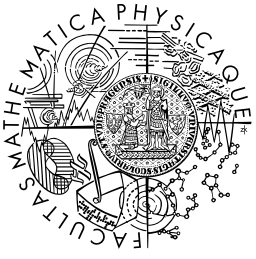

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Mathematical Statistics 4

NMST 545

Course notes

Marek Omelka and Stanislav Nagy

Last updated: January 3, 2025

Contents

In several places in this course, using the stochastic o_P and O_P notation will be essential. Those not familiar with these concepts can find definitions and a few basic algebraic rules for these symbols in the Appendix, Definition A9.

1 Bootstrap and other resampling methods

Suppose we observe independent and identically distributed k-dimensional random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution F_X and let $\boldsymbol{\theta}_X = \boldsymbol{\theta}(F_X)$ be the quantity of interest. Let $\mathbf{R}_n = \mathbf{g}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\theta}_X)$ be a p-dimensional random vector that we want to use for doing inference about $\boldsymbol{\theta}_X$, e.g.

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

where $\widehat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_0$ is a known value.

To infer about parameter $\boldsymbol{\theta}$, one needs to know the distribution of $\mathbf{R}_n$. Usually, we are not able to derive the exact distribution of $\mathbf{R}_n$ analytically. For instance consider the distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)$, where $\widehat{\boldsymbol{\theta}}_n$ is a maximum likelihood estimator for n fixed. In such situations, the inference is often based on the asymptotic distribution of $\mathbf{R}_n$. For example by Nagy (2023a, Theorem 25) or Omelka (2023, Theorem 5), for a maximum likelihood estimator in regular models one has $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\left(\mathbf{0}_p, I^{-1}(\boldsymbol{\theta}_X)\right)$, for $I(\boldsymbol{\theta}_X)$ the Fisher information matrix of $\boldsymbol{\theta}_X$. 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'*.

1.1 Monte Carlo principle

Sometimes one knows the distribution of $\boldsymbol{X}_i$ and thus also of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ so one is (at least theoretically) able to derive the distribution of $\mathbf{R}_n = (R_{n,1}, \ldots, R_{n,p})^{\mathsf{T}}$. 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 Nagy (2023a, Section 3.3.1) or Omelka (2023, Chapter 2.4) when the null hypothesis holds.

If one knows the distribution of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, then one can generate random vectors $\mathbb{X}^*$, which have the same distribution as $\mathbb{X}$. The Monte Carlo principle runs as follows.

- Choose B sufficiently large and for each $b \in \{1, \ldots, B\}$ independently generate random samples $\mathbb{X}_b^* = (\boldsymbol{X}_{1,b}^*, \ldots, \boldsymbol{X}_{n,b}^*)$ such that the distribution of $\mathbb{X}_b^*$ is the same as the distribution of $\mathbb{X}$. We get B independent samples $\mathbb{X}_1^*, \ldots, \mathbb{X}_B^*$.

- For each b, compute $\mathbf{R}_{n,b}^*$ as the quantity $\mathbf{R}_n$ calculated from the b-th sample $\mathbb{X}_b^*$.
- The unknown distribution function

$$H_n(\mathbf{x}) = \mathsf{P}\left(\mathbf{R}_n \leq \mathbf{x}\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^p$$

of $\mathbf{R}_n$ can now be estimated by the empirical distribution function of $\mathbf{R}_{n,b}^*$, $b = 1, \ldots, B$

$$\widehat{H}_{n,B}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{\mathbf{R}_{n,b}^* \leq \mathbf{x}\} \quad \text{for } \mathbf{x} \in \mathbb{R}^p.$$

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 A10) implies

$$\sup_{\mathbf{x}\in\mathbb{R}^p} \left|\widehat{H}_{n,B}(\mathbf{x}) - H_n(\mathbf{x})\right| \xrightarrow[B\to\infty]{\text{a.s.}} 0. \tag{1}$$

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 (1) 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. We only need that **we can generate independent copies of $\mathbf{R}_n$**. Also, it is interesting to see that in (1), we do not estimate the asymptotic distribution of $\mathbf{R}_n$ as $n \to \infty$, but rather the exact, finite sample distribution with the sample size n fixed.

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 1. * Consider the Neyman-Pearson test (Nagy, 2023a, Section 3.1), or any likelihood-based test introduced in Nagy (2023a, Section 3.3.1) or Omelka (2023, Chapter 2.4).

* Examples designated by R are accompanied by R codes.

We test the null hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$, for $\boldsymbol{\theta}_0$ given. The test statistic R_n is one-dimensional and explicitly given, but its exact distribution under H_0 does not have to be simple to determine. Under H_0, however, we know that $R_n = g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\theta}_0)$ for $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ a random sample from distribution with parameter $\boldsymbol{\theta}_0$, which is completely specified. The significance of the test statistic R_n can thus be directly assessed using the Monte Carlo principle.

Example 2. We observe a random vector with a multinomial distribution $\mathsf{M}_K(n; p_1, \ldots, p_K)$. Denote $\mathbf{p} = (p_1, \ldots, p_K)^\mathsf{T}$ and let $\mathbf{p}_X$ be the true value of the parameter $\mathbf{p}$. 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 $\mathbf{p}^{(0)} = (p_1^{(0)}, \ldots, p_K^{(0)})^\mathsf{T}$ 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.

The Monte Carlo principle does not have to be used only if the distribution of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is completely specified. In the following examples, we will utilise that, in fact, it is not necessary to know the data-generating mechanism of $\mathbb{X}$ exactly, provided we can generate independent copies of $\mathbf{R}_n$.

Example 3. 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. We are interested in testing the hypothesis

$$H_0 : \rho_X = \rho_0, \qquad \text{vs.} \qquad H_1 : \rho_X \neq \rho_0, \tag{2}$$

for $\rho_0 \in (-1, 1)$ given. Our intention is to use the test statistic $R_n = \sqrt{n}\,(\widehat{\rho}_n - \rho_0)$ for $\widehat{\rho}_n$ the sample correlation coefficient, see also Kulich and Omelka (2022, Section 10.1.2). A direct use of a Monte Carlo approach to assess the significance of R_n seems impossible, since we deal with a problem of testing with nuisance parameters (means and variances of the marginal variables X_i and Y_i). For simulating Monte Carlo replicates from R_n, one would thus need to choose the values of these nuisance parameters for sampling $\mathbb{X}^*$, which could affect the distribution of R_n.

The use of Monte Carlo is, however, still possible. For every $a, b, c, d \in \mathbb{R}$, $a\,c \neq 0$ we have for the correlation coefficient $\rho(X, Y)$ between random variables X and Y that

$$\rho(a\,X + b, c\,Y + d) = \operatorname{sgn}(a\,c)\rho(X, Y).$$

From this expression it is easy to see that both the correlation coefficient and the distribution of the sample correlation coefficient depend only on the single parameter ρ_X of the bivariate

normal distribution. Thus, one should be able (at least theoretically) to calculate the distribution of R_n when the null hypothesis holds. But this distribution is rather complicated.*

The same observation, however, shows that also the distribution of R_n depends only on ρ_X, and it is the same for any choice of the nuisance parameters. Thus, when generating random variables $(Y_1^*, X_1^*)^\mathsf{T}, \ldots, (Y_n^*, X_n^*)^\mathsf{T}$, one can choose any values of the nuisance parameters, as long as $\rho(X_i^*, Y_i^*) = \rho_X$ for all i. The resulting distribution of $R_n^* = \sqrt{n}\,(\widehat{\rho}_n^* - \rho_0)$ has, under the null hypothesis, necessarily the same distribution as R_n. Think how this can be used to calculate (estimate) the p-value of the test of the hypothesis (2).

Example 4. Let $\mathbb{X} = (X_1, \ldots, X_n)$ be a random sample from the distribution F_X in $\mathbb{R}$. We want to test the hypothesis

$$H_0 : F_X(x) = F_0(x), \forall x \in \mathbb{R}, \quad \text{vs.} \quad H_1 : \exists x \in \mathbb{R} \quad F_X(x) \neq F_0(x)$$

for F_0 a given distribution function, using the Kolmogorov-Smirnov test statistic

$$R_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|.$$

Here, $\widehat{F}_n$ is the empirical cumulative distribution function of $\mathbb{X}$. Compared to the classical approach based on the asymptotic distribution of the statistic R_n (Kulich and Omelka, 2022, Section 5.1), Monte Carlo approximation has two major advantages: (i) the approximation is non-asymptotic, and works well also for small sample size n, and (ii) does not require the assumption of continuity of F_0, i.e. works well also for discrete distributions. Furthermore, using a Monte Carlo approach, analogous goodness-of-fit tests can be considered also in the setup of multivariate distributions in $\mathbb{R}^k$.

Example 5. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from 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. We want to use the Monte Carlo principle to test the following hypothesis

$$H_0 : \lambda_X = \lambda_Y, \quad \text{vs.} \quad H_1 : \lambda_X \neq \lambda_Y. \tag{3}$$

Again, we deal with the problem of a nuisance parameter, as under H_0 there is still the common parameter $\lambda_X = \lambda_Y$ to be specified.

We base the test on the best unbiased (and efficient) point estimators $\widehat{\lambda}_{X,n_1} = \left(\bar{X}_{n_1}\right)^{-1}$ and $\widehat{\lambda}_{Y,n_2} = \left(\bar{Y}_{n_2}\right)^{-1}$. There are several test statistics to consider. The first to consider might assess the difference $R_{1,n_1,n_2} = \widehat{\lambda}_{X,n_1} - \widehat{\lambda}_{Y,n_2}$. Take the distribution of R_{1,n_1,n_2} under

* https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Using_the_exact_distribution

H_0, given that the common value of the parameter is $\lambda = \lambda_X = \lambda_Y$. We know that for X with exponential distribution with density $f(x;\lambda)$, the distribution of λX is the standard exponential with density $f(x;1)$. Thus,

$$R_{1,n_1,n_2} = \widehat{\lambda}_{X,n_1} - \widehat{\lambda}_{Y,n_2} = \frac{n_1}{\sum_{i=1}^{n_1} X_i} - \frac{n_2}{\sum_{i=1}^{n_2} Y_i} = \lambda \left(\frac{n_1}{\sum_{i=1}^{n_1} \lambda X_i} - \frac{n_2}{\sum_{i=1}^{n_2} \lambda Y_i} \right).$$

On the right-hand side, the expression in the brackets has a distribution that depends only on the sample sizes n_1 and n_2. But, the factor λ in front of the brackets makes the distribution of R_{1,n_1,n_2} depend also on λ. The distribution of the test statistic thus depends on a nuisance parameter, and a Monte Carlo approach is not appropriate.

A second choice of a test statistic might be the ratio $R_{2,n_1,n_2} = \widehat{\lambda}_{X,n_1} / \widehat{\lambda}_{Y,n_2}$. We have

$$R_{2,n_1,n_2} = \frac{\bar{Y}_{n_2}}{\bar{X}_{n_1}} = \frac{n_1}{n_2} \frac{\sum_{i=1}^{n_2} Y_i}{\sum_{i=1}^{n_1} X_i} = \frac{n_1}{n_2} \frac{\sum_{i=1}^{n_2} \lambda Y_i}{\sum_{i=1}^{n_1} \lambda X_i},$$

and again because all $\lambda X_1, \ldots, \lambda X_{n_1}, \lambda Y_1, \ldots, \lambda Y_{n_2}$ are independent with standard exponential distribution, the distribution of R_{2,n_1,n_2} does not depend on the unknown λ. The test statistic R_{2,n_1,n_2} is thus under H_0 pivotal, and can be used for a Monte Carlo test of the hypothesis (3), similarly as we did in Example 3.

Application to confidence intervals

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

$$\widehat{H}_{n,B}^{-1}(u) \xrightarrow[B\to\infty]{\text{a.s.}} H_n^{-1}(u),$$

provided that H_n is continuous and increasing in u.* Thus one can use the quantile $\widehat{H}_{n,B}^{-1}(u)$ as an estimate (approximation) of the quantile $H_n^{-1}(u)$.

Let $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,p}\right)^{\mathsf{T}}$ be an estimator of $\boldsymbol{\theta}_X = \left(\theta_{X,1}, \ldots, \theta_{X,p}\right)^{\mathsf{T}}$ and $\mathbf{R}_n = \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X$. Suppose the distribution of $\mathbf{R}_n$ does not depend on the parameter $\boldsymbol{\theta}_X$, i.e. that one is able to generate random variables $\mathbf{R}_n^*$ with the same distribution as $\mathbf{R}_n$. Further, suppose that we want to find the confidence interval for $\theta_{X,k}$ (the k-th component of $\boldsymbol{\theta}_X$). Denote H_n the distribution function of $\widehat{\theta}_{n,k} - \theta_{X,k}$ and $\widehat{H}_{n,B}$ 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

$$\lim_{B\to\infty} \mathsf{P}\left(\widehat{H}_{n,B}^{-1}(\alpha/2) < \widehat{\theta}_{n,k} - \theta_{X,k} < \widehat{H}_{n,B}^{-1}(1-\alpha/2) \right) = 1 - \alpha.$$

Thus the approximate confidence interval for $\theta_{X,k}$ can be calculated as

$$\left(\widehat{\theta}_{n,k} - \widehat{H}_{n,B}^{-1}(1-\alpha/2), \widehat{\theta}_{n,k} - \widehat{H}_{n,B}^{-1}(\alpha/2) \right).$$

* 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. the main theorem of Serfling (1980, Section 2.3.1).

Observe that on the left-hand side of this interval is the upper sample quantile $\widehat{H}_{n,B}^{-1}(1-\alpha/2)$, and on the right-hand side the lower sample quantile $\widehat{H}_{n,B}^{-1}(\alpha/2)$.

Example 6. Let $X_1, \ldots, X_n$ be a random sample from a distribution F_X that belongs to a location family, i.e.

$$F_X \in \mathcal{F} = \big\{F(\cdot - \theta), \theta \in \mathbb{R}\big\}, \tag{4}$$

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 $R_n = \widehat{\theta}_n - \theta_X = \widehat{\theta}_n(X_1, \ldots, X_n) - \theta_X = \widehat{\theta}_n(X_1 - \theta_X, \ldots, X_n - \theta_X)$ depends only on the distribution of $X_i - \theta_X$, $i = 1, \ldots, n$. We have

$$\mathsf{P}\left(X_i - \theta_X \leq x\right) = \mathsf{P}\left(X_i \leq x + \theta_X\right) = F(x + \theta_X - \theta_X) = F(x) \quad \text{for all } x \in \mathbb{R}.$$

Thus, the distribution of all $X_i - \theta_X$, and consequently also the distribution of R_n, depends only on the known function F but it does not depend on θ_X. The distribution of R_n can be thus approximated using a Monte Carlo procedure by simulating from the distribution with a given θ_0 (e.g. $\theta_0 = 0$) and calculating $R_{n,b}^* = \widehat{\theta}_n^* - \theta_0$, in the same way as we did in Example 3.

Use this approach to find a Monte Carlo confidence interval for

- parameter $\theta \in \mathbb{R}$ given a random sample $X_1, \ldots, X_n$ from the logistic distribution with density
$$f(x) = \frac{\exp\{-(x-\theta)\}}{\big(1 + \exp\{-(x-\theta)\}\big)^2}, \quad x \in \mathbb{R},$$
- the median $\theta \in \mathbb{R}$ of F_X, given a random sample $X_1, \ldots, X_n$ from a location family (4) with F having a symmetric density.

The end of lecture 1 (2.10.2024)

1.2 Standard nonparametric bootstrap

In the Monte Carlo principle, we leveraged the fact that the data-generating process of $\mathbf{R}_n$ was known completely, and we were able to sample from the distribution of $\mathbf{R}_n$ directly. This is in practice quite rare, and Monte Carlo per se is thus of relatively limited interest.

A generalisation of the Monte Carlo principle is the bootstrap. In that case, instead of knowing the distribution of $\mathbb{X}$ precisely, we estimate it, and sample $\mathbb{X}_b^*$ from that estimated distribution. Depending on whether this distribution is estimated parametrically or nonparametrically, we distinguish parametric or nonparametric bootstrap.

Throughout this section, we suppose that 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 presentation purposes, it will be instructive to write the estimator as $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\widehat{F}_n)$, with $\widehat{F}_n$ the empirical distribution

$$\widehat{F}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \leq \mathbf{x}\} \quad \text{for } \mathbf{x} \in \mathbb{R}^k.$$

We are interested in the distribution of a p-dimensional random vector

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

where $\mathbf{g}_n$ is a deterministic and known function that depends only on n.

1.2.1 Idea of nonparametric bootstrap

In nonparametric bootstrap* the unknown F_X is estimated by the empirical distribution function $\widehat{F}_n$. The empirical distribution puts mass $1/n$ to each observation $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus, generating independent random vectors $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ from $\widehat{F}_n$ 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.

Our intention is to approximate/estimate the unknown distribution function H_n of $\mathbf{R}_n$, i.e.

$$H_n(\mathbf{x}) = \mathsf{P}\left(\mathbf{R}_n \leq \mathbf{x}\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^p. \tag{5}$$

The nonparametric bootstrap algorithm runs as follows.

(i) Choose B sufficiently large. 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) using a simple random sample with replacement from $\mathbb{X}$.

(ii) 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 the empirical distribution function based on $\mathbb{X}_b^*$.

(iii) The distribution function H_n of $\mathbf{R}_n$ is now (by the combination of the Monte Carlo and plug-in principle) estimated by

$$\widehat{H}_{n,B}^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big\{\mathbf{R}_{n,b}^* \leq \mathbf{x}\big\} \quad \text{for } \mathbf{x} \in \mathbb{R}^p. \tag{6}$$

* *neparametrický bootstrap* † *prostý náhodný výběr s vracením*

It is important to observe that $\widehat{H}^*_{n,B}$ from (6) is, in fact, a two-step approximation of the true distribution function H_n from (5):

- **Plug-in.** First, 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 $\mathbf{R}^*_n$, and conditionally on $\mathbb{X}$ (that is, if the original random sample $\mathbb{X}$ is taken as fixed). In the first approximation of H_n, the distribution of $\mathbf{R}_n$ is approximated by the conditional distribution of $\mathbf{R}^*_n$ given $\mathbb{X}$. Its (conditional) distribution function is

$$\begin{aligned} \widehat{H}_n(\mathbf{x}) &= \mathsf{P}\left(\mathbf{R}^*_n \leq \mathbf{x} \,|\, \mathbb{X}\right) = \mathsf{P}\left(\mathbf{g}_n\big(\boldsymbol{\theta}(\widehat{F}^*_n), \boldsymbol{\theta}(\widehat{F}_n)\big) \leq \mathbf{x} \,|\, \mathbb{X}\right) \\ &= \mathsf{P}\left(\mathbf{g}_n\big(\widehat{\boldsymbol{\theta}}^*_n, \widehat{\boldsymbol{\theta}}_n\big) \leq \mathbf{x} \,|\, \mathbb{X}\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^p. \end{aligned} \tag{7}$$

  That is, for fixed $\mathbb{X}$, $\widehat{H}_n$ is the true distribution of $\mathbf{R}^*_n$ if only the randomness in sampling $\mathbb{X}^*_b$ is involved. Because $\widehat{H}_n$ still depends on the random sample $\mathbb{X}$, it is itself random.

- **Monte Carlo.** In the second approximation of H_n from (5), the distribution function (7) is estimated by the empirical distribution function of the B bootstrap replicates $\mathbf{R}^*_{n,1}, \ldots, \mathbf{R}^*_{n,B}$ from $\mathbf{R}^*_n$, i.e. using $\widehat{H}^*_{n,B}$ from (6).

Because $\mathbf{R}^*_{n,1}, \ldots, \mathbf{R}^*_{n,B}$ are a random sample from the conditional distribution of $\mathbf{R}^*_n$ given $\mathbb{X}$, by the Glivenko-Cantelli Theorem (Theorem A10) we know that

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left|\widehat{H}^*_{n,B}(\mathbf{x}) - \widehat{H}_n(\mathbf{x})\right| \xrightarrow[B\to\infty]{\text{a.s.}} 0,$$

Taking B sufficiently large, we see that $\widehat{H}^*_{n,B}$ can be made arbitrarily close to $\widehat{H}_n$. For this reason, the second approximation (**Monte Carlo**) is always valid, as long as B is large enough. Consequently, in the theory of bootstrap, the second approximation of $\widehat{H}_n$ by $\widehat{H}^*_{n,B}$ is usually ignored, and only the first, **plug-in** approximation (of H_n by $\widehat{H}_n$) is studied. If $\widehat{H}_n$ is a 'good' estimator of H_n, then the nonparametric bootstrap is said to 'work', or to be consistent.

1.2.2 Convergence of conditional distributions

The distribution function $\widehat{H}_n$ depends on the random sample $\mathbb{X}$ and thus it is random, and can be viewed as an 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

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

Suppose that we have a sequence of random vectors $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and $\boldsymbol{Y}$ with distribution functions $G_1, G_2, \ldots$, and G. Lemma A2 given in the Appendix states that if the limiting distribution function G is **continuous**, then ρ_∞ can be used for metrizing the convergence in distribution, meaning that $\rho_\infty(G_n, G) \xrightarrow[n\to\infty]{} 0$ if and only if $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{Y}$.

Recall the random vector $\mathbf{R}_n^*$ whose distribution function $\widehat{H}_n$ is given by (7). We saw that the distribution of $\mathbf{R}_n^*$ depends on (the realisations of our data) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus the distribution $\mathbf{R}_n^*$ is conditionally on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. We need to define a notion of convergence for conditional distributions.

Let $\mathbf{R}$ be a candidate for the limiting random vector $\mathbf{R}_n^*$, and let H be the distribution function of $\mathbf{R}$. Let ρ be a metric on the space of distribution functions that can be used for metrizing weak convergence (for instance the supremum metric ρ_∞ if the limiting distribution is continuous, but in literature other metrics can be found). Since $\widehat{H}_n$ given by (7) depends on $\mathbb{X}$, $\rho(\widehat{H}_n, H)$ is a random variable (also depending on $\mathbb{X}$).

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

$$\rho(\widehat{H}_n, H) \xrightarrow[n\to\infty]{P} 0 \quad \Big(\text{i.e. for each } \varepsilon > 0 \lim_{n\to\infty} \mathsf{P}\left[\omega \in \Omega : \rho\big(\widehat{H}_n(\omega), H\big) \geq \varepsilon\right] = 0\Big).$$

- **conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$** the random variable $\mathbf{R}_n^*$ **converges in distribution** to **$\mathbf{R}$ almost surely** if

$$\rho(\widehat{H}_n, H) \xrightarrow[n\to\infty]{\text{a.s.}} 0 \quad \Big(\text{i.e. } \mathsf{P}\left[\omega \in \Omega : \lim_{n\to\infty} \rho\big(\widehat{H}_n(\omega), H\big) = 0\right] = 1\Big).$$

In the following theorem, we formulate the conditions needed for the nonparametric bootstrap to work.

