\widehat{\theta}_{nj} + \tfrac{r^*_{n,B}(\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{nj} + \tfrac{r^*_{n,B}(1-\alpha/2)}{\sqrt{n}}\Big).$$

Thus when using the percentile confidence interval one hopes that (taking $B = \infty$)

$$\begin{aligned}\lim_{n\to\infty} \mathsf{P}\Big[\Big(\widehat{\theta}_{nj} + \tfrac{r^*_n(\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{nj} + \tfrac{r^*_n(1-\alpha/2)}{\sqrt{n}}\Big) \ni \theta_{Xj}\Big] \\ = \lim_{n\to\infty} \mathsf{P}\Big[-r^*_n(1-\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{nj} - \theta_{Xj}\right) < -r^*_n(\alpha/2)\Big] = 1-\alpha.\end{aligned}$$

Thus the use of the percentile interval can be justified if the limiting distribution of R_{nj} is symmetric, because then

$$r^*_n(1-\alpha/2) \xrightarrow[n\to\infty]{P} r_j(1-\alpha/2) = -r_j(\alpha/2)$$

and analogously $r^*_n(\alpha/2) \xrightarrow[n\to\infty]{P} -r_j(1-\alpha/2)$. As the limiting distribution of $\mathbf{R}_n$ is typically zero mean Gaussian distribution, the assumption of the symmetry of R_j is typically satisfied.

Note that the practical advantage of the percentile confidence is that it is always contained in the parametric space.

Remark 19. Suppose for simplicity that $\theta_X \in \mathbb{R}$. Then using $R_n = \sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$ is natural for location estimators. But sometimes it may be of interest to consider for instance $R_n = \sqrt{n}\left(\frac{\widehat{\theta}_n}{\theta_X} - 1\right)$ or $R_n = \sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta_X)\right)$, where g is a function that stabilises the asymptotic variance (see Chapter 1.4).

8.3.2 Studentized bootstrap confidence interval

Usually it is recommended to 'bootstrap' a variable whose limit distribution is **pivotal** (i.e. does not depend on the unknown parameters).

Suppose that (98) holds and consider $\widetilde{R}_{nj} = \frac{\sqrt{n}\,(\widehat{\theta}_{nj} - \theta_{Xj})}{\sqrt{v_{n,jj}}}$, where $v_{n,jj}$ is a consistent estimate of the j-th diagonal element of $\mathbb{V}$. Let $\tilde{r}^*_n(\alpha)$ be the α-th quantile of the distribution $\widetilde{R}^*_{nj} = \frac{\sqrt{n}\,(\widehat{\theta}^*_{nj} - \widehat{\theta}_{nj})}{\sqrt{v^*_{n,jj}}}$, where $v^*_{n,jj}$ is an estimate of the j-th diagonal element of $\mathbb{V}$ but calculated from the bootstrap sample. Thus if 'bootstrap works' (i.e. Theorem 13 holds), then

$$\lim_{n\to\infty} \mathsf{P}\Big[\tilde{r}^*_n(\alpha/2) < \tfrac{\sqrt{n}\,(\widehat{\theta}_{nj} - \theta_{Xj})}{\sqrt{v_{n,jj}}} < \tilde{r}^*_n(1-\alpha/2)\Big] = 1-\alpha,$$

which yields an asymptotic confidence interval

$$\left(\widehat{\theta}_{nj} - \frac{\tilde{r}^*_{n,B}(1-\alpha/2)\sqrt{v_{n,jj}}}{\sqrt{n}}, \widehat{\theta}_{nj} - \frac{\tilde{r}^*_{n,B}(\alpha/2)\sqrt{v_{n,jj}}}{\sqrt{n}}\right), \tag{103}$$

where $\tilde{r}^*_{n,B}(\alpha)$ is a Monte-Carlo approximation of $\tilde{r}^*_n(\alpha)$. The confidence interval in (103) is usually called the *studentized bootstrap confidence interval*.

Note that in comparison with (99) we replace the quantiles $u_{\alpha/2}$ and $u_{1-\alpha/2}$ with $-\tilde{r}^*_{n,B}(1-\alpha/2)$ and $-\tilde{r}^*_{n,B}(\alpha/2)$. There are theoretical results that state that the studentized confidence interval (103) is (for finite sample sizes) more accurate than the asymptotic confidence interval (99) as well as (97).

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

8.4 Parametric bootstrap

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

Example 64. Suppose we are in a situation of Example 62. Then it is possible to show that if one uses the parametric bootstrap, i.e. if $X^*_{1,b}, \ldots, X^*_{n,b}$ is generated as a random sample from the uniform distribution on $(0, \widehat{\theta}_n)$, then bootstrap works. Note also that it is more natural to resample $R_n = n\left(\frac{\widehat{\theta}_n}{\theta_X} - 1\right)$, as its asymptotic distribution is pivotal.

Example 65. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ 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

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

where $r^*_{n,B}(\alpha)$ is the estimate of the α-quantile of R^*_n.

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

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

By combining Δ-theorem and bootstrap one can also use

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

Example 66. Bootstrap estimation of the distribution of estimators of parameters in $AR(p)$ process.

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.

$$H_0 : F \in \mathcal{F} = \big\{F(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\big\}, \qquad H_1 : F \notin \mathcal{F}.$$

As a test statistic one can use for instance

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

where $\widehat{F}_n$ 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
$$KS^*_{n,b} = \sup_{\mathbf{x}\in\mathbb{R}^k} \big|F^*_{n,b}(\mathbf{x}) - F(\mathbf{x}; \widehat{\boldsymbol{\theta}}^*_{n,b})\big|,$$
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
$$\frac{1 + \sum_{b=1}^{B} \mathbb{I}\big\{KS^*_{n,b} \geq KS_n\big\}}{B+1},$$
where B is usually chosen as 999 or 9 999.

Remark 20. Sometimes people are ignoring the fact that the value of the parameter $\boldsymbol{\theta}_X$ is not fixed in advance and assess the significance of the Kolmogorov-Smirnov test statistic with the help of the (asymptotic) distribution of

$$Z_n = \sqrt{n} \sup_{\mathbf{x}\in\mathbb{R}^k} \big|\widehat{F}_n(\mathbf{x}) - F(\mathbf{x}; \boldsymbol{\theta}_X)\big|,$$

where $F_X(\mathbf{x}) = F(\mathbf{x}; \boldsymbol{\theta}_X)$ is the true distribution function. The problem is that under the null hypothesis the (asymptotic) distribution of

$$\widetilde{Z}_n = \sqrt{n} \sup_{\mathbf{x} \in \mathbb{R}^k} \left| \widehat{F}_n(\mathbf{x}) - F(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n) \right|$$

is rather different from the (asymptotic) distribution of Z_n. The simulation studies show that if the significance of $\sqrt{n}\, KS_n$ is assessed with the help of the distribution Z_n, then the true level of the test is much smaller than the prescribed value α. The intuitive reason is that as $\widehat{\boldsymbol{\theta}}_n$ is estimated from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the empirical distribution function $\widehat{F}_n(\mathbf{x})$ is closer to its parametric estimate $F(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)$ than to the true distribution $F(\mathbf{x}; \boldsymbol{\theta}_X)$.

To conclude, using the (asymptotic) distribution of Z_n to assess the significance of the test statistic $\sqrt{n}\, KS_n$ results in a huge loss of power.

Remark 21. Instead of the test statistic KS_n it is usually recommended to use one of the following statistics. The reason is that the tests based on these statistics have usually more power against the alternatives that seem to be natural.

Cramér-von-Mises:

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

Anderson-Darling:

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

Example 67. Testing goodness-of-fit of multinomial distribution with estimated parameters.

8.5 Testing hypotheses and bootstrap

First of all note that provided the parameter of interest is one-dimensional and one can construct a confidence interval for this parameter (see Section 8.3), then one can use the duality of confidence intervals and testing hypotheses. But in many situations the approach based on an appropriate test statistic is more straightforward.

Suppose that we have a test statistic $T_n = T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and that large values of T_n speak against the null hypothesis. Let $\mathbb{X}_1^* = (\boldsymbol{X}_{1,1}^*, \ldots, \boldsymbol{X}_{n,1}^*)^\mathsf{T}$, $\ldots$, $\mathbb{X}_B^* = (\boldsymbol{X}_{1,B}^*, \ldots, \boldsymbol{X}_{n,B}^*)^\mathsf{T}$ 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

$$\widehat{p}_B = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\{T_{n,b}^* \geq T_n\}}{B + 1}. \tag{104}$$

Example 68. Let $X_1, \ldots, X_n$ be a random sample such that $\mathsf{var}\, X_1 \in (0, \infty)$ and $H_0 : \mathsf{E}\, X_1 = \mu_0$. 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.

Note that in this situation no permutation test (permutation tests are introduced in Chapter 8.6) is available.

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 hypotheses

$$H_0 : \mathsf{E}\, X_1 = \mathsf{E}\, Y_1, \qquad H_1 : \mathsf{E}\, X_1 \neq \mathsf{E}\, Y_1.$$

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

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

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 $X^*_{1,b}, \ldots, X^*_{n_1,b}$ and $Y^*_{1,b}, \ldots, Y^*_{n_2,b}$ 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 $|\widetilde{T}^*_{n,1}|, \ldots, |\widetilde{T}_{n,B^*}|$. 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 $X^*_{1,b}, \ldots, X^*_{n_1,b}$ and $Y^*_{1,b}, \ldots, Y^*_{n_2,b}$ as independent random samples with replacement from $X_1 - \overline{X}_{n_1}, \ldots, X_{n_1} - \overline{X}_{n_1}$ and $Y_1 - \overline{Y}_{n_2}, \ldots, Y_{n_2} - \overline{Y}_{n_2}$ respectively.

Example 69. Suggest a test that would compare medians in two-sample problems.

