Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS

Flexibility, Robustness and Discontinuity in Nonparametric Regression Approaches

Mgr. Matúš Maciak, M.Sc.

Department of Probability and Mathematical Statistics

Supervised by
Prof. RNDr. Marie Hušková, CSc.

Study program
Mathematics

Study branch
Probability and Mathematical Statistics

Est modus in rebus...

(Horatius)

"All I am, or can be, I owe to my angel parents..."

Abraham Lincoln (1809 – 1865)

ACKNOWLEDGEMENT

I would like to express my thanks to my supervisor Prof. RNDr. Marie Hušková, DrSc. for her valuable help, many pieces of advise and very useful suggestions during my whole work on this thesis.
I appreciate it very much.

Many thanks are also addressed to the professors and teachers from the Department of Probability and Mathematical Statistics at Charles University in Prague where I have spent nine good years while obtaining an interesting training in mathematics, probability and statistics. Thank you for introducing me into the world of real mathematics. I am grateful for any inspiration, pieces of advice and new ideas I could learn from you and I am very proud to count some of you among my friends.

"It is the supreme art of the teacher to awaken joy in creative expression and knowledge."
Albert Einstein (1879 – 1955)

I also want to express my thanks to my parents and my brothers and sisters. You are those who were giving me enough courage and confidence during the pursuit of my (any) degree and the whole composition of this thesis as well. You were still willing to provide my with encouragement, which helps more than many people can even think...

"You don't choose your family. They are God's gift to you, as you are to them."
Desmond Mpilo Tutu (1931 – . . .)

Last but not least, I want to express my thanks to all my friends, especially those who have spent most of time with me here in Prague. It was always nice to enjoy time and fun with you. One can never stay focussed on one thing only. He needs a variety of interests in order to survive and to be happy.
You have provided my with this and I hope you always will...

"Without friends no one would choose to live, though he had all other goods."
Artistotle (384 BC – 322 BC)

Matúš (ka,ik)

STATEMENT OF HONESTY

I hereby declare that I have written this doctoral thesis separately, independently and entirely with using the quoted resources. I agree that the University Library shall make it available to borrowers under the rules of the Library.

Prague, March 29, 2011

signature

ANNOTATIONS

❒ Title

Flexibility, Robustness and Discontinuity
in Nonparametric Regression Approaches

✍ Author

Mgr. Matúš Maciak, M.Sc.
✉ maciak@karlin.mff.cuni.cz

✎ Department

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

☞ Supervisor

Prof. RNDr. Marie Hušková, CSc.
✉ huskova@karlin.mff.cuni.cz

✉ Mailing address

KPMS MFF UK
Sokolovská 83
186 75 Prague 8
Czech Republic

Thesis title: **Flexibility, Robustness and Discontinuity in Nonparametric Regression Approaches**

Author: Mgr. Matúš Maciak, M.Sc.

Department: Department of Probability and Mathematical Statistics,
Charles University in Prague

Supervisor: Prof. RNDr. Marie Hušková, DrSc.
huskova@karlin.mff.cuni.cz

Abstract:

In this thesis we focus on local polynomial estimation approaches of an unknown regression function while taking into account also some robust issues like a presence of outlying observations or heavy-tailed distributions of random errors as well. We will discuss the most common method used for such settings, so called local polynomial M-smoothers and we will present the main statistical properties and asymptotic inference for this method. The M-smoothers method is especially suitable for such cases because of its natural robust flavour, which can nicely deal with outliers as well as heavy-tailed distributed random errors.

Another important quality we will focus in this thesis on is a discontinuity issue where we allow for sudden changes (discontinuity points) in the unknown regression function or its derivatives respectively.

We will propose a discontinuity model with different variability structures for both independent and dependent random errors while the discontinuity points will be treated in a proper statistical way using one-sided M-smoothers estimates. We will propose a statistical test to decide if an estimated jump and its location are significant for the model they are not. Given the asymptotic distribution for the test statistic under the null hypothesis, which depends on some unknown quantities we will propose some bootstrap algorithms, which can be used to mimic the unknown distribution of interest. The appropriate bootstrap algorithms will be proposed for every considered model scenario and all necessary proofs will be provided.

Finally, the proposed methods and the stated results are tested through out an extensive simulation study presented at the end. Similarly, we also apply the proposed testing and estimating methods to a real data case and the finite sample performance will be compared and discussed at the very end of this thesis.

Keywords:

Local polynomial M-smoothers, flexibility in modelling, robustness, discontinuity in nonparametric regression, Change-points, residual based bootstrap, block-bootstrap, α-mixing dependence.

"Every accomplishment starts with the decision to try."

Anonymous author

1 Introduction to Nonparametric Regression and Robustness

1.1 Preface

An idea of a regression analysis was used by Francis Galton[1] for the first time in the 19^{th} century. He studied a dependence of average heights between parents and their sons using a classical linear regression approach where he also found a significant relationship, which was described as a half-way inheriting reduction in excessive heights from parents to their sons.

Since that time regression modelling techniques attract a lot of attention not only on the field of theoretical or applied statistics but also in almost all areas of real life. There were many new proposals and adaptations introduced later on in order to make them more suitable for almost any possible scenario. In this thesis we will focus on a class of nonparametric regression techniques while we also want to stay less restricted with respect to a considered family of distribution functions of random errors. Specifically, we will point our attention to a class of nonparametric robust regression approaches, which allow for a presence of outlying observations in a model and what is even of more importance they also allow us to consider heavy-tailed random error distributions, which was definitely not the case in classical regression techniques based on a regular least squares method introduced by Carl Friedrich Gauss in 1795.

Local polynomial M-smoothers are described in detail in this thesis and they are further discussed as a statistical method, which can provide us with both a reasonably good level of flexibility in case of modelling on one hand and a set of less restricted distributional assumptions in case of an inference on the other hand. From the technical point of view the local polynomial M-smoothers can be thought of as a combination of two at the first glance slightly unrelated statistical estimation methods: nonparametric regression techniques and M-estimation procedures. More precisely, by nonparametric regression we will refer in this thesis to a generalization of classical polynomial regression techniques into a local polynomial regression, which brings in a huge additional flexibility in modelling by adopting a local approach. This is a common fact for a nonparametric regression in general.

On the other hand the robust M-estimation methods introduce a parameter estimation under some less restricted distributional assumptions, which allow us to incorporate outlying observations into consideration and to apply the proposed models and techniques for heavy-tailed distributions as well.

[1] Francis Galton (1822–1911) was an English statistician, explorer and anthropologist who was knighted in 1909. He was the first who used a notion "regression" in the same sense as we use it today. He is therefore considered to be a pioneer of regression and correlation analysis.

In the next two sections we will briefly mention both, the nonparametric regression techniques and M-estimation procedures and we will shortly describe their main statistical properties and important issues in order to offer to a reader some small statistical background, which will be necessary for a better understanding of the whole M-smoothers concept that is discussed in later chapters.

1.2 Local polynomial regression

Let us start with a classical nonparametric regression, which is used later on to build a corresponding M-smoothers regression method. We will discuss only the main outlines of the nonparametric regression and we will state only the results, which are important for our further investigation of the M-smoothers inference. For all further details we refer to a nice summary book of Fan and Gijbels (1996).

Nonparametric regression methods can be seen as an extension of classical regression techniques where one a priori assumes more flexibility in hands as there is no parametric shape of the unknown regression function considered in advance. The idea is to apply classical regression techniques however, in small regions only – locally. This local approach is introduced via a combination of classical regression methods and a kernel function together with a bandwidth parameter where the kernel function assigns weights in a local region which is targeted by the estimation procedure and the bandwidth parameter specifies this local region, its boundaries respectively. By adopting a nonparametric regression approach one uses data driven procedures to come up with the final shape of the model estimate rather than using any parametric assumptions, which are rather too much strict however, standard for any parametric regression modelling.