Theorem 1. *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*

$$\rho_\infty(\widehat{H}_n, H_n) \xrightarrow[n\to\infty]{P} 0 \quad (\textit{or } \xrightarrow[n\to\infty]{a.s.} 0), \tag{B}$$

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, (B), and Lemma A2 we have

$$\rho_\infty(\widehat{H}_n, H) \leq \rho_\infty(\widehat{H}_n, H_n) + \rho_\infty(H_n, H) \xrightarrow[n\to\infty]{P} 0 \quad (\text{or } \xrightarrow[n\to\infty]{\text{a.s.}} 0).$$

This is precisely what we wanted to prove. □

The first thing worth noting in Theorem 1 is the assumption 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 use of normal approximation (to which using bootstrap is an alternative), where we also normalise the random vector so that it asymptotically has a non-degenerate distribution. In our analysis, the assumption of non-degeneracy of $\mathbf{R}$ appears to be a remnant of the use of the supremum norm ρ_∞, but this limitation can be shown to be fundamental. Thus, for the nonparametric bootstrap to work appropriately asymptotically, it must be assumed that the **limiting distribution $\mathbf{R}$ of $\mathbf{R}_n$ is non-degenerate**. Typically we choose $\mathbf{R}_n$ so that it converges to a multivariate normal distribution, thus also the continuity of the limit distribution of $\mathbf{R}$ is satisfied.

The crucial condition (B) from Theorem 1 is precisely what is needed for the first approximation (**plug-in**) of H_n by $\widehat{H}_n$ in nonparametric bootstrap to be valid. Thus, we say that nonparametric bootstrap is consistent (or simply that it 'works') if (B) is true. In what follows, we will explore when this is the case.

The end of lecture 2 (9.10.2024)

1.2.3 Consistent nonparametric bootstrap

In view of Theorem 1, the crucial question is whether the convergence in (B) holds. Our first answer is the next theorem, which states that (B) holds for a sample mean.

Theorem 2. *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent identically distributed random vectors such that $\mathsf{E}\,\|\boldsymbol{X}_1\|^2 < \infty$ and the variance matrix $\mathbb{\Sigma} = \mathsf{var}(\boldsymbol{X}_1)$ is positive definite. Consider $\mathbf{R}_n = \sqrt{n}\left(\overline{\boldsymbol{X}}_n - \mathsf{E}\,\boldsymbol{X}_1\right)$ and $\mathbf{R}_n^* = \sqrt{n}\left(\overline{\boldsymbol{X}}_n^* - \overline{\boldsymbol{X}}_n\right)$. Then*

$$\rho_\infty(\widehat{H}_n, H_n) \xrightarrow[n\to\infty]{a.s.} 0. \tag{8}$$

Proof. By the standard central limit theorem for random vectors, we have $\rho_\infty(H_n, H) \xrightarrow[n\to\infty]{} 0$, for H the distribution function of the k-variate normal distribution $\mathsf{N}_k(\mathbf{0}, \mathbb{\Sigma})$. We can use the triangle inequality for ρ_∞ and write

$$\rho_\infty(\widehat{H}_n, H_n) \le \rho_\infty(\widehat{H}_n, H) + \rho_\infty(H, H_n).$$

As $n \to \infty$, the second summand above vanishes, so it remains to prove

$$\rho_\infty(\widehat{H}_n, H) \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{9}$$

We proceed conditionally on the sequence $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, i.e., the values $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are considered to be fixed, and the only randomness in $\widehat{H}_n$ comes from the bootstrap resampling in $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*$ from the fixed distribution given by $\widehat{F}_n$. This means that the random variables

$\boldsymbol{X}_i^*$ are all independent and identically distributed for $i = 1, \ldots, n$, but have a different distribution $\widehat{F}_n$ for each $n = 1, 2, \ldots$. Thus, it is more appropriate to write $\boldsymbol{X}_{n,i}^*$ to emphasise that the (conditional) distribution of $\boldsymbol{X}_i^*$ depends on n. We want to use a central limit theorem to prove that conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, the quantity $\mathbf{R}_n^*$ converges in distribution to $\mathsf{N}_k(\mathbf{0}, \mathbb{\Sigma})$ almost surely. That would be enough to conclude that also (9) is true.

The conditional mean and variance of $\boldsymbol{X}_{n,i}^*$ are for each $i = 1, \ldots, n$

$$\begin{aligned}
\mathsf{E}\left(\boldsymbol{X}_{n,i}^* \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\right) &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i = \overline{\boldsymbol{X}}_n, \\
\mathsf{var}\left(\boldsymbol{X}_{n,i}^* \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\right) &= \mathsf{E}\left(\left(\boldsymbol{X}_{n,i}^* - \overline{\boldsymbol{X}}_n\right)\left(\boldsymbol{X}_{n,i}^* - \overline{\boldsymbol{X}}_n\right)^{\mathsf{T}} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n\right)^{\mathsf{T}} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{\Sigma},
\end{aligned}$$

the final limit following from the usual strong law of large numbers. Define

$$\boldsymbol{Y}_{n,i} = \frac{\boldsymbol{X}_{n,i}^*}{\sqrt{n}} \quad \text{for } i = 1, \ldots, n.$$

We use the Lindeberg-Feller central limit theorem for the triangular array $\boldsymbol{Y}_{n,i}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$, see Theorem A11 in the Appendix. We have

$$\sum_{i=1}^n \mathsf{var}(\boldsymbol{Y}_{n,i} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots) = \sum_{i=1}^n \frac{\mathsf{var}(\boldsymbol{X}_{n,i}^* \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots)}{n} = \mathsf{var}\left(\boldsymbol{X}_{n,1}^* \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\right) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{\Sigma}.$$

It remains to check the 'Lindeberg-Feller condition' (A108); for $\varepsilon > 0$ we have

$$\begin{aligned}
\sum_{i=1}^n \mathsf{E}^*\left[\|\boldsymbol{Y}_{n,i}\|^2 \mathbb{I}\{\|\boldsymbol{Y}_{n,i}\| > \varepsilon\}\right] &= \frac{1}{n}\sum_{i=1}^n \mathsf{E}^*\left[\left\|\boldsymbol{X}_{n,i}^*\right\|^2 \mathbb{I}\{\left\|\boldsymbol{X}_{n,i}^*\right\| > \varepsilon\sqrt{n}\}\right] \\
&= \mathsf{E}^*\left[\left\|\boldsymbol{X}_{n,1}^*\right\|^2 \mathbb{I}\{\left\|\boldsymbol{X}_{n,1}^*\right\| > \varepsilon\sqrt{n}\}\right]
\end{aligned}$$

where E^* is a shortcut for the conditional expectation given $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$. If $M > 0$ is any constant so that $M \le \varepsilon\sqrt{n}$, we know that $\left\|\boldsymbol{X}_{n,1}^*\right\| > \varepsilon\sqrt{n}$ implies $\left\|\boldsymbol{X}_{n,1}^*\right\| > M$, and thus for any $M < \infty$ we find that for all n large enough we can bound

$$\begin{aligned}
\mathsf{E}^*\left[\left\|\boldsymbol{X}_{n,1}^*\right\|^2 \mathbb{I}\{\left\|\boldsymbol{X}_{n,1}^*\right\| > \varepsilon\sqrt{n}\}\right] &\le \mathsf{E}^*\left[\left\|\boldsymbol{X}_{n,1}^*\right\|^2 \mathbb{I}\{\left\|\boldsymbol{X}_{n,1}^*\right\| > M\}\right] \\
= \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{X}_i\|^2 \mathbb{I}\{\|\boldsymbol{X}_i\| > M\} &\xrightarrow[n\to\infty]{\text{a.s.}} \mathsf{E}\left[\|\boldsymbol{X}_1\|^2 \mathbb{I}\{\|\boldsymbol{X}_1\| > M\}\right].
\end{aligned}$$

Because this is true for any $M \in (0, \infty)$, we can take M sufficiently large so that the right-hand side is arbitrarily small. Necessarily, $\sum_{i=1}^n \mathsf{E}^*\left[\|\boldsymbol{Y}_{n,i}\|^2 \mathbb{I}\{\|\boldsymbol{Y}_{n,i}\| > \varepsilon\}\right] \xrightarrow[n\to\infty]{\text{a.s.}} 0$. The Lindeberg-Feller central limit theorem (Theorem A11) now gives that, conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$,

$$\sum_{i=1}^n \left(\boldsymbol{Y}_{n,i} - \mathsf{E}\left(\boldsymbol{Y}_{n,i} | \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\right)\right) = \mathbf{R}_n^* \xrightarrow[n\to\infty]{d} \mathsf{N}_k(\mathbf{0}, \mathbb{\Sigma}),$$

and this convergence is true almost surely. We conclude the proof. □

Note that for $\boldsymbol{X}_1$ 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 1 and 2 imply that conditionally on $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ also

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

Thus one can say that $\widehat{H}_n$ estimates the distribution function of $\mathsf{N}_p\big(\mathbf{0}_p, \mathsf{var}(\boldsymbol{X}_1)\big)$, and bootstrap works.

Example 7. Let $X_1, \dots, X_n$ be independent and identically distributed random variables and we are interested in the expectation $\mathsf{E}\, X_i$. The usual approach to find the confidence interval for $\mathsf{E}\, X_i$ is to use the convergence

$$\frac{\sqrt{n}\,\big(\overline{X}_n - \mathsf{E}\, X_i\big)}{S_n} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1), \tag{10}$$

which holds provided that $\mathsf{var}(X_i) \in (0, \infty)$. Here, of course, S_n^2 is the sample variance of $X_1, \dots, X_n$.

In view of the theory presented above we want to approximate/estimate the distribution function

$$H_n(x) = \mathsf{P}(R_n \le x), \quad \text{where } R_n = \sqrt{n}\,\big(\overline{X}_n - \mathsf{E}\, X_i\big).$$

With the help of (10) the estimate of this distribution based on the normal approximation is

$$\widehat{H}_n^{(\mathrm{norm})}(x) = \Phi\big(\tfrac{x}{S_n}\big). \tag{11}$$

Alternatively one can use the nonparametric bootstrap resulting in an estimator $\widehat{H}_{n,B}^*$ from (6).

Figure 1 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 $R_n = \sqrt{n}\,\big(\overline{X}_n - \mathsf{E}\, X_i\big)$ (black solid), the normal approximation (11) (blue solid) and the limit distribution which is $\mathsf{N}(0,1)$ (dotted, the variance is 1 because of $\mathsf{Exp}(1)$ chosen in the simulations). 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 $H_n(x) - \widehat{H}_n^{(\mathrm{norm})}(x)$ 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. However, none of the approximations seems evidently preferable in this example.

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

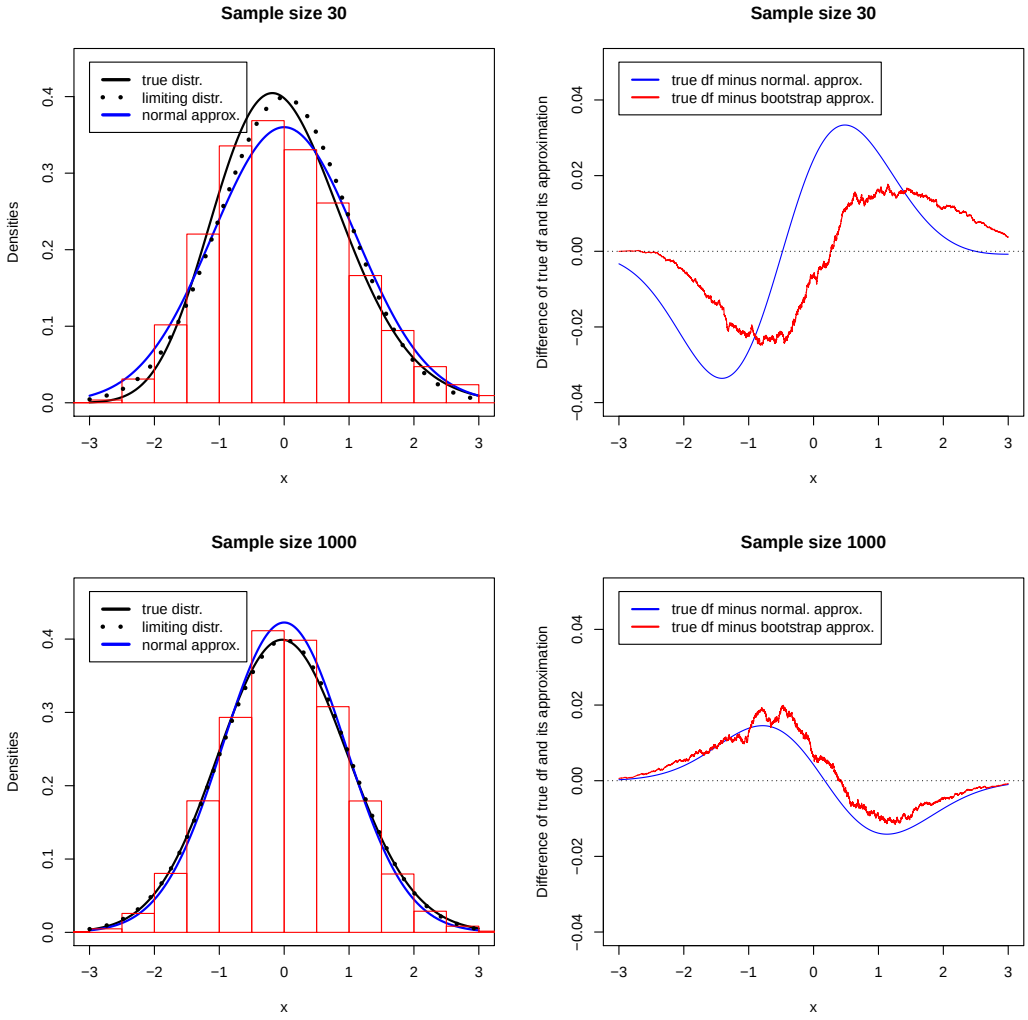

Figure 1: Comparison of the normal and bootstrap approximations of the distribution of the random variable $R_n = \sqrt{n}\,(\overline{X}_n - \mathsf{E}\ X_i)$.

Theorem 3. *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 (8) and (B) both hold, i.e. nonparametric bootstrap is consistent.*

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, ...).

Finally, the are also plenty of situations when the bootstrap works with statistics that are not (smooth transformations of) sample means. Roughly speaking, it can be shown that (B) holds provided that $\widehat{\boldsymbol{\theta}}_n$ satisfies the following asymptotic representation

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_X + \frac{1}{n}\sum_{i=1}^{n} IF(\boldsymbol{X}_i) + o_P\left(\tfrac{1}{\sqrt{n}}\right), \tag{12}$$

where $IF(\mathbf{x})$ is a given function. In this case, $\widehat{\boldsymbol{\theta}}_n$ can be well approximated by a 'sample mean' of variables $IF(\boldsymbol{X}_i)$, $i = 1, \ldots, n$, and thus a variant of Theorem 3 can still be stated. The 'linearization' of the statistic $\widehat{\boldsymbol{\theta}}_n$ from (12) can be formalised through the concept of influence functions, and Fréchet, or Hadamard-differentiability of the functional $F \mapsto \boldsymbol{\theta}(F)$ at F_X. That is, however, out of the scope of this course; for details and references one can see e.g. Nagy (2023b, Section 2.2).

In summary, we have found that nonparametric bootstrap works when

- the random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are independent and identically distributed,
- the moment assumption $\mathsf{E}\,\|\boldsymbol{X}_1\|^2 < \infty$ is true,
- $\mathbf{R}_n = \sqrt{n}\left(\mathbf{g}(\overline{\boldsymbol{X}}_n) - \mathbf{g}(\boldsymbol{\mu})\right)$ for a sufficiently smooth function $\mathbf{g}$ and $\boldsymbol{\mu} = \mathsf{E}\,\boldsymbol{X}_1$, or more generally, an expansion such as (12) holds true for $\mathbf{R}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)$, and
- the limiting distribution $\mathbf{R}$ of $\mathbf{R}_n$ exists, and is non-degenerate.

1.2.4 Comparison of nonparametric bootstrap and normal approximation

Theorems 1 and 3 imply the asymptotic validity of bootstrap provided that (B) holds. A most interesting question is whether the bootstrap estimate $\widehat{H}_n$ can be a better estimate of H_n than the asymptotic distribution H (with estimated unknown parameters).

To answer the above question, consider $p = 1$ and the case of the sample mean. The following result can be found in Shao and Tu (1996, Theorem 3.11).

Theorem 4. *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with a continuous distribution such that $\mathsf{E}\, X_1^4 < \infty$. Denote $\gamma_1 = \mathsf{E}\left(\frac{X_1-\mu}{\sigma}\right)^3$, where $\mu = \mathsf{E}\, X_1, \sigma^2 = \mathsf{var}(X_1)$, and let φ and Φ be the density and the distribution function of the standard normal distribution, respectively. Then*

$$H_n(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}_n-\mu)}{S_n} \le x\right) = \Phi(x) + \frac{\gamma_1}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O\big(\tfrac{1}{n}\big), \tag{13}$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Further, for $\widehat{H}_n(x)$ we have

$$\widehat{H}_n(x) = \mathsf{P}\left(\frac{\sqrt{n}\,(\overline{X}_n^*-\overline{X}_n)}{S_n^*} \le x \,\middle|\, \mathbb{X}\right) = \Phi(x) + \frac{\gamma_{1,n}}{6\sqrt{n}}\,(2x^2+1)\varphi(x) + O_P\big(\tfrac{1}{n}\big), \tag{14}$$

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

First, observe that Theorem 4 is stated for the **studentized** sample mean statistic

$$R_n = \frac{\sqrt{n}\,(\overline{X}_n - \mathsf{E}\, X_i)}{S_n}.$$

Suppose now that $\mathsf{E}\, X_1^6 < \infty$. Then the standard central limit theorem and the Δ-method give $\gamma_1 - \gamma_{1,n} = O_P\big(\frac{1}{\sqrt{n}}\big)$, and comparing (13) and (14) 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 from the Berry-Esséen inequality only

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

Thus if $\gamma_1 \neq 0$, one can expect that for R_n based on the studentized sample mean, in comparison with Φ, the bootstrap estimator $\widehat{H}_n$ is closer to the true distribution H_n of R_n. Without studentization, it can be shown that this advantage of using bootstrap disappears. We observed this in Example 7.

Example 8. We are in the same situation as in Example 7. But instead of approximating/estimating the distribution of $\sqrt{n}\,(\overline{X}_n - \mathsf{E}\, X_i)$, 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 $\widehat{H}_n^{(\mathrm{norm})}(x) = \Phi(x)$. The comparison of the true distribution function with its either normal or bootstrap approximation is found in Figure 2. Similarly as in Example 7, the results are for the random sample from the standard exponential distribution. In agreement with Theorem 4, the bootstrap approximation is better than the normal approximation.

The end of lecture 3 (16.10.2024)

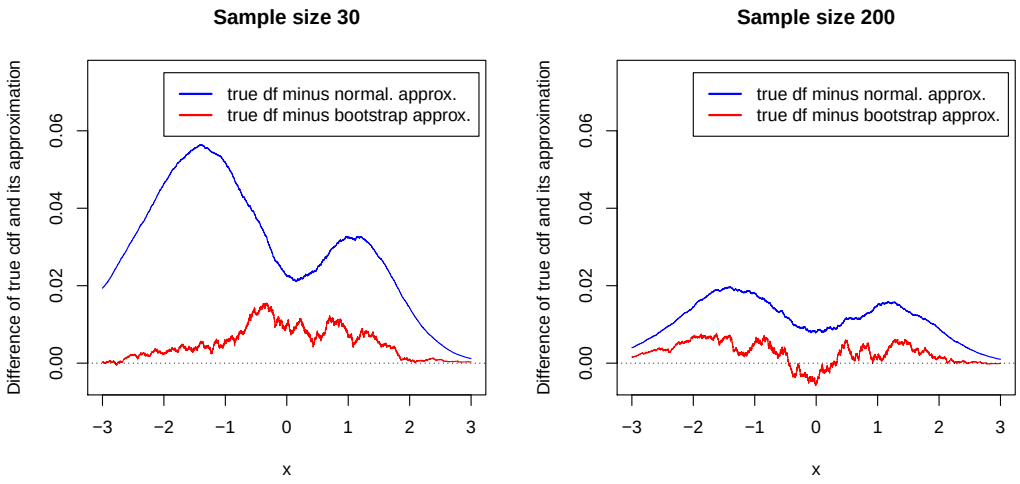

Figure 2: Comparison of the normal and bootstrap approximation of the distribution of the random variable $R_n = \frac{\sqrt{n}\,(\overline{X}_n - \mathsf{E}\,X_i)}{S_n}$.

1.2.5 Bootstrap-based 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 are interested in finding the confidence interval for $\theta_{X,j}$ (the j-th component of $\boldsymbol{\theta}_X$).

Suppose for a moment that the distribution of $R_{n,j}$ (the j-th component of $\mathbf{R}_n$) is known and continuous. Then one has

$$\mathsf{P}\left[r_n(\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{n,j} - \theta_{X,j}\right) < r_n(1-\alpha/2)\right] = 1-\alpha,$$

where $r_n(\alpha)$ is the α-quantile of $R_{n,j}$. Thus one would get a 'theoretical' confidence interval

$$\left(\widehat{\theta}_{n,j} - \tfrac{r_n(1-\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{n,j} - \tfrac{r_n(\alpha/2)}{\sqrt{n}}\right). \tag{15}$$

The problem is that the distribution of $R_{n,j}$ is not known and thus also the quantiles $r_n(\alpha/2)$ and $r_n(1-\alpha/2)$ are not known.

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 1 are satisfied, i.e., bootstrap works. Let $r^*_n(\alpha)$ be the quantile of the bootstrap distribution of $R^*_{n,j} = \sqrt{n}\left(\widehat{\theta}^*_{n,j} - \widehat{\theta}_{n,j}\right)$. Then Theorem 1 and Lemma A3 from the Appendix imply 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

$$\lim_{n\to\infty} \mathsf{P}\left[r_n^*(\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{n,j} - \theta_{X,j}\right) < r_n^*(1-\alpha/2)\right] = 1-\alpha. \tag{16}$$

Now with the help of (16) one can construct an asymptotic confidence interval for $\theta_{X,j}$ as

$$\left(\widehat{\theta}_{n,j} - \frac{r_{n,B}^*(1-\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{n,j} - \frac{r_{n,B}^*(\alpha/2)}{\sqrt{n}}\right), \tag{17}$$

where $r_{n,B}^*(\alpha) = \left(\widehat{H}_{n,B}^*\right)^{-1}(\alpha)$ is a Monte-Carlo approximation (estimate) of $r_n^*(\alpha)$. The confidence interval in (17) is usually called *basic bootstrap confidence interval*.

The formula for the confidence interval (17) mimics the formula for the theoretical confidence interval (15). 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').

Typically

$$\mathbf{R}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_p(\mathbf{0}_p, \mathbb{V}). \tag{18}$$

Then, the advantage of the confidence interval given by (16) 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 (such as Theorem 4) stating that the bootstrap confidence interval is more accurate require that the asymptotic distribution of $R_{n,j}$ is pivotal (i.e., it does not depend on unknown parameters). If this is not the case, then the basic bootstrap confidence interval (17) can be (for finite sample sizes) less accurate than the standard asymptotic confidence interval

$$\left(\widehat{\theta}_{n,j} - \frac{u_{1-\alpha/2}\sqrt{v_{n,j,j}}}{\sqrt{n}}, \widehat{\theta}_{n,j} + \frac{u_{1-\alpha/2}\sqrt{v_{n,j,j}}}{\sqrt{n}}\right), \tag{19}$$

where $v_{n,j,j}$ is a consistent estimate of the j-th diagonal element of the matrix $\mathbb{V}$. Consider the following example.