8.6 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 $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples with the distribution functions F and G respectively. Let the null hypothesis state that the distribution 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 $\mathbb{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$ 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 $\mathbb{Z}_{(\cdot)} = (Z_{(1)}, \ldots, Z_{(n)})^\mathsf{T}$ be the ordered sample, that is $Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)}$. 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 $z_{(1)} \le z_{(2)} \le \ldots \le z_{(n)}$.

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

$$\left(X^*_{1,b}, \ldots, X^*_{n_1,b}\right) = \left(Z^*_{1,b}, \ldots, Z^*_{n_1,b}\right), \qquad \left(Y^*_{1,b}, \ldots, Y^*_{n_2,b}\right) = \left(Z^*_{n_1+1,b}, \ldots, Z^*_{n,b}\right)$$

and the p-value is estimated by (104).

Remark 22. Note that in fact for two-samples there are only $\binom{n}{n_1}$ permutations which can give rise to different values of the test statistic. So if n_1 and n_2 are small then one can calculate the p-value exactly, where exactly means with respect to the permutation distribution of the test

statistic. But usually the number $\binom{n}{n_1}$ is already too big and one generates only B random permutations to estimate the p-value.

Example 70. The permutation test approach can be used to assess for instance the significance of the two-sample Kolmogorov-Smirnov test statistic

$$K_{n_1,n_2} = \sup_{x\in\mathbb{R}} \left|\widehat{F}_{n_1}(x) - \widehat{G}_{n_2}(x)\right|.$$

Note that the standard inference is based on asymptotic distribution of K_n that is derived in case that the distribution function F (and under the null hypothesis also G) is continuous. Thus using the permutation test can be of interest in particular in the presence of ties (e.g. due to rounding).

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 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

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

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 71. Permutation version of χ^2-test of independence.

Remark 23. Generally K-sample problem can be viewed as the testing of independence problem. The reason is that one can view the data as random vectors $\binom{Z_1}{I_1}, \ldots, \binom{Z_n}{I_n}$, where $I_i = k$ (for $i = 1, \ldots, n$, $k = 1, \ldots, K$), if the observation Z_i belongs to the k-th sample. Thus independence of Z_1 and G_1 is equivalent to the fact that all the random samples have the same distribution function.

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

8.7 Bootstrap in linear models

Suppose we observe $\binom{\boldsymbol{X}_1}{Y_1}, \ldots, \binom{\boldsymbol{X}_n}{Y_n}$ a random sample, where, $\boldsymbol{X}_i$ is a p-dimensional random vector. The standard nonparametric bootstrap generates $\binom{\boldsymbol{X}_1^*}{Y_1^*}, \ldots, \binom{\boldsymbol{X}_n^*}{Y_n^*}$ as a simple random sample with replacement from the vectors $\binom{\boldsymbol{X}_1}{Y_1}, \ldots, \binom{\boldsymbol{X}_n}{Y_n}$. Note that one can usually assume that this bootstrap method works, provided the estimator $\widehat{\boldsymbol{\beta}}_n$ is asymptotically normal.

In linear models we usually assume a more specific structure

$$Y_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{105}$$

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

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

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

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

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 $\binom{\boldsymbol{X}_1}{Y_1^*}, \ldots, \binom{\boldsymbol{X}_n}{Y_n^*}$.

The advantage of the nonparametric bootstrap is that it does not require model (105) to hold. On the other hand if model (105) 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.

8.8 Variance estimation and bootstrap

Often one knows that

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

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

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

Note that

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

Thus for a valid inference we need that

$$n\ \mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}^*_n \,|\, \mathbb{X}\big) \xrightarrow[n\to\infty]{P} \mathbb{V}. \tag{106}$$

Note that $\sqrt{n}\,(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{d} \mathsf{N}(\mathbf{0}, \mathbb{V})$ almost surely (or in probability) conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ generally **does not imply** that (106) 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 $\frac{1}{n}\mathbb{V}$.

Example 72. 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 $\mathsf{var}\,\big(g(\overline{X}^*_n)|\mathbb{X}\big) \xrightarrow[n\to\infty]{\text{a.s.}} \infty$.

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

8.9 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 $\widehat{\boldsymbol{\theta}}^{(bc)}_n := \widehat{\boldsymbol{\theta}}_n - \mathbf{b}^*_n$.

* Not done at the lecture nor exercise class.

Example 73. 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

$$\begin{aligned} \mathsf{E}\, g(\overline{X}_n) - g(\mu) &= \mathsf{E}\left\{ g'(\mu)(\overline{X}_n - \mu) + \tfrac{g''(\mu)}{2}(\overline{X}_n - \mu)^2 \right\} + \frac{R_n}{3!} \\ &= \frac{g''(\mu)\,\sigma^2}{2\,n} + O\Big(\tfrac{1}{n^{3/2}}\Big), \end{aligned} \tag{107}$$

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

$$\begin{aligned} b_n^* = \mathsf{E}\,[g(\overline{X}_n^*) \,|\, \mathbb{X}] - g(\overline{X}_n) &= \frac{g''(\overline{X}_n)}{2n}\ \mathsf{var}[X_1^* \,|\, \mathbb{X}] + O_P\Big(\tfrac{1}{n^{3/2}}\Big) \\ &= \frac{g''(\overline{X}_n)\,\widehat{\sigma}_n^2}{2n} + O_P\Big(\tfrac{1}{n^{3/2}}\Big). \end{aligned} \tag{108}$$

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

Now by comparing (107) and (108) 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.

8.10 Jackknife*

Jackknife can be considered as an ancestor of bootstrap. It was originally suggested to reduce the 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

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

* Not done at the lecture nor exercise class.

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

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

and the 'bias-corrected' estimator is defined as

$$\mathbf{T}_n^{(bc)} = \mathbf{T}_n - \widehat{\mathbf{b}}_n. \tag{109}$$

Remark 24. The rationale of the estimator (109) is as follows. For simplicity let θ_X be a one-dimensional parameter and suppose that the bias of estimator T_n is given by

$$\mathsf{E}\, T_n - \theta_X = \frac{a}{n} + \frac{b}{n^{3/2}} + \frac{c}{n^2} + o\Big(\tfrac{1}{n^{5/2}}\Big). \tag{110}$$

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

$$\begin{aligned}
\mathsf{E}\,\widehat{b}_n &= (n-1)\Big(\frac{a}{n-1} + \frac{b}{(n-1)^{3/2}} + \frac{c}{(n-1)^2} + o\Big(\tfrac{1}{(n-1)^{5/2}}\Big) \\
&\qquad - \frac{a}{n} - \frac{b}{n^{3/2}} - \frac{c}{n^2} - o\Big(\tfrac{1}{n^{5/2}}\Big)\Big), \\
&= (n-1)\left(\frac{a}{n(n-1)} + \frac{b(1-\frac{1}{n})^{3/2}}{(n-1)^{3/2}} + \frac{c(1-\frac{1}{n})^2}{(n-1)^2}\right) + O\Big(\tfrac{1}{n^{3/2}}\Big) \\
&= \frac{a}{n} + O\Big(\tfrac{1}{n^{3/2}}\Big).
\end{aligned} \tag{111}$$

Now combining (110) and (111) 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 and indentically distributed, i.e.

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

Literature: Shao and Tu (1996) Chapter 1.3.

The end of the self study for the week (27. 4. - 1. 5. 2020)

9 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

$$\widetilde{f}_n(x) = \frac{\widehat{F}_n(x+h_n) - \widehat{F}_n(x-h_n)}{2\,h_n} = \frac{1}{2\,h_n} \sum_{i=1}^{n} \frac{\mathbb{I}\{X_i \in (x-h_n, x+h_n]\}}{n}, \tag{112}$$

where $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\}$ is the empirical distribution function and (the bandwidth) h_n is a sequence of positive constants going to zero.

It is straightforward to show

$$\mathsf{E}\ \widetilde{f}_n(x) = \frac{F(x+h_n) - F(x-h_n)}{2\,h_n} \xrightarrow[n\to\infty]{} f(x)$$

and

$$\begin{aligned} \mathsf{var}\big(\widetilde{f}_n(x)\big) &= \frac{\big[F(x+h_n) - F(x-h_n)\big]\big[1 - F(x+h_n) + F(x-h_n)\big]}{4\,h_n^2\,n} \\ &= \frac{F(x+h_n) - F(x-h_n)}{2\,h_n} \frac{1 - F(x+h_n) + F(x-h_n)}{2\,n\,h_n} \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

provided that $h_n \to 0$ and at the same time $(n\,h_n) \to \infty$.

Note that the estimator (112) can be rewritten as

$$\widetilde{f}_n(x) = \frac{1}{2\,n\,h_n} \sum_{i=1}^{n} \mathbb{I}\big\{-1 \le \tfrac{x-X_i}{h_n} < +1\big\} = \frac{1}{n\,h_n} \sum_{i=1}^{n} w\big(\tfrac{x-X_i}{h_n}\big), \tag{113}$$

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 (113) we define the kernel estimator of a density as

$$\widehat{f}_n(x) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K\big(\tfrac{x-X_i}{h_n}\big), \tag{114}$$

where the function K is called a kernel function and h_n is usually called bandwidth† or smoothing parameter. Usually the function K is taken as a symmetric density of a probability distribution. The common choices of K are summarised in Table 1.

Remark 25. Note that:

* *Jádrové odhady hustoty* † *V češtině se mluví o šířce vyhlazovací okna nebo jednodušeji o vyhlazovacním prametru.*

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.

(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) Note that $\widehat{f}_n(x)$ is continuous (has a continuous derivative) if K is continuous (has a continuous derivative). That is why usually a continuous function K is preferred.

(iii) If K is a density of a probability distribution, then $\int \widehat{f}_n(x)\,\mathrm{d}x = 1$.