Firstly, let us consider a random sample $\{(X_i, Y_i);\ i = 1, \ldots, N \in \mathbb{N}\}$ given from some unknown joint distribution function $F_{(X,Y)}(x, y)$, where we assume the decomposition

$$Y_i = m(X_i) + \varepsilon_i \sigma(x), \quad \text{for } \ i = 1, \ldots, N, \tag{1.1}$$

to hold. Under this notation function $m(\cdot)$ is called the *regression function* and function $\sigma(\cdot)$ is called the *variance function*. Quantities $\{\varepsilon_i\}_{i=1}^N$ stand here for random error terms, which are assumed to be independent and identically distributed ($i.i.d.$) with a zero mean and a unit variance. Given such model we are interested in estimation of the unknown regression function $m(\cdot)$ (or $\sigma(\cdot)$ optionally), which is unknown moreover, no other properties except some smoothness are further assumed.

This is the key difference between the nonparametric and parametric methods (semi-parametric methods respectively), where one has to assume a specific parametric shape of the unknown regression function in order to make sure that asymptotic properties are really satisfied. On the other hand, given no parametric flavour in nonparametric modelling one gets models, which are more difficult to interpret and they also become to much complex to by applied for some further prediction purposes. Even thought, nonparametric methods are widely used in statistical modelling approaches and we will consider them as the main building block for constructing the robust modelling approaches.

The most common estimate which assumes no parametric shape of the estimated regression function was proposed by Nadaraya (1964) and Watson (1964) and it is defined by the minimization problem

$$\widehat{m}(x) = \underset{b \in \mathbb{R}}{Argmin} \sum_{i=1}^{N} (Y_i - b)^2 \cdot K\left(\frac{X_i - x}{h_N}\right), \tag{1.2}$$

which can be equivalently rewritten as an explicit solution for $b \in \mathbb{R}$ of the equation

$$\widehat{m}(x) = \frac{\sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) Y_i}{\sum_{j=1}^{N} K\left(\frac{X_j - x}{h_N}\right)}, \tag{1.3}$$

where function $K(\cdot)$ is a kernel function, which is usually a symmetric probability density function on some compact support and $h_N > 0$ is the bandwidth parameter, which controls the amount of smoothness in the final fit and it satisfies that $h_N \to 0$ as $N \to \infty$, such that $Nh_N \to \infty$. Let us mention that the original Nadaraya-Watson estimate was introduced for $X_1, \dots X_N$ to be fixed and known constants while $Y_1, \dots, Y_N$ were random quantities however, a generalization for $X_1, \dots X_N$ to be random variables as well was proposed afterwards. We will further consider such generalization only. The Nadaraya-Watson estimators have been extensively studied by many authors, for example Rosenblatt (1969), Mack and Silverman (1982) or Härdle (1990).

Another well-known proposal for an estimator of an unknown regression function $m(\cdot)$, which does not assume a parametric flavour was proposed by Gasser and Müller (1979). This estimator can be thought of as an alternative to the Nadaraya-Watson estimator as it has a lower bias term however, it suffers from larger variance instead (see Fan and Gijbels (1996)). On the other hand, a serious drawback of both proposed estimators (Nadaraya-Watson and Gasser-Müller) is given by so called boundary issues. This reflects the fact that both estimators tend to produce inconsistent estimates at boundary regions of the domain of interest. There were many proposals discussed and some modifications (like adaptive kernel methods) were introduced to avoid these problems however, we will rather go for another approach based on a local polynomial estimation, which as was shown in Fan and Gijbels (1996) can handle boundary issues in a far more convenient and straightforward way.

In order to inherit nice properties from both estimators (a smaller bias term from the Gasser-Müller estimator and a smaller variance term from the Nadaraya-Watson estimator) a generalization of the Nadaraya-Watson estimator has been proposed where one incorporates a local linear modelling instead of the local constant, which was used in both estimators before. Indeed, as one can see in Table 1.1 below the local linear approach improves the bias imprecision involved in the Nadaraya-Watson estimate and the variability inefficiency given within the Gasser-Müller estimate too.

Estimation method	Bias term approximation	Variance term approximation
■ Nadaraya-Watson	$\frac{1}{2}\left(\int_{\mathbb{R}} u^2 K(u)\mathrm{d}u\right) \cdot \left[m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right]$	$\frac{\sigma(x)}{f(x)Nh_N} \int_{\mathbb{R}} K^2(u)\mathrm{d}u$
■ Gasser-Müller	$\frac{1}{2}\left(\int_{\mathbb{R}} u^2 K(u)\mathrm{d}u\right) \cdot m''(x)$	$\frac{3\sigma(x)}{2f(x)Nh_N} \int_{\mathbb{R}} K^2(u)\mathrm{d}u$
■ Local linear	$\frac{1}{2}\left(\int_{\mathbb{R}} u^2 K(u)\mathrm{d}u\right) \cdot m''(x)$	$\frac{\sigma(x)}{f(x)Nh_N} \int_{\mathbb{R}} K^2(u)\mathrm{d}u$

Table 1.1: Pointwise asymptotic bias and variance approximations of kernel regression smoothers taken from Fan (1992) given for some point $x \in (h, 1-h)$ where the interval $(0,1)$ is assumed to be the domain of interest for $m(\cdot)$.

The most important advantage of the local linear approach however, is the fact that no boundary issues occur for this method any more therefore, no modifications or adapted kernels are needed for this regression setting. Indeed, it was shown in Fan and Gijbels (1996) that the local linear estimation can avoid boundary problems, which are common for local constant fits and it can produce fully consistent estimates also in regions arbitrarily close to the boundary points[2].

[2] The consistency result of the local polynomial fit in boundary regions is given asymptotically only so the real data result depends on finite sample properties of the given estimator.

Just a straightforward extension of the local polynomial fit into higher orders of approximation gives us a local polynomial modelling technique. The local polynomial estimator of an unknown regression function $m(\cdot)$ at the given point from the domain of interest is defined by the minimization problem

$$\widehat{\boldsymbol{\beta}}_x = \underset{(b_0,\ldots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \quad \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j \right)^2 \cdot K\left(\frac{X_i - x}{h_N} \right), \tag{1.4}$$

where $\widehat{\boldsymbol{\beta}}_x = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top \in \mathbb{R}^{p+1}$ is a vector of parameter estimates[3] at some given $x \in \mathbb{R}$ from the domain of interest and for the estimate $\widehat{m}(x)$ of the unknown regression function $m(\cdot)$ at the point $x \in (0,1)$ it holds that $\widehat{m}(x) = \widehat{\beta}_0$. Given a classical L_2 norm used for the minimization one can easily obtain the parameter estimates from (1.4) in explicit forms in a similar way as in the case of Nadaraya-Watson estimator in (1.3). Moreover, for the local polynomial regression one obtains not only an estimate for the regression function itself but also for its consecutive derivatives and it holds that $\widehat{\beta}_j = \widehat{m}(x)/j!$.
This approach however requires a smoothness assumptions for the unknown regression function $m(\cdot)$ up to the order $p+1$ at least. It is also possible to use an adaptation of (1.4) and by adopting an appropriate kernel function one can directly get an estimate for the corresponding derivative only (see e.g. an approach discussed in Zelinka and Horová (2001)).