Example 9. We have a random sample $X_1, \ldots, X_n$ from a normal distribution $\mathsf{N}(\lambda, \lambda^2)$, with $\lambda > 0$ an unknown parameter. We construct a basic bootstrap confidence interval (17) for λ based on $R_n = \sqrt{n}(\overline{X}_n - \lambda_X)$; we know that the assumptions of both Theorems 1 and 2 are valid with the limiting random variable R distributed as $\mathsf{N}(0, \lambda_X^2)$, and the bootstrap is thus consistent. We approximate the α-quantiles of R by $r_{n,B}^*(\alpha)$, and set the confidence interval (17) to be

$$\left(\overline{X}_n - \frac{r_{n,B}^*(1-\alpha/2)}{\sqrt{n}}, \overline{X}_n - \frac{r_{n,B}^*(\alpha/2)}{\sqrt{n}}\right). \tag{20}$$

Suppose now first that the true value of λ is $\lambda_X = 1$. Then R has distribution $\mathsf{N}(0,1)$, and the estimated quantiles $r^*_{n,B}(\alpha)$ approximate u_α, the quantiles of $\mathsf{N}(0,1)$. Our confidence interval (20) is thus, for n and B large, approximately

$$\left(\overline{X}_n - \tfrac{u_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n - \tfrac{u_{\alpha/2}}{\sqrt{n}}\right).$$

A confidence interval should, however, cover the true value of the parameter λ_X with high probability **for any** $\lambda_X > 0$. Taking a different $\lambda_X > 0$ and $X_1, \ldots, X_n$ a random sample from $\mathsf{N}(\lambda_X, \lambda_X^2)$, we get for the confidence interval from (20) (and the quantiles $r^*_{n,B}(\alpha)$ computed with $\lambda_X = 1$ fixed)

$$\begin{aligned}
&\mathsf{P}\left(\lambda_X \in \left(\overline{X}_n - \tfrac{r^*_{n,B}(1-\alpha/2)}{\sqrt{n}}, \overline{X}_n - \tfrac{r^*_{n,B}(\alpha/2)}{\sqrt{n}}\right) \mid r^*_{n,B}(\alpha/2), r^*_{n,B}(1-\alpha/2)\right) \\
&= \mathsf{P}\left(\frac{r^*_{n,B}(\alpha/2)}{\lambda_X} < \sqrt{n}\left(\frac{\overline{X}_n - \lambda_X}{\lambda_X}\right) < \frac{r^*_{n,B}(1-\alpha/2)}{\lambda_X} \mid r^*_{n,B}(\alpha/2), r^*_{n,B}(1-\alpha/2)\right) \\
&\approx \Phi\left(\frac{u_{1-\alpha/2}}{\lambda_X}\right) - \Phi\left(\frac{u_{\alpha/2}}{\lambda_X}\right).
\end{aligned}$$

For $\lambda_X = 10$ and $\alpha = 0.05$, the coverage on right-hand side is approximately only 0.155, very far from the desired $1 - \alpha = 0.95$. This shows that in the confidence interval (17), also the quantiles r^*_n (or $r^*_{n,B}$) must be considered random, as the distribution of the limiting quantity $\mathbf{R}$ can still depend on the unknown parameter $\boldsymbol{\theta}$.

These difficulties, of course, disappear if we choose a quantity $R_{n,j}$ for the construction of confidence interval (17) pivotal, i.e., not depending on the parameter $\boldsymbol{\theta}$.

We see that if possible, it is beneficial to use asymptotically pivotal $R_{n,j}$, or quantities $R_{n,j}$ that at least 'less dependent' on the unknown parameters (see Remark 2 and Section 1.2.5 below).

Example 10. 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, generalised 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 $\beta_{X,j}$ (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 (Omelka, 2023, Section 4.3.2). The bootstrap can thus present an interesting alternative. In this situation, the nonparametric 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 formula for a confidence interval different from (17) can be found. To explain this formula note that $r^*_{n,B}(\alpha)$ is the sample α-quantile of $R^*_{n,j,1},\dots,R^*_{n,j,B}$, where $R^*_{n,j,b} = \sqrt{n}\left(\widehat{\theta}^*_{n,j,b} - \widehat{\theta}_{n,j}\right)$. Further let $q^*_{n,B}(\alpha)$ be a sample α-quantile calculated from the values $\widehat{\theta}^*_{n,j,1},\dots,\widehat{\theta}^*_{n,j,B}$. Then

$$r^*_{n,B}(\alpha) = \sqrt{n}\left(q^*_{n,B}(\alpha) - \widehat{\theta}_{n,j}\right) \tag{21}$$

and because

$$\widehat{\theta}_{n,j} - \frac{r^*_{n,B}(\alpha)}{\sqrt{n}} = \widehat{\theta}_{n,j} - \left(q^*_{n,B}(\alpha) - \widehat{\theta}_{n,j}\right) = 2\,\widehat{\theta}_{n,j} - q^*_{n,B}(\alpha),$$

the basic bootstrap confidence interval (17) can also be rewritten in an equivalent form as

$$\left(2\,\widehat{\theta}_{n,j} - q^*_{n,B}(1-\alpha/2), 2\,\widehat{\theta}_{n,j} - q^*_{n,B}(\alpha/2)\right). \tag{22}$$

Thus in practice it is sufficient to calculate $\widehat{\theta}^*_{n,j,b}$ instead of $R^*_{n,j,b}$ and then use formula (22). On the other hand, the approach based on calculating $R^*_{n,j,b}$ is more appropriate from the theoretical point of view. The reason is that to justify the bootstrap, one needs (among others) that the limiting distribution $R_{n,j}$ has a continuous distribution function (see Theorem 1).

Remark 1. Sometimes, in literature, one can find a bootstrap confidence interval of the form

$$\left(q^*_{n,B}(\alpha/2), q^*_{n,B}(1-\alpha/2)\right), \tag{23}$$

which is usually called the *percentile confidence interval*. With the help of (21) this confidence interval can be rewritten as

$$\left(\widehat{\theta}_{n,j} + \tfrac{r^*_{n,B}(\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{n,j} + \tfrac{r^*_{n,B}(1-\alpha/2)}{\sqrt{n}}\right).$$

Thus, when using the percentile confidence interval, one hopes that (taking $B=\infty$)

$$\begin{aligned}
&\lim_{n\to\infty} \mathsf{P}\left[\left(\widehat{\theta}_{n,j} + \tfrac{r^*_n(\alpha/2)}{\sqrt{n}}, \widehat{\theta}_{n,j} + \tfrac{r^*_n(1-\alpha/2)}{\sqrt{n}}\right) \ni \theta_{X,j}\right] \\
&\quad = \lim_{n\to\infty} \mathsf{P}\left[-r^*_n(1-\alpha/2) < \sqrt{n}\left(\widehat{\theta}_{n,j} - \theta_{X,j}\right) < -r^*_n(\alpha/2)\right] = 1-\alpha.
\end{aligned}$$

The use of the percentile interval can thus be justified if the limiting distribution of $R_{n,j}$ 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 a zero mean Gaussian distribution, the assumption of the symmetry of R_j is often satisfied.

The practical advantage of the percentile confidence is that it is always contained in the parametric space.

Remark 2. 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 Omelka, 2023, Chapter 1.4). This might be useful especially if one can guarantee that the limiting distribution of R_n is pivotal.

Studentized bootstrap confidence interval

We saw that it is recommended to 'bootstrap' a variable whose limit distribution is **pivotal** (i.e. does not depend on the unknown parameters).

Suppose that the asymptotic normality (18) holds and consider $\widetilde{R}_{n,j} = \frac{\sqrt{n}\,(\widehat{\theta}_{n,j} - \theta_{X,j})}{\sqrt{v_{n,j,j}}}$, where $v_{n,j,j}$ 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 $\widetilde{R}^*_{n,j} = \frac{\sqrt{n}\,(\widehat{\theta}^*_{n,j} - \widehat{\theta}_{n,j})}{\sqrt{v^*_{n,j,j}}}$, where $v^*_{n,j,j}$ 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 1 holds), then

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

which yields an asymptotic confidence interval

$$\left(\widehat{\theta}_{n,j} - \frac{\tilde{r}^*_{n,B}(1-\alpha/2)\sqrt{v_{n,j,j}}}{\sqrt{n}}, \widehat{\theta}_{n,j} - \frac{\tilde{r}^*_{n,B}(\alpha/2)\sqrt{v_{n,j,j}}}{\sqrt{n}}\right), \tag{24}$$

where $\tilde{r}^*_{n,B}(\alpha)$ is a Monte-Carlo approximation of $\tilde{r}^*_n(\alpha)$. The confidence interval in (24) is usually called the *studentized bootstrap confidence interval*.

Note that in comparison with (19) we replace the quantiles $-u_{1-\alpha/2}$ and $u_{1-\alpha/2}$ with $-\tilde{r}^*_{n,B}(1-\alpha/2)$ and $-\tilde{r}^*_{n,B}(\alpha/2)$, respectively. There are theoretical results that state that the studentized confidence interval (24) is (for finite sample sizes) more accurate than the standard asymptotic confidence interval (19) based on asymptotic normality, as well as the basic bootstrap confidence interval (17).

Ⓡ Example 11. Consider $X_1, \ldots, X_n$ a random sample from exponential distribution $\mathsf{Exp}(\lambda)$ with an unknown parameter $\lambda > 0$. We are interested in the expectation $\theta = \mathsf{E}\, X_1 = 1/\lambda$. Consider different types of confidence intervals for θ:

- The exact interval using the assumption of exponential distribution;
- The standard asymptotic confidence interval using the central limit theorem;
- The asymptotic confidence interval based on the variance-stabilising transformation;
- The standard bootstrap confidence interval and its studentized variant; and
- The bootstrap confidence interval based on the variance-stabilising transformation.

Compare the performance of all these confidence intervals both under the validity of the exponentiality assumption, and under model misspecification (i.e., when the true model is not exponential, but the confidence intervals are based on the assumption of exponential distribution).

Literature: Davison and Hinkley (1997, Chapters 5.1–5.3), Efron and Tibshirani (1993, Chapters 12 and 13).

1.2.6 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 estimator 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 applying the standard law of large numbers conditionally on $\mathbb{X}$ we get

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

Thus, for a valid inference we need a condition analogous to (B) saying that

$$n\;\mathsf{var}\,\big(\widehat{\boldsymbol{\theta}}^*_n \,|\, \mathbb{X}\big) \xrightarrow[n\to\infty]{P} \mathbb{V}. \tag{25}$$

Condition (B) and Theorem 1 in this situation give 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 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$. This, however, generally does not imply that (25) 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 $\frac{1}{n}\mathbb{V}$; we know that convergence in distribution does not generally imply convergence of moments.

Example 12. 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 (Omelka, 2023, 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, Section 3.2.2) .

1.2.7 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 $\mathbf{b}^*_n = \mathsf{E}\,[\widehat{\boldsymbol{\theta}}^*_n|\mathbb{X}] - \widehat{\boldsymbol{\theta}}_n$. The *bias-corrected estimator* of $\boldsymbol{\theta}$ is then defined as $\widehat{\boldsymbol{\theta}}^{(bc)}_n := \widehat{\boldsymbol{\theta}}_n - \mathbf{b}^*_n$.

Example 13. 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 neighbourhood U 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 generally not unbiased for $g(\mu)$. Put $\sigma^2 = \mathsf{var}(X_1)$. Then, one can use Taylor's expansion of g, and subsequently apply an expected value, to approximate the bias of $g(\overline{X}_n)$

$$\begin{aligned} b_n = \mathsf{E}\,g(\overline{X}_n) - g(\mu) &= \mathsf{E}\left\{g'(\mu)(\overline{X}_n - \mu) + \tfrac{g''(\mu)}{2}(\overline{X}_n - \mu)^2\right\} + \frac{R_n}{3!} \\ &= \frac{g''(\mu)\,\sigma^2}{2\,n} + O\Big(\tfrac{1}{n^{3/2}}\Big). \end{aligned} \tag{26}$$

To bound the remainder term R_n we have used that for n large enough, $\overline{X}_n \in U$ almost surely and thus

$$\begin{aligned} |R_n| &\le \sup_{x\in U}\big|g'''(x)\big|\;\mathsf{E}\,\big|\overline{X}_n - \mu\big|^3 \le \sup_{x\in U}\big|g'''(x)\big|\,\Big[\mathsf{E}\,\big(\overline{X}_n - \mu\big)^4\Big]^{3/4} \\ &= \sup_{x\in U}\big|g'''(x)\big|\,\Big[\frac{1}{n^4}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{\ell=1}^n \mathsf{E}\,\big(X_i-\mu\big)\big(X_j-\mu\big)\big(X_k-\mu\big)\big(X_\ell-\mu\big)\Big]^{3/4}. \end{aligned}$$

We used the Jensen inequality with the concave function $t \mapsto t^{3/4}$ in the second inequality above. Now, it is enough to realise that the last expectation will be zero unless (i) either all i, j, k, ℓ are the same, in which case we have

$$\mathsf{E}\,\big(X_i-\mu\big)\big(X_j-\mu\big)\big(X_k-\mu\big)\big(X_\ell-\mu\big) = \mathsf{E}\,\big(X_1-\mu\big)^4,$$

or (ii) if two of the indices i, j, k, ℓ are the same, and the other two are also equal, but not the same as the first two. In case (ii) we have

$$\mathsf{E}\,\big(X_i-\mu\big)\big(X_j-\mu\big)\big(X_k-\mu\big)\big(X_\ell-\mu\big) = \Big(\mathsf{E}\,\big(X_1-\mu\big)^2\Big)^2 = \sigma^4.$$

Case (i) appears in n summands; it is not hard to calculate that case (ii) appears in $\binom{n}{2}\binom{4}{2} = O(n^2)$ out of the total number of n^4 summands. We can thus bound

$$|R_n| \le \sup_{x\in U}\big|g'''(x)\big|\,\left[\frac{1}{n^3}\,\mathsf{E}\,\big(X_1-\mu\big)^4 + O\Big(\frac{1}{n^2}\Big)\sigma^4\right]^{3/4} = \Big[O\big(\tfrac{1}{n^2}\big)\Big]^{3/4} = O\Big(\tfrac{1}{n^{3/2}}\Big),$$

as we claimed in (26).

Analogously, one can expand also $g(\overline{X}_n^*)$ around $g(\overline{X}_n)$, and apply the conditional expectation $\mathsf{E}^*[\cdot] = \mathsf{E}[\cdot|\mathbb{X}]$ to obtain

$$\begin{aligned} b_n^* &= \mathsf{E}\left[g(\overline{X}_n^*) \,|\, \mathbb{X}\right] - g(\overline{X}_n) = \frac{g''(\overline{X}_n)}{2n}\, \mathsf{var}[X_1^* \,|\, \mathbb{X}] + O_P\Big(\tfrac{1}{n^{3/2}}\Big) \\ &= \frac{g''(\overline{X}_n)\,\widehat{\sigma}_n^2}{2n} + O_P\Big(\tfrac{1}{n^{3/2}}\Big). \end{aligned} \tag{27}$$

where $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. To obtain the stochastic bound $O_P\Big(\frac{1}{n^{3/2}}\Big)$ we considered, similarly as above,

$$\mathsf{E}^*\left[\left|\overline{X}_n^* - \overline{X}_n\right|^3\right] \leq \mathsf{E}^*\left[\left(\overline{X}_n^* - \overline{X}_n\right)^4\right]^{3/4} = \left[\mathsf{E}^*\left[\left(\overline{Y}_n^*\right)^4\right]\right]^{3/4},$$

where we denote $Y_i^* = X_i^* - \overline{X}_n$ and $\overline{Y}_n^* = \frac{1}{n}\sum_{i=1}^n Y_i^*$. Conditionally on $\mathbb{X}$, we expand $\left(\overline{Y}_n^*\right)^4$ as we did before for $\left(\overline{X}_n - \mu\right)^4$, and find that

$$\mathsf{E}^*\left[\left(\overline{Y}_n^*\right)^4\right] = \frac{1}{n^3}\,\mathsf{E}^*\left[(Y_1^*)^4\right] + O\left(\frac{1}{n^2}\right)\widehat{\sigma}_n^4.$$

Finally, since $\widehat{\sigma}_n^4 \xrightarrow[n\to\infty]{\text{a.s.}} \sigma^4$ and

$$\mathsf{E}^*\left[(Y_1^*)^4\right] = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^4 \xrightarrow[n\to\infty]{\text{a.s.}} \mathsf{E}\,(X_1 - \mathsf{E}\,X_1)^4,$$

we have that both $\hat{\sigma}_n^4$ and $\mathsf{E}^*\left[(Y_1^*)^4\right]$ are $O_P(1)$, and the final expression in (27) is correct.

Now, comparing (26) and (27) one gets that

$$b_n - b_n^* = \frac{1}{2n}\Big(g''(\mu)\sigma^2 - g''(\overline{X}_n)\widehat{\sigma}_n^2\Big) + O_P\Big(\tfrac{1}{n^{3/2}}\Big) = O_P\Big(\tfrac{1}{n^{3/2}}\Big), \tag{28}$$

where we used the Δ-theorem (Omelka, 2023, Theorem 3) for the sample mean with the function g'', and the fact that $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2)$ is asymptotically normal (Kulich and Omelka, 2022, Theorem 2.6), to get

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

Suppose that $g''(\mu) \neq 0$. The the bias of the estimator $\widehat{\theta}_n^{(bc)} = g(\overline{X}_n) - b_n^*$ is given by

$$\mathsf{E}\,\widehat{\theta}_n^{(bc)} - g(\mu) = \mathsf{E}\; g(\overline{X}_n) - g(\mu) - \mathsf{E}\; b_n^* = b_n - \mathsf{E}\; b_n^*.$$

We saw in (28) that $b_n - b_n^* = O_P(n^{-3/2})$. This does not generally imply $b_n - \mathsf{E}\; b_n^* = O(n^{-3/2})$ (convergence in distribution does not imply convergence of moments), but under an appropriate uniform integrability assumption, it does. We conclude that typically, while the bias of the original estimator $\widehat{\theta}_n = g(\overline{X}_n)$ is of order $O(n^{-1})$ by (26), the bias-corrected estimator $\widehat{\theta}_n^{(bc)} = g(\overline{X}_n) - b_n^*$ will typically have bias only of order $O\Big(n^{-3/2}\Big)$.

Literature: Efron and Tibshirani (1993, Chapter 10).

1.2.8 Jackknife

Jackknife can be considered to be an ancestor of bootstrap; its history goes back to 1949 and the work of Quenouille. Jackknife was originally proposed as a method to reduce the bias of an estimator. Later, it was found that it can often be also used to estimate the variance of an estimator.

Suppose that $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is a random sample, and denote $\mathbf{T}_n = \mathbf{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ the estimator of the parameter of interest $\boldsymbol{\theta}_X$. The jackknife is based on 'bootstrapping' $\mathbb{X}$ by erasing single observations, that is the i-th jackknife sample from $\mathbb{X}$ is given by

$$\mathbb{X}_i^* = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_n) \qquad \text{for } i = 1, \ldots, n,$$

of sample size $n-1$. The i-th jackknifed estimator is given by

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

The quantity

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

with $\overline{\mathbf{T}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{T}_{n-1,i}$ is then used as an estimator of the bias of $\mathbf{T}_n$. The scaling factor $(n-1)$ comes from a Taylor expansion similar to that performed in Example 13. It is meant to guarantee that the bias-corrected jackknife estimator

$$\mathbf{T}_n^{(bc)} = \mathbf{T}_n - \widehat{\mathbf{b}}_n. \tag{29}$$

achieves bias of order $O(n^{-3/2})$, while the original estimator $\mathbf{T}_n$ has bias of order $O(n^{-1})$.

Literature: (Shao and Tu, 1996, Section 1.3).

The end of lecture 4 (23.10.2024)

1.2.9 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 can also fail. This happens, for example, in situations when the asymptotic normality of $\mathbf{R}_n$ does not hold, for extremal statistics, or non-smooth transformations of sample means. 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 inappropriate in regression problems with fixed design or time series problems.

We give two examples when nonparametric bootstrap fails for independent and identically distributed data.

Example 14. We are in the situation with smooth transforms of the sample mean from Theorem 3. 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 (B) holds (by Theorem 3), 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 twice continuously differentiable at μ_X, such that $g'(\mu_X) = 0$ and $g''(\mu_X) \neq 0$. Then by the delta theorem (Omelka, 2023, 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 3 convergence (B) holds, one cannot say whether bootstrap works as the limiting distribution $\mathbf{R}$ of $\mathbf{R}_n$ is not continuous (the assumptions of Theorem 1 are not satisfied).

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

$$\widetilde{R}_n = 2\,n\left(g(\overline{X}_n) - g(\mu_X)\right) \xrightarrow[n\to\infty]{d} \left[g''(\mu_X)\right] \sigma^2\, \chi^2_1.$$

So the bootstrap would work if the convergence (B) holds also for $\widetilde{R}^*_n = 2\,n\left(g(\overline{X}^*_n) - g(\overline{X}_n)\right)$, 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, it can be shown that (B) does not hold (see Example 3.6 of Shao and Tu, 1996). The standard nonparametric bootstrap thus, in this situation, fails to be consistent.

Ⓡ Example 15. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta_X)$ with distribution function F_{X_1}. Then the maximum likelihood estimator of θ_X is given by $\widehat{\theta}_n = \max_{1\leq i\leq n} X_i =: X_{(n)}$. For $x < 0$

$$\begin{aligned}\mathsf{P}\left(n(X_{(n)} - \theta_X) \leq x\right) &= \mathsf{P}\left(X_{(n)} \leq \theta_X + \tfrac{x}{n}\right) = F^n_{X_1}\left(\theta_X + \tfrac{x}{n}\right) \\ &= \left[\frac{\theta_X + \frac{x}{n}}{\theta_X}\right]^n = \left[1 + \frac{x}{n\,\theta_X}\right]^n \xrightarrow[n\to\infty]{} \mathrm{e}^{\frac{x}{\theta_X}}.\end{aligned}$$

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

$$\mathsf{P}(Y \leq x) = \begin{cases} \mathrm{e}^{\frac{x}{\theta_X}}, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

On the other hand, for $R^*_n = n\left(X^*_{(n)} - X_{(n)}\right)$ we have

$$\begin{aligned}\mathsf{P}\left(R^*_n = 0 \,|\, \mathbb{X}\right) &= \mathsf{P}(X^*_{(n)} = X_{(n)} \,|\, \mathbb{X}) = 1 - \mathsf{P}\left(X_{(n)} \notin \{X^*_1, \ldots, X^*_n\} \,|\, \mathbb{X}\right) \\ &= 1 - \left(\tfrac{n-1}{n}\right)^n \xrightarrow[n\to\infty]{} 1 - \mathrm{e}^{-1}\end{aligned}$$

and thus (B) cannot hold for R^*_n.

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

1.3 Parametric bootstrap

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors, each with distribution $F(\cdot; \boldsymbol{\theta}_X)$ that is known 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 the estimated distribution $F(\cdot; \widehat{\boldsymbol{\theta}}_n)$, where $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_X$.