Example 74. Consider a random sample of size 200 from the distribution with the distribution function

$$F(x) = \tfrac{1}{2}\,\Phi\big(x\big) + \tfrac{1}{2}\,\Phi\big(\tfrac{x-4}{2}\big),$$

i.e. the distribution is given by the normal mixture $\frac{1}{2}\,\mathsf{N}(0,1) + \frac{1}{2}\,\mathsf{N}(4,4)$. The kernel estimates with different bandwidth choices of h_n and the Gaussian kernel is found Figure 74. For reasons of comparison also the corresponding histogram with the width of the columns given by $2\,h_n$ is included.

The true density is indicated by the black solid line. Note that a reasonable bandwidth seems to be between 0.5 and 1. The bandwidth smaller than 0.5 results in a estimate that is too wiggly (the variance dominates). On the other hand the bandwidth greater than 1 results in an estimate that is too biased.

Unfortunately in practice we do not know what is the true density so it is much more difficult to guess what a reasonable bandwidth should be. Note that for the histogram the problem of the choice of the bandwidth h_n correspond to the choice of the width of the columns.

9.1 Consistency and asymptotic normality

Theorem 16 (Bochner's theorem)**.** *Let the function K satisfy*

$$\text{(B1)}\ \int_{-\infty}^{+\infty} |K(y)|\,dy < \infty, \qquad \text{(B2)}\ \lim_{|y|\to\infty} |y\,K(y)| = 0. \tag{115}$$

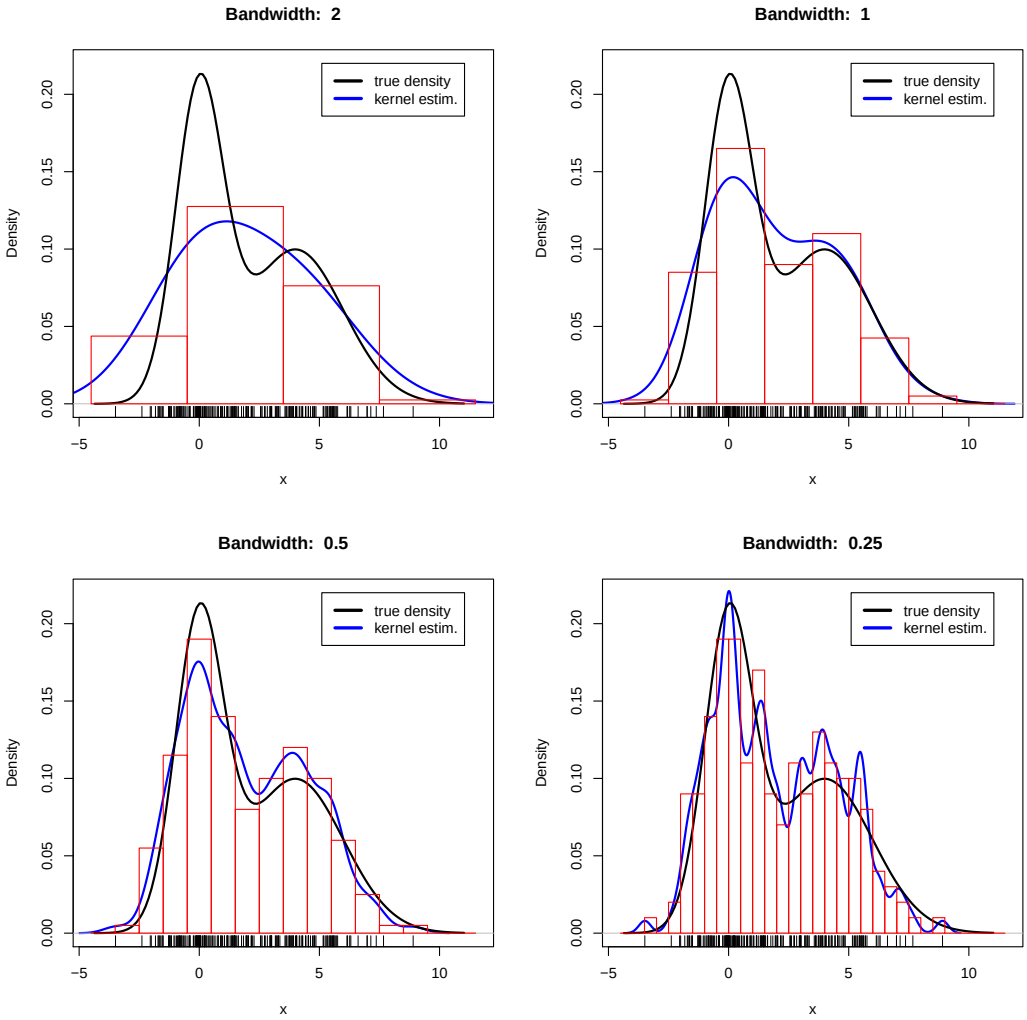

Figure 4: Kernel estimates vs. histograms for different bandwidth choices.

Further let the function g satisfy $\int_{-\infty}^{+\infty} |g(y)|\,dy < \infty$. Put

$$g_n(x) = \frac{1}{h_n} \int_{-\infty}^{+\infty} g(z)\, K\big(\tfrac{x-z}{h_n}\big)\,\mathrm{d}z,$$

where $h_n \searrow 0$ as $n \to \infty$. Then in each point x of the continuity of g it holds that

$$\lim_{n\to\infty} g_n(x) = g(x) \int_{-\infty}^{+\infty} K(y)\,\mathrm{d}y. \tag{116}$$

Proof. * Let x be the point of continuity g. We need to show that

$$\lim_{n\to\infty} \left| g_n(x) - g(x) \int K(y)\,\mathrm{d}y \right| = 0.$$

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 $\delta > 0$:

$$\frac{\delta}{h_n} \to \infty \qquad \text{and} \qquad \frac{1}{\delta} \sup_{t:|t|\geq \frac{\delta}{h_n}} \big|t\,K(t)\big| \to 0, \qquad \text{as} \quad n \to \infty.$$

Thus there exists a sequence of positive constants $\{\delta_n\}$ such that

$$\delta_n \to 0, \quad \frac{\delta_n}{h_n} \to \infty \quad \text{and} \quad \frac{1}{\delta_n} \sup_{t:|t|\geq \frac{\delta_n}{h_n}} \big|t\,K(t)\big| \to 0, \quad \text{as} \quad n \to \infty. \tag{117}$$

Taking δ_n satisfying (117) one can bound

$$\begin{aligned} \left| g_n(x) - g(x) \int K(y)\,\mathrm{d}y \right| \leq{}& \underbrace{\frac{1}{h_n} \int_{-\delta_n}^{\delta_n} \big|g(x-y) - g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:A_n} \\ &+ \underbrace{\frac{1}{h_n} \int_{y:|y|\geq\delta_n} \big|g(x-y) - g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_n}. \end{aligned} \tag{118}$$

Dealing with A_n. As g is continuous in the point x

$$A_n \leq \sup_{y:|y|\leq\delta_n} \big|g(x-y) - g(x)\big| \int_{-\delta_n}^{\delta_n} \tfrac{1}{h_n} \big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y \leq o(1) \underbrace{\int_{\mathbb{R}} |K(t)|\,\mathrm{d}t}_{<\infty;\,\text{(B1)}} = o(1), \tag{119}$$

as $n \to \infty$.

* Not done at the lecture.

Dealing with B_n. Further one can bound B_n with

$$B_n \leq \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x-y)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{1n}} + \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{2n}}. \tag{120}$$

Using the substitution $t = \frac{y}{h_n}$ and (117) one gets

$$B_{2n} = \big|g(x)\big| \int_{y:|y|\geq\delta_n} \tfrac{1}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y = \big|g(x)\big| \int_{t:|t|\geq\frac{\delta_n}{h_n}} |K(t)|\,\mathrm{d}t \xrightarrow[n\to\infty]{} 0. \tag{121}$$

Finally using (117)

$$\begin{aligned} B_{1n} &= \int_{y:|y|\geq\delta_n} \underbrace{\tfrac{|y|}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|}_{\leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}}|t\,K(t)|} \frac{\big|g(x-y)\big|}{|y|}\,\mathrm{d}y \leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \int_{y:|y|\geq\delta_n} \frac{\big|g(x-y)\big|}{|y|}\,\mathrm{d}y \\ &\leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \frac{1}{\delta_n} \underbrace{\int \big|g(x-y)\big|\,\mathrm{d}y}_{=\int|g(y)|\,\mathrm{d}y<\infty} \xrightarrow[n\to\infty]{} 0. \end{aligned} \tag{122}$$

Now combining (118), (119), (120), (121) and (122) yields the statement of the theorem. □

Remark 26. Note that:

(i) If K is a density, then $\int |K(y)|\,\mathrm{d}y = \int K(y)\,\mathrm{d}y = 1$ 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 16 (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 16.

(iii) If K is a density but with an unbounded support, then assumption (B2) is satisfied if there exists a finite constant $c > 0$ such that K is non-decreasing on $(-\infty, -c)$ and non-increasing on (c, ∞).

(iv) If g is uniformly continuous then one can show that also the convergence in (116) is uniform.

(v) Note that the kernel $K(x) = \sum_{n=1}^{\infty} \frac{1}{2^n}\mathbb{I}\{x \in (2^n - 1, 2^n + 1)\}$ meets assumption (B1), but (B2) is not satisfied.