We will now formulate two crucial theorems, which will play an important role for deriving a proper statistical inference for the M-smoothers estimates. These theorems state the asymptotic conditional bias and variance terms and the corresponding limit distribution of the local polynomial estimate, both under some common model assumptions.
Let us however, firstly introduce a notation, which will be used in the next theorems but also in the following chapters as well. For some $K(\cdot)$ to be a probability density function defined on interval $[-1,1]$ such that $K(\cdot)$ is positive on this support we define quantities

$$\begin{aligned} \mathsf{S}_1 &= \left(\int_{-1}^{1} u^{j+l} K(u) \mathrm{d}u \right)_{\substack{\mathrm{j}=0,\ldots,\mathrm{p};\\ \mathrm{l}=0,\ldots,\mathrm{p};}}, & \mathsf{S}_2 &= \left(\int_{-1}^{1} u^{j+l} K^2(u) \mathrm{d}u \right)_{\substack{\mathrm{j}=0,\ldots,\mathrm{p};\\ \mathrm{l}=0,\ldots,\mathrm{p};}}, \\ \mathbf{e}_\nu &= (\underbrace{0,\ldots,0}_{\nu-times}, 1, \underbrace{0,\ldots,0}_{(p-\nu)-times})^\top, & \boldsymbol{\mu}_p &= \left(\int_{-1}^{1} u^{p+1} K(u) \mathrm{d}u, \ldots, \int_{-1}^{1} u^{2p+2} K(u) \mathrm{d}u \right)^\top, \end{aligned} \tag{1.5}$$

for some $p \in \mathbb{N}_0$ to be an order of the local polynomial approximation and $\nu \in \{0,\ldots,p\}$ where it holds that $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let the following assumptions[4] are satisfied:

- C1 Random errors $\varepsilon_1, \ldots, \varepsilon_N$ are assumed to be independent and identically distributed, such that $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{V}\mathrm{ar}\ \varepsilon_i = 1$, for $i = 1, \ldots, N$;
- C2 The unknown regression function $m(\cdot)$ is assumed to be $(p+1)$-times Lipschitz over the domain of interest, which is generally assumed to be the interval $(0,1)$;
- C3 The variance function $\sigma(\cdot)$ is Lipschitz and positive on the domain of interest;
- C4 The density function $f(\cdot)$ of random variables X_i, for $i = 1, \ldots, N$ is assumed to be positive and Lipschitz over the domain of interest;

[3] Using notation $\boldsymbol{\beta}_x$ for the vector of true parameters $(\beta_0, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ we just want to emphasize the relationship between the vector of parameter estimates and the given point of interest $x \in (0,1)$, where the vector of parameter estimates is defined at. For further simplicity we will omit the "x" notation for single elements of the vector $\boldsymbol{\beta}_x$.

[4] In case we are interested in estimation of the unknown regression function $m(\cdot)$ at the point $x \in (0,1)$ only, it is sufficient to consider assumptions C2, C3 and C4 to be satisfied in some small neighbourhood of $x \in (0,1)$ only.

C5 The bandwidth parameter h_N satisfies $h_N \to 0$ as $N \to \infty$, such that $Nh_N^3 \to \infty$;

Given the notation above and the stated assumptions we can formulate now the following theorems, which describe the main statistical properties of the local polynomial estimate.

THEOREM 1.1 (Asymptotic bias and variance)
Let us assume model (1.1) and let conditions C1 – C5 be all satisfied. Then the following holds:

(i) the asymptotic conditional bias for $p - \nu$ odd is given by

$$\mathbb{A}\text{s.}\mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right] = \frac{\nu!\ m^{(p+1)}(x)}{(p+1)!\ h_N^{\nu-p-1}} \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1} \boldsymbol{\mu}_p + o_{\mathbf{P}}\left(h_N^{p+1-\nu}\right);$$

(ii) the asymptotic conditional bias for $p - \nu$ even is given by

$$\mathbb{A}\text{s.}\mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right] = \frac{\nu!\ \left[m^{(p+2)}(x) + (p+2)m^{(p+1)}(x)\frac{f'(x)}{f(x)}\right]}{(p+2)!\ h_N^{\nu-p-2}} \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1} \boldsymbol{\mu}_{p+1} + o_{\mathbf{P}}\left(h_N^{p+2-\nu}\right);$$

(iii) the asymptotic conditional variance is given by

$$\mathbb{A}\text{s.}\mathbb{V}\text{ar}\left[\widehat{m}^{(\nu)}(x)\right] = \frac{\nu!^2\sigma^2(x)}{f(x)Nh_N^{1+2\nu}} \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}\mathbf{e}_\nu + o_{\mathbf{P}}\left(\frac{1}{Nh_N^{1+2\nu}}\right);$$

where $p \in \mathbb{N}_0$ is a degree of the local polynomial approximation and $\nu \in \{0, 1, \ldots, p\}$.

THEOREM 1.2 (Asymptotic normality)
Let us assume model (1.1) and conditions C1 – C5 and some fixed $x \in (0,1)$ from the domain of interest. Then the following convergence holds true

$$\sqrt{Nh_N^{1+2\nu}}\left(\widehat{m}^{(\nu)}(x) - m^{(\nu)}(x) - \mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right]\right) \xrightarrow[N\to\infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{\nu!^2\sigma^2(x)}{f(x)} \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}\mathbf{e}_\nu\right),$$

for any $\nu \in \{0, 1, \ldots, p\}$ and $p \in \mathbb{N}_0$ to be the order of the local polynomial approximation.

Proofs. The proofs of both theorems can be found in J.Fan and Huang (1993) together with some further discussion on this topic so, we will omit it in this thesis. ■

Both results in theorems above are important in order to derive proper statistical properties for the local polynomial M-smoothers, which are of the key interest in our research.

Finally, we would like to mention a few additional issues related to the choice of the order of the local polynomial approximation – the degree $p \in \mathbb{N}_0$. One has to be aware of some irremissible trade here where the accuracy and a small bias term on one hand stand against a small variance and good interpretation options given a reasonable smoothness assumptions on the other hand. The higher the order of the local polynomial approximation the lower the bias term however, this is balanced on the other hand with a higher imprecision involved in a larger variability and somehow more strict smoothness assumptions that we have to consider. And vice versa. Therefore, a reasonable balance has to be always found in order to provide a suitable final fit.

From the other point of view, it was shown in Ruppert and Wand (1994) that there is no loss in terms of asymptotic variance by doing a local linear approach instead of a local constant fit, just the bias term improves a little. The same reasoning also applies to a comparison of any even order of approximation with its consecutive odd order of approximation (see Figure 1.1). Moreover, as we have already mentioned the boundary issues in the case of local constant fits and their successful elimination in the case of local linear fits, this also generally applies for any odd orders of the local polynomial approximation. To be specific, serious boundary problems appear when using even order fits, which contrasts with odd order approximations that have a nice boundary adaptive property similarly as we have discussed in the case of local linear regression.

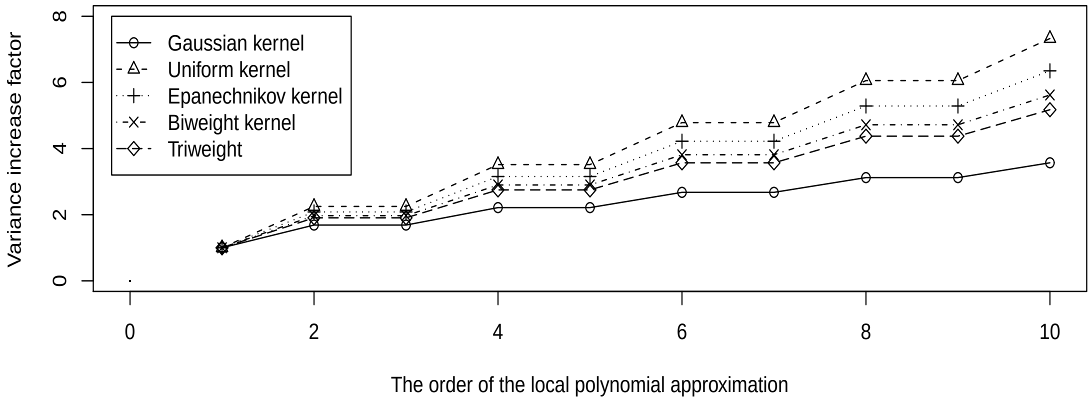


Figure 1.1: The behaviour of an increase factor of the variance term given a specific order of the local polynomial approximation ($p \in \mathbb{N}$) taken from Fan and Gijbels (1995) – see p. 79.