R Example 16. Suppose we are in the situation from Example 15, i.e. X_i are distributed uniformly on $(0, \theta_X)$. Apply now the parametric bootstrap, by generating, conditionally on $\mathbb{X}$, a random sample $X^*_1, \ldots, X^*_n$ from the uniform distribution on $(0, X_{(n)})$. Then, for $\widehat{H}_n$ the (conditional) distribution function of $R^*_n = n\,(X^*_{(n)} - X_{(n)})$ with $X^*_{(n)} = \max_{1 \le i \le n} X^*_i$ we have for $x < 0$

$$\begin{aligned}\widehat{H}_n(x) &= \mathsf{P}\left(n\,(X^*_{(n)} - X_{(n)}) \le x \,|\, \mathbb{X}\right) = \mathsf{P}\left(X^*_{(n)} \le X_{(n)} + x/n \,|\, \mathbb{X}\right) \\ &= \left[\frac{X_{(n)} + x/n}{X_{(n)}}\right]^n = \left[1 + \frac{x}{n\,X_{(n)}}\right]^n \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{e}^{\frac{x}{\theta_X}}.\end{aligned}$$

In the final limit, we used that $X_{(n)} \xrightarrow[n\to\infty]{\text{a.s.}} \theta_X$. Comparing the previous expression with Example 15, we see that in this situation, the parametric bootstrap works,* since we found that for $H(x) = \mathrm{e}^{\min\left\{\frac{x}{\theta_X}, 0\right\}}$ we have

$$\mathsf{P}\left(\widehat{H}_n(x) \xrightarrow[n\to\infty]{} H(x) \text{ for each } x \in \mathbb{R}\right) = 1.$$

Example 17. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from 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 a confidence interval for $\frac{\lambda_X}{\lambda_Y}$.

Solution. The maximum likelihood estimators are given by $\widehat{\lambda}_X = \frac{1}{\overline{X}_{n_1}}, \widehat{\lambda}_Y = \frac{1}{\overline{Y}_{n_2}}$. Now generate $X^*_1, \ldots, X^*_{n_1}$ and $Y^*_1, \ldots, Y^*_{n_2}$ as two independent random samples from the exponential distributions with the parameters $\widehat{\lambda}_X$ and $\widehat{\lambda}_Y$, respectively. Put

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

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

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

where $r^*_{n,B}(\alpha)$ is the estimate of the α-quantile of R^*_n.

* Note also that it would be more natural to resample $R_n = n\left(\frac{\widehat{\theta}_n}{\theta_X} - 1\right)$, as its asymptotic distribution is pivotal.

Alternatively instead of bootstrap one can use the Δ-theorem (Omelka, 2023, Theorem 3), which implies that

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

Combining Δ-theorem and bootstrap one can also use

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

1.3.1 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. We are interested in testing whether 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 difficult to obtain, the significance of the test statistic is derived as follows.

(i) For $b = 1, \ldots, B$ generate an independent 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 $F(\mathbf{x};\widehat{\boldsymbol{\theta}}_n)$.

(ii) Calculate

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

where $\widehat{F}_{n,b}^*$ 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 $\mathbb{X}_b^*$.

(iii) 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 high, e.g. 999 or 9 999.

Remark 3. Sometimes, people ignore 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 KS_n with the help of the (asymptotic) distribution of

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

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 quite different from the (asymptotic) distribution of Z_n. Simulation studies show that if the significance of $\sqrt{n}\, KS_n$ is assessed with the help of the distribution of Z_n, the true level of the test is much smaller than the prescribed value α. The test is thus very conservative. The intuitive reason is that $\widehat{\boldsymbol{\theta}}_n$ is estimated from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Thus, 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 4. Instead of the test statistic KS_n in $\mathbb{R}$, one of the following statistics is usually recommended. The reason is that the tests based on these statistics usually have more power against the alternatives that seem to be natural.

Cramér-von-Mises:

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

Anderson-Darling:

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

Example 18. Testing goodness-of-fit of multinomial distribution with estimated parameters.

1.4 Testing hypotheses and bootstrap

Provided the parameter of interest is one-dimensional and one can construct a confidence interval for this parameter (see Section 1.2.5), 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 $\mathbb{X}_1^* = (\boldsymbol{X}_{1,1}^*, \ldots, \boldsymbol{X}_{n,1}^*)$, $\ldots$, $\mathbb{X}_B^* = (\boldsymbol{X}_{1,B}^*, \ldots, \boldsymbol{X}_{n,B}^*)$ 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 can be estimated as

$$\widehat{p}_B = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\{T_{n,b}^* \geq T_n\}}{B+1}. \tag{30}$$

Example 19. Let $X_1, \ldots, X_n$ be a random sample such that $\mathsf{var}\, X_1 \in (0, \infty)$ and $H_0 : \mathsf{E}\, X_1 = \mu_0$. One can use nonparametric bootstrap and generate $X_{1,b}^*, \ldots, X_{n,b}^*$ as a simple random sample with replacement from $X_1 - \overline{X}_n + \mu_0, \ldots, X_n - \overline{X}_n + \mu_0$. A possible test statistic is then

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

and $T_{n,b}^* = \frac{\sqrt{n}\,(\overline{X}_{n,b}^* - \mu_0)}{S_{n,b}^*}$, where $\overline{X}_{n,b}^*$ and $S_{n,b}^*$ are the sample mean and sample standard deviation calculated from the bootstrap sample. Observe that this procedure is equivalent with sampling $X_{1,b}^*, \ldots, X_{n,b}^*$ from $X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n$, and consequently using $T_{n,b}^* = \frac{\sqrt{n}\,(\overline{X}_{n,b}^* - 0)}{S_{n,b}^*}$.

As an alternative, one could also use parametric bootstrap. What procedure do we obtain in this situation?

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 distributions F and G, respectively. We are interested in testing the hypothesis

$$H_0 : \mathsf{E}\, X_1 = \mathsf{E}\, Y_1, \quad \text{vs.} \quad H_1 : \mathsf{E}\, X_1 \neq \mathsf{E}\, Y_1.$$

There are several options for how to test for the above 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

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

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, $T_n \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1)$ (Kulich and Omelka, 2022, Section 6.3).

2. Welch t-test is based on the test statistics

$$\widetilde{T}_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}. \tag{31}$$

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 also use a test statistic such as $T_{n,0} = \left|\overline{X}_{n_1} - \overline{Y}_{n_2}\right|$, but it is recommended to use a test statistic whose asymptotic distribution under the null hypothesis does not depend on the 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.

A further alternative to how to approach a two-sample problem is the use of an appropriate permutation test.

Example 20. Suggest a test that would compare medians in two-sample problems.

1.5 Permutation tests

Permutation tests present an interesting alternative to nonparametric bootstrap. They are particularly useful in two situations:

- in two (or more generally K) sample problems, and
- when testing for independence.

1.5.1 Permutation tests in two-sample problems

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples with 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 $Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)}$. 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 $\mathbb{Z}_{(\cdot)}$ (see, e.g., Kulich and Omelka, 2022, Theorem 2.15). More formally, let M be the cardinality of the set

$$\{(z_{i_1}, \ldots, z_{i_n}) : \text{where } (i_1, \ldots, i_n) \text{ is a permutation of the set } (1, \ldots, n)\}.$$

If there are no ties, i.e., if all values $z_1, \ldots, z_n$ are distinct, then $M = n!$.* Now the conditional distribution of $\mathbb{Z}$ given $\mathbb{Z}_{(\cdot)}$ is given by

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

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

In permutation tests, the 'bootstrap' samples $\mathbb{Z}_1^*, \ldots, \mathbb{Z}_B^*$ are generated by randomly permuting the joint sample $\mathbb{Z}$. For each $b \in \{1, \ldots, B\}$ the test statistic $T_{n,b}^*$ is recalculated from

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

and the p-value is estimated by (30).

Remark 5. For two-sample problems, there are only $\binom{n}{n_1}$ permutations which can give rise to different values of the test statistic (provided that the test statistic is symmetric with respect to the permutations within $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, respectively). So, if n_1 and n_2 are small, then one can calculate the permutation p-value exactly, where exactly means with respect to the exact 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 of $\mathbb{Z}$ to estimate the p-value.

Ⓡ Example 21. The permutation test approach can be used to assess the significance of the two-sample Kolmogorov-Smirnov test statistic

$$K_{n_1,n_2} = \sup_{x \in \mathbb{R}} \big|\widehat{F}_{n_1}(x) - \widehat{G}_{n_2}(x)\big|,$$

where $\widehat{F}_{n_1}$ is the empirical distribution function of $X_1, \ldots, X_{n_1}$, and analogously for $\widehat{G}_{n_2}$ and the sample $Y_1, \ldots, Y_{n_2}$. For this test, the standard inference is based on the asymptotic distribution of K_{n_1,n_2} that is derived in case the distribution function F (under the null hypothesis equal to G) is continuous. Using the permutation test, we do not have this restriction. Permutation testing can thus be of interest especially in the presence of ties (e.g. due to rounding), or when F is discontinuous, or when $\boldsymbol{X}_i$ are k-dimensional random vectors.

* If there are ties, let $a_1, \ldots, a_J$ be the distinct values of $(z_1, \ldots, z_n)$. Put $r_j = \sum_{i=1}^n \mathbb{I}\{z_i = a_j\}$. Then $M = \frac{n!}{r_1! \ldots r_J!}$.

All permutation tests above assumed that **under the null hypothesis, the distribution functions F and G coincide**, or more generally, an **exchangeability** condition under H_0. Such permutation tests are called *exact*. In practice, it is also of interest to know whether the permutation test is useful 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 the null hypothesis has a limiting distribution that does not depend on the unknown parameters, then the permutation test holds the prescribed level asymptotically. A permutation test is called *approximate* in that situation. It was shown by simulations in many different settings that the level of approximate permutation tests is usually closer to the prescribed value α than the level of a test that directly uses the asymptotic distribution of the test statistic T_n. This is quite similar to what we saw in Theorem 4 in the situation with bootstrapping studentized averages.

Example 22. It can be shown that the permutation version of the Welch t-test, see (31), is asymptotically valid also in models where the null hypothesis holds (i.e., $\mathsf{E}\, X_i = \mathsf{E}\, Y_i$), but the distributions of X_i and Y_i are different.

The end of lecture 5 (30.11.2024)

1.5.2 Permutation tests of 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 of Y_1. Then, under the null hypothesis, we have

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

where M is analogously as above the cardinality of

$$\{(y_{i_1}, \ldots, y_{i_n}) : \text{where } (i_1, \ldots, i_n) \text{ is a permutation of the set } (1, \ldots, n)\}.$$

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. This permutation scheme can be used for assessing the significance of the test statistic based on a correlation coefficient, or of the χ^2-test of independence.

Example 23. Consider a contingency table with J rows and K columns. The row vectors can be represented as realisations of independent multinomial random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J$,

for $\boldsymbol{X}_j$ with distribution $\mathsf{Mult}_K(n_j, \mathbf{p}_j)$. Here, n_j is the sum of the elements of $\boldsymbol{X}_j$, and the parameter $\mathbf{p}_j \in [0,1]^K$, whose entries sum to one, is unknown. We want to test the independence of rows and columns in this contingency table, which is equivalent to testing

$$H_0\colon \mathbf{p}_1 = \cdots = \mathbf{p}_J \quad \text{vs} \quad H_1\colon H_0 \text{ is not true.}$$

This is commonly done using the χ^2-test of independence (Kulich and Omelka, 2022, Section 8.2) with test statistic χ^2. To get a permutation version of this test, one decomposes each $\boldsymbol{X}_j$ into $\boldsymbol{X}_j = \sum_{i=1}^{n_j} \boldsymbol{Y}_{j,i}$ with $\boldsymbol{Y}_{j,i}$ independent and distributed as $\mathsf{Mult}_K(1, \mathbf{p}_j)$. Each $\boldsymbol{Y}_{j,i}$ represents a single observation in the table. Under H_0, the random vectors $\boldsymbol{Y}_{1,1}, \ldots, \boldsymbol{Y}_{1,n_1}, \boldsymbol{Y}_{2,1}, \ldots, \boldsymbol{Y}_{J,n_J}$ are independent and identically distributed. One thus permutes these $n = \sum_{j=1}^{J} n_j$ vectors to get $\boldsymbol{Y}^*_{1,1}, \ldots, \boldsymbol{Y}^*_{1,n_1}, \boldsymbol{Y}^*_{2,1}, \ldots, \boldsymbol{Y}^*_{J,n_J}$, sets $\boldsymbol{X}^*_j = \sum_{i=1}^{n_j} \boldsymbol{Y}^*_{j,i}$, and uses $\boldsymbol{X}^*_1, \ldots, \boldsymbol{X}^*_J$ for a Monte Carlo approximation of the distribution of the test statistic χ^2 under H_0. There are two advantages of the permutation test. For n small, it might be possible to determine also the exact permutation distribution of χ^2, by considering all possible $n!$ permutations of the observations. That leads to an exact testing procedure. Second, it is known that the convergence of χ^2 to its asymptotic distribution is slow if some elements of $\mathbf{p}_j$ are close to zero (Kulich and Omelka, 2022, Section 8.1). This is not a problem for a permutation test, as it does not involve asymptotics.

Remark 6. Generally, any K-sample problem can be viewed as a 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, the independence of Z_1 and I_1 is equivalent to all the random samples having the same distribution function.

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

1.6 Model-based bootstrap

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. We assume the structure of a linear model

$$Y_i = \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{32}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed zero-mean random variables independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and $\boldsymbol{\beta}$ is an unknown parameter. We are interested in the distribution of an estimator $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$. In Example 10, we considered the standard nonparametric bootstrap that generates $\binom{\boldsymbol{X}^*_1}{Y^*_1}, \ldots, \binom{\boldsymbol{X}^*_n}{Y^*_n}$ as a simple random sample with replacement from the vectors

$\binom{\boldsymbol{X}_1}{Y_1}, \ldots, \binom{\boldsymbol{X}_n}{Y_n}$. Provided the estimator $\widehat{\boldsymbol{\beta}}_n$ is asymptotically normal, one can usually assume that this bootstrap method works.

Another possibility is to use the **model-based bootstrap** that runs as follows. Calculate the standardised 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

$$Y_i^* = \boldsymbol{X}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}_n + \varepsilon_i^*, \quad i = 1, \ldots, n, \tag{33}$$

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 (32) to hold. On the other hand, if model (32) 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 also be used with a fixed design regression. On the other hand, a model-based bootstrap is inappropriate, for instance, under heteroscedasticity.

Model-based bootstrap can be successfully used also in time series analysis.

R Example 24. Take the autoregressive process $AR(1)$ given by $X_t = a\,X_{t-1} + \varepsilon_t$, $t = 1, \ldots, n$ for $X_0 = 0$, $a \in (-1, 1)$ an unknown parameter, and each ε_t with distribution $\mathsf{N}(0, 1)$, independent of the remaining quantities. Since the observed data $X_1, \ldots, X_n$ are not independent, one cannot apply nonparametric bootstrap directly. In a model-based approach, one estimates a by $\widehat{a}_n$ and considers the residuals $\hat{\varepsilon}_t = X_t - \widehat{a}_n\,X_{t-1}$, $t = 1, \ldots, n$. These are the counterparts of the independent and identically distributed errors ε_t; we thus resample the residuals. We take $\varepsilon_1^*, \ldots, \varepsilon_n^*$ a simple random sample with replacement from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$, and with $X_0^* = 0$ define

$$X_t^* = \widehat{a}_n\,X_{t-1}^* + \varepsilon_t^* \quad \text{for } t = 1, \ldots, n.$$

Bootstrap can be performed analogously for estimators in an $AR(p)$ process.

R Example 25. We have a linear model as in (32), but we suppose that the error terms $\varepsilon_1, \ldots, \varepsilon_n$ might form a time series. For the validity of the classical least-squared inference, we thus need to test the independence of the errors. For that, we adapt a model, and assume that $\varepsilon_1, \ldots, \varepsilon_n$ form an $AR(1)$ process as in Example 24, with (conditional) autocorrelation

$\rho = \mathsf{corr}\,(\varepsilon_i, \varepsilon_{i-1} \mid \mathbb{X}) = \rho_X \in (-1, 1)$. Here, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ are the regressors. We want to test

$$H_0 \colon \rho_X = 0 \quad \text{vs.} \quad H_1 \colon \rho_X \neq 0.$$

We calculate the least-squares fit $\widehat{\boldsymbol{\beta}}_n$, the residuals $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}_n$, and use the test statistic

$$T_n = \frac{\sum_{i=2}^n (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^n \hat{\varepsilon}_i^2}.$$

This test statistic can be shown to estimate $2\,(1 - \rho_X)$, so H_0 is rejected if T_n is far from 2. The distribution of T_n under H_0, however, depends on the model matrix $\mathbb{X}$, and thus no universally valid asymptotic inference is possible. This problem can be solved using bootstrap. Under H_0, the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed. We can thus resample $\varepsilon_1^*, \ldots, \varepsilon_n^*$ as a simple random sample with replacement from the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Then we generate the bootstrap sample $\binom{\boldsymbol{X}_1}{Y_1^*}, \ldots, \binom{\boldsymbol{X}_n}{Y_n^*}$ as in (33), recalculate the fit $\widehat{\boldsymbol{\beta}}_n^*$, and evaluate the test statistic T_n with the new residuals $Y_i^* - \boldsymbol{X}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}_n^*$, $i = 1, \ldots, n$. The model-based bootstrap now allows us to approximate the conditional distribution of T_n given $\mathbb{X}$ under H_0.

This test is called the Durbin-Watson test in linear models; in R, it is implemented in function durbinWatsonTest in package car. For more details, see Komárek (2021, Section 9.5.1).

Literature: Davison and Hinkley (1997, Chapter 6.3).

2 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 the Lebesgue measure**. 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

$$\widetilde{f}_n(x) = \frac{\widehat{F}_n(x + h_n) - \widehat{F}_n(x - h_n)}{2\,h_n} = \frac{1}{2\,h_n} \sum_{i=1}^n \frac{\mathbb{I}\{X_i \in (x - h_n, x + h_n]\}}{n}, \tag{34}$$

where $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq x\}$ is the empirical distribution function and (the bandwidth) h_n is a sequence of positive constants going to zero.

It follows from the Lebesgue differentiation theorem† that for almost every point $x \in \mathbb{R}$ we have

$$\mathsf{E}\ \widetilde{f}_n(x) = \frac{F(x + h_n) - F(x - h_n)}{2\,h_n} \xrightarrow[n \to \infty]{} f(x)$$

* *Jádrové odhady hustoty* † https://en.wikipedia.org/wiki/Lebesgue_differentiation_theorem

and

$$\mathsf{var}(\widetilde{f}_n(x)) = \frac{\big[F(x+h_n) - F(x-h_n)\big]\big[1 - F(x+h_n) + F(x-h_n)\big]}{4\,h_n^2\,n}$$
$$= \frac{F(x+h_n) - F(x-h_n)}{2\,h_n}\,\frac{1 - F(x+h_n) + F(x-h_n)}{2\,n\,h_n} \xrightarrow[n\to\infty]{} 0$$

provided that $h_n \to 0$ and at the same time $(n\,h_n) \to \infty$.

The estimator (34) can be rewritten as

$$\widetilde{f}_n(x) = \frac{1}{2\,n\,h_n}\sum_{i=1}^{n} \mathbb{I}\big\{-1 < \tfrac{X_i - x}{h_n} \le +1\big\} = \frac{1}{n\,h_n}\sum_{i=1}^{n} w\big(\tfrac{X_i - x}{h_n}\big), \tag{35}$$

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 (35) we define the *kernel density estimator* as

$$\widehat{f}_n(x) = \frac{1}{n\,h_n}\sum_{i=1}^{n} K\big(\tfrac{X_i - x}{h_n}\big) \quad \text{for } x \in \mathbb{R}, \tag{36}$$

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.

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.

The end of lecture 6 (6.11.2024)

Remark 7. Note that:

(i) The estimator (36) can be interpreted as an average of n terms of the form $\frac{1}{h_n}K\big(\frac{X_i - x}{h_n}\big)$ for $i = 1, \dots, n$. If K is a symmetric density with unit variance (without loss of generality), each of these terms is a density in $x \in \mathbb{R}$; it corresponds to a random variable centred at X_i with variance $h_n^2 > 0$, see Figure 3.

(ii) When compared to a histogram, none of the estimators $\widetilde{f}_n(x)$ and $\widehat{f}_n(x)$ require to specify the starting point to calculate the intervals.

* *V češtině se mluví o šířce vyhlazovací okna nebo jednodušeji o vyhlazovacním prametru.*

(iii) The function $\widehat{f}_n(x)$ is continuous (has a continuous derivative) if K is continuous (has a continuous derivative). That is why usually a continuous function K is preferred.

(iv) If K is a density of a probability distribution, then $\widehat{f}_n(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{\mathbb{R}} \widehat{f}_n(x)\,\mathrm{d}x = 1$.

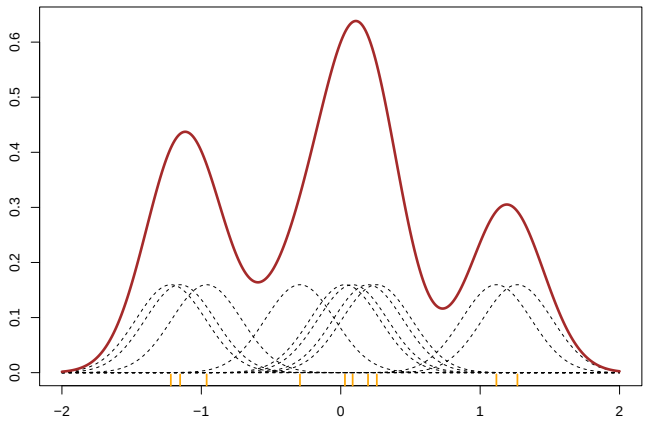

Figure 3: A kernel density estimator (thick brown curve) based on $n = 10$ observations (orange ticks on the horizontal axis). We used the Gaussian kernel K with bandwidth $h = 1/4$. The resulting estimator is a sum of n functions (thin dashed lines), each centred at one X_i, with a scale proportional to the bandwidth h.

Example 26. Consider a random sample of size 200 from the distribution with distribution function

$$F(x) = \tfrac{1}{2}\,\Phi\big(x\big) + \tfrac{1}{2}\,\Phi\big(\tfrac{x-4}{2}\big) \quad \text{for } x \in \mathbb{R},$$

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 corresponding kernel estimates with different bandwidth choices h_n and the Gaussian kernel K are found in Figure 4. For reasons of comparison, also the associated histogram with the width of the columns given by $2\,h_n$ is included.

The true density $f = F'$ is indicated by the black solid line. Note that a reasonable bandwidth seems to be between 0.5 and 1. Bandwidths smaller than 0.5 result in a 'too wiggly' estimate (the variance term of the estimator dominates). On the other hand, bandwidths greater than 1 result in an estimate that is too biased.

Unfortunately, in practice, we do not know what is the true density, which makes it much more difficult to guess what a reasonable bandwidth should be. For the histogram, the problem of the choice of the bandwidth h_n corresponds to the choice of the width of the bars.