Theorem 17 (Variance and consistency of $\widehat{f}_n(x)$)**.** *Let the estimator $\widehat{f}_n(x)$ be given by (114) and the function K satisfies* (B1) *and* (B2) *introduced in (115). 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) $\lim_{n\to\infty} n\,h_n\ \mathsf{var}\left(\widehat{f}_n(x)\right) = f(x)\int K^2(y)\,\mathrm{d}y$;

(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

$$\begin{aligned}\mathsf{var}\left(\widehat{f}_n(x)\right) &= \mathsf{var}\left[\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\tfrac{x-X_i}{h_n}\right)\right] = \frac{1}{n\,h_n^2}\,\mathsf{var}\left[K\left(\tfrac{x-X_1}{h_n}\right)\right]\\ &= \frac{1}{n\,h_n^2}\left[\mathsf{E}\,K^2\left(\tfrac{x-X_1}{h_n}\right) - \left(\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right)\right)^2\right]. \qquad (123)\end{aligned}$$

Now using Theorem 16

$$\frac{1}{h_n}\,\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right) = \int \frac{1}{h_n}K\left(\tfrac{x-y}{h_n}\right)f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} f(x)\int K(y)\,\mathrm{d}y = f(x). \qquad (124)$$

Analogously

$$\frac{1}{h_n}\,\mathsf{E}\,K^2\left(\tfrac{x-X_1}{h_n}\right) = \frac{1}{h_n}\int K^2\left(\tfrac{x-y}{h_n}\right)f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} f(x)\int K^2(y)\,\mathrm{d}y, \qquad (125)$$

where we have used again Theorem 16 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 (123), (124) and (125) 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 (124)

$$\mathsf{E}\,\widehat{f}_n(x) = \frac{1}{h_n}\,\mathsf{E}\,K\left(\tfrac{x-X_1}{h_n}\right) \xrightarrow[n\to\infty]{} f(x). \qquad (126)$$

Now with the help of (i) and (126)

$$\mathsf{E}\left[\widehat{f}_n(x) - f(x)\right]^2 = \mathsf{var}\left[\widehat{f}_n(x)\right] + \left[\mathsf{E}\,\widehat{f}_n(x) - f(x)\right]^2 \xrightarrow[n\to\infty]{} 0,$$

which implies the consistency of $\widehat{f}_n(x)$. □

Remark 27. Note that Theorem 17 implies only pointwise consistency. It would be much more difficult to show that $\sup_{x\in\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \xrightarrow[n\to\infty]{P} 0$.

Remark 28. Note that one cannot use the standard law of large numbers to prove the consistency, as one would need a law of large numbers for a triangular array.

Theorem 18 (Asymptotic normality of $\widehat{f}_n(x)$)**.** *Let the assumptions of Theorem 17 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 17 we know that

$$\frac{\mathsf{var}\left(\widehat{f}_n(x)\right)}{\frac{f(x)R(K)}{n\,h_n}} \xrightarrow[n\to\infty]{} 1,$$

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

$$X_{n,i} = \frac{1}{\sqrt{n\,h_n}} \frac{K\left(\frac{x-X_i}{h_n}\right) - \mathsf{E}\ K\left(\frac{x-X_i}{h_n}\right)}{\sqrt{f(x)R(K)}}, \quad i = 1,\dots,n,$$

are independent and identically distributed 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 $i = 1,\dots,n$:

$$\begin{aligned}\mathbb{I}\left\{|X_{n,i}| \geq \varepsilon\right\} &= \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \left|\frac{K\left(\frac{x-X_i}{h_n}\right) - \mathsf{E}\ K\left(\frac{x-X_i}{h_n}\right)}{\sqrt{f(x)R(K)}}\right| \geq \varepsilon\right\} \\ &\leq \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \frac{2\sup_y |K(y)|}{\sqrt{f(x)R(K)}} \geq \varepsilon\right\} = 0,\end{aligned}$$

which implies that the 'Feller-Lindeberg condition'

$$\lim_{n\to\infty} \sum_{i=1}^n \mathsf{E}\left[X_{n,i}^2\, \mathbb{I}\left\{|X_{n,i}| \geq \varepsilon\right\}\right] = 0$$

is satisfied. □

Remark 29. Note that Theorem 18 implies

$$\frac{\widehat{f}_n(x) - f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1), \tag{127}$$

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 (127). But this would require that $h_n = o\big(n^{-1/5}\big)$ which would exclude the optimal bandwidth choice (see the next section).

9.2 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 17

$$\mathsf{var}\left(\widehat{f}_n(x)\right) = \frac{f(x)R(K)}{n\,h_n} + o\big(\tfrac{1}{n\,h_n}\big), \tag{128}$$

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

$$\text{bias}\big(\widehat{f}_n(x)\big) = \mathsf{E}\ \widehat{f}_n(x) - f(x) = \tfrac{1}{2}\, h_n^2\, f''(x)\, \mu_{2K} + o(h_n^2),$$

which together with (128) implies

$$\text{MSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\, f(x)\, R(K) + \tfrac{1}{4}\, h_n^4\, [f''(x)]^2 \mu_{2K}^2 + o\big(\tfrac{1}{n\,h_n}\big) + o\big(h_n^4\big). \tag{129}$$

Ignoring the remainder $o(\cdot)$ terms in (129), AMSE (asymptotic mean squared error) of $\widehat{f}_n(x)$ is given by

$$\text{AMSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\, f(x)\, R(K) + \tfrac{1}{4}\, h_n^4\, [f''(x)]^2 \mu_{2K}^2. \tag{130}$$

Minimising (130) one gets *asymptotically optimal local bandwidth* (i.e. bandwidth that minimises the AMSE)

$$h_n^{(opt)}(x) = n^{-1/5} \left[\frac{f(x)\, R(K)}{[f''(x)]^2\, \mu_{2K}^2}\right]^{1/5}. \tag{131}$$

To get a global bandwidth it is useful to define **(A)MISE - (asymptotic) mean integrated squared error**. Introduce

$$\text{MISE}\big(\widehat{f}_n\big) = \int \text{MSE}\big(\widehat{f}_n(x)\big)\, \mathrm{d}x = \int \mathsf{E}\, \big[\widehat{f}_n(x) - f(x)\big]^2\, \mathrm{d}x,$$

and its asymptotic approximation

$$\begin{aligned} \text{AMISE}\big(\widehat{f}_n\big) &= \int \text{AMSE}\big(\widehat{f}_n(x)\big)\, \mathrm{d}x = \int \tfrac{1}{n\,h_n} f(x) R(K) + \tfrac{[f''(x)]^2 \mu_{2K}^2}{4}\, h_n^4\, \mathrm{d}x \\ &= \frac{R(K)}{n\,h_n} + h_n^4\, \frac{R(f'')\, \mu_{2K}^2}{4}, \end{aligned} \tag{132}$$

where $R(f'') = \int \big[f''(x)\big]^2\, \mathrm{d}x$.

Minimising (132) one gets *asymptotically optimal global bandwidth* (i.e. bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K)}{R(f'')\, \mu_{2K}^2}\right]^{1/5}. \tag{133}$$

Remark 30. Note that after substitution of the optimal bandwidth (133) into (132) one gets that the optimal AMISE is given by

$$\frac{5\, \big[R(f'')\big]^{1/5}}{4\, n^{4/5}} \left\{\big[R(K)\big]^2\, \mu_{2K}\right\}^{2/5}.$$

It can be shown that if we consider kernels that are densities of probability distributions then $\big[R(K)\big]^2\, \mu_{2K}$ is minimised for K being 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:	$K(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) \mathbb{I}\left\{\lvert x\rvert \le \sqrt{5}\right\}$
Triangular kernel:	$K(x) = \frac{1}{\sqrt{6}}(1 - \lvert x\rvert) \mathbb{I}\left\{\lvert x\rvert \le \sqrt{6}\right\}$
Uniform kernel:	$K(x) = \frac{1}{2\sqrt{3}} \mathbb{I}\left\{\lvert x\rvert \le \sqrt{3}\right\}$
Biweight kernel:	$K(x) = \frac{15}{16\sqrt{7}}(1 - x^2)^2 \mathbb{I}\left\{\lvert x\rvert \le \sqrt{7}\right\}$
Tricube kernel:	$K(x) = \frac{70\sqrt{243}}{81\sqrt{35}}(1 - \lvert x\rvert^3)^3 \mathbb{I}\left\{\lvert x\rvert \le \sqrt{\frac{35}{243}}\right\}$
Gaussian kernel:	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$

Table 2: Some kernel functions standardised so that $\mu_{2K} = 1$.

9.2.1 Normal reference rule

The problem of asymptotically optimal bandwidths given in (131) and (133) 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

$$f'(x) = \tfrac{1}{\sigma^2}\varphi'\left(\tfrac{x-\mu}{\sigma}\right), \qquad f''(x) = \tfrac{1}{\sigma^3}\varphi''\left(\tfrac{x-\mu}{\sigma}\right),$$

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 (131) one gets

$$\widehat{h}_n(x) = n^{-\frac{1}{5}}\,\widehat{\sigma}\,\left[\frac{R(K)}{\mu_{2K}^2}\frac{1}{\left[\frac{(x-\widehat{\mu})^2}{\widehat{\sigma}^2} - 1\right]^2\varphi\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)}\right]^{\frac{1}{5}},$$

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

$$\begin{aligned}
R(f'') &= \int \left[f''(x)\right]^2 \mathrm{d}x = \int \left\{\frac{1}{\sigma^3}\left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]\varphi\left(\tfrac{x-\mu}{\sigma}\right)\right\}^2 \mathrm{d}x \\
&= \frac{1}{\sigma^6}\int \left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]^2 \varphi^2\left(\tfrac{x-\mu}{\sigma}\right) \mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \left| \begin{array}{rcl} t &=& \frac{x-\mu}{\sigma} \\ \mathrm{d}t &=& \frac{\mathrm{d}x}{\sigma} \end{array} \right| = \frac{1}{\sigma^5} \int (t^2-1)^2 \varphi^2(t)\,\mathrm{d}t \\
&= \frac{1}{\sigma^5} \int (t^4 - 2t^2 + 1) \frac{1}{2\pi} \mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{1}{\sigma^5 2\sqrt{\pi}} \int (t^4 - 2t^2 + 1) \underbrace{\frac{1}{\sqrt{\pi}} \mathrm{e}^{-t^2}}_{\sim \mathsf{N}\left(0, \frac{1}{2}\right)}\,\mathrm{d}t \\
&= \frac{1}{2\sigma^5\sqrt{\pi}}\, \mathsf{E}\left(Y^4 - 2Y^2 + 1\right) = \frac{1}{2\sigma^5\sqrt{\pi}} \left[3 \cdot \left(\tfrac{1}{2}\right)^2 - 2 \cdot \tfrac{1}{2} + 1\right] = \frac{3}{8\sigma^5\sqrt{\pi}},
\end{aligned}
$$

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

$$
h_n = 1.06\, n^{-1/5}\, \min\left\{S_n, \widetilde{IQR}_n\right\}, \tag{134}
$$

where

$$
S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}, \qquad \text{and} \qquad \widetilde{IQR}_n = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{1.34}.
$$

It was found out that the bandwidth selector (134) 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

$$
h_n = 0.9\, n^{-1/5}\, \min\left\{S_n, \widetilde{IQR}_n\right\}.
$$

For a more detailed argumentation see Silverman (1986), page 48.