Given the previous argumentation, it is more common to consider odd orders of approximation where one gets a better estimate with respect to a smaller bias term given the same variability (considering the closest smaller even order) while also eliminating the boundary issues in the final model. In real applications the most common choice of the degree of the local polynomial approximation is $p = 3$, which introduces a local cubic approach but other choices can be also commonly found in practice.
It is worth to mention here that from Figure 1.1 one can also get an idea what could be an ideal kernel function to be used for estimation as the variance increase is the slowest for the Gaussian kernel. However, this is not as much crucial therefore, we will not rely on this in our work.

1.3 Robust M-estimates

Let us now turn our attention to M-estimation methods, which are later on used together with the local polynomial regression as a key part of the whole concept of a construction of M-smoothers. Again, we will only state the main results, which are of essential importance for developing a proper statistical background required for the M-smoothers inference.

Let us consider a random sample $X_1, \ldots, X_N$, for some $N \in \mathbb{N}$, which comes from some unknown distribution function $F_\theta(\cdot)$, where $\theta \in \Theta \subseteq \mathbb{R}$ is a location parameter and $\Theta \subseteq \mathbb{R}$ is an open set. The distribution function $F_\theta(\cdot)$ is assumed to be symmetric around $\theta \in \Theta$. Under this setting the true but unknown location parameter $\theta_0 \in \Theta$ of the distribution function $F_{\theta_0}(\cdot)$ can be expressed as

$$\theta_0 = M(F_{\theta_0}) = \underset{t\in\Theta}{Argmin} \int_{\mathbb{R}} (x-t)^2 \,\mathrm{d}F_{\theta_0}(x).$$

Given the random sample $\{X_i;\ i = 1, \ldots, N\}$ it is quite natural to define an estimate $\widehat{\theta}_N$ of the true parameter $\theta_0 \in \Theta$ as

$$\widehat{\theta}_N = M(F_N) = \underset{t\in\Theta}{Argmin} \int_{\mathbb{R}} (x-t)^2 \,\mathrm{d}F_N(x) = \underset{t\in\Theta}{Argmin} \sum_{i=1}^{N} (X_i - t)^2,$$

where F_N denotes the empirical distribution function of $X_1, \ldots, X_N$. However, such solution defined by the least squares method requires a normality assumption for distribution F_θ to be satisfied in order to obtain nice statistical properties. Even a small number of outlying observations can cause a failure of this estimator and moreover, if one considers a heavy-tailed distribution of random errors the least squares estimator is completely inappropriate for such cases.

In order to avoid these problems Huber (1964) proposed to use a generalization of the least squares methods where the empirical estimate $\widetilde{\theta}_N$ of the true parameter $\theta_0 \in \Theta$ is defined as

$$\widetilde{\theta}_N = \widetilde{M}(F_N) = \underset{t\in\Theta}{Argmin} \int_{\mathbb{R}} \rho\,(x-t) \,\mathrm{d}F_N(x) = \underset{t\in\Theta}{Argmin} \sum_{i=1}^{N} \rho\,(X_i - t), \tag{1.6}$$

for some general loss function $\rho(\cdot)$, which is assumed to be symmetric and convex. In general, the choice of the loss function $\rho(\cdot)$ effects how much robust with respect to outliers or heavy-tailed distributions the given M-estimator will be. In an analogous way the expression (1.6) can be thought of as an empirical counterpart of the theoretical functional

$$\theta_0 = \widetilde{M}(F_{\theta_0}) = \underset{t\in\Theta}{Argmin} \int_{\mathbb{R}} \rho\,(x-t) \,\mathrm{d}F_{\theta_0}(x),$$

given for the true value of the parameter $\theta_0 \in \Theta$. Let us now assume that the general loss function $\rho(\cdot)$ has a derivative (one sided derivatives at least) and it holds that $\rho' = \psi$ almost everywhere. Under this assumption it is easy to see that if $\widetilde{\theta}_N = \widetilde{M}(F_N)$ exists it than solves the equation

$$\sum_{i=1}^{N} \psi\,(X_i - t) = 0, \tag{1.7}$$

with respect to $t \in \Theta$. However, the solution of (1.7) could be also not well-defined (none or more than just one solution exists) therefore, it is usual to work with M-estimates defined by

$$\widetilde{\theta}_N = \frac{1}{2}\left[\widetilde{\theta}_N^+ + \widetilde{\theta}_N^-\right], \quad \text{where} \quad \begin{cases} \widetilde{\theta}_N^+ &= \sup_{t\in\Theta} \quad \left\{\sum_{i=1}^N \psi(X_i - t) > 0\right\}; \\ \widetilde{\theta}_N^- &= \inf_{t\in\Theta} \quad \left\{\sum_{i=1}^N \psi(X_i - t) < 0\right\}; \end{cases}$$

We will further assume the definition (1.7). For more details we refer to Jurečková and Sen (1982) or Huber (1981). Some modifications can be also found in Hampel (1986) or Hampel (1974).

In the next step, there is a convention in the M-estimation theory to define function $\lambda_{F_\theta}(\cdot)$ where

$$\lambda_{F_\theta}(t) \overset{def.}{=} -\int_{\mathbb{R}} \psi\,(x-t) \,\mathrm{d}F_\theta(x), \tag{1.8}$$

and to assume that such function is differentiable in $t \in \Theta$, which is mostly the case indeed. This allows us to investigate statistical properties of M-estimates using the Taylor expansion of the smooth function $\lambda_{F_\theta}(t)$ for $t \in \Theta$, rather than investigating properties of M-estimates via a non-smooth or even discontinuous function $\psi(\cdot)$. Using this notation it is easy to see that 1.8 is (up to the sign) just a theoretical version of the empirical equation 1.7 therefore, it holds that $\lambda_{F_\theta}(t) = 0$, for the true value $t = \theta_0$ of the parameter $\theta \in \Theta$.

Given the notation above we can now formulate two theorems, which are crucial for deriving results in the case of M-smoothers inference.

THEOREM 1.3 (Existence of the solution)
Let $\theta_0 \in \Theta$ be an isolated root of the equation $\lambda_{F_\theta}(t) = 0$ and let $\psi(x - t)$ be monotone in $t \in \mathbb{R}$. Then $\theta_0 \in \Theta$ is unique and any solution sequence $\{\widetilde{\theta}_N\}_{N \in \mathbb{N}}$ of the empirical equation $\lambda_{F_N}(t) = 0$ converges to θ_0 almost surely. If moreover, $\psi(t)$ is continuous in t in a small neighbourhood of $\theta_0 \in \Theta$ at least then there exists such a solution sequence.

THEOREM 1.4 (Asymptotic normality of the solution)
Let $\theta_0 \in \Theta$ be an isolated root of the equation $\lambda_{F_\theta}(t) = 0$ and let $\psi(x - t)$ be monotone in $t \in \mathbb{R}$. Suppose that $\lambda_{F_\theta}(t)$ is differentiable at $t = 0$ such that $\lambda'_{F_\theta}(0) \neq 0$. Let $\int_{\mathbb{R}} \psi^2(x - t) dF_\theta(x)$ is finite for t in some small neighbourhood of $\theta_0 \in \Theta$ and it is continuous at $t = 0$. Then any solution sequence $\{\widetilde{\theta}_N\}_{N \in \mathbb{N}}$ of the empirical equation $\lambda_{F_N}(t) = 0$ is asymptotically normal with the expectation to be equal to θ_0 and the variance term equal to

$$\sigma^2(\psi, F_{\theta_0}) = \frac{\int_{\mathbb{R}} \psi^2(x - \theta_0) dF_{\theta_0}(x)}{\left[\lambda'_{F_{\theta_0}}(\theta_0)\right]}.$$

Proofs. Complete and detailed proofs of both theorems can be found in Serfling (1980). ■

M-estimates are frequently referred to as robust estimation methods because of their robust flavour, which is given by the choice of the general loss function $\rho(\cdot)$. There are many different choices for this function (see a detailed discussion in Huber (1981) or a brief overview in Chapter 5) however, in this thesis we will focus mainly on three specific ones: a classical L_2 norm which brings M-estimates back to the least squares estimation, L_1 norm which is one of the most popular choices in case of robust M-estimates and finally, the Huber function, which is defined as

$$\rho(x) = \begin{cases} x^2/2, & \text{for } |x| \leq T; \\ T\left(|x| - \frac{T}{2}\right), & \text{for } |x| > T; \end{cases}$$

where $T > 0$ is some trimming constant, which effects the amount of sensitivity with respect to outlying observations. The Huber function is considered to be the main representative of the whole class of loss functions usually considered for the robust M-estimations procedures[5].