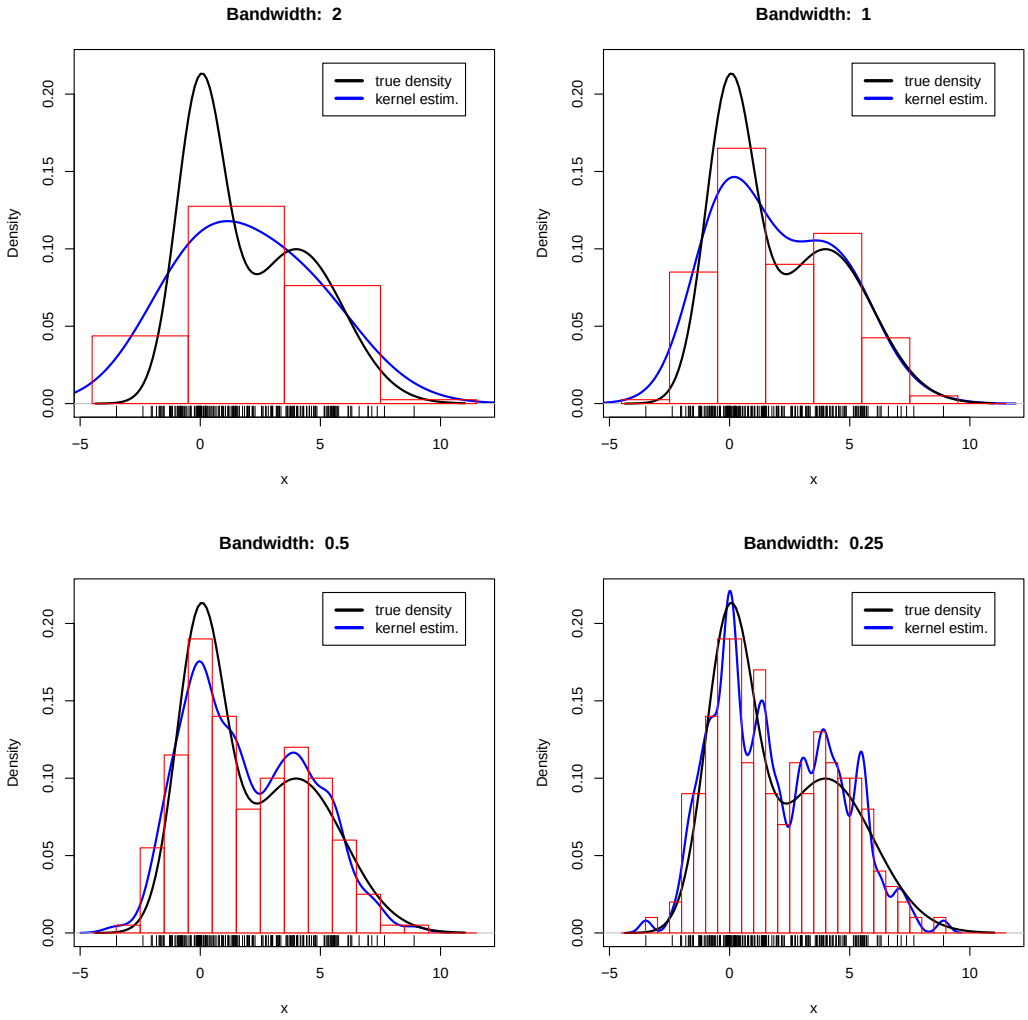

Figure 4: Kernel estimates vs. histograms for different bandwidth choices.

For a general kernel density estimate, the bandwidth corresponds to the width (scaling) of the individual summands, as seen in Figure 3.

2.1 Consistency and asymptotic normality

In what follows, we study the properties of the kernel density estimator (36). Observe that because $X_1, \ldots, X_n$ are independent and identically distributed, the expected value of $\widehat{f}_n(x)$ takes the form

$$\mathsf{E}\ \widehat{f}_n(x) = \mathsf{E}\ \frac{1}{h_n} K\left(\frac{X_1 - x}{h_n}\right) = \frac{1}{h_n} \int_{\mathbb{R}} f(z)\, K\left(\frac{z - x}{h_n}\right) \mathrm{d}\, z, \tag{37}$$

for f the density of X_1. The following theorem will be essential for understanding the behaviour of integrals such as that on the right-hand of (37).

Theorem 5 (Bochner's theorem)**.** *Let the function K satisfy*

(B1) $\int_{\mathbb{R}} |K(y)|\,\mathrm{d}y < \infty$, *and*

(B2) $\lim_{|y|\to\infty} |y\,K(y)| = 0$.

Further let the function g satisfy $\int_{\mathbb{R}} |g(y)|\,\mathrm{d}y < \infty$. Put

$$g_n(x) = \frac{1}{h_n}\int_{\mathbb{R}} g(z)\,K\big(\tfrac{z-x}{h_n}\big)\,\mathrm{d}z,$$

where $h_n \searrow 0$ as $n \to \infty$. Then, in each point x of continuity of g it holds that

$$\lim_{n\to\infty} g_n(x) = g(x)\int_{\mathbb{R}} K(y)\,\mathrm{d}y. \tag{38}$$

Proof. Let x be a point of continuity g. We need to show that

$$\lim_{n\to\infty}\left|g_n(x) - g(x)\int_{\mathbb{R}} K(y)\,\mathrm{d}y\right| = 0.$$

Using the two substitutions $y = z - x$ and $z = \frac{y}{h_n}$ one can write

$$\begin{aligned} g_n(x) - g(x)\int_{\mathbb{R}} K(z)\,\mathrm{d}z &= \frac{1}{h_n}\int_{\mathbb{R}} g(x+y)K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y - \frac{g(x)}{h_n}\int_{\mathbb{R}} K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y \\ &= \frac{1}{h_n}\int_{\mathbb{R}} [g(x+y) - g(x)]K\big(\tfrac{y}{h_n}\big)\,\mathrm{d}y. \end{aligned}$$

Before we proceed, note that for each fixed $\delta > 0$ we have because of $h_n \searrow 0$ and (B2) that

$$\frac{\delta}{h_n} \to \infty \qquad \text{and} \qquad \frac{1}{\delta}\sup_{t:|t|\geq\frac{\delta}{h_n}} \big|t\,K(t)\big| \to 0, \qquad \text{as} \quad n \to \infty. \tag{39}$$

Thus, it is possible to find a sequence of positive constants $\{\delta_n\}_{n=1}^{\infty}$ that converges to zero so slowly so that

$$\delta_n \to 0, \quad \frac{\delta_n}{h_n} \to \infty \quad \text{and} \quad \frac{1}{\delta_n}\sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \to 0, \quad \text{as} \quad n \to \infty. \tag{40}$$

This can be seen as follows. Take any sequence of positive constants $\{a_m\}_{m=1}^{\infty}$ such that $a_m \searrow 0$. Thanks to (39), for each $m \geq 1$ there exists an index $n_m \geq 1$ such that for all $n \geq n_m$ we have

$$a_m/h_n > m, \qquad \text{and} \qquad \frac{1}{a_m}\sup_{t:|t|\geq\frac{a_m}{h_n}} \big|t\,K(t)\big| < \frac{1}{m}. \tag{41}$$

It is surely possible to choose this so that the sequence of integers $\{n_m\}_{m=1}^{\infty}$ is strictly increasing. Define

$$\delta_n = \begin{cases} 1 & \text{if } n < n_1, \\ a_m & \text{if } n \in [n_m, n_{m+1}). \end{cases}$$

We see that for any $n \geq n_1$ we have that $\delta_n = a_m$ implies $n \geq n_m$, meaning that by (41) we have

$$\delta_n/h_n > m, \qquad \text{and} \qquad \frac{1}{\delta_n} \sup_{t:|t|\geq \frac{\delta_n}{h_n}} \big|t\,K(t)\big| < \frac{1}{m}.$$

As $n \to \infty$, also $m \to \infty$ in the definition of δ_n, and we see that we get the sequence $\{\delta_n\}_{n=1}^{\infty}$ as required.

Now, taking δ_n satisfying (40) one can bound

$$\Big|g_n(x) - g(x)\int_{\mathbb{R}} K(y)\,\mathrm{d}y\Big| \leq \underbrace{\frac{1}{h_n}\int_{-\delta_n}^{\delta_n} \big|g(x+y)-g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:A_n} + \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x+y)-g(x)\big|\,\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_n}. \tag{42}$$

Dealing with A_n. As g is continuous in the point x

$$A_n \leq \sup_{y:|y|\leq\delta_n} \big|g(x+y)-g(x)\big| \int_{-\delta_n}^{\delta_n} \tfrac{1}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y \leq o(1) \underbrace{\int_{\mathbb{R}} |K(t)|\,\mathrm{d}t}_{<\infty;\ \text{by (B1)}} = o(1), \tag{43}$$

as $n \to \infty$.

Dealing with B_n. Further, one can bound B_n with

$$B_n \leq \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x+y)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{1,n}} + \underbrace{\frac{1}{h_n}\int_{y:|y|\geq\delta_n} \big|g(x)\big|\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y}_{=:B_{2,n}}. \tag{44}$$

Using the substitution $t = \frac{y}{h_n}$ and (40) one gets

$$B_{2,n} = \big|g(x)\big| \int_{y:|y|\geq\delta_n} \tfrac{1}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|\,\mathrm{d}y = \big|g(x)\big| \int_{t:|t|\geq\frac{\delta_n}{h_n}} |K(t)|\,\mathrm{d}t \xrightarrow[n\to\infty]{} 0. \tag{45}$$

because $\delta_n/h_n \to \infty$ and (B1).

Finally using (40) again, we have

$$\begin{aligned} B_{1,n} &= \int_{y:|y|\geq\delta_n} \underbrace{\tfrac{|y|}{h_n}\big|K\big(\tfrac{y}{h_n}\big)\big|}_{\leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} |t\,K(t)|} \frac{\big|g(x+y)\big|}{|y|}\,\mathrm{d}y \leq \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \int_{y:|y|\geq\delta_n} \frac{\big|g(x+y)\big|}{|y|}\,\mathrm{d}y \\ &\leq \frac{1}{\delta_n} \sup_{t:|t|\geq\frac{\delta_n}{h_n}} \big|t\,K(t)\big| \underbrace{\int_{\mathbb{R}} \big|g(x+y)\big|\,\mathrm{d}y}_{=\int_{\mathbb{R}}|g(y)|\,\mathrm{d}y<\infty} \xrightarrow[n\to\infty]{} 0. \end{aligned} \tag{46}$$

Now combining (42), (43), (44), (45) and (46) yields the statement of the theorem. □

Remark 8. Note that:

(i) If K is a density, then $\int_{\mathbb{R}} |K(y)|\,\mathrm{d}y = \int_{\mathbb{R}} K(y)\,\mathrm{d}y = 1$ 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 5 (dealing with $B_{1,n}$) it follows that for K with bounded support one can even drop the assumption $\int_{\mathbb{R}} |g(y)|\,\mathrm{d}y < \infty$ from Theorem 5. This observation will be useful later when dealing with kernel regression estimators.

(iii) If K is a density but with 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) A direct modification of the proof of Theorem 5 shows that if g is uniformly continuous, then the convergence in (38) is uniform.

(v) 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 6 (Variance and consistency of $\widehat{f}_n(x)$)**.** *Let the estimator $\widehat{f}_n(x)$ be given by (36) and the function K satisfies (B1) and (B2) introduced in Theorem 5. Further, let $\int_{\mathbb{R}} K(y)\,\mathrm{d}y = 1$, $\sup_{y\in\mathbb{R}} |K(y)| < \infty$, $h_n \searrow 0$ as $n \to \infty$ and $(n\,h_n) \to \infty$ as $n \to \infty$. Then at each point $x \in \mathbb{R}$ of continuity of f:*

(i) $\lim_{n\to\infty} n\,h_n\, \mathsf{var}\left(\widehat{f}_n(x)\right) = f(x) \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y$;

(ii) $\widehat{f}_n(x) \xrightarrow[n\to\infty]{P} f(x)$.

Proof. Let x be a point of continuity of f.

Showing (i). Because $X_1, \ldots, X_n$ are assumed to be independent and identically distributed, we can calculate

$$\begin{aligned} \mathsf{var}\left(\widehat{f}_n(x)\right) &= \mathsf{var}\left[\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\tfrac{X_i - x}{h_n}\right)\right] = \frac{1}{n\,h_n^2}\,\mathsf{var}\left[K\left(\tfrac{X_1 - x}{h_n}\right)\right] \\ &= \frac{1}{n\,h_n^2}\left[\mathsf{E}\, K^2\left(\tfrac{X_1 - x}{h_n}\right) - \left(\mathsf{E}\, K\left(\tfrac{X_1 - x}{h_n}\right)\right)^2\right]. \end{aligned} \tag{47}$$

Now using Theorem 5

$$\frac{1}{h_n}\,\mathsf{E}\, K\left(\tfrac{X_1 - x}{h_n}\right) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\tfrac{y - x}{h_n}\right) f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} f(x) \int_{\mathbb{R}} K(y)\,\mathrm{d}y = f(x). \tag{48}$$

Analogously

$$\frac{1}{h_n}\,\mathsf{E}\, K^2\left(\tfrac{X_1 - x}{h_n}\right) = \frac{1}{h_n}\int_{\mathbb{R}} K^2\left(\tfrac{y - x}{h_n}\right) f(y)\,\mathrm{d}y \xrightarrow[n\to\infty]{} f(x) \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y, \tag{49}$$

where we have used again Theorem 5 with the kernel K replaced by K^2. Assumptions (B1) and (B2) are satisfied also for K^2 as

$$\text{ad (B1)}: \int_{\mathbb{R}} |K^2(y)|\,\mathrm{d}y \leq \underbrace{\sup_{y\in\mathbb{R}} |K(y)|}_{<\infty} \underbrace{\int_{\mathbb{R}} |K(y)|\,\mathrm{d}y}_{<\infty;\text{ by (B1) for } K} < \infty$$

and

$$\text{ad (B2)}: \lim_{|y|\to\infty} |yK^2(y)| \leq \underbrace{\sup_{y\in\mathbb{R}} |K(y)|}_{<\infty} \underbrace{\lim_{|y|\to\infty} |yK(y)|}_{=0;\text{ by (B2) for } K} = 0.$$

Now combining (47), (48) and (49) yields

$$n\,h_n\,\mathsf{var}\left(\widehat{f}_n(x)\right) = \underbrace{\frac{1}{h_n}\,\mathsf{E}\,K^2\left(\tfrac{X_1-x}{h_n}\right)}_{\to f(x)\int_{\mathbb{R}} K^2(y)\,\mathrm{d}y} - \underbrace{\left[\frac{1}{h_n}\,\mathsf{E}\,K\left(\tfrac{X_1-x}{h_n}\right)\right]^2}_{\to [f(x)]^2} h_n \xrightarrow[n\to\infty]{} f(x)\int_{\mathbb{R}} K^2(y)\,\mathrm{d}y. \quad (50)$$

Showing (ii). With the help of (48)

$$\mathsf{E}\,\widehat{f}_n(x) = \frac{1}{h_n}\,\mathsf{E}\,K\left(\tfrac{X_1-x}{h_n}\right) \xrightarrow[n\to\infty]{} f(x). \quad (51)$$

Now with the help of (i) and (51)

$$\mathsf{E}\left[\widehat{f}_n(x) - f(x)\right]^2 = \mathsf{var}\left[\widehat{f}_n(x)\right] + \left[\mathsf{E}\,\widehat{f}_n(x) - f(x)\right]^2 \xrightarrow[n\to\infty]{} 0,$$

which implies the consistency of $\widehat{f}_n(x)$ (Kulich and Omelka, 2022, Theorem 3.1). □

Remark 9. Note that Theorem 6 implies only point-wise consistency. It is much more difficult to show that $\sup_{x\in\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \xrightarrow[n\to\infty]{P} 0$, see, e.g., Wied and Weißbach (2012, Theorem 2).

Remark 10. It is not possible to prove the consistency of $\widehat{f}_n(x)$ using the standard law of large numbers, as one would need a law of large numbers for a triangular array.

In the proof of the following theorem and subsequently in the text, we use the notation $R(K) = \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y$ for a square-integrable function $K\colon \mathbb{R} \to \mathbb{R}$. The letter R stands for 'roughness' of K.

The end of lecture 7 (13.11.2024)

Theorem 7 (Asymptotic normality of $\widehat{f}_n(x)$)**.** *Let the assumptions of Theorem 6 be satisfied and further let $x \in \mathbb{R}$ be such 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 6 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 $R(K) = \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y$. Thus thanks to CS (Omelka, 2023, Theorem 2) it is sufficient to consider

$$\frac{\widehat{f}_n(x) - \mathsf{E}\ \widehat{f}_n(x)}{\sqrt{\frac{f(x)R(K)}{n\,h_n}}} = \frac{\frac{1}{\sqrt{n\,h_n}}\sum_{i=1}^n \left[K\left(\frac{X_i-x}{h_n}\right) - \mathsf{E}\ K\left(\frac{X_i-x}{h_n}\right)\right]}{\sqrt{f(x)R(K)}} = \sum_{i=1}^n X_{n,i},$$

where

$$X_{n,i} = \frac{1}{\sqrt{n\,h_n}} \frac{K\left(\frac{X_i-x}{h_n}\right) - \mathsf{E}\ K\left(\frac{X_i-x}{h_n}\right)}{\sqrt{f(x)R(K)}}, \quad i = 1,\ldots,n,$$

are independent and identically distributed random variables (with the distribution depending on n). Thus, the statement follows from the Lindeberg-Feller central limit theorem (Theorem A11), provided its assumptions are satisfied. To apply it, we have to verify its assumptions. We have

$$\mathsf{E}\ X_{n,1} = \cdots = \mathsf{E}\ X_{n,n} = 0.$$

As for the condition for the variance in Theorem A11, we have by part (i) of Theorem 6 that

$$\begin{aligned} n\,h_n\ \mathsf{var}\left(\widehat{f}_n(x)\right) &= n\,h_n\ \mathsf{var}\left(\frac{1}{n\,h_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)\right) \\ &= \frac{1}{h_n}\,\mathsf{var}\left(K\left(\frac{X_1-x}{h_n}\right)\right) \xrightarrow[n\to\infty]{} f(x)\,R(K). \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{i=1}^n \mathsf{var}(X_{n,i}) &= \sum_{i=1}^n \mathsf{var}\left(\frac{1}{\sqrt{n\,h_n}} \frac{K\left(\frac{X_i-x}{h_n}\right) - \mathsf{E}\ K\left(\frac{X_i-x}{h_n}\right)}{\sqrt{f(x)R(K)}}\right) \\ &= n\ \mathsf{var}\left(\frac{1}{\sqrt{n\,h_n}} \frac{K\left(\frac{X_1-x}{h_n}\right)}{\sqrt{f(x)R(K)}}\right) = \frac{1}{h_n\,f(x)\,R(K)}\,\mathsf{var}\left(K\left(\frac{X_1-x}{h_n}\right)\right) \xrightarrow[n\to\infty]{} 1. \end{aligned}$$

Finally for each $\varepsilon > 0$ for all sufficiently large n it holds that uniformly in $i = 1,\ldots,n$

$$\begin{aligned} \mathbb{I}\left\{|X_{n,i}| > \varepsilon\right\} &= \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \left|\frac{K\left(\frac{X_i-x}{h_n}\right) - \mathsf{E}\ K\left(\frac{X_i-x}{h_n}\right)}{\sqrt{f(x)R(K)}}\right| > \varepsilon\right\} \\ &\le \mathbb{I}\left\{\frac{1}{\sqrt{n\,h_n}} \frac{2\sup_{y\in\mathbb{R}}|K(y)|}{\sqrt{f(x)R(K)}} > \varepsilon\right\} = 0, \end{aligned}$$

which implies that the 'Lindeberg-Feller condition' from (A108)

$$\lim_{n\to\infty}\sum_{i=1}^n \mathsf{E}\left[X_{n,i}^2\,\mathbb{I}\left\{|X_{n,i}| > \varepsilon\right\}\right] = 0$$

is satisfied. □

Remark 11. Note that in Theorem 7, we have in the numerator $\widehat{f}_n(x) - \mathsf{E}\,\widehat{f}_n(x)$, but not the usual expression $\widehat{f}_n(x) - f(x)$ that one might expect. In fact, Theorem 7 implies

$$\frac{\widehat{f}_n(x) - f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1), \tag{52}$$

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. We already saw in (50) that $\mathsf{var}\left(\widehat{f}_n(x)\right) = O\left(\frac{1}{n\,h_n}\right)$. As we will see later in (55), typically we have $\mathrm{bias}(\widehat{f}_n(x)) = O(h_n^2)$, which together gives

$$\frac{\mathsf{E}\,\widehat{f}_n(x) - f(x)}{\sqrt{\mathsf{var}\left(\widehat{f}_n(x)\right)}} = \frac{O(h_n^2)}{\sqrt{O\left(\frac{1}{n\,h_n}\right)}} = O\left(\sqrt{n\,h_n^5}\right)$$

and thus $\lim_{n\to\infty} n\,h_n^5 = 0$ is needed to show (52). But this would require that $h_n = o\big(n^{-1/5}\big)$ which would exclude the optimal bandwidth choice (see the next section).

2.2 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 6

$$\mathsf{var}\left(\widehat{f}_n(x)\right) = \frac{f(x)R(K)}{n\,h_n} + o\big(\tfrac{1}{n\,h_n}\big), \tag{53}$$

where $R(K) = \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y$.