9.2.2 Least-squares cross-validation*

By this method we choose the bandwith as

$$
h_n^{(LSCV)} = \arg\min_{h_n > 0} \mathcal{L}(h_n),
$$

* 'cross-validation' se střídavě překládá jako metoda křížového ověřování, metoda křížové validace nebo prostě jako krosvalidace.

where

$$\mathcal{L}(h_n) = \int \big[\widehat{f}_n(x)\big]^2 \,\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i),$$

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 rationale 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

$$\begin{aligned}\mathrm{MISE}\big(\widehat{f}_n\big) &= \int \mathsf{E}\,\big(\widehat{f}_n(x) - f(x)\big)^2 \,\mathrm{d}x \overset{\text{Fub.}}{=} \mathsf{E}\int \widehat{f}_n^2(x) - 2\widehat{f}_n(x)f(x) + f^2(x)\,\mathrm{d}x \\ &= \mathsf{E}\int \widehat{f}_n^2(x)\,\mathrm{d}x - 2\,\mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x + \int f^2(x)\,\mathrm{d}x.\end{aligned}$$

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 depend on h_n. Thus it remains to estimate $\mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x$. 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 (124) and (126)

$$\begin{aligned}\mathsf{E}\,\hat{f}_{-i}(X_i) &= \mathsf{E}\left[\frac{1}{(n-1)h_n}\sum_{j=1,j\neq i}^{n} K\big(\tfrac{X_i-X_j}{h_n}\big)\right] = \frac{1}{h_n}\,\mathsf{E}\,K\big(\tfrac{X_1-X_2}{h_n}\big) \\ &= \frac{1}{h_n}\int\int K\big(\tfrac{x-y}{h_n}\big)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = \int \underbrace{\left[\int \frac{1}{h_n}K\big(\tfrac{x-y}{h_n}\big)f(y)\,\mathrm{d}y\right]}_{=\mathsf{E}\,\widehat{f}_n(x)} f(x)\,\mathrm{d}x \qquad (135) \\ &= \int \mathsf{E}\,\widehat{f}_n(x)f(x)\,\mathrm{d}x \overset{\text{Fub}}{=} \mathsf{E}\int \widehat{f}_n(x)f(x)\,\mathrm{d}x.\end{aligned}$$

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 31. Stone (1984) has proved that

$$\frac{\mathrm{ISE}\left(h_n^{(LSCV)}\right)}{\min_{h_n} \mathrm{ISE}(h_n)} \xrightarrow[n\to\infty]{\text{a.s.}} 1,$$

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.

9.2.3 Biased cross-validation*

This method aims at minimizing the AMISE given by (132). 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 32. 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 (131).

9.3 Higher order kernels†

By a formal calculation (for sufficiently large n, sufficiently smooth f and x an interior point of the support) one gets

$$\begin{aligned}
\mathsf{E}\ \widehat{f}_n(x) &= \int K(t) f(x - t h_n)\,\mathrm{d}t \\
&= f(x)\int K(t)\,\mathrm{d}t - f'(x) h_n \int t K(t)\,\mathrm{d}t \\
&\quad + \frac{f''(x)}{2} h_n^2 \int t^2 K(t)\,\mathrm{d}t - \frac{f'''(x)}{3!} h_n^3 \int t^3 K(t)\,\mathrm{d}t + \dots .
\end{aligned}$$

The kernel of order p is such that $\int K(t)\,dt = 1$ and

$$\int t^j\, K(t)\,dt = 0, \qquad j = 1, \dots, p-1, \qquad \text{and} \qquad \int t^p\, K(t)\,dt \neq 0.$$

But note that if the above equations hold for $p > 2$, then (among others) $\int t^2 K(t)\,\mathrm{d}t = 0$, 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

$$K(y) = \tfrac{1}{2}\,(3 - y^2)\,\tfrac{1}{\sqrt{2\,\pi}}\, e^{-y^2/2}.$$

* Not done at the lecture. † Not done at the lecture.

9.4 Mirror-reflection*

The standard kernel density estimator (114) 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

$$\widehat{f}_n^{(MR)}(x) = \begin{cases} \frac{1}{n\,h_n} \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) + \frac{1}{n\,h_n} \sum_{i=1}^n K\left(\frac{x+X_i}{h_n}\right), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{136}$$

Note that the first term on the right-hand side of (136) (for $x \geq 0$) is the standard kernel density estimator $\widehat{f}_n(x)$. The second term on the right-hand side of (136) 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.

The end of the self study for the week (4. 5. -8. 5. 2020)

10 Kernel regression†

Suppose that one observes independent and identically distributed bivariate random vectors $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$. Our primary interest in this section is to estimate the conditional mean function of Y_1 given $X_1 = x$, i.e.

$$m(x) = \mathsf{E}\left[Y_1 \,|\, X_1 = x\right]$$

without assuming any parametric form of $m(x)$.

In what follows it is useful to denote the conditional variance function as

$$\sigma^2(x) = \mathsf{var}\left[Y_1 \,|\, X_1 = x\right].$$

10.1 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

$$m(X_i) \doteq m(x) + m'(x)\,(X_i - x) + \ldots + \frac{m^{(p)}(x)}{p!}\,(X_i - x)^p. \tag{137}$$

* Not done at the lecture. † *Jádrové regresní odhady*

Thus 'locally' one can view and estimate the function $m(x)$ as a polynomial. This motivates defininition of the local polynomial estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(x) &= \big(\widehat{\beta}_0(x), \dots, \widehat{\beta}_p(x)\big)^{\mathsf{T}} \\ &= \underset{b_0,\dots,b_p}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1(X_i - x) - \dots - b_p(X_i - x)^p\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big), \end{aligned} \tag{138}$$

where K is a given kernel function and h_n is a smoothing parameter (bandwidth) going to zero as $n \to \infty$.

Comparing (137) and (138) 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

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \qquad \mathbb{X}_p(x) = \begin{pmatrix} 1 & (X_1 - x) & \dots & (X_1 - x)^p \\ 1 & (X_2 - x) & \dots & (X_2 - x)^p \\ \dots & \dots & \dots & \dots \\ 1 & (X_n - x) & \dots & (X_n - x)^p \end{pmatrix}$$

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 (138) can be written as the weighted least squares problem

$$\widehat{\boldsymbol{\beta}}(x) = \underset{\mathbf{b} \in \mathbb{R}^{p+1}}{\arg\min} \Big\{ \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big)^{\mathsf{T}} \mathbb{W}(x) \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big) \Big\}, \tag{139}$$

where $\mathbf{b} = (b_0, b_1, \dots, b_p)^{\mathsf{T}}$. The solution of (139) 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 non-singular.

The following technical lemma will be useful in deriving the properties of the local polynomial estimator.

Lemma 8. *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 18) 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 33. Note that Lemma 8 implies that

$$S_{n,0}(x) = f_X(x) + \tfrac{h_n^2}{2}\, f_X''(x)\, \mu_{2K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = f_X(x) + o_P(1), \tag{140}$$

$$S_{n,1}(x) = h_n\, f'(x)\, \mu_{2K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = o_P(1), \tag{141}$$

$$S_{n,2}(x) = f(x)\, \mu_{2K} + o_P(1), \tag{142}$$

$$S_{n,3}(x) = h_n\, f'(x) \int t^4\, K(t)\,\mathrm{d}t + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n\,h_n}}\big) = o_P(1). \tag{143}$$

10.2 Nadaraya-Watson estimator

For $p = 0$ the local polynomial estimator given by (138) 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

$$\widehat{\beta}_0(x) = \sum_{i=1}^{n} w_{ni}(x)\, Y_i =: \widehat{m}_{NW}(x),$$

where

$$w_{ni}(x) = \frac{K\big(\frac{X_i - x}{h_n}\big)}{\sum_{j=1}^{n} K\big(\frac{X_j - x}{h_n}\big)} = \frac{\frac{1}{n\,h_n} K\big(\frac{X_i - x}{h_n}\big)}{S_{n,0}(x)}.$$

This estimator is in the context of the local polynomial regression also called **a locally constant estimator**.

Note that for each x for which the weights are defined

$$\sum_{i=1}^{n} w_{ni}(x) = 1.$$

Moreover if the kernel K is non-negative function then also the weights are non-negative.

Remark 34. Let us consider the kernel with the support $[-1, 1]$. Then the $w_{ni}(x)$ is zero if $X_i \notin [x - h_n, x + h_n]$.

Further, if we assume the uniform kernel, i.e. $K(x) = \frac{1}{2}\mathbb{I}\{|x| \leq 1\}$, then all the weights $w_{ni}(x)$ for which $X_i \in [x - h_n, x + h_n]$ are equal. Thus for this kernel the Nadaraya-Watson estimator $\widehat{m}_{NW}(x)$ is given simply by the sample mean calculated from the observation Y_i for which $X_i \in [x - h_n, x + h_n]$, i.e.