Although, we have described M-estimates related to the estimation of a location parameter only straightforward extensions were introduced for regression approaches as well. One can easily consider M-estimates within classical linear regression models (see McKean (2004) or Yohai and Maronna (1979)) and further generalizations can be also found for nonparametric approaches and the robust local linear

[5] Some other common choices of the loss function $\rho(\cdot)$ are Tuckey's function, Andrew's function or many others. For a short discussion and a list of the most common loss functions used for M-estimation we refer to Chapter 5.

regression especially (He and Shao, 1996). We will extend the robust local linear approach into higher orders of approximation and we will discuss local polynomial M-smoothers[6].

The main idea of the M-smoothers approaches is to bring together the flexibility property gained from adopting the local polynomial approach and the robustness property coming from the M-estimation theory. One expects that combining these two statistical theories will bring new results, which can be also seen in terms of a combination of results from these two theories separately. We will discuss the M-smoothers regression techniques in more detail in the next chapters.

1.4 The aim of the thesis

The aim of this thesis is to develop a proper statistical theory for a flexible and robust estimation of an unknown regression function while taking into account some possible structural breaks – discontinuity points. The main idea is to generalize the local linear M-smoothers approach proposed by H.Rue et al. (1998) and to consider local polynomial M-smoothers instead. Local linear M-smoothers were proposed as a straightforward generalization of the local polynomial regression techniques to offer a new and flexible method for image processing analysis where flexibility was the key issue with respect to frequent outliers in one dimensional image data sequences. Local linear M-smoothers are also well discussed in Antoch and Janssen (1989) and some brief overview can be also found in a nice summary book on nonparametric regression by Fan and Gijbels (1996).

The main advantage of such generalization is to bring more flexibility into the final model and to weak some distributional assumptions required for the statistical inference to hold. Adopting higher orders of estimation (polynomial approach) allows us to model more complex regression functions while also obtaining estimates for consecutive derivatives of the function.
On the other hand, using change-point techniques and regression methods together brings a possibility to consider and to model regression functions with sudden changes like jumps, changes in direction or changes in curvature as well. One can be also interested in higher order changes however, in such cases the interpretation of results gets much more difficult and it is not straightforward any more[7].

The change-point methods were firstly introduced within a noparametric regression framework by Müller (1992) who considered one-sided kernel estimates in order to detect location changes.
One-sided kernel estimates were also considered by Loader (1996) who investigated changes in location and direction at the same time. We will also base our approach on the one-sided kernel estimates and we will consider local polynomial M-smoothers together with a change-point problem similarly as it was also considered by Antoch et al. (2006) for the local linear scenario and the L_2 norm.

A proper statistical theory is to be developed under a variety of different conditions, which includes homoscedastic and heteroscedastic variance structures given for independent random errors and also for some models with dependent structures of random error sequences. Specifically, we will considerer an α-mixing dependence structure proposed by Rosenblatt (1956), which involves most common dependent random error models including a very common auto-regressive (AR) model.

In order to make qualitative decisions regarding some possible change-point occurrences there will be some statistical tests proposed in this thesis and the main statistical properties for such tests are investigated under the conditions posed for the models under consideration.

[6] We will equivalently refer to the local polynomial M-smoothers as to M-smoothers only.

[7] It is said in general that a human naked eye is able to distinguish changes in behaviour of a function up to the third order at most (jumps, direction and curvature) and all higher order structural changes are not visible any more.

Once the testing theory for structural changes in the model is developed one will get into many difficulties that make the application of such tests and direct methods rather inconvenient maybe even infeasible. In order to avoid such difficulties we will propose an alternative way based on resampling methods and bootstrapping and a proper statistical inference will be derived. We will use a small adaptation of the bootstrap technique proposed by Bickel and Freedman (1981). An extensive study on some bootstrap approaches applied to nonparametric regression can be also found in Belyaev (1995). In case of block bootstrap techniques, which are required for dependent data cases we refer to Politis and Romano (1992).

This thesis is structured as follows: in the next chapter we will mainly discuss the local polynomial M-smoothers approaches under the variety of conditions and the main results will be also proved there. In Chapter 3 we will introduce a model with structural breaks, change-points respectively. As far as we are primary interested in estimation of regression functions with structural changes an appropriate statistical theory will be developed in this chapter and some statistical tests for testing of significancy of such changes will be proposed as well.

In Chapter 4 we will discuss alternative bootstrap approaches, which are used to mimic the unknown distribution of interest for the test statistic under the null hypothesis. Three different approaches will be shown and all proofs of the main results will be also presented. In Chapter 5 where we will shortly mention some additional however, important key issues (e.g. bandwidth parameter selection, scale function estimations or loss function choice), which are unfortunately not discussed in more details in the thesis however, one can easily find all required details in references.

Finally, in Chapter 6 an extensive simulation study is presented in order to see the actual performance of the proposed M-smoothers and the bootstrap algorithms. We will also use the modelling and testing tools developed in this thesis on a real data sample and a detailed comparison and discussion will be given. At the very end of this thesis a detailed discussion and conclusion is given in Chapter 7.

1.5 State of Arts

Before we start with a description of the main results in the thesis let us firstly clearly state here in a few items what are the results derived by other authors before and what was derived directly in this thesis as a new contribution to a statistical nonparametric modelling.

- we have proposed to extend the local linear M-smoothers approaches to the local polynomial M-smoothers methods and the corresponding statistical inference and asymptotic is provided;
- a generalization of a homoscedastic model into a heteroscedastic variance structure is introduced and all necessary results are proved;
- a generalization of some statistical tests and testing approaches is adopted in order to fit them on considered model scenarios including homoscedastic and heteroscedastic variance structures with both independent as well as dependent random error terms;
- we have proposed new bootstrap algorithms and all their adaptations, which are used to mimic the unknown distribution functions of the test statistics under the null hypothesis and the considered model scenario and all proofs are given in detail as well;
- an adaptation of the moving blocks bootstrap algorithm is discussed for models with dependent random error terms for both, the homoscedastic as well as heteroscedastic model scenarios;
- we have investigated computational aspects of the proposed methods and an extensive simulation study and a real data example are presented in this thesis together with a detailed discussion;

"A pleasure in the job
puts perfection in the work."

Aristotle
(384 BC – 322 BC)

2 Local Polynomial M-smoothers

2.1 Introduction to M-smoothers

Let $\{(X_i, Y_i);\ i = 1, \ldots, N \in \mathbb{N}\}$ be a finite two-dimensional random sample from some unknown distribution function $F_{(X,Y)}(x, y)$. We would like to investigate a dependence structure of the random variable Y given the value of the random variable X in sense of a classical nonparametric regression based on a conditional expectation function $m(x) = \mathbb{E}[Y|X = x]$. We will consider two different scenarios here: firstly, a simpler case when one is interested in an estimation of the unknown regression function $m(\cdot)$ at some pre-defined point $x \in (0, 1)$ only, where we assume interval the $(0, 1)$ to be the domain of interest[8] of the random variable X. On the other hand, one can be also interested in estimation of the whole regression function at once, which involves estimation of $m(x)$ for all $x \in (0, 1)$. This represents a slightly more complex approach. As far as a proper statistical inference slightly differs for these two cases we will discuss both of them. Firstly, we will point our attention to an estimation of $m(\cdot)$ at some pre-specified point and afterwards, we will generalize this simpler approach to estimate the whole regression curve at once.