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 such that $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1$, $\int_{\mathbb{R}} t\,K(t)\,\mathrm{d}t = 0$ and $\int_{\mathbb{R}} |t^2 K(t)|\,\mathrm{d}t < \infty$. This is true, e.g., if K is a bounded even function with a bounded

support. Then, for all sufficiently large n

$$\begin{aligned}\mathsf{E}\,\widehat{f}_n(x) &= \tfrac{1}{h_n}\,\mathsf{E}\,K\big(\tfrac{X_1-x}{h_n}\big) = \int_{\mathbb{R}} \frac{1}{h_n} K\big(\tfrac{y-x}{h_n}\big) f(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}} K(t) f(x+th_n)\,\mathrm{d}t = \int_{\mathbb{R}} K(t)\left[f(x) + th_n f'(x) + \tfrac{1}{2}\,t^2\,h_n^2\,f''(x) + o(t^2h_n^2)\right]\,\mathrm{d}t \\ &= f(x) + \tfrac{1}{2}\,h_n^2\,f''(x)\,\mu_{2,K} + o(h_n^2),\end{aligned} \tag{54}$$

where $\mu_{2,K} = \int_{\mathbb{R}} y^2\,K(y)\,\mathrm{d}y$. Thus one gets

$$\mathrm{bias}\big(\widehat{f}_n(x)\big) = \mathsf{E}\,\widehat{f}_n(x) - f(x) = \tfrac{1}{2}\,h_n^2\,f''(x)\,\mu_{2,K} + o(h_n^2), \tag{55}$$

which together with (53) implies

$$\mathrm{MSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\,f(x)\,R(K) + \tfrac{1}{4}\,h_n^4\,[f''(x)]^2\mu_{2,K}^2 + o\big(\tfrac{1}{n\,h_n}\big) + o\big(h_n^4\big). \tag{56}$$

Ignoring the remainder $o(\cdot)$ terms in (56), AMSE (asymptotic mean squared error) of $\widehat{f}_n(x)$ is given by

$$\mathrm{AMSE}\big(\widehat{f}_n(x)\big) = \tfrac{1}{n\,h_n}\,f(x)\,R(K) + \tfrac{1}{4}\,h_n^4\,[f''(x)]^2\mu_{2,K}^2. \tag{57}$$

We want to minimise (57) to get an optimal bandwidth choice. Taking a derivative of AMSE with respect to h we get

$$\frac{\mathrm{d}}{\mathrm{d}\,h}\mathrm{AMSE} = -\frac{f(x)\,R(K)}{n\,h^2} + h^3\,[f''(x)]^2\mu_{2,K}^2.$$

Setting this equation equal zero and solving for h one gets, provided that $f''(x) \neq 0$, the *asymptotically optimal local bandwidth* (i.e., bandwidth that minimises the AMSE)

$$h_n^{(opt)}(x) = n^{-1/5}\left[\frac{f(x)\,R(K)}{[f''(x)]^2\,\mu_{2,K}^2}\right]^{1/5}. \tag{58}$$

To get a global bandwidth, it is useful to define **(A)MISE - (asymptotic) mean integrated squared error**. Introduce

$$\mathrm{MISE}\big(\widehat{f}_n\big) = \int_{\mathbb{R}} \mathrm{MSE}\big(\widehat{f}_n(x)\big)\,\mathrm{d}x = \int_{\mathbb{R}} \mathsf{E}\,\big[\widehat{f}_n(x) - f(x)\big]^2\,\mathrm{d}x, \tag{59}$$

and its asymptotic approximation

$$\begin{aligned}\mathrm{AMISE}\big(\widehat{f}_n\big) &= \int_{\mathbb{R}} \mathrm{AMSE}\big(\widehat{f}_n(x)\big)\,\mathrm{d}x = \int_{\mathbb{R}} \tfrac{1}{n\,h_n} f(x)R(K) + \tfrac{[f''(x)]^2\mu_{2,K}^2}{4}\,h_n^4\,\mathrm{d}x \\ &= \frac{R(K)}{n\,h_n} + h_n^4\,\frac{R(f'')\,\mu_{2,K}^2}{4},\end{aligned} \tag{60}$$

where $R(f'') = \int_{\mathbb{R}} \big[f''(x)\big]^2\,\mathrm{d}x$.

Minimising (60) one gets the *asymptotically optimal global bandwidth* (i.e., bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K)}{R(f'')\,\mu_{2,K}^2} \right]^{1/5}. \tag{61}$$

Remark 12. After substitution of the optimal bandwidth (61) into (60) one gets that the optimal AMISE is given by

$$\frac{5\left[R(f'')\right]^{1/5}}{4\,n^{4/5}} \left\{ \left[R(K)\right]^2 \mu_{2,K} \right\}^{2/5}.$$

It can be shown that if we consider kernels that are densities of probability distributions, then $\left[R(K)\right]^2 \mu_{2,K}$ is minimised for K being Epanechnikov kernel, as proved by Müller (1984). Further, note that for $\widetilde{K}(x) = \sqrt{\mu_{2,K}}\, K\left(\sqrt{\mu_{2,K}}\, x\right)$ one has

$$\mu_{2,\widetilde{K}} = 1 \quad \text{and} \quad \left[R(\widetilde{K})\right]^{4/5} = \left[R(K)\right]^{4/5} \mu_{2,K}^{2/5}$$

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 $\mu_{2,K} = 1$. 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:	$K(x) = \frac{1}{\sqrt{6}}\left(1 - \lvert x\rvert\right) \mathbb{I}\left\{\lvert x\rvert \le \sqrt{6}\right\}$
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 $\mu_{2,K} = 1$.

2.2.1 Normal reference rule

The problem of asymptotically optimal bandwidths given in (58) and (61) 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 the 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

$$\begin{aligned} \varphi'(x) &= \tfrac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}(-x) = \tfrac{-x}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} = -x\,\varphi(x), \\ \varphi''(x) &= \tfrac{-1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} + \tfrac{x^2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} = (x^2 - 1)\,\varphi(x). \end{aligned}$$

Thus, with the help of (58) one gets

$$\widehat{h}_n(x) = n^{-\frac{1}{5}} \left[\frac{R(K)}{\mu_{2,K}^2} \frac{\widehat{\sigma}^{-1}\varphi\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)}{\widehat{\sigma}^{-6}\left[\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)^2 - 1\right]^2 \varphi\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)^2} \right]^{\frac{1}{5}}$$
$$= n^{-\frac{1}{5}}\,\widehat{\sigma} \left[\frac{R(K)}{\mu_{2,K}^2} \frac{1}{\left[\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)^2 - 1\right]^2 \varphi\left(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\right)} \right]^{\frac{1}{5}},$$

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

$$\begin{aligned}
R(f'') &= \int_{\mathbb{R}} \left[f''(x)\right]^2 \mathrm{d}x = \int_{\mathbb{R}} \left\{ \frac{1}{\sigma^3}\left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]\varphi\left(\tfrac{x-\mu}{\sigma}\right)\right\}^2 \mathrm{d}x \\
&= \frac{1}{\sigma^6}\int_{\mathbb{R}} \left[\left(\tfrac{x-\mu}{\sigma}\right)^2 - 1\right]^2 \varphi^2\left(\tfrac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \left| \begin{array}{rcl} t &=& \frac{x-\mu}{\sigma} \\ \mathrm{d}t &=& \frac{\mathrm{d}x}{\sigma} \end{array} \right| = \frac{1}{\sigma^5}\int_{\mathbb{R}} (t^2-1)^2 \varphi^2(t)\,\mathrm{d}t \\
&= \frac{1}{\sigma^5}\int_{\mathbb{R}} (t^4 - 2t^2 + 1)\frac{1}{2\pi}\mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{1}{\sigma^5 2\sqrt{\pi}}\int_{\mathbb{R}} (t^4 - 2t^2 + 1)\underbrace{\frac{1}{\sqrt{\pi}}\mathrm{e}^{-t^2}}_{\sim \mathsf{N}\left(0,\frac{1}{2}\right)}\,\mathrm{d}t \\
&= \frac{1}{2\sigma^5\sqrt{\pi}}\,\mathsf{E}\left(Y^4 - 2Y^2 + 1\right) = \frac{1}{2\sigma^5\sqrt{\pi}}\left[3\cdot\left(\tfrac{1}{2}\right)^2 - 2\cdot\tfrac{1}{2} + 1\right] = \frac{3}{8\sigma^5\sqrt{\pi}},
\end{aligned}$$

where $Y \sim \mathsf{N}\left(0, \frac{1}{2}\right)$. Thus the asymptotically optimal global bandwidth would be

$$h_n^{(opt)} = \sigma\, n^{-1/5}\left[\frac{8\sqrt{\pi}\,R(K)}{3\,\mu_{2,K}^2}\right]^{1/5}.$$

Further, if one uses the Gaussian kernel $K(y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}$, one gets

$$\begin{aligned}
\mu_{2,K} &= \int_{\mathbb{R}} y^2 K(y)\,\mathrm{d}y = 1, \\
R(K) &= \int_{\mathbb{R}} K^2(y)\,\mathrm{d}y = \tfrac{1}{2\sqrt{\pi}}\int_{\mathbb{R}} \tfrac{1}{\sqrt{\pi}}\mathrm{e}^{-y^2}\,\mathrm{d}y = \tfrac{1}{2\sqrt{\pi}},
\end{aligned}$$

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

$$h_n = 1.06\, n^{-1/5}\,\min\left\{S_n, \widetilde{IQR}_n\right\}, \tag{62}$$

where

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \qquad \text{and} \qquad \widetilde{IQR}_n = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{1.34}.$$

Here, the constant 1.34 approximately equals $\Phi^{-1}(3/4) - \Phi^{-1}(1/4)$, the inter-quartile range of the standard normal distribution function Φ. The bandwidth (62) is in R implemented as the function bw.nrd.

It was found out that the bandwidth selector (62) works well if the true distribution is 'very close' to the normal distribution. But simultaneously, the bandwidth is usually too large for distributions 'moderately' deviating from a normal distribution. That is why some authors prefer to use

$$h_n = 0.9\, n^{-1/5}\, \min\left\{S_n, \widetilde{IQR}_n\right\}.$$

This choice is in R implemented as bw.nrd0. See Silverman (1986, page 48) for a more detailed argumentation.

The end of lecture 8 (20.11.2024)

2.2.2 Least-squares cross-validation*

Our intention is to find the bandwidth h_n by minimising $\mathrm{MISE}\big(\widehat{f}_n\big)$ from (59). That can be rewritten as

$$\begin{aligned}\mathrm{MISE}\big(\widehat{f}_n\big) &= \int_{\mathbb{R}} \mathsf{E}\,\big(\widehat{f}_n(x) - f(x)\big)^2\,\mathrm{d}x \overset{\text{Fub.}}{=} \mathsf{E}\int_{\mathbb{R}} \Big[\widehat{f}_n(x)\Big]^2 - 2\widehat{f}_n(x)f(x) + [f(x)]^2\,\mathrm{d}x \\ &= \mathsf{E}\int_{\mathbb{R}} \Big[\widehat{f}_n(x)\Big]^2\,\mathrm{d}x - 2\,\mathsf{E}\int_{\mathbb{R}} \widehat{f}_n(x)f(x)\,\mathrm{d}x + \int_{\mathbb{R}} [f(x)]^2\,\mathrm{d}x.\end{aligned}$$

An unbiased estimator of $\mathsf{E}\int_{\mathbb{R}}\Big[\widehat{f}_n(x)\Big]^2\,\mathrm{d}x$ is simply given by $\int_{\mathbb{R}}\Big[\widehat{f}_n(x)\Big]^2\,\mathrm{d}x$. Further, the term $\int_{\mathbb{R}}[f(x)]^2\,\mathrm{d}x$ does not depend on h_n. Thus it remains to estimate $\mathsf{E}\int_{\mathbb{R}}\widehat{f}_n(x)f(x)\,\mathrm{d}x$. The last formula can be interpreted as $\mathsf{E}\,\widehat{f}_n(X)$, for X with the same distribution as the sample variables X_i, but independent from them.

In the sequel we show that an unbiased estimator of $A_n = \mathsf{E}\,\widehat{f}_n(X)$ is

$$\widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i), \tag{63}$$

where

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)h_n}\sum_{j=1,j\neq i}^{n} K\big(\tfrac{X_j - x}{h_n}\big)$$

is the estimate of $f(x)$ that is based the sample without the i-th observation X_i.

* 'cross-validation' se střídavě překládá jako metoda křížového ověřování, metoda křížové validace nebo prostě jako krosvalidace.

Overall, using least-squares cross-validation, we choose the global bandwidth as

$$h_n^{(LSCV)} = \operatorname*{arg\,min}_{h_n>0} \mathcal{L}(h_n),$$

where

$$\mathcal{L}(h_n) = \int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i), \tag{64}$$

with $\widehat{f}_{-i}(x)$ as above.

As our first observation, note that the integral in (64) can be computed directly from the random sample $X_1, \ldots, X_n$, as

$$\begin{aligned}
\int_{\mathbb{R}} \left[\widehat{f}_n(x)\right]^2 \mathrm{d}x &= \int_{\mathbb{R}} \left[\frac{1}{n\,h_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)\right]^2 \mathrm{d}x \\
&= \frac{1}{(n\,h_n)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}} K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \mathrm{d}x \\
&= \frac{1}{n^2\,h_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}} K\left(u\right) K\left(u + \frac{X_j - X_i}{h_n}\right) \mathrm{d}u = \frac{1}{n^2\,h_n}\sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{K}\left(\frac{X_i - X_j}{h_n}\right).
\end{aligned}$$

Here, $\widetilde{K}$ is the so-called convolution kernel of K. It is given by $\widetilde{K}(t) = \int_{\mathbb{R}} K(u)\,K(u-t)\,\mathrm{d}u$, which can for K symmetric be written also as

$$\widetilde{K}(t) = \int_{\mathbb{R}} K(u)\,K(t-u)\,\mathrm{d}u \quad \text{for } t \in \mathbb{R}.$$

If K is seen as a density of a random variable, $\widetilde{K}$ is the density of the sum $Z + Z'$ with Z and Z' independent variables with density K. Thus, $\widetilde{K}$ is usually easy to calculate explicitly.

It remains to show that $\widehat{A}_n$ from (63) is an unbiased estimator of $A_n = \int_{\mathbb{R}} \widehat{f}_n(x) f(x)\,\mathrm{d}x$. For that, we have

$$\mathsf{E}\ \widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\ \widehat{f}_{-i}(X_i).$$

Now with the help of (48) and (51)

$$\begin{aligned}
\mathsf{E}\ \widehat{f}_{-i}(X_i) &= \mathsf{E}\left[\frac{1}{(n-1)h_n}\sum_{j=1, j\neq i}^{n} K\left(\tfrac{X_j - X_i}{h_n}\right)\right] = \frac{1}{h_n}\,\mathsf{E}\,K\left(\tfrac{X_1 - X_2}{h_n}\right) \\
&= \frac{1}{h_n}\int_{\mathbb{R}}\int_{\mathbb{R}} K\left(\tfrac{y-x}{h_n}\right) f(x) f(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}} \underbrace{\left[\int_{\mathbb{R}} \frac{1}{h_n} K\left(\tfrac{y-x}{h_n}\right) f(y)\,\mathrm{d}y\right]}_{=\mathsf{E}\,\widehat{f}_n(x)} f(x)\,\mathrm{d}x \qquad (65) \\
&= \int_{\mathbb{R}} \mathsf{E}\,\widehat{f}_n(x) f(x)\,\mathrm{d}x \overset{\text{Fub.}}{=} \mathsf{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x)\,\mathrm{d}x.
\end{aligned}$$

Thus $\widehat{A}_n$ is an unbiased estimator of $\mathsf{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x)\,\mathrm{d}x$ and $\mathcal{L}(h_n)$ is an unbiased estimator of $\mathsf{E}\int_{\mathbb{R}} [\widehat{f}_n(x)]^2\,\mathrm{d}x - 2\,\mathsf{E}\int_{\mathbb{R}} \widehat{f}_n(x) f(x)\,\mathrm{d}x$.

In R, this method can be found implemented as `bw.ucv` (unbiased cross-validation).

Remark 13. Stone (1984) 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 $\mathrm{ISE}(h_n) = \int_{\mathbb{R}} (\widehat{f}_n(x) - f(x))^2 \,\mathrm{d}x$. But, simulations show that the variance of $h_n^{(LSCV)}$ (for not too large sample sizes) is rather large. Thus, this method cannot be used blindly.

2.2.3 Biased cross-validation

This method minimises the AMISE given by (60), that is

$$\mathrm{AMISE}\left(\widehat{f}_n\right) = \frac{R(K)}{n\,h_n} + h_n^4\,\frac{R(f'')\,\mu_{2,K}^2}{4}.$$

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

$$\mathcal{B}(h_n) = \frac{R(K)}{n\,h_n} + \tfrac{1}{4}\,h_n^4\,\mu_{2,K}^2\left[R\big(\widehat{f}_n''\big) - \frac{R(K'')}{n\,h_n^5}\right]$$

is the estimated counterpart of AMISE. In R, this method can be found implemented as bw.bcv (biased cross-validation).

Remark 14. It can be proved that $\frac{h_n^{(BCV)}}{h_n^{(opt)}} \xrightarrow[n\to\infty]{P} 1$, where $h_n^{(opt)}$ is given by (61).

2.3 Higher order kernels

In the same way as when we evaluated the bias of $\widehat{f}_n(x)$ in (54), a formal application of Taylor's expansion (for sufficiently large n, sufficiently smooth f and x an interior point of the support) one gets

$$\begin{aligned}
\mathsf{E}\,\widehat{f}_n(x) &= \int_{\mathbb{R}} K(t) f(x+th_n)\,\mathrm{d}t \\
&= f(x)\int_{\mathbb{R}} K(t)\,\mathrm{d}t + f'(x)h_n\int_{\mathbb{R}} tK(t)\,\mathrm{d}t \\
&\quad + \frac{f''(x)}{2}h_n^2\int_{\mathbb{R}} t^2K(t)\,\mathrm{d}t + \frac{f'''(x)}{3!}h_n^3\int_{\mathbb{R}} t^3K(t)\,\mathrm{d}t + \ldots.
\end{aligned}$$

The kernel of order p is such that $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1$ and

$$\int_{\mathbb{R}} t^j\,K(t)\,\mathrm{d}t = 0, \qquad j = 1,\ldots,p-1, \qquad \text{and} \qquad \int_{\mathbb{R}} t^p\,K(t)\,\mathrm{d}t \neq 0.$$

Considering a kernel K of order $p > 2$, we can thus conclude that for the bias of $\widehat{f}_n(x)$ we have

$$\mathsf{bias}\left(\widehat{f}_n(x)\right) = O(h_n^p),$$

which converges to zero faster than $O(h_n^2)$ that we obtained for bias in (54). Thus, it might seem that higher-order kernels might be preferable to the standard choice of the second-order kernel from before.

However, if we have a kernel of order $p > 2$, then (among others) necessarily $\int_{\mathbb{R}} t^2 K(t)\,\mathrm{d}t = \mu_{2,K} = 0$, which implies that K cannot be non-negative. As a consequence, with a kernel of order $p > 2$ it might happen that the estimator $\widehat{f}_n(x)$ is negative.

One possible modification of a Gaussian kernel to get a kernel of order 4 is given by

$$K(y) = \tfrac{1}{2}\,(3 - y^2)\,\tfrac{1}{\sqrt{2\,\pi}}\,\mathrm{e}^{-y^2/2} \qquad \text{for } y \in \mathbb{R}.$$

2.4 Mirror-reflection

The standard kernel density estimator (36) is usually not consistent at 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 to improve the performance of $\widehat{f}_n(x)$ close to the boundary points. The most straightforward is the *mirror-reflection method*.

Suppose for simplicity that we know that the support of the distribution with the density f is $[0, \infty)$, and let K be an even function. The modified kernel density estimator that uses mirror-reflection is given by

$$\widehat{f}_n^{(MR)}(x) = \begin{cases} \frac{1}{n\,h_n}\sum_{i=1}^n K\big(\frac{X_i - x}{h_n}\big) + \frac{1}{n\,h_n}\sum_{i=1}^n K\big(\frac{X_i + x}{h_n}\big), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{66}$$

The first term on the right-hand side of (66) (for $x \geq 0$) is the standard kernel density estimator $\widehat{f}_n(x)$. The second term on the right-hand side of (66) 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 $[0, \infty)$. The mirror-reflected density estimator $\widehat{f}_n^{(MR)}(x)$ can be written also in a more compact form

$$\widehat{f}_n^{(MR)}(x) = \left(\widehat{f}_n(x) + \widehat{f}_n(-x)\right) \mathbb{I}\{x \geq 0\}.$$

2.5 Multivariate kernel density estimation

Suppose now that we observe multivariate ($\mathbb{R}^d$-valued) random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that are independent and sampled from a distribution with density $f\colon \mathbb{R}^d \to [0, \infty)$. The role of a

kernel is now played by a function $K\colon \mathbb{R}^d \to \mathbb{R}$, which is typically chosen to be a d-dimensional density, e.g., the d-variate standard Gaussian density.

For one-dimensional data, the bandwidth h_n was interpreted as a factor multiplying the random variable Z with density K. Using the random variable Z, we saw in Theorem 6 that the expected value $\mathsf{E}\,\widehat{f}_n(x)$ could be interpreted as the density of the random variable $X + h_n\,Z$ for X and Z independent, evaluated at x.

For multivariate data, we proceed analogously. Let K be a density, and let $\mathbf{Z}$ be a d-variate random vector with this density. This time, however, we multiply $\mathbf{Z}$ by a matrix of constants $\mathbf{A} \in \mathbb{R}^{d\times d}$, obtaining a random variable $\mathbf{AZ}$ whose density is

$$K_{\mathbf{A}}(\mathbf{x}) = \frac{1}{\det(\mathbf{A})} K\left(\mathbf{A}^{-1}\mathbf{x}\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^d.$$

For simplicity, suppose that K is a standard Gaussian density. The kernel $K_{\mathbf{A}}$ then corresponds to the d-variate Gaussian distribution with zero mean and variance $\mathbf{H} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$. The role of the bandwidth is now played by $\mathbf{H}$, which is assumed to be positive definite. A natural extension of the univariate kernel density estimator to $\mathbb{R}^d$ is

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\det(\mathbf{H})^{1/2}} K\left(\mathbf{H}^{-1/2}\left(\boldsymbol{X}_i - \mathbf{x}\right)\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^d.$$

The reason why we consider as a bandwidth any positive definite matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ is the flexibility this choice borrows, see Figure 5. Considering different matrices $\mathbf{H}$, we are not restricting only to kernels associated with (multiples of) the standard normal distribution $\mathbf{Z}$ (as we see on the left-hand panel of Figure 5), but also kernels of different shapes represented by $\mathbf{H}$, and associated with the elliptically symmetric distributions $\mathbf{H}^{1/2}\mathbf{Z}$ (right-hand panel of Figure 5). On the other hand, the choice of the bandwidth parameters represented by the matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ becomes more involved.

Using analysis similar to what we did for $d = 1$ in Section 2.2, it is possible to show that for kernel density estimators in $\mathbb{R}^d$, the mean integrated squared error is of order

$$\mathrm{MISE}\big(\widehat{f}_n\big) = \int_{\mathbb{R}^d} \mathrm{MSE}\big(\widehat{f}_n(\mathbf{x})\big)\,\mathrm{d}\mathbf{x} = O(h_n^4) + O\left(\frac{1}{n\,h_n^d}\right), \tag{67}$$

where $h_n > 0$ measures the "size" of the bandwidth matrix $\mathbf{H} = h_n\mathbf{H}_0$, for some $\mathbf{H}_0 \in \mathbb{R}^{d\times d}$ fixed. In particular, compared to the density estimation with $d = 1$ and formula (59), the exponent d in the variance term in (67) says that with growing dimension d, kernel density estimation becomes much more difficult.

Literature: Wand and Jones (1995, Sections 2.5, 3.2, 3.3), Scott (2015).

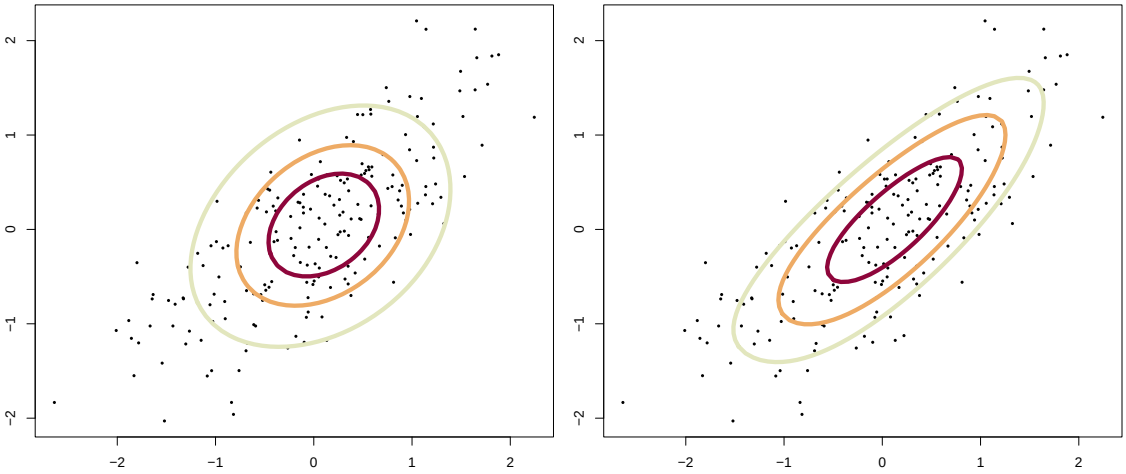

Figure 5: Several contours of two bivariate kernel density estimates with $\mathbf{H}$ the identity matrix (left), and $\mathbf{H}$ chosen to be $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ (right). As can be seen, the form of the bandwidth matrix changes the shape of the resulting estimator profoundly.