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i \, \mathbb{I}\{|X_i - x| \leq h_n\}}{\sum_{i=1}^{n} \mathbb{I}\{|X_i - x| \leq h_n\}}$$

Thus one can view $\widehat{m}_{NW}(x)$ as a 'moving average' in the covariate direction.

To formulate some theoretic properties of $\widehat{m}_{NW}(x)$ put $\mathbb{X} = (X_1, \ldots, X_n)$. Further 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 19. *Suppose that the assumptions of Lemma 8 are satisfied and further that* $(n\, h_n^3) \xrightarrow[n\to\infty]{} \infty$, *the density* $f_X(\cdot)$ *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*

$$\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = h_n^2\, \mu_{2K} \left(\frac{m'(x)\, f_X'(x)}{f_X(x)} + \frac{m''(x)}{2} \right) + o_P(h_n^2), \tag{144}$$

$$\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\, R(K)}{f_X(x)\, n\, h_n} + o_P\big(\tfrac{1}{n\, h_n}\big), \tag{145}$$

where

$$R(K) = \int K^2(x)\, \mathrm{d}x \qquad \textit{and} \qquad \mu_{2K} = \int x^2\, K(x)\, \mathrm{d}x. \tag{146}$$

Proof. Showing (144). Let us calculate

$$\begin{aligned}
\mathsf{E}\left[\widehat{m}_{NW}(x)|\mathbb{X}\right] &= \sum_{i=1}^{n} w_{ni}(x)\ \mathsf{E}\left[Y_i|\mathbb{X}\right] = \sum_{i=1}^{n} w_{ni}(x)\ \mathsf{E}\left[Y_i|X_i\right] = \sum_{i=1}^{n} w_{ni}(x) m(X_i) \\
&= \sum_{i=1}^{n} w_{ni}(x) \left[m(x) + (X_i - x)\, m'(x) + \tfrac{(X_i-x)^2}{2}\, m''(x) + (X_i - x)^2\, \tilde{R}(X_i)\right] \\
&= m(x) \sum_{i=1}^{n} w_{ni}(x) + m'(x) \sum_{i=1}^{n} w_{ni}(x)(X_i - x) + \frac{m''(x)}{2} \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \\
&\quad + \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \tilde{R}(X_i) \\
&= m(x) + m'(x)\, A_n + \frac{m''(x)}{2}\, B_n + C_n, \qquad (147)
\end{aligned}$$

where $\tilde{R}(z) \to 0$ as $z \to x$ and

$$A_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x),\ B_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2,\ C_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \tilde{R}(X_i). \quad (148)$$

Now with the help of (140) and (141)

$$\begin{aligned}
A_n &= \sum_{i=1}^{n} w_{ni}(x)(X_i - x) = \frac{h_n \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)(X_i - x)\frac{1}{h_n^2}}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n}\right)\frac{1}{h_n}} = \frac{h_n S_{n,1}(x)}{S_{n,0}(x)} \\
&= \frac{h_n \left[h_n f'_X(x)\, \mu_{2K} + o(h_n^2) + O_P\left(\frac{1}{\sqrt{nh_n}}\right)\right]}{f_X(x) + o_P(1)} = \frac{h_n^2 f'_X(x)\mu_{2K} + o(h_n^3) + O_P\left(\frac{h_n}{\sqrt{nh_n}}\right)}{f_X(x) + o_P(1)} \\
&= \frac{h_n^2 f'_X(x)\mu_{2K}}{f_X(x)} + o_P(h_n^2) + O_P\Big(\frac{h_n^2}{\sqrt{nh_n^3}}\Big) = \frac{h_n^2 f'_X(x)\mu_{2K}}{f_X(x)} + o_P(h_n^2), \qquad (149)
\end{aligned}$$

as $(nh_n^3) \to \infty$. Further with the help of (140) and (142)

$$\begin{aligned}
B_n &= \sum_{i=1}^{n} w_{ni}(X_i - x)^2 = \ldots = \frac{h_n^2 S_{n,2}(x)}{S_{n,0}(x)} \\
&= \frac{h_n^2 \left[f_X(x)\, \mu_{2K} + o_P(1)\right]}{f_X(x) + o_P(1)} = h_n^2 \mu_{2K} + o_P(h_n^2). \qquad (150)
\end{aligned}$$

Concerning C_n thanks to (150) and the fact that the support of K is $(-1, 1)$ one can bound

$$\begin{aligned}
|C_n| &\le \left|\sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \tilde{R}(X_i)\right| \le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)| \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^2 \\
&= o(1)\, O_P(h_n^2) = o_P(h_n^2). \qquad (151)
\end{aligned}$$

Now combining (149), (150) and (151) one gets

$$\mathsf{E}\left[\widehat{m}_{NW}(x)|\mathbb{X}\right] = m(x) + m'(x) h_n^2 \frac{f'_X(x)}{f_X(x)} \mu_{2K} + \frac{m''(x)}{2} h_n^2 \mu_{2K} + o_P(h_n^2),$$

which implies (144).

Showing (145). Let us calculate

$$\begin{aligned}\mathsf{var}[\widehat{m}_{NW}(x)|\mathbb{X}] &= \sum_{i=1}^{n} w_{ni}^2(x)\,\mathsf{var}[Y_i|X_i] = \sum_{i=1}^{n} w_{ni}^2(x)\sigma^2(X_i)\\ &= \frac{\sum_{i=1}^{n} K^2\big(\frac{X_i-x}{h_n}\big)\sigma^2(X_i)}{\Big[\sum_{j=1}^{n} K\big(\frac{X_j-x}{h_n}\big)\Big]^2} = \frac{1}{nh_n}\frac{V_n}{[S_{n,0}(x)]^2},\end{aligned}$$

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 17 it is proved that $\widehat{f}_n(x) \xrightarrow[n\to\infty]{P} f(x)$, we will show that

$$V_n \xrightarrow[n\to\infty]{P} f_X(x)\,\sigma^2(x)\,R(K), \tag{152}$$

which combined with (140) implies (145).

Showing (152). First with the help of Bochner's theorem (Theorem 16)

$$\begin{aligned}\mathsf{E}\,V_n &= \frac{1}{h_n}\,\mathsf{E}\,\Big[K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\Big]\\ &= \int \frac{1}{h_n}K^2\big(\tfrac{z-x}{h_n}\big)\,\sigma^2(z)f_X(z)\,\mathrm{d}z \xrightarrow[n\to\infty]{} \sigma^2(x)f_X(x)\int K^2(t)\,\mathrm{d}t.\end{aligned}$$

Now it remains to show that $\mathsf{var}(V_n) \xrightarrow[n\to\infty]{} 0$. Using again Bochner's theorem (Theorem 16)

$$\begin{aligned}\mathsf{var}(V_n) &= \frac{1}{nh_n^2}\left[\mathsf{E}\,K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1) - \Big(\mathsf{E}\,K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\Big)^2\right]\\ &= \frac{1}{nh_n}\left[\frac{1}{h_n}\,\mathsf{E}\,K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1)\right] - \frac{1}{n}\left[\frac{1}{h_n}\,\mathsf{E}\,K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\right]^2\\ &= \frac{1}{nh_n}\left[\sigma^4(x)f_X(x)\int K^4(t)\,\mathrm{d}t + o(1)\right] - \frac{1}{n}\left[\sigma^2(x)f_X(x)\int K^2(t)\,\mathrm{d}t + o(1)\right]^2\\ &\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

□

10.3 Local linear estimator

For $p = 1$ the local polynomial estimator given by (138) simplifies to

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \operatorname*{arg\,min}_{b_0,b_1} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i-x}{h_n}\big).$$

By solving the above optimisation task one gets

$$\widehat{\beta}_0(x) = \sum_{i=1}^{n} w_{ni}(x)\,Y_i =: \widehat{m}_{LL}(x),$$

where the (local linear) weights can be written in the form

$$w_{ni}(x) = \frac{\frac{1}{n\,h_n}\,K\big(\frac{X_i-x}{h_n}\big)\big(S_{n,2}(x) - \frac{X_i-x}{h_n}\,S_{n,1}(x)\big)}{S_{n,0}(x)\,S_{n,2}(x) - S_{n,1}^2(x)}, \qquad i = 1, \ldots, n. \tag{153}$$

It is easy to check (see also Remark 35 below) that the weights satisfy (for each x that the weights are defined)

$$\sum_{i=1}^{n} w_{ni}(x) = 1, \qquad \sum_{i=1}^{n} w_{ni}(x)(X_i - x) = 0. \tag{154}$$

On the other hand it might happen that the weights are negative. In practice this happens if x is either 'close' to to the minimal or maximal value of the covariate.

Remark 35. To see (154) note that

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{S_{n,0}(x)\,S_{n,2}(x) - S_{n,1}^2(x)}{S_{n,0}(x)\,S_{n,2}(x) - S_{n,1}^2(x)} = 1$$

and

$$\begin{aligned}\sum_{i=1}^{n} w_{ni}(x)(X_i - x) &= \frac{\sum_{i=1}^{n} \frac{1}{nh_n} K\big(\frac{X_i-x}{h_n}\big)(X_i - x)S_{n,2}(x) - \frac{1}{nh_n^2}\sum_{i=1}^{n} K\big(\frac{X_i-x}{h_n}\big)(X_i - x)^2}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} \\ &= \frac{S_{n,1}(x)\,S_{n,2}(x) - S_{n,2}(x)\,S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} = 0.\end{aligned}$$

Theorem 20. *Suppose that the assumptions of Theorem 19 hold. Then*

$$\mathsf{bias}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2K}\,\frac{m''(x)}{2} + o_P(h_n^2), \tag{155}$$

$$\mathsf{var}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big), \tag{156}$$

where $R(K)$ and μ_{2K} are given in (146).