Unlike the classical nonparametric approaches we would like to adopt methods, which would allow us to model the unknown regression function in a less restricted way specifically, we would like to weaken some distributional assumptions to cover heavy-tailed distributions of random errors and to drop some smoothness assumptions at the same time (implementing discontinuity points into the model).
Basically, we will proceed in three steps in order to propose a final method, which will take into account all stated requirements.

❏ Using higher order approximation methods one gains much better results especially with respect to a bias term, which necessarily decreases every time we increase the order of the polynomial approximation. On the other hand, using the local estimation techniques (kernel methods) rather than some global ones (classical linear regression or some piece-wise estimation techniques) brings into the final estimate a huge amount of additional flexibility. By combining these two approaches one obtains a very flexible class of estimators, which are referred to as the local polynomial estimators.

❏ In order to deal with some possible outlying observations or even with heavy-tailed distributions (as in the case of classical M-estimates discussed in Chapter 1) we combine the local polynomial regression techniques and M-estimation methods to introduce the local polynomial M-smoothers approaches. Specifically, we will further consider three different situations here: a model with a homoscedastic and heteroscedastic variance structure, both under the assumption of independent and identically distributed random errors and we will also discuss an α-mixing model for a kind

[8] We consider interval $(0, 1)$ to be the domain of interest for simplicity however, one can also consider any general interval (a, b) for $a, b \in \mathbb{R}$ and the results will follow analogously.

of dependence structure of a sequence of random errors, which include most dependence forms that can be commonly considered in regression scenarios.
We will assume a heteroscedastic variance structure for this model in general.

❑ Finally, in the third step we would like to use the proposed local polynomial M-smoothers and to admit even higher level of flexibility especially in cases where one wants to model a regression function with some possible discontinuity points, so called jumps[9]. Moreover, we will accept an occurrence of discontinuity points not only in the original regression function itself but also in all order derivatives, up to the order of the local polynomial approximation.

Given the three steps above we are now interested in the local polynomial M-smoothers estimates while also assuming some discontinuity points and we will discuss their main statistical properties and the corresponding inference in the next sections.

For the given random sample $\{(X_i, Y_i);\ i = 1, \ldots, N \in \mathbb{N}\}$ we want to estimate the unknown regression function $m(\cdot)$ in sense of a conditional expectation $m(x) = \mathbb{E}\,[Y|X = x]$ given at some chosen point of interest $x \in (0,1)$ or over the whole interval $(0,1)$, which is the domain of interest.
In general, the local polynomial M-smoothers estimate is defined by the minimization problem

$$\widehat{\boldsymbol{\beta}}_x = \underset{(b_0,\ldots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \quad \sum_{i=1}^{N} \rho \left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j \right) \cdot K\left(\frac{X_i - x}{h_N} \right), \tag{2.1}$$

where $\widehat{\boldsymbol{\beta}}_x = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$ is a $(p+1)$-dimensional vector of parameter estimates given at the point $x \in (0,1)$ and $p \in \mathbb{N}$ is a degree of the local polynomial approximation[10]. Function $K(\cdot)$ stands for a classical kernel function common for nonparametric regression modelling and h_N is an appropriate bandwidth parameter. Function $\rho(\cdot)$ stands for a general loss function, which is assumed to be symmetric and convex such that for its derivative (or one-sided derivatives at least) it holds that $\rho' = \psi$ almost everywhere (a.e.).

Under such representation the M-smoothers estimate $\widehat{m}(x)$ of the regression function $m(\cdot)$ at the fixed point $x \in (0,1)$ is given by the assignment $\widehat{m}(x) = \widehat{\beta}_0$, where $\widehat{\beta}_0$ is just the first element of the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x$ defined by (2.1). Moreover, if one is interested in an estimate of the first derivative of the regression function $m(\cdot)$ at $x \in (0,1)$ rather than the regression function itself an analogous identity $\widehat{m}'(x) = \widehat{\beta}_1$ holds as well. In general, we can write $\widehat{m}^{(\nu)}(x) = \nu! \widehat{\beta}_\nu$, for any $\nu = 0, \ldots, p$, in order to get estimates of all other consecutive orders of derivatives of the regression function $m(\cdot)$ up to the order $p \in \mathbb{N}$, all given at the same point $x \in (0,1)$. However, unlike the classical nonparametric approach presented in (1.4) the functional representation of $m(\cdot)$ where we had $m(x) = \mathbb{E}\,[Y|X = x]$ is not so much straightforward for (2.1) as the functional representation of $m(\cdot)$ heavily depends now on the loss function $\rho(\cdot)$ we choose for the minimization[11].

In case we want to estimate the whole regression function $m(\cdot)$ at once, which is by our assumption defined on interval $(0,1)$, we need to run the estimator defined by (2.1) for all $x \in (0,1)$.

[9] We will refer to jumps as to change-points or structural breaks respectively as we allow for discontinuity points (jumps) not only in the regression function itself but in its derivatives as well. The term "change-point" is commonly used in times series analysis and sequential methods however, we will also adopt this notation for our purposes.

[10] From now on we will only assume the local polynomial approach given for the orders $p \geq 1$ however, all results immediately hold for $p = 0$ (the local constant approach) as well

[11] Some more discussion on the functional representation of the unknown regression function $m(\cdot)$ is addressed in Chapter 5 in the section on different choices of the loss function $\rho(\cdot)$.

Looking now at equation (2.1) one can easily see there a combination of both, the local polynomial regression approach and M-estimation approach as well. Local polynomial regression is implemented through the kernel function $K(\cdot)$, which yields an estimation in a small neighbourhood of $x \in (0,1)$ only, which increases a local flexibility and also a polynomial approximating term $\sum_{j=0}^{p} b_j (X_i - x)^j$, which is implemented in order to improve a local precision of the model fit. On the other hand robust M-estimation methods are introduced in (2.1) via a general loss function $\rho(\cdot)$ that generally takes a form of the L_2 norm however, in the M-smoothers approach another loss functions like L_1 norm or Huber's function are used preferably. Apropos, the Huber function is considered to be the most traditional choice in the case of robust estimation[12] in general.

Given the finite sample size[13] the convex property of ρ and the fact that $\rho' = \psi$ a.e. we can consider a problem equivalent to (2.1), which is now expressed as a set of normal equations

$$\sum_{i=1}^{N} \psi\left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j\right) \cdot (X_i - x)^l \cdot K\left(\frac{X_i - x}{h_N}\right) = 0, \qquad (2.2)$$
$$\text{for} \quad l = 0, \ldots, p,$$

which are solved for $\widehat{\boldsymbol{\beta}}_x = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top \in \mathbb{R}^{p+1}$. However, unlike the classical regression techniques based on the L_2 norm the solution of the set of equations (2.2) is not given in an explicit form any more. This also involves some issues related to an optimal bandwidth selection as the Asymptotic Mean Squared Error term (AMSE), which is commonly used to determine the asymptotically optimal bandwidth parameter in nonparametric regression settings is not expressible in an explicit form either therefore, one has to implement iterative procedures or other methods in order to get close to the solution[14]. Another common usage of the AMSE quantity is to judge the performance of the given estimator but using the same argumentation as before this can not be so easily accessible for the M-smoothers methods either. We will focus on some common approaches to these issues and we will also discuss an optimal bandwidth parameter selection in Section 5.1.

Let us now start with the simplest model case we can consider under this scenario – a model with a homoscedastic variance structure with independent random errors – and we will discuss the main properties for such model in the next section. We will prove the main results and afterwards this homoscedastic variance model will be generalized into a model with heteroscedastic variance and a model with dependent random errors as well.

2.2 Homoscedastic models

Let us remind the given random sample $\{(X_i, Y_i);\ i = 1, \ldots, N \in \mathbb{N}\}$, which comes from some unknown distribution function $F_{(X,Y)}(x, y)$ where the dependence of the random variable Y given the value of the random variable X can be in general described by the model

$$Y_i = m(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim G, \qquad i.i.d, \qquad i = 1, \ldots, N, \qquad (2.3)$$

[12] One can also consider many other choices of the loss function $\rho(\cdot)$ for example Tuckey's function or Andrew's function or many others (see Section 5.3 some brief discussion on loss functions). However, in this text we will only consider L_2 and L_1 norms and the Huber function respectively.