3 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.

$$m(x) = \mathsf{E}\left[Y_1 \,|\, X_1 = x\right] \quad \text{for } x \in \mathbb{R},$$

without assuming any parametric form of $m(x)$.

In what follows, it is useful to denote the conditional variance function as

$$\sigma^2(x) = \mathsf{var}\left[Y_1 \,|\, X_1 = x\right] \quad \text{for } x \in \mathbb{R}.$$

3.1 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 using the Taylor polynomial as

$$m(X_i) \doteq m(x) + m'(x)\,(X_i - x) + \ldots + \frac{m^{(p)}(x)}{p!}\,(X_i - x)^p. \tag{68}$$

* *Jádrové regresní odhady*

Thus 'locally', one can view and estimate the function $m(x)$ as a polynomial. This motivates the definition of the *local polynomial estimator* as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(x) &= \big(\widehat{\beta}_0(x), \dots, \widehat{\beta}_p(x)\big)^{\mathsf{T}} \\ &= \underset{b_0,\dots,b_p\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1(X_i - x) - \dots - b_p(X_i - x)^p\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big), \end{aligned} \tag{69}$$

where K is a given kernel function and h_n is a smoothing parameter (bandwidth) going to zero as $n \to \infty$.

The end of lecture 9 (27.11.2024)

Comparing (68) and (69) 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 write $\mathbb{W}(x)$ for the diagonal matrix with the i-th element of the diagonal given by $K\big(\frac{X_i - x}{h_n}\big)$.

The optimisation problem in (69) can be written as a weighted least squares problem

$$\widehat{\boldsymbol{\beta}}(x) = \underset{\mathbf{b}\in\mathbb{R}^{p+1}}{\arg\min} \Big\{ \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big)^{\mathsf{T}} \mathbb{W}(x) \big(\mathbb{Y} - \mathbb{X}_p(x)\,\mathbf{b}\big) \Big\}, \tag{70}$$

where $\mathbf{b} = (b_0, b_1, \dots, b_p)^{\mathsf{T}}$. This is very similar to the situation with general linear models considered in Komárek (2021, Chapter 15); the only difference is that here, the matrix of weights $\mathbb{W}(x)$ depends on x. The solution of (70) can be explicitly written as

$$\widehat{\boldsymbol{\beta}}(x) = \Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{Y}, \tag{71}$$

provided that the matrix $\Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p\Big)$ is non-singular.

The estimator of $m(x)$ is $\widehat{\beta}_0(x)$, the first element of the vector $\widehat{\boldsymbol{\beta}}(x)$. From (71) we get that if we denote by $(w_{n,1}(x), \dots, w_{n,n}(x))^{\mathsf{T}} \in \mathbb{R}^n$ the first row of the matrix

$$\mathbb{H}(x) = \Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x), \tag{72}$$

then $\widehat{\beta}_0(x)$ can be written in the form

$$\widehat{\beta}_0(x) = \sum_{i=1}^{n} w_{n,i}(x)\, Y_i.$$

Because

$$\mathbb{H}(x)\mathbb{X}_p(x) = \Big(\mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_p(x) = \mathbb{I}_{p+1}$$

is the identity matrix of size $p+1$, looking at the first row of the last formula, we get that

$$\sum_{i=1}^{n} w_{n,i}(x) = 1,$$

and

$$\sum_{i=1}^{n} w_{n,i}(x)\,(X_i - x)^{\ell} = 0 \quad \text{for all } \ell = 1, \ldots, p. \tag{73}$$

In particular, $\widehat{\beta}_0(x)$ is a special weighted average of the responses Y_i, $i = 1, \ldots, n$.

The following technical lemma will be useful in deriving the properties of the local polynomial estimator.

Lemma 1. *Let*

- *the kernel K be bounded, symmetric around zero, positive on its support $(-1, 1)$, and such that $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 1$,*
- *$h_n \to 0$ and $(n\,h_n) \to \infty$, and*
- *suppose that the density f_X of X_1 is positive and twice differentiable at x.*

For $\ell \in \mathbb{N} \cup \{0\}$ put

$$S_{n,\ell}(x) = \frac{1}{n} \sum_{i=1}^{n} \tfrac{1}{h_n} K\big(\tfrac{X_i - x}{h_n}\big) \big(\tfrac{X_i - x}{h_n}\big)^{\ell}.$$

Then

$$S_{n,\ell}(x) = \begin{cases} f_X(x) \int_{\mathbb{R}} K(t)\,t^{\ell}\,\mathrm{d}t + \frac{h_n^2}{2} f_X''(x) \int_{\mathbb{R}} K(t)\,t^{\ell+2}\,\mathrm{d}t + o(h_n^2) + O_P\big(\frac{1}{\sqrt{n\,h_n}}\big), & \ell \text{ even}, \\ h_n\, f_X'(x) \int_{\mathbb{R}} K(t)\,t^{\ell+1}\,\mathrm{d}t + o(h_n^2) + O_P\big(\frac{1}{\sqrt{n\,h_n}}\big), & \ell \text{ odd}. \end{cases}$$

Proof. For $\widehat{f}_n(x) = S_{n,0}(x)$ we proved in Theorems 6 and 7 that

$$\sqrt{n\,h_n}\,(S_{n,0}(x) - \mathsf{E}\; S_{n,0}(x)) \xrightarrow[n\to\infty]{d} \mathsf{N}\left(0, f_X(x) \int_{\mathbb{R}} K^2(t)\,\mathrm{d}t\right) = \mathsf{N}\left(0, f_X(x) R(K)\right).$$

For other $S_{n,\ell}(x)$ we apply the same theorems, but with kernel functions $\widetilde{K}(t) = K(t)\,t^{\ell}$, $\ell = 1, 2, \ldots$. Because we assume that the support of K is bounded, conditions (B1) and (B2) from Theorem 5 are trivially satisfied for $\widetilde{K}$. We can thus apply Theorems 6 and 7 also to the kernel $\widetilde{K}$, and get that

$$\sqrt{nh_n}\,\big(S_{n,\ell}(x) - \mathsf{E}\; S_{n,\ell}(x)\big) \xrightarrow[n\to\infty]{d} \mathsf{N}\big(0, \sigma^2(x)\big),$$

where

$$\sigma^2(x) = f_X(x) R(\widetilde{K}) = f_X(x) \int_{\mathbb{R}} t^{2\ell}\,K^2(t)\,\mathrm{d}t.$$

Thus

$$S_{n,\ell}(x) = \mathsf{E}\ S_{n,\ell}(x) + (S_{n,\ell}(x) - \mathsf{E}\ S_{n,\ell}(x)) = \mathsf{E}\ S_{n,\ell}(x) + O_P\big(\tfrac{1}{\sqrt{n h_n}}\big)$$

and it remains to calculate $\mathsf{E}\ S_{n,\ell}(x)$. 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

$$\begin{aligned}
\mathsf{E}\ S_{n,\ell}(x) &= \mathsf{E}\ \tfrac{1}{h_n} K\big(\tfrac{X_1-x}{h_n}\big)\big(\tfrac{X_1-x}{h_n}\big)^\ell = \int_{\mathbb{R}} \tfrac{1}{h_n} K\big(\tfrac{y-x}{h_n}\big)\big(\tfrac{y-x}{h_n}\big)^\ell f_X(y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}} K(t)\,t^\ell\, f_X(x + t h_n)\,\mathrm{d}t \\
&= f_X(x)\int_{\mathbb{R}} K(t) t^\ell\,\mathrm{d}t + h_n f_X'(x)\int_{\mathbb{R}} K(t) t^{\ell+1}\,\mathrm{d}t + \tfrac{h_n^2}{2} f_X''(x)\int_{\mathbb{R}} K(t) t^{\ell+2}\,\mathrm{d}t + o(h_n^2).
\end{aligned}$$

As K is symmetric, we get that $\int_{\mathbb{R}} K(t)t^{\ell+1}\,\mathrm{d}t = 0$ for ℓ even and $\int_{\mathbb{R}} K(t)t^{\ell+2}\,\mathrm{d}t = 0$ for ℓ odd. □

Remark 15. Lemma 1 implies that

$$S_{n,0}(x) = f_X(x) + \tfrac{h_n^2}{2} f_X''(x)\,\mu_{2,K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n h_n}}\big) = f_X(x) + o_P(1), \tag{74}$$

$$S_{n,1}(x) = h_n\, f_X'(x)\,\mu_{2,K} + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n h_n}}\big) = o_P(1), \tag{75}$$

$$S_{n,2}(x) = f_X(x)\,\mu_{2,K} + o_P(1), \tag{76}$$

$$S_{n,3}(x) = h_n\, f_X'(x)\int_{\mathbb{R}} t^4\,K(t)\,\mathrm{d}t + o(h_n^2) + O_P\big(\tfrac{1}{\sqrt{n h_n}}\big) = o_P(1). \tag{77}$$

The first expression (74) agrees with the bias for the kernel density estimator that we derived in (54).

3.2 Nadaraya-Watson estimator

For $p = 0$ the local polynomial estimator given by (69) simplifies to

$$\widehat{\beta}_0(x) = \underset{b_0\in\mathbb{R}}{\arg\min} \sum_{i=1}^n \Big[Y_i - b_0\Big]^2 K\big(\tfrac{X_i-x}{h_n}\big),$$

and solving this optimisation task one gets

$$\widehat{\beta}_0(x) = \sum_{i=1}^n w_{n,i}(x)\,Y_i =: \widehat{m}_{NW}(x),$$

where

$$w_{n,i}(x) = \frac{K\big(\tfrac{X_i-x}{h_n}\big)}{\sum_{j=1}^n K\big(\tfrac{X_j-x}{h_n}\big)} = \frac{\tfrac{1}{n\,h_n} K\big(\tfrac{X_i-x}{h_n}\big)}{S_{n,0}(x)}.$$

This estimator is in the context of the local polynomial regression also called the **locally constant estimator**.

For each x for which the weights are defined we have

$$\sum_{i=1}^{n} w_{n,i}(x) = 1.$$

Moreover, if the kernel K is a non-negative function then also the weights are non-negative.

Remark 16. Let us consider the kernel with the support $[-1, 1]$. Then $w_{n,i}(x)$ is zero if $X_i \notin [x - h_n, x + h_n]$. Further, if we assume the uniform kernel, i.e. $K(x) = \frac{1}{2}\,\mathbb{I}\{|x| \le 1\}$, then all the weights $w_{n,i}(x)$ 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 those observations Y_i for which $X_i \in [x - h_n, x + h_n]$, i.e.

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i\,\mathbb{I}\{|X_i - x| \le h_n\}}{\sum_{j=1}^{n} \mathbb{I}\{|X_j - x| \le h_n\}}.$$

Thus one can view $\widehat{m}_{NW}(x)$ as a 'moving average' in the covariate direction.

To formulate theoretic properties of the estimator $\widehat{m}_{NW}(x)$ 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 8. *Suppose that the assumptions of Lemma 1 are satisfied and further suppose that*

- $(n\,h_n^3) \xrightarrow[n\to\infty]{} \infty$,
- *the function $m(\cdot)$ is twice differentiable at the point x, and*
- *the function $\sigma^2(\cdot)$ is continuous at the point x.*

Then

$$\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2,K}\left(\frac{m'(x)\,f_X'(x)}{f_X(x)} + \frac{m''(x)}{2}\right) + o_P(h_n^2), \tag{78}$$

$$\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big), \tag{79}$$

where

$$R(K) = \int_{\mathbb{R}} K^2(t)\,\mathrm{d}t \qquad \text{and} \qquad \mu_{2,K} = \int_{\mathbb{R}} t^2\,K(t)\,\mathrm{d}t. \tag{80}$$

Proof. Showing (78). Let us calculate

$$\begin{aligned}
\mathsf{E}\left[\widehat{m}_{NW}(x)|\mathbb{X}\right] &= \sum_{i=1}^{n} w_{n,i}(x)\ \mathsf{E}\left[Y_i|\mathbb{X}\right] = \sum_{i=1}^{n} w_{n,i}(x)\ \mathsf{E}\left[Y_i|X_i\right] = \sum_{i=1}^{n} w_{n,i}(x) m(X_i) \\
&= \sum_{i=1}^{n} w_{n,i}(x) \left[m(x) + (X_i - x)\, m'(x) + \tfrac{(X_i-x)^2}{2}\, m''(x) + (X_i - x)^2\, \tilde{R}(X_i)\right] \\
&= m(x) \sum_{i=1}^{n} w_{n,i}(x) + m'(x) \sum_{i=1}^{n} w_{n,i}(x)(X_i - x) + \frac{m''(x)}{2} \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2 \\
&\quad + \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2 \tilde{R}(X_i) \\
&= m(x) + m'(x)\, A_n + \frac{m''(x)}{2}\, B_n + C_n, \qquad (81)
\end{aligned}$$

where $\tilde{R}(z) \to 0$ as $z \to x$ and

$$A_n = \sum_{i=1}^{n} w_{n,i}(x)(X_i - x),\ B_n = \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2,\ C_n = \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2 \tilde{R}(X_i). \quad (82)$$

Now with the help of (74) and (75)

$$\begin{aligned}
A_n = \sum_{i=1}^{n} w_{n,i}(x)(X_i - x) &= \frac{h_n \sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right)\left(\frac{X_i - x}{h_n}\right)}{\sum_{j=1}^{n} \frac{1}{h_n} K\left(\frac{X_j - x}{h_n}\right)} = \frac{h_n S_{n,1}(x)}{S_{n,0}(x)} \\
&= \frac{h_n \left[h_n f'_X(x)\, \mu_{2,K} + o(h_n^2) + O_P\left(\frac{1}{\sqrt{nh_n}}\right)\right]}{f_X(x) + o_P(1)} = \frac{h_n^2 f'_X(x) \mu_{2,K} + o(h_n^3) + O_P\left(\frac{h_n}{\sqrt{nh_n}}\right)}{f_X(x) + o_P(1)} \\
&= \frac{h_n^2 f'_X(x) \mu_{2,K}}{f_X(x)} + o_P(h_n^2) + O_P\Big(\frac{h_n^2}{\sqrt{nh_n^3}}\Big) = \frac{h_n^2 f'_X(x) \mu_{2,K}}{f_X(x)} + o_P(h_n^2), \qquad (83)
\end{aligned}$$

as $(nh_n^3) \to \infty$. Further, with the help of (74) and (76)

$$\begin{aligned}
B_n = \sum_{i=1}^{n} w_{n,i}(X_i - x)^2 &= \frac{h_n^2 S_{n,2}(x)}{S_{n,0}(x)} \\
&= \frac{h_n^2 \left[f_X(x)\, \mu_{2,K} + o_P(1)\right]}{f_X(x) + o_P(1)} = h_n^2 \mu_{2,K} + o_P(h_n^2). \qquad (84)
\end{aligned}$$

Concerning C_n, thanks to (84) and the fact that the support of K is $(-1, 1)$ one can bound

$$\begin{aligned}
|C_n| = \left|\sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2 \tilde{R}(X_i)\right| &\le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)| \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^2 \\
&= o(1)\, B_n = o(1)\, O_P(h_n^2) = o_P(h_n^2). \qquad (85)
\end{aligned}$$

Now combining (83), (84) and (85) one gets

$$\mathsf{E}\left[\widehat{m}_{NW}(x)|\mathbb{X}\right] = m(x) + m'(x) h_n^2 \frac{f'_X(x)}{f_X(x)} \mu_{2,K} + \frac{m''(x)}{2} h_n^2\, \mu_{2,K} + o_P(h_n^2),$$

which implies (78).

Showing (79). Let us calculate

$$\begin{aligned}\mathsf{var}[\widehat{m}_{NW}(x)|\mathbb{X}] &= \sum_{i=1}^{n} w_{n,i}^2(x)\,\mathsf{var}[Y_i|X_i] = \sum_{i=1}^{n} w_{n,i}^2(x)\sigma^2(X_i)\\ &= \frac{\sum_{i=1}^{n} K^2\big(\frac{X_i-x}{h_n}\big)\sigma^2(X_i)}{\Big[\sum_{j=1}^{n} K\big(\frac{X_j-x}{h_n}\big)\Big]^2} = \frac{1}{nh_n}\frac{V_n}{[S_{n,0}(x)]^2},\end{aligned}$$

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 6 was proved that $\widehat{f}_n(x) \xrightarrow[n\to\infty]{P} f(x)$, in the rest of this proof we will show that

$$V_n \xrightarrow[n\to\infty]{P} f_X(x)\,\sigma^2(x)\,R(K), \tag{86}$$

which combined with (74) implies (79).

Showing (86). First, with the help of Bochner's theorem (Theorem 5)

$$\begin{aligned}\mathsf{E}\,V_n &= \frac{1}{h_n}\,\mathsf{E}\left[K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\right]\\ &= \int_{\mathbb{R}} \frac{1}{h_n}K^2\big(\tfrac{z-x}{h_n}\big)\,\sigma^2(z)f_X(z)\,\mathrm{d}z \xrightarrow[n\to\infty]{} \sigma^2(x)f_X(x)\int_{\mathbb{R}} K^2(t)\,\mathrm{d}t.\end{aligned}$$

Now it remains to show that $\mathsf{var}(V_n) \xrightarrow[n\to\infty]{} 0$. Using again Bochner's theorem (Theorem 5)

$$\begin{aligned}\mathsf{var}(V_n) &= \frac{1}{nh_n^2}\left[\mathsf{E}\,K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1) - \Big(\mathsf{E}\,K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\Big)^2\right]\\ &= \frac{1}{nh_n}\left[\frac{1}{h_n}\,\mathsf{E}\,K^4\big(\tfrac{X_1-x}{h_n}\big)\sigma^4(X_1)\right] - \frac{1}{n}\left[\frac{1}{h_n}\,\mathsf{E}\,K^2\big(\tfrac{X_1-x}{h_n}\big)\sigma^2(X_1)\right]^2\\ &= \frac{1}{nh_n}\left[\sigma^4(x)f_X(x)\int_{\mathbb{R}} K^4(t)\,\mathrm{d}t + o(1)\right] - \frac{1}{n}\left[\sigma^2(x)f_X(x)\int_{\mathbb{R}} K^2(t)\,\mathrm{d}t + o(1)\right]^2\\ &\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

□

The end of lecture 10 (04.12.2024)

3.3 Local linear estimator

For $p = 1$ the local polynomial estimator given by (69) simplifies to

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \underset{b_0,b_1\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i-x}{h_n}\big).$$

To solve this optimisation task, one needs to find the first row of the matrix

$$\Big(\mathbb{X}_1^{\mathsf{T}}(x)\,\mathbb{W}(x)\,\mathbb{X}_1(x)\Big)^{-1}\mathbb{X}_1^{\mathsf{T}}(x)\,\mathbb{W}(x) \tag{87}$$

from (72) with

$$\mathbb{X}_1(x) = \begin{pmatrix} 1 & (X_1 - x) \\ 1 & (X_2 - x) \\ \dots & \dots \\ 1 & (X_n - x) \end{pmatrix},$$

and $\mathbb{W}(x)$ the diagonal matrix with the i-th element of the diagonal given by $K\big(\frac{X_i - x}{h_n}\big)$. We have

$$\begin{aligned} \mathbb{X}_1^{\mathsf{T}}(x)\, \mathbb{W}(x)\, \mathbb{X}_1(x) &= \begin{pmatrix} \sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big) & \sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)\,(X_i - x) \\ \sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)\,(X_i - x) & \sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)\,(X_i - x)^2 \end{pmatrix} \\ &= n\, h_n \begin{pmatrix} S_{n,0}(x) & h_n\, S_{n,1}(x) \\ h_n\, S_{n,1}(x) & h_n^2\, S_{n,2}(x) \end{pmatrix}. \end{aligned}$$

Inverting this matrix and plugging into (87), one gets

$$\widehat{\beta}_0(x) = \sum_{i=1}^n w_{n,i}(x)\, Y_i =: \widehat{m}_{LL}(x),$$

where the (local linear) weights can be written in the form

$$w_{n,i}(x) = \frac{\frac{1}{n\,h_n}\, K\big(\frac{X_i-x}{h_n}\big)\big(S_{n,2}(x) - \frac{X_i-x}{h_n}\, S_{n,1}(x)\big)}{S_{n,0}(x)\, S_{n,2}(x) - S_{n,1}^2(x)}, \qquad i = 1, \dots, n. \tag{88}$$

From (73) it follows (see also Remark 17 below) that the weights satisfy (for each x so that the weights are defined)

$$\sum_{i=1}^n w_{n,i}(x) = 1, \qquad \sum_{i=1}^n w_{n,i}(x)(X_i - x) = 0. \tag{89}$$

On the other hand, it might happen that the weights are negative. In practice, this happens if x is either 'close' to the minimal or maximal value of the covariate.

Remark 17. Formula (89) is possible to be seen also directly, as

$$\sum_{i=1}^n w_{n,i}(x) = \frac{S_{n,0}(x)\, S_{n,2}(x) - S_{n,1}^2(x)}{S_{n,0}(x)\, S_{n,2}(x) - S_{n,1}^2(x)} = 1$$

and

$$\begin{aligned} \sum_{i=1}^n w_{n,i}(x)(X_i - x) &= \frac{1}{nh_n} \frac{\sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)(X_i - x)S_{n,2}(x) - \sum_{i=1}^n K\big(\frac{X_i-x}{h_n}\big)\frac{(X_i-x)^2}{h_n} S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} \\ &= h_n \frac{S_{n,1}(x)\, S_{n,2}(x) - S_{n,2}(x)\, S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} = 0. \end{aligned}$$

Theorem 9. *Suppose that the assumptions of Theorem 8 hold. Then*

$$\mathsf{bias}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2,K}\,\frac{m''(x)}{2} + o_P(h_n^2), \tag{90}$$

$$\mathsf{var}\left(\widehat{m}_{LL}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big), \tag{91}$$

where $R(K)$ and $\mu_{2,K}$ are given in (80).

By Theorem 8 for the Nadaraya-Watson estimator one has

$$\mathsf{bias}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = h_n^2\,\mu_{2,K}\left(\frac{m'(x)\,f_X'(x)}{f_X(x)} + \frac{m''(x)}{2}\right) + o_P(h_n^2),$$

$$\mathsf{var}\left(\widehat{m}_{NW}(x)|\mathbb{X}\right) = \frac{\sigma^2(x)\,R(K)}{f_X(x)\,n\,h_n} + o_P\big(\tfrac{1}{n\,h_n}\big).$$

We see that the main terms in the approximation of the conditional variances of $\widehat{m}_{NW}(x)$ and $\widehat{m}_{LL}(x)$ are the same, 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 $h_n^2\,\mu_{2,K}\,\frac{m'(x)\,f_X'(x)}{f_X(x)}$. This is the reason why the authors usually prefer $\widehat{m}_{LL}(x)$ to $\widehat{m}_{NW}(x)$.