Note that by Theorem 19 for the Nadaraya-Watson estimator one has

$$\begin{aligned}\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) &= h_n^2\,\mu_{2K}\left(\frac{m'(x)\,f_X'(x)}{f_X(x)} + \frac{m''(x)}{2}\right) + o_P(h_n^2), \\ \mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) &= \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big).\end{aligned}$$

It is worth noting that the main terms in the approximation of the conditional variances of $\widehat{m}_{NW}(x)$ and $\widehat{m}_{LL}(x)$, i.e.

$$\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = \mathsf{var}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) + o_P\big(\tfrac{1}{n\,h_n}\big).$$

Also the conditional biases are of the same order. But the conditional bias of $\widehat{m}_{LL}(x)$ in comparison to $\widehat{m}_{NW}(x)$ has 'a simple structure', as it does not contain the term $h_n^2\,\mu_{2K}\,\frac{m'(x)\,f_X'(x)}{f_X(x)}$. This is the reason why usually the authors usually prefer $\widehat{m}_{LL}(x)$ to $\widehat{m}_{NW}(x)$.

Proof of Theorem 20. Showing (155). Completely analogously as in the proof of Theorem 19 one can arrive at (147) with the only difference that now the weights $w_{ni}(x)$ are given by (153). Now with the help of (154)

$$A_n = \sum_{i=1}^{n} w_{ni}(x)(X_i - x) = 0. \tag{157}$$

Further using (140), (141), (142) and (143)

$$\begin{aligned} B_n &= \sum_{i=1}^{n} w_{ni}(x) \tfrac{(X_i-x)^2}{h_n^2}\, h_n^2 = h_n^2\, \frac{S_{n,2}^2(x) - S_{n,3}(x)S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} \\ &= h_n^2\, \frac{\big[f_X(x) \int t^2 K(t)\,\mathrm{d}t + o_P(1)\big]^2 - o_P(1)o_P(1)}{(f_X(x) + o_P(1))\,\big[f_X(x) \int t^2 K(t)\,\mathrm{d}t + o_P(1)\big] - (o_P(1))^2} \\ &= h_n^2\, \mu_{2K} + o_P(h_n^2). \end{aligned} \tag{158}$$

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 (140)–(142) one gets

$$D_n = f_X^2(x)\, \mu_{2K}^2 + o_P(1). \tag{159}$$

Now with the help (159) and Lemma 8 one can bound

$$\begin{aligned} |C_n| &\le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)|\, h_n^2 \sum_{i=1}^{n} |w_{ni}(x)|\, \tfrac{(X_i-x)^2}{h_n^2} \\ &\le h_n^2\, o(1)\, \frac{S_{n,2}^2(x) + |S_{n,1}(x)| \sum_{i=1}^{n} \frac{1}{nh_n} K\big(\frac{X_i-x}{h_n}\big) \big|\frac{X_i-x}{h_n}\big|^3}{|D_n(x)|} \\ &= o(h_n^2)\, \frac{f_X^2(x)\mu_{2K}^2 + o_P(1) + o_P(1)\,\big[f_X(x) \int K(t)\,|t|^3\,\mathrm{d}t + o_P(1)\big]}{f_X^2(x)\,\mu_{2K} + o_P(1)} = o_P(h_n^2), \end{aligned}$$

which together with (148), (157) and (158) yields (155).

Showing (156). With the help of (141), (142), (152) and (159) one can calculate

$$\begin{aligned} \mathsf{var}[\widehat{m}_{LL}(x)|\mathbb{X}] &= \sum_{i=1}^{n} w_{ni}^2(x)\sigma^2(X_i) \\ &= \frac{1}{D_n^2(x)} \left[\frac{1}{n^2h_n^2} \sum_{i=1}^{n} K^2\big(\tfrac{X_i-x}{h_n}\big) \Big(S_{n,2}(x) - \tfrac{X_i-x}{h_n} S_{n,1}(x)\Big)^2 \sigma^2(X_i)\right] \\ &= \frac{1}{nh_n} \frac{1}{D_n^2(x)} \big[S_{n,2}^2(x) + o_P(1)\big]\, \frac{1}{nh_n} \sum_{i=1}^{n} K^2\big(\tfrac{X_i-x}{h_n}\big)\sigma^2(X_i) \\ &= \frac{1}{nh_n} \frac{1}{f_X^4(x)\,\mu_{2K}^2 + o_P(1)} \big[f_X^2(x)\,\mu_{2K}^2 + o_P(1)\big]\,\big[f_X(x)\,\sigma^2(x)\,R(K) + o_P(1)\big], \end{aligned}$$

which implies (156). □

10.4 Locally polynomial regression (general p)*

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

$$\left(\mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)\, \mathbb{X}_p(x)\right)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)$$

and satisfy that

$$\sum_{i=1}^{n} w_{ni}(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} w_{ni}(x)(X_i - x)^{\ell} = 0, \quad \ell = 1, \ldots, p.$$

With the help of this property one can show (analogously as in Theorems 19 and 20) that if p is **even** then the conditional 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 $\widehat{m}_p(x)$.

Further, it can be proved that conditional variances are of the same order for each p and it holds

$$\mathsf{var}\left(\widehat{m}_p(x)|\mathbb{X}\right) = \frac{V_p\, \sigma^2(x)}{f_X(x)\, n\, h_n} + o_P\left(\tfrac{1}{n\, h_n}\right),$$

where $V_0 = V_1 < V_2 = V_3 < V_4 = V_5 < \ldots$ and so on.

To sum it up, for p **even** increasing the order of polynomial to $p + 1$ does not increase the asymptotic variance but it has a potential to reduce the bias. On the other hand if p is **odd** then increasing the order of polynomial to $p + 1$ increases the asymptotic variance.

That is why in practice usually odd choices of p are preferred.

Literature: Fan and Gijbels (1996) Chapters 3.1 and 3.2.1.

10.5 Bandwidth selection

10.5.1 Asymptotically optimal bandwidths

In what follows we will consider $p = 1$. With the help of Theorem 20 one can approximate the conditional MSE (mean squared error) of $\widehat{m}_{LL}(x)$ as

$$\mathrm{MSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\, h_n} \tfrac{\sigma^2(x)\, R(K)}{f_X(x)} + \tfrac{1}{4}\, h_n^4\, [m''(x)]^2 \mu_{2K}^2 + o_P\big(\tfrac{1}{n\, h_n}\big) + o_P\big(h_n^4\big). \tag{160}$$

Ignoring the remainder $o_P(.)$ terms in (160), we get that AMSE (asymptotic mean squared error) of $\widehat{m}_{LL}(x)$ is given by

$$\mathrm{AMSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\, h_n} \tfrac{\sigma^2(x)\, R(K)}{f_X(x)} + \tfrac{1}{4}\, h_n^4\, [m''(x)]^2 \mu_{2K}^2. \tag{161}$$

* Not done at the lecture.

Minimising (161) one gets asymptotically optimal *local bandwidth* (i.e. bandwidth that minimises the AMSE)

$$h_n^{(opt)}(x) = n^{-1/5} \left[\frac{\sigma^2(x)\, R(K)}{f_X(x)\, [m''(x)]^2\, \mu_{2K}^2} \right]^{1/5}.$$

The integrated mean squared error (MISE) is usually defined as

$$\mathrm{MISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) = \int \mathrm{MSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \tag{162}$$

where $w_0(x)$ is a given weight function which is introduced in order to guarantee that the integral is hopefully finite (for instance $w_0(x) = \mathbb{I}\{x \in [a,b]\}$).

Now with the help of (161) and (162) the asymptotic integrated mean squared error (AMISE) is defined as

$$\begin{aligned} \mathrm{AMISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) &= \int \mathrm{AMSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \\ &= \frac{R(K)}{n\, h_n} \int \sigma^2(x)\, w_0(x)\, \mathrm{d}x + \tfrac{1}{4}\, h_n^4\, \mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x. \end{aligned} \tag{163}$$

Minimising (163) one gets asymptotically optimal *global bandwidth* (i.e. the bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K) \int \sigma^2(x)\, w_0(x)\, \mathrm{d}x}{\mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}. \tag{164}$$

10.5.2 Rule of thumb for bandwidth selection

Suppose that $\sigma(x)$ is constant. Then the asymptotically optimal global bandwidth (164) is for $\widehat{m}_{LL}$ given by

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K)\, \sigma^2 \int w_0(x)\, \mathrm{d}x}{\mu_{2K}^2 \int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}.$$

Now let $\widetilde{m}(x)$ be an estimated mean function fitted by the (global) polynomial regression of order 4 (generally $p+3$ is recommended) through the standard least squares method.

Now in (164) 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 $\int [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x = \mathsf{E}_X [m''(X_1)]^2\, w_0(X_1)$, which can be estimated by

$$\frac{1}{n} \sum_{i=1}^n [\widetilde{m}''(X_i)]^2\, w_0(X_i).$$

This results in the bandwidth selector

$$h_n^{(ROT)} = n^{-1/5} \left[\frac{R(K)\, \widetilde{\sigma}^2 \int w_0(x)\, \mathrm{d}x}{\mu_{2K}^2\, \frac{1}{n} \sum_{i=1}^n [\widetilde{m}''(X_i)]^2\, w_0(X_i)} \right]^{1/5}.$$

10.5.3 Cross-validation

$$h_n^{(CV)} = \arg\min_{h_n>0} \mathcal{CV}(h_n),$$

where

$$\mathcal{CV}(h_n) = \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - \widehat{m}_p^{(-i)}(X_i)\right]^2 w_0(X_i)$$

with $\widehat{m}_p^{(-i)}$ being the estimator based on a sample that leaves out the i-th observation.