[13] For an infinite sum one would need to satisfy some necessary assumptions in order to chance the sum and derivative operators however, such assumptions trivially hold given the set of assumptions required for the model anyway.

[14] One can not even use a classical Newton-Raphson iterative algorithm here as it assumes an existence of a derivative of a function under the minimization with respect to an argument which is of interest. However, this is usually not satisfied in case of M-smoothers for most choices of the loss function $\rho(\cdot)$.

where $i.i.d.$ stands for independent and identically distributed random error terms with a symmetric distribution function $G(\cdot)$, such that $G(e) = 1 - G(-e)$, for all $e \in \mathbb{R}$.

Using this model definition we want to avoid any finite moments assumptions for random error terms and to also account for heavy tailed distributions such as the Cauchy distribution or others. Given this requirement we will refer to a scale of random errors rather than referring to their variance, which is under our model scenario allowed to be infinite as well.

In order to quantify the scale of some random variable such that the definition will remain free of any finite moment assumptions one can use many proposed methods, which have been extensively discussed in literature (we refer to Bickel and Lehmann (1976a) and Bickel and Lehmann (1976b) for some further details). However, in the following text we will only consider an interquartile range of a random variable defined as

$$\Delta(\varepsilon_1) \equiv \Delta(G) \overset{def.}{=} G^{-1}\left(\frac{3}{4}\right) - G^{-1}\left(\frac{1}{4}\right),$$

for G to be a distribution function of ε_1 and

$$G^{-1}(u) = \inf\{e \in \mathbb{R};\ G(e) \geq u\}, \text{ for } 0 < u < 1,$$

stands for a quantile function, which corresponds to the distribution function $G(\cdot)$. It was proved in Bickel and Lehmann (1976a) or Bickel and Lehmann (1976b) respectively that $\Delta(G)$ when defined for some function $G(\cdot)$ to be a symmetric distribution function then it satisfies all necessary assumptions, which are required in general for a measure of dispersion (spread respectively) to be satisfied.

For model (2.3) the estimate of the regression function $m(\cdot)$ at the given point $x \in (0,1)$ or the estimates of the derivatives at the same point up to the order $p \in \mathbb{N}$ are defined by the minimization (2.1) same as they can be given in sense of an iterative solution of the set of equations (2.2).

In the following lines we will discuss the main statistical properties for such estimates and we will investigate their asymptotic inference namely, the bias and variance terms and their distributional properties too. Given the knowledge of an asymptotic behaviour of an estimate we can easily apply the proposed M-smoothers method to deal with real data problems as we can easily construct confidence bounds or critical regions for a corresponding hypothesis testing problems, which are mostly of the main statistical interest.

2.2.1 Assumptions for homoscedastic model

Before we state the main results derived for the homoscedastic model (2.3) let us firstly discuss a set of assumptions, which are required for the results to hold[15].

A1 The marginal density function $f(\cdot)$ of the random variables $X_1, \dots, X_N$, which are $i.i.d.$ is assumed to be absolutely continuous, positive and bounded on $[0,1]$;

A2 Random errors $\varepsilon_1, \dots, \varepsilon_N$, are assumed to be $i.i.d.$, mutually independent of X_i, for $i = 1, \dots, N$, with a symmetric and continuous distribution function $G(\cdot)$;

[15] Once we are interested in an estimation of the unknown regression function or its derivatives respectively at some pre-specified point $x \in (0,1)$ only, it is sufficient to consider assumptions A1 and A3 to be satisfied locally only – for some small neighbourhood of $x \in (0,1)$.

A3 The regression function $m(\cdot)$ is assumed to be $(p+1)$-times Lipschitz on interval $(0,1)$, where $p \in \mathbb{N}$ is a degree of the local polynomial approximation;

A4 The loss function $\rho(\cdot)$ is symmetric and convex moreover, we assume it is absolutely continuous and it holds that $\rho' = \psi$ almost everywhere (a.e.);

A5 For function $\lambda_G(t) = -\int \psi(e-t)\mathrm{d}G(e)$, where $t \in \mathbb{R}$ we assume that the derivative $\lambda'_G(t)$ exists and it is continuous at some neighbourhood of $t = 0$. Moreover, $\int \psi^2(e)\mathrm{d}G(e) < \infty$ and $\lambda'_G(0) \neq 0$. It is also satisfied that $\int (\psi(e-\epsilon_N) - \psi(e))^2 \mathrm{d}G(e) < \mathcal{K} \cdot |\epsilon_N|$, for some $\mathcal{K} > 0$ and any sequence $\{\epsilon_N\}_{N=1}^{\infty}$ such that $\epsilon_N \to 0$ as $N \to \infty$;

A6 Function $K(\cdot)$ is a kernel function, which is assumed to be a symmetric[16] density function with a common support on interval $[-1,1]$, such that $\int_{-1}^{1} K^2(u)\mathrm{d}u < \infty$;

A7 The bandwidth parameter h_N satisfies the following: $h_N \overset{N\to\infty}{\longrightarrow} 0$, such that $Nh_N \overset{N\to\infty}{\longrightarrow} \infty$, more precisely $h_N \asymp N^{\iota}$, where $\iota \in \left(-\frac{1+\delta}{(1+2p)}, -\delta\right)$, for some $\delta > 0$ small enough;

Most of the assumptions above are just usual conditions required for the M-smoothers method to work and they are generally satisfied for most common situations including all practical choices of the loss function $\rho(\cdot)$, the form of the distribution function $G(\cdot)$ or the shape of the kernel function $K(\cdot)$, which is used for the estimation procedure. Assumption A1 requires that $0 < f(x) \leq \mathcal{K}$, for some $0 < \mathcal{K} < \infty$ and any $x \in [0,1]$. It is also important to assume a mutual independence of measurements and random errors in the assumption A2 however, under the given model it is a quite natural condition. Similarly, it is quite reasonable to assume a symmetric property of the distribution function $G(\cdot)$ in A2 and moreover, it also simplifies the proofs a lot.

One could think of assumption A3 to be slightly restrictive as it requires the Lipschitz property of the order $p+1$, where $p \in \mathbb{N}$ is the given degree of the local polynomial approximation however, this is directly related to the fact that we claim the M-smoothers estimate to be a smooth estimate of the order p. Once we do not want to assume the Lipschitz property for such high orders we do also have to decrease the degree of the local polynomial approximation.

Referring to assumptions A4 and A5, they both come directly as straightforward modifications of the assumptions posed in case of simple M-estimation techniques discussed in Section 1.3 specifically, those are assumptions required for both Serfling's theorems to hold[17]. The last two assumptions refer to the kernel function $K(\cdot)$ and the bandwidth parameter $h_N > 0$. They both come from the theory on nonparametric kernel estimates (for some further discussion see Section 5.1).

The robust property of the M-smoothers estimates is implicitly given within the assumption A2 where we allow for any symmetric distributions with no explicit conditions posed on an existence of any finite moments of random errors. This of course includes many kinds of heavy-tailed distributions including the Cauchy distribution that usually causes serious problems (even inconsistency of an estimate) when they go together with the classical regression estimation methods based on the L_2 norm.

On the other hand, this also accounts for models with random errors that could be possibly comprised even of an inconsiderable amount of outlying observations. Unlike the classical least squares regression methods, which would fail providing a correct inference using the M-smoothers approach one will get fully consistent estimates with a proper inference in such cases as well.

[16] The symmetric property of the kernel function $K(\cdot)$ is not necessarily required however, it is quite common and natural to assume so. We will also discuss non-symmetric kernels in Chapter 3.

[17] For more details on Serfling's theorems see Chapter 1, specifically Section 1.3 or the paper by Serfling (1980).

2.2.2 The main asymptotic results

In this section we will derive the main statistical properties for the M-smoothers estimates under the homoscedastic model 2.3 and we will discuss their most important qualities. All proofs of the results will be also given here.