Proof of Theorem 9. Showing (90). Completely analogously as in the proof of Theorem 8 one can arrive at (81) with the only difference that now the weights $w_{n,i}(x)$ are given by (88). Now with the help of (89)

$$A_n = \sum_{i=1}^{n} w_{n,i}(x)(X_i - x) = 0. \tag{92}$$

Further using (74), (75), (76) and (77)

$$\begin{aligned} B_n &= \sum_{i=1}^{n} w_{n,i}(x)\tfrac{(X_i-x)^2}{h_n^2}\,h_n^2 = h_n^2\,\frac{S_{n,2}^2(x) - S_{n,3}(x)S_{n,1}(x)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)} \\ &= h_n^2\,\frac{[f_X(x)\mu_{2,K} + o_P(1)]^2 - o_P(1)o_P(1)}{(f_X(x) + o_P(1))\,[f_X(x)\mu_{2,K} + o_P(1)] - (o_P(1))^2} \\ &= h_n^2\,\mu_{2,K} + o_P(h_n^2). \end{aligned} \tag{93}$$

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 (74)–(76) one gets

$$D_n = f_X^2(x)\,\mu_{2,K} + o_P(1). \tag{94}$$

Now using (94) and Lemma 1 we can bound

$$
\begin{aligned}
|C_n| &\le \sup_{z:|z-x|\le h_n} |\tilde{R}(z)|\, h_n^2 \sum_{i=1}^n |w_{n,i}(x)|\, \tfrac{(X_i-x)^2}{h_n^2} \\
&\le h_n^2\, o(1)\, \frac{S_{n,2}^2(x) + |S_{n,1}(x)|\, \sum_{i=1}^n \frac{1}{nh_n} K\big(\frac{X_i-x}{h_n}\big) \big|\frac{X_i-x}{h_n}\big|^3}{|D_n|} \\
&= o(h_n^2)\, \frac{f_X^2(x)\mu_{2,K}^2 + o_P(1) + o_P(1)\left[f_X(x)\int_{\mathbb{R}} K(t)\,|t|^3\,\mathrm{d}t + o_P(1)\right]}{f_X^2(x)\,\mu_{2,K} + o_P(1)} = o_P(h_n^2),
\end{aligned}
$$

which together with (82), (92) and (93) yields (90).

Showing (91). With the help of (75), (76), (86) and (94) one can calculate

$$
\begin{aligned}
\mathsf{var}[\widehat{m}_{LL}(x)|\mathbb{X}] &= \sum_{i=1}^n w_{n,i}^2(x)\sigma^2(X_i) \\
&= \frac{1}{D_n^2}\left[\frac{1}{n^2h_n^2}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(S_{n,2}(x) - \tfrac{X_i-x}{h_n}S_{n,1}(x)\Big)^2\sigma^2(X_i)\right] \\
&= \frac{1}{nh_n}\frac{1}{D_n^2}\left[S_{n,2}^2(x)\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\sigma^2(X_i) + o_P(1)\right] \qquad (95) \\
&= \frac{1}{nh_n}\frac{1}{f_X^4(x)\,\mu_{2,K}^2 + o_P(1)}\big[f_X^2(x)\,\mu_{2,K}^2 + o_P(1)\big]\big[f_X(x)\,\sigma^2(x)\,R(K) + o_P(1)\big],
\end{aligned}
$$

which implies (91). In the equality (95) above we used the fact that

$$
\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(\tfrac{X_i-x}{h_n}\Big)\sigma^2(X_i) = O_P(1)
$$

and

$$
\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(\tfrac{X_i-x}{h_n}\Big)^2\sigma^2(X_i) = O_P(1).
$$

Both these formulas follow in the same way as

$$
V_n = \frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\sigma^2(X_i) = O_P(1),
$$

that was shown in (86) in the proof of Theorem 8 (that is, using Bochner's Theorem 5). Now, by (75) we have $S_{n,1}(x) = o_P(1)$ and by (76) we have $S_{n,2}(x) = O_P(1)$. Thus, we can write

$$
\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(\tfrac{X_i-x}{h_n}\Big)S_{n,1}(x)S_{n,2}(x)\sigma^2(X_i) = o_P(1)O_P(1) = o_P(1),
$$

and

$$
\frac{1}{nh_n}\sum_{i=1}^n K^2\big(\tfrac{X_i-x}{h_n}\big)\Big(\tfrac{X_i-x}{h_n}\Big)^2 S_{n,1}^2(x)\sigma^2(X_i) = o_P(1)O_P(1) = o_P(1),
$$

meaning that our simplification in (95) was correct. This concludes the proof. □

3.4 Locally polynomial regression (general p)

Analogously as for $p \in \{0, 1\}$ one gets the estimator of $m(x)$ in the form

$$\widehat{m}_p(x) = \sum_{i=1}^{n} w_{n,i}(x)\, Y_i,$$

where the weights $w_{n,i}(x)$ are given by the first row of the matrix

$$\Big(\mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)\, \mathbb{X}_p(x)\Big)^{-1} \mathbb{X}_p^{\mathsf{T}}(x)\, \mathbb{W}(x)$$

and satisfy that by (73) we have

$$\sum_{i=1}^{n} w_{n,i}(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} w_{n,i}(x)(X_i - x)^{\ell} = 0, \quad \ell = 1, \ldots, p.$$

With the help of this property one can show (analogously as in Theorems 8 and 9) 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

$$\mathsf{var}\left(\widehat{m}_p(x)|\mathbb{X}\right) = \frac{V_p\, \sigma^2(x)}{f_X(x)\, n\, h_n} + o_P\big(\tfrac{1}{n\, h_n}\big),$$

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 the 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, Sections 3.1 and 3.2.1).

3.5 Bandwidth selection

3.5.1 Asymptotically optimal bandwidths

In what follows, we will consider $p = 1$. With the help of Theorem 9, one can approximate the conditional MSE (mean squared error) of $\widehat{m}_{LL}(x)$ as

$$\mathrm{MSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\, h_n} \tfrac{\sigma^2(x)\, R(K)}{f_X(x)} + \tfrac{1}{4}\, h_n^4\, [m''(x)]^2 \mu_{2,K}^2 + o_P\big(\tfrac{1}{n\, h_n}\big) + o_P\big(h_n^4\big). \tag{96}$$

Ignoring the remainder $o_P(.)$ terms in (96), we get that AMSE (asymptotic mean squared error) of $\widehat{m}_{LL}(x)$ is given by

$$\mathrm{AMSE}\big(\widehat{m}_{LL}(x)\,|\,\mathbb{X}\big) = \tfrac{1}{n\, h_n} \tfrac{\sigma^2(x)\, R(K)}{f_X(x)} + \tfrac{1}{4}\, h_n^4\, [m''(x)]^2 \mu_{2,K}^2. \tag{97}$$

Minimising (97) one gets the asymptotically optimal *local bandwidth* (i.e. bandwidth that minimises the AMSE)

$$h_n^{(opt)}(x) = n^{-1/5} \left[\frac{\sigma^2(x)\, R(K)}{f_X(x)\, [m''(x)]^2\, \mu_{2,K}^2} \right]^{1/5}.$$

The mean integrated squared error (MISE) is usually defined as

$$\mathrm{MISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) = \int_{\mathbb{R}} \mathrm{MSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \tag{98}$$

where $w_0(x)$ is a given weight function which is introduced in order to guarantee that the integral is — hopefully — finite (for instance $w_0(x) = \mathbb{I}\{x \in [a,b]\}$ can be used).

Now with the help of (97) and (98), the asymptotic mean integrated squared error (AMISE) is defined as

$$\begin{aligned} \mathrm{AMISE}\big(\widehat{m}_{LL} \,|\, \mathbb{X}\big) &= \int_{\mathbb{R}} \mathrm{AMSE}\big(\widehat{m}_{LL}(x) \,|\, \mathbb{X}\big)\, w_0(x)\, f_X(x)\, \mathrm{d}x \\ &= \frac{R(K)}{n\, h_n} \int_{\mathbb{R}} \sigma^2(x)\, w_0(x)\, \mathrm{d}x + \tfrac{1}{4}\, h_n^4\, \mu_{2,K}^2 \int_{\mathbb{R}} [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x. \end{aligned} \tag{99}$$

Minimising (99) one gets the asymptotically optimal *global bandwidth* (i.e., the bandwidth that minimises the AMISE)

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K) \int_{\mathbb{R}} \sigma^2(x)\, w_0(x)\, \mathrm{d}x}{\mu_{2,K}^2 \int_{\mathbb{R}} [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}. \tag{100}$$

3.5.2 Rule of thumb for bandwidth selection

Suppose that $\sigma^2(x) = \sigma^2 > 0$ is constant. Then, the asymptotically optimal global bandwidth (100) is for $\widehat{m}_{LL}$ given by

$$h_n^{(opt)} = n^{-1/5} \left[\frac{R(K)\, \sigma^2 \int_{\mathbb{R}} w_0(x)\, \mathrm{d}x}{\mu_{2,K}^2 \int_{\mathbb{R}} [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x} \right]^{1/5}.$$

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. In (100), 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 $\int_{\mathbb{R}} [m''(x)]^2\, w_0(x)\, f_X(x)\, \mathrm{d}x = \mathsf{E}_X [m''(X)]^2\, w_0(X)$ 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

$$h_n^{(ROT)} = n^{-1/5} \left[\frac{R(K)\, \widetilde{\sigma}^2 \int_{\mathbb{R}} w_0(x)\, \mathrm{d}x}{\mu_{2,K}^2\, \frac{1}{n} \sum_{i=1}^n [\widetilde{m}''(X_i)]^2\, w_0(X_i)} \right]^{1/5}.$$

In `R`, it is implemented in function `thumbBw` in package `locpol`.

The end of lecture 11 (11.12.2024)

3.5.3 Cross-validation

Similarly as for the unbiased cross-validation for the kernel density estimator, we set

$$h_n^{(CV)} = \arg\min_{h_n>0} \mathcal{CV}(h_n),$$

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.

$$\begin{aligned} \mathrm{ISE}\big(\widehat{m}_p(x)\big) &= \int_{\mathbb{R}} \big(\widehat{m}_p(x) - m(x)\big)^2 f_X(x)\, w_0(x)\, \mathrm{d}x \\ &= \mathsf{E}_{X'} \big(\widehat{m}_p(X') - m(X')\big)^2 w_0(X'), \end{aligned} \tag{101}$$

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 (101). 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 (101), 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 18. 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 19. Another view of the cross-validation procedure is that we aim at finding the bandwidth h_n that minimises 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 from 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 $\widehat{m}_p^{(-i)}$.

3.5.4 Nearest-neighbour bandwidth choice

Suppose that the support of the kernel function K is the interval $(-1, 1)$. Then $w_{n,i}(x) = 0$ 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 we have $|X_i - x| \leq h_n$. 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

$$h_n^{(NN)}(x) = d_{(k)}(x). \tag{102}$$

Note that (102) presents a local bandwidth choice.

To get an insight into the bandwidth choice (102), let us approximate

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{|X_i - x| \leq h\} \doteq \widehat{F}_n(x+h) - \widehat{F}_n(x-h) \doteq F_X(x+h) - F_X(x-h) \doteq f_X(x) 2h. \tag{103}$$

By plugging $h = d_{(k)}(x) = h_n(x)$ into (103), 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 20. To derive the asymptotic properties of $\widehat{m}_{LL}$ when the bandwidth h_n is chosen as (102), one needs to consider $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$.

Remark 21. Using $h_n^{(NN)}(x)$ usually makes the problem more computationally expensive, as one is using a local bandwidth. Further, there is no guarantee that the estimator $\widehat{m}_p(x)$ is for instance continuously differentiable even if K is continuously differentiable. To prevent those difficulties, some authors recommend transforming the covariates to

$$X_i' = \widehat{F}_n(X_i), \quad i = 1, \ldots, n,$$

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., using the cross-validation procedure described in Section 3.5.3). 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

$$\mathsf{E}\left[Y \,|F_X(X) = x\right] = \mathsf{E}\left[Y \,|\, X = F_X^{-1}(x)\right] = m\big(F_X^{-1}(x)\big).$$

3.6 Conditional variance estimation

The most straightforward estimate of the conditional variance $\sigma^2(x) = \mathsf{var}\left[Y_1 \,|\, X_1 = x\right]$ is given by

$$\widehat{\sigma}_n^2(x) = \sum_{i=1}^n w_{n,i}(x)\, Y_i^2 - \widehat{m}_p^2(x), \tag{104}$$

where $\widehat{m}_p(x) = \sum_{i=1}^n w_{n,i}(x)\, Y_i$ is an estimator of $m(x) = \mathsf{E}\left[Y_1 \,|\, X_1 = x\right]$. This estimator is based on the expression

$$\sigma^2(x) = \mathsf{E}\left[Y_1^2 \,|\, X_1 = x\right] - \big(\mathsf{E}\left[Y_1 \,|\, X_1 = x\right]\big)^2.$$

The estimator (104) is usually preferred in theoretical papers as its properties can be derived completely analogously as for $\widehat{m}_p(x)$.

In practice, it is usually recommended to start from

$$\sigma^2(x) = \mathsf{E}\left[\,\big(Y_1 - \mathsf{E}\left[Y_1 \,|\, X_1 = x\right]\big)^2 \,|\, X_1 = x\right],$$

and use the following estimator

$$\widetilde{\sigma}_n^2(x) = \sum_{i=1}^n w_{n,i}(x)\big(Y_i - \widehat{m}_p(X_i)\big)^2. \tag{105}$$

If the weights $w_{n,i}(x)$ are not guaranteed to be non-negative, then there is generally no guarantee that either of the estimators (104) or (105) is positive.

3.7 Robust locally weighted regression (LOWESS)

LOWESS is an algorithm for 'LOcally WEighted Scatterplot Smoothing'. It is used among others in regression diagnostics; in R, it is implemented in function lowess in package stats. The algorithm runs as follows.

The end of lecture 12 (18.12.2024)

Literature: Fan and Gijbels (1996, Sections 2.4.1, 3.2.3, 4.2, 4.10.1, 4.10.2).

* In case there are no ties in covariate values, one gets $\{X_1', \ldots, X_n'\} = \{\frac{1}{n}, \ldots, \frac{n}{n}\}$.

LOWESS: Locally weighted scatterplot smoothing.

Input: A dataset $\binom{X_1}{Y_1}, \binom{X_2}{Y_2}, \ldots, \binom{X_n}{Y_n}$.

Output: A robust local linear regression estimator $\widehat{m}$.

0. Set
   - $K(t) = \frac{70}{81}\big(1 - |t|^3\big)^3 \mathbb{I}\{|t| \leq 1\}$ the tricube kernel,
   - h_n the k-nearest neighbour bandwidth with $k = \lfloor n\, f \rfloor$ and $f = 2/3$, and
   - $\delta_i = 1$ for each $i = 1, \ldots, n$.

1. Fit $\widehat{m}(x)$ as a weighted local linear estimator with a kernel K and a bandwidth h_n. That is, $\widehat{m}(x) = \widehat{\beta}_0(x)$, where

$$\big(\widehat{\beta}_0(x), \widehat{\beta}_1(x)\big) = \underset{b_0, b_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \Big[Y_i - b_0 - b_1\,(X_i - x)\Big]^2 K\big(\tfrac{X_i - x}{h_n}\big)\, \delta_i.$$

   In the fist loop with $\delta_i = 1$ for all $i = 1, \ldots, n$, we obtain the usual local linear estimator $\widehat{m}(x) = \widehat{m}_{LL}(x)$.

2. Consider the residuals $r_i = Y_i - \widehat{m}(X_i)$ of the current fit, $i = 1, \ldots, n$.

3. For $B(t) = (1 - t^2)^2\, \mathbb{I}\{|t| \leq 1\}$, calculate the 'measures of outlyingness'

$$\delta_i = B\Big(\frac{r_i}{6\,\mathsf{med}\,\big(|r_1|, \ldots, |r_n|\big)}\Big), \qquad i = 1, \ldots, n,$$

   that assess how much 'extreme' is the residual r_i compared to the other residuals.

4. Repeat steps 1–3 three times.

Appendix

Stochastic o_P and O_p symbols

This section is identical to parts of Omelka (2023, Section 1.1).

Definition A9. Let $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ be a sequence of random vectors in $\mathbb{R}^k$ 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 $\mathbf{0}_k = (0,\ldots,0)^\mathsf{T}$ is a zero point in $\mathbb{R}^k$;

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

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

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

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

Remark 22. Note that

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

(ii) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{P} \mathbf{0}_k$ 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). 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

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

The statement now follows from the fact that

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

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

* *omezená v pravděpodobnosti*

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

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

Remark 23. 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 $\{\boldsymbol{X}_n\}_{n=1}^{\infty}, \{\boldsymbol{Y}_n\}_{n=1}^{\infty}$ be such that $\boldsymbol{X}_n = O_P(1), \boldsymbol{Y}_n = o_P(1)$ and $\boldsymbol{Y}_n \boldsymbol{X}_n$ makes sense. Let $\varepsilon > 0$ be given and consider for instance the Euclidean norm (for other norms the proof would go through up to a multiplicative constant in some of the arguments). Then one can find $K < \infty$ 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 Jiang (2010, Chapter 3.4).

Uniform consistency of the empirical distribution function

The following theorem can be found in Serfling (1980, Section 2.1.4) as Theorem A.

Theorem A10. *(**Glivenko-Cantelli theorem**) Suppose we observe independent and identically distributed random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ (in $\mathbb{R}^k$) from a distribution with the cumulative distribution function F. Let*

$$\widehat{F}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i \leq \mathbf{x}\} \quad \textit{for } \mathbf{x} \in \mathbb{R}^k$$

be the cumulative empirical distribution function. Then

$$\sup_{\mathbf{x} \in \mathbb{R}^k} \left| \widehat{F}_n(\mathbf{x}) - F(\mathbf{x}) \right| \xrightarrow[n \to \infty]{a.s.} 0.$$

Supremum metric and convergence in distribution

Lemma A2. *Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and $\boldsymbol{Y}$ are random vectors (with values in $\mathbb{R}^k$) 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 $\sup_{\mathbf{y} \in \mathbb{R}^k} |G_n(\mathbf{y}) - G(\mathbf{y})| \to 0$ implies that $G_n(\mathbf{y}) \to G(\mathbf{y})$ for each $\mathbf{y} \in \mathbb{R}^k$.

The implication $\Leftarrow$ is slightly 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 $\mathbf{y} \in \mathbb{R}^k$ one can find $\mathbf{y}_L, \mathbf{y}_U \in B_\varepsilon$ that

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

By an inequality of p-dimensional vectors above we mean that the inequality is true for all their components. For each $\mathbf{y} \in \mathbb{R}^k$ one can bound

$$G_n(\mathbf{y}) - G(\mathbf{y}) \leq G_n(\mathbf{y}_U) - G(\mathbf{y}) \leq G_n(\mathbf{y}_U) - G(\mathbf{y}_L) \leq G_n(\mathbf{y}_U) - G(\mathbf{y}_U) + \tfrac{\varepsilon}{2} \tag{A106}$$

and analogously also

$$G_n(\mathbf{y}) - G(\mathbf{y}) \geq G_n(\mathbf{y}_L) - G(\mathbf{y}) \geq G_n(\mathbf{y}_L) - G(\mathbf{y}_U) \geq G_n(\mathbf{y}_L) - G(\mathbf{y}_L) - \tfrac{\varepsilon}{2}. \tag{A107}$$

Now combining (A106) and (A107) together with $G_n \xrightarrow[n \to \infty]{w} G$ one gets that for all sufficiently large n

$$\sup_{\mathbf{y} \in \mathbb{R}^k} \left| G_n(\mathbf{y}) - G(\mathbf{y}) \right| \leq \max_{\mathbf{y} \in B_\varepsilon} \left| G_n(\mathbf{y}) - G(\mathbf{y}) \right| + \tfrac{\varepsilon}{2} \leq \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon,$$

which implies the statement of the lemma. □

Linderberg-Feller central limit theorem

The following result is a variant of the classical Linderberg-Feller central limit theorem for triangular sequences of independent random vectors. It can be found in van der Vaart (2000, Proposition 2.27).

Theorem A11. *For each $n = 1, 2, \ldots$, let $\boldsymbol{Y}_{n,1}, \ldots, \boldsymbol{Y}_{n,k_n}$ be independent random vectors (in $\mathbb{R}^k$) with finite variances such that*

$$\sum_{i=1}^{k_n} \mathsf{E}\left[\|\boldsymbol{Y}_{n,i}\|^2 \mathbb{I}\{\|\boldsymbol{Y}_{n,i}\| > \varepsilon\}\right] \xrightarrow[n\to\infty]{} 0 \quad \text{for every } \varepsilon > 0, \tag{A108}$$

and

$$\sum_{i=1}^{k_n} \mathsf{var}\left(\boldsymbol{Y}_{n,i}\right) \xrightarrow[n\to\infty]{} \mathbb{\Sigma},$$

for a positive definite matrix $\mathbb{\Sigma} \in \mathbb{R}^{k\times k}$. Then

$$\sum_{i=1}^{k_n} \left(\boldsymbol{Y}_{n,i} - \mathsf{E}\, \boldsymbol{Y}_{n,i}\right) \xrightarrow[n\to\infty]{d} \mathsf{N}_k\left(\mathbf{0}, \mathbb{\Sigma}\right).$$

Equivalence of convergence in distribution and convergence of quantiles

The following result can be found as van der Vaart (2000, Lemma 21.2).

Lemma A3. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and F_{X_n} be the cumulative distribution function of X_n. Then $X_n \xrightarrow[n\to\infty]{d} X$ if and only if $F_{X_n}^{-1}(u) \xrightarrow[n\to\infty]{} F_X^{-1}(u)$ for each $u \in (0, 1)$.*

References

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.

Jiang, J. (2010). *Large sample techniques for statistics*. Springer Texts in Statistics. Springer, New York.

Komárek, A. (2021). *NMST407: Linear Regression*. https://www.karlin.mff.cuni.cz/~kulich/vyuka/linreg/doc/2021-NMSA407-notes.pdf.

Kulich, M. and Omelka, M. (2022). *NMSA331: Matematická statistika 1.* https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf.

Müller, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.*, 12(2):766–774.

Nagy, S. (2023a). *NMSA332: Mathematical statistics 2.* https://www.karlin.mff.cuni.cz/~nagy/NMSA332/NMSA332.pdf.

Nagy, S. (2023b). *NMSA444: Robust statistical methods.* https://www.karlin.mff.cuni.cz/~nagy/NMST444/NMST444.pdf.

Omelka, M. (2023). *NMST424: Mathematical Statistics 3.* https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmst424/nmst424_course-notes.pdf.

Prášková, Z. (2004). Metoda bootstrap. In Antoch, J. and Dohnal, G., editors, *Sborník prací 13. letní školy JČMF ROBUST 2004*, pages 299–314. JČMF, Praha. ISBN 80-7015-972-3.

Scott, D. W. (2015). *Multivariate density estimation.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition. Theory, practice, and visualization.

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.

Wied, D. and Weißbach, R. (2012). Consistency of the kernel density estimator: a survey. *Statist. Papers*, 53(1):1–21.