The rationale of the above procedure is that one aims at minimising the estimated integrated squared error, i.e.

$$\begin{aligned} \mathrm{ISE}\big(\widehat{m}_p(x)\big) &= \int \big(\widehat{m}_p(x) - m(x)\big)^2 f_X(x)\, w_0(x)\, \mathrm{d}x \\ &= \mathsf{E}_{X'} \big(\widehat{m}_p(X') - m(X')\big)^2 w_0(X'), \end{aligned} \tag{165}$$

where X' is independent of observations $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$.

To illustrate that put $\varepsilon_i = Y_i - m(X_i)$ and calculate

$$\begin{aligned} \mathcal{CV}(h_n) &= \frac{1}{n}\sum_{i=1}^{n} \Big[\varepsilon_i + m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w_0(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, w_0(X_i) + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big] w_0(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \Big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\Big]^2 w_0(X_i). \end{aligned}$$

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 $\frac{1}{n}\sum_{i=1}^{n} \big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big]^2 w_0(X_i)$ can be considered as a reasonable estimate of (165).

Finally $\frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big] w_0(X_i)$ does not 'bias' the estimate of (165), as

$$\begin{aligned} \mathsf{E}\left[\varepsilon_i \big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big] w_0(X_i)\right] &= \mathsf{E}\left\{\mathsf{E}\left[\varepsilon_i \big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big] w_0(X_i) \,|\, \mathbb{X}\right]\right\} \\ &= \mathsf{E}\left\{\mathsf{E}\,[\varepsilon_i | X_i]\ \mathsf{E}\left[\big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big] w_0(X_i) \,|\, \mathbb{X}\right]\right\} = 0, \end{aligned}$$

where we have used that $\mathsf{E}\,[\varepsilon_i | X_i] = 0$ and that ε_i and $\big[m(X_i) - \widehat{m}_p^{(-i)}(X_i)\big] w_0(X_i)$ are independent conditionally on X_i (and thus also conditionally on $\mathbb{X}$).

Remark 36. Note that it would not make much sense to search for h_n that minimises the residual sum of squares. $RSS(h_n) = \frac{1}{n}\sum_{i=1}^{n} \big[Y_i - \widehat{m}(X_i)\big]^2 w_0(X_i)$. The reason is that $RSS(h_n)$ is minimised if $Y_i = \widehat{m}(X_i)$, which would result in a very low bandwidth h_n.

Remark 37. Another view of the cross-validation procedure is that we aim at finding the bandwidth h_n that minimizes the prediction error. More precisely, suppose that $\binom{X'}{Y'}$ is a random vector that has the same distribution as $\binom{X_1}{Y_1}$ and that is independent with our random sample $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$. Then the prediction error (viewed as a function of h_n) is given by

$$\mathcal{R}(h_n) = \mathsf{E}_{X',Y'} \left(Y' - \widehat{m}_p(X')\right)^2 w(X'),$$

where the expectation is taken only with respect to the random vector $\binom{X'}{Y'}$. Now $\mathcal{CV}(h_n)$ presents a natural estimator of $\mathcal{R}(h_n)$ as $\binom{X_i}{Y_i}$ is independent of $\widehat{m}_p^{(-i)}$.

10.5.4 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 $|X_i - x| \leq h_n$. 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

$$h_n^{(NN)}(x) = d_{(k)}(x). \tag{166}$$

Note that (166) presents **a local bandwidth choice**.

To get an insight into the bandwidth choice (166) let us approximate

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{|X_i - x| \leq h\} \doteq \widehat{F}_n(x+h) - \widehat{F}_n(x-h) \doteq F_X(x+h) - F_X(x-h) \doteq f_X(x) 2h. \tag{167}$$

By plugging $h = d_{(k)}(x) = h_n(x)$ into (167) one gets $\frac{k}{n} \doteq f_X(x) 2h_n(x)$ which further implies that

$$h_n^{(NN)}(x) \doteq \frac{k}{2n f_X(x)}.$$

Remark 38. To derive the asymptotic properties of $\widehat{m}_{LL}$ when bandwidth h_n is chosen as (166) one needs to consider $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$.

Remark 39. Using $h_n^{(NN)}(x)$ makes usually the problem more computational intensive as one is using a local bandwidth. Further there is no guarantee that the estimator $\widehat{m}_p(x)$ is continuous even if K is continuous. To prevent those difficulite some authors recommend to transform the covariates as

$$X_i' = \widehat{F}_n(X_i), \quad i = 1, \ldots, n,$$

where $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \le x\}$ is the empirical distribution function of the covariates. Then the transformed covariates are 'approximately uniformly spread' on $(0,1)$* and one can use a global bandwidth choice (e.g. by the cross-validation procedure described in Section 10.5.3). As F_n is a consistent estimator of F_X one should keep in mind that when using the transformed covariates X_i' one estimates

$$\mathsf{E}\left[Y \,|F_X(X) = x\right] = \mathsf{E}\left[Y \,|\, X = F_X^{-1}(x)\right] = m\big(F_X^{-1}(x)\big).$$

10.6 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, $K(t) = \frac{70}{81}\big(1 - |t|^3\big)^3\,\mathbb{I}\{|t| \le 1\}$, 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'

$$\delta_i = B\Big(\frac{r_i}{6\,\mathsf{med}\left(|r_1|,\ldots,|r_n|\right)}\Big), \qquad i = 1, \ldots, n,$$

where $B(t) = (1-t^2)^2\,\mathbb{I}\{|t| \le 1\}$. 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

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \operatorname*{arg\,min}_{b_0,b_1} \sum_{i=1}^n \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big)\,\delta_i.$$

By default there are 3 iterations.

10.7 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

$$\widehat{\sigma}_n^2(x) = \sum_{i=1}^n w_{ni}(x)\,Y_i^2 - \widehat{m}_p^2(x), \tag{168}$$

* Note that in case there are no ties in covariate values one gets $\{X_1', \ldots, X_n'\} = \{\frac{1}{n}, \ldots, \frac{n}{n}\}$.

where $\widehat{m}_p(x) = \sum_{i=1}^n w_{ni}(x)\, Y_i$ is an estimator of $m(x) = \mathsf{E}\left[Y_1 \,|\, X_1 = x\right]$. 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

$$\widetilde{\sigma}_n^2(x) = \sum_{i=1}^n w_{ni}(x)\big(Y_i - \widehat{m}_p(X_i)\big)^2. \tag{169}$$

Note that if the weights $w_{ni}(x)$ are not guaranteed to be non-negative, then there is generally no guarantee that either of the estimators (168) or (169) is positive.

Literature: Fan and Gijbels (1996) Chapters 2.4.1, 3.2.3, 4.2, 4.10.1, 4.10.2.

The end of the self study for the week (11. 5. - 15. 5. 2020)

Appendix

Inverse function theorem

The following theorem is sometimes also called the theorem about the local diffeomorphism. It follows easily from the implicit function theorem applied to the function $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{f}(\mathbf{y})$.

Theorem A1. *Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ have continuous first order partial derivatives in a neighbourhood of the point $\mathbf{a} \in \mathbb{R}^n$ and the Jacobi matrix $\mathbb{D}_{\mathbf{f}}(\mathbf{a})$ is a non-singular matrix. Then there exist open neighbourhoods U of the point $\mathbf{a}$ and V of the point $\mathbf{f}(\mathbf{a})$ such that $\mathbf{f}$ is a bijection of U on V. Further there exists an inverse function $\mathbf{f}^{-1}$ on V with continuous first order partial derivatives.*

Banach fixed point theorem

Definition. Let (P, ρ) be a metric space. Then a map $T : P \mapsto P$ is called a *contraction mapping* on P if there exists $q \in [0, 1)$ such that for all $x, y \in P$

$$\rho\big(T(x), T(y)\big) \leq q\rho(x, y).$$

Theorem A2. *Let (P, ρ) be a non-empty complete metric space with a contraction mapping $T : P \mapsto P$. Then T admits a unique fixed-point $x^* \in P$ (i.e. $T(x^*) = x^*$).*

Uniform consistency of the empirical distribution function

The following theorem can be found for instance in Section 2.1.4 of Serfling (1980) as Theorem A.

Theorem A3. *(**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 empirical cumulative distribution function F. Let*

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

be the cumulative empirical distribution function. Then

$$\sup_{\mathbf{x} \in \mathbb{R}^k} \left|\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})\right| \xrightarrow[n \to \infty]{a.s.} 0.$$

Bayes theorem for densities

Theorem A4. *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_k)^\mathsf{T}$ and $\mathbf{Z} = (Z_1, \ldots, Z_G)^\mathsf{T}$ be random vectors defined on the same probability space. Let $f_{\boldsymbol{X}}$ and $f_{\mathbf{Z}}$ be the densities of $\boldsymbol{X}$ and $\mathbf{Z}$ respectively*

and $f_{\boldsymbol{X}|\mathbf{Z}}$ *be the conditional density of* $\boldsymbol{X}$ *given* $\mathbf{Z}$*. Then the conditional density of* $\mathbf{Z}$ *given* $\boldsymbol{X}$ *equals*

$$f_{\mathbf{Z}|\boldsymbol{X}}(\mathbf{z}|\mathbf{x}) = \begin{cases} \frac{f_{\boldsymbol{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})}{f_{\boldsymbol{X}}(\mathbf{x})}, & \text{for } f_{\boldsymbol{X}}(\mathbf{x}) > 0, \\ 0, & \text{for } f_{\boldsymbol{X}}(\mathbf{x}) = 0. \end{cases}$$

Proof. The proof follows from the fact that $f_{\boldsymbol{X},\mathbf{Z}}(\mathbf{x},\mathbf{z}) = f_{\boldsymbol{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$ is the joint density of $\binom{\boldsymbol{X}}{\mathbf{Z}}$ and then by the definition of the conditional density. For details see e.g. Chapter 3.5 of Anděl (2007). □

References

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.

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.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, 142(1):201–211.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer, 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*, 48:817–838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103.

Zvára, K. (2008). *Regrese*. MATFYZPRESS.