For sake of simplicity let us firstly introduce the following notation:

$$\mathsf{X}_N = \left(\left[\frac{X_i - x}{h_N} \right]^j \right)_{\substack{i=1,\ldots,N;\\ j=0,\ldots,p;}}, \quad \mathsf{S}_1 = \left(\int_{-1}^{1} u^{j+l} K(u) \mathrm{d}u \right)_{\substack{j=0,\ldots,p;\\ l=0,\ldots,p;}}, \quad \mathsf{S}_2 = \left(\int_{-1}^{1} u^{j+l} K^2(u) \mathrm{d}u \right)_{\substack{j=0,\ldots,p;\\ l=0,\ldots,p;}}$$

$$\mathsf{W}_N = \operatorname{diag}\left\{ K\left(\frac{X_1 - x}{h_N} \right), \ldots, K\left(\frac{X_N - x}{h_N} \right) \right\}, \quad \text{and} \quad \mathsf{H}_N = \operatorname{diag}\left\{ 1, h_N^{-1}, \ldots, h_N^{-p} \right\},$$

are given matrices of the appropriate types and

$$\mathbf{e}_\nu = (\underbrace{0,\ldots,0}_{\nu-times}, 1, \underbrace{0,\ldots,0}_{(p-\nu)-times})^\top \qquad \text{and} \qquad \boldsymbol{\psi}(\mathbf{c}) = (\psi(c_1), \psi(c_2), \ldots, \psi(c_N))^\top$$

are both real-valued vectors where $\nu \in \{0,\ldots,p\}$ and $\mathbf{c} = (c_1,\ldots,c_N) \in \mathbb{R}^N$ is some arbitrary N-dimensional real-valued (or random respectively) vector.
Moreover, $\lambda_G'(t) = -\frac{\partial}{\partial t} \int_{-\infty}^{\infty} \psi(e-t) \mathrm{d}G(e)$ stands for a derivative of $\lambda_G(\cdot)$ defined by 1.8.

Given the assumption A6 both matrices S_1 and S_2 are just $(p+1) \times (p+1)$ type matrices with all finite elements. Moreover, the definition of function $\lambda_G(t)$ comes from the M-estimates theory as we have described in Section 1.3 and matrices X_N and W_N respectively, can be thought as nonparametric counterparts of a classical design matrix and a diagonal matrix of weights respectively, which are commonly used in parametric regression. Under this notation and the assumptions above we can formulate now the main results for the local polynomial M-smoothers estimates under the homoscedastic model 2.3.

THEOREM 2.1 (Consistency for homoscedastic M-smoothers)
For model 2.3 and assumptions A1 – A7 the M-smoothers estimates of the regression function $m(\cdot)$ and its derivatives respectively are consistent, in other words it holds that

$$\sqrt{N h_N^{1+2\nu}} \cdot \left(\widehat{\beta}_\nu - \frac{m^{(\nu)}(x)}{\nu!} \right) = O_{\mathbf{P}}(1),$$

for $N \to \infty$, any $\nu \in \{0,\ldots,p\}$ and the given point $x \in (0,1)$.

Proof. See the proof of this theorem in Section 2.2.3 below. ■

The consistency result is one of the most important statistical properties for any estimation technique as it can be interpreted as a property of converging to an unknown but the true quantity that is of interest. Without assuring the consistency result firstly there is not too much sense in investigating any further statistical properties. Once we have derived the consistency result we can focus on additional statistical properties, which may be in hand when dealing with real data problems.

THEOREM 2.2 (Asymptotic conditional bias and variance for $\widehat{\beta}_x$)
For model 2.3 and assumptions A1 – A7 we can express the asymptotic conditional bias and variance terms for the vector of parameter estimates as

- *Asymptotic conditional bias*[18] *term:*

$$\mathsf{H}_N^{-1}\left(\widehat{\boldsymbol{\beta}}_x-\boldsymbol{\beta}_x\right)=\left(\mathsf{X}_N^\top\mathsf{W}_N\mathsf{X}_N\right)^{-1}\cdot\mathsf{X}_N^\top\mathsf{W}_N\left(\frac{m^{(p+1)}(x)}{(p+1)!}\cdot(\mathbf{X}-x)^{(p+1)}\right)+o_{\mathbf{P}}\left(h_N^{p+1}\right)\tag{2.4}$$

- *Asymptotic conditional variance term for* $\mathsf{H}_N\widehat{\boldsymbol{\beta}}_x$*:*

$$\frac{\mathbb{E}\left[\psi(\varepsilon_1)\right]^2}{[\lambda_G'(0)]^2}\left(\mathsf{X}_N^\top\mathsf{W}_N\mathsf{X}_N\right)^{-1}\mathsf{X}_N^\top\mathsf{W}_N^2\mathsf{X}_N\cdot\left(\mathsf{X}_N^\top\mathsf{W}_N\mathsf{X}_N\right)^{-1}+o_{\mathbf{P}}\left(\frac{1}{Nh_N}\right)\tag{2.5}$$

Proof. The proof of this theorem will be given in Section 2.2.3 below. ■

In Theorem 2.2 the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x$ is defined by the minimization problem (2.1) and $(\mathbf{X}-x)^{(p+1)}=\left((X_1-x)^{p+1},\ldots,(X_N-x)^{p+1}\right)^\top$ as $\mathbf{X}=(X_1,\ldots,X_N)^\top$. Moreover, by the assumption A5 it holds that $\mathbb{E}\psi^2(\varepsilon_1)=\int_{\mathbb{R}}\psi^2(\varepsilon_1)\mathrm{d}G(\varepsilon_1)<\infty$.

One has to be aware of the matrix notation used in Theorem 2.2 as it expresses the bias term and the variance not only for the estimate of the unknown regression function at the given point $x\in(0,1)$ but it also specifies the bias and variance terms for estimates off all order derivatives of the regression function $m(\cdot)$ at $x\in(0,1)$ up to the order of the local polynomial approximation $p\in\mathbb{N}$.

However, rather than relying on the conditional bias and variance terms one is in statistic mostly interested in an asymptotic distributional properties of the given estimate. Therefore, we will investigate the asymptotic distribution of the M-smoothers estimates and we will also formulate two additional theorems to state the asymptotic bias and variance terms as well.

THEOREM 2.3 (Asymptotic bias term for $\widehat{m}^{(\nu)}(x)$)
For model (2.3), assumptions A1 – A7 and the given notation the asymptotic bias term for the M-smoothers estimate of the unknown regression function or its derivatives respectively is equal to

$$\mathbb{A}\mathsf{s}.\mathbb{B}\mathsf{ias}\left[\widehat{m}^{(\nu)}(x)\right]=\nu!h_N^{p+1-\nu}\cdot\left(\frac{m^{(p+1)}(x)}{(p+1)!}\right)\cdot\mathbf{e}_\nu^\top\mathsf{S}_1^{-1}\boldsymbol{\mu}_p+o_{\mathbf{P}}\left(h_N^{p+1-\nu}\right),$$

where $\boldsymbol{\mu}_p=\left(\int_{-1}^1u^{p+1}K(u)du,\int_{-1}^1u^{p+2}K(u)du,\ldots,\int_{-1}^1u^{2p+2}K(u)du\right)^\top\in\mathbb{R}^{p+1}$ *and* $\nu\in\{0,\ldots,p\}$ *stands for the order of the corresponding derivative of the unknown regression function* $m(\cdot)$*.*

Proof. See Section 2.2.3 below. ■

THEOREM 2.4 (Asymptotic variance term for $\widehat{m}^{(\nu)}(x)$)
For model (2.3), assumptions A1 – A7 and the notation above the asymptotic variance term for the M-smoothers estimate of the unknown regression function or its derivatives respectively is equal to

$$\mathbb{A}\mathsf{s}.\mathbb{V}\mathsf{ar}\left[\widehat{m}^{(\nu)}(x)\right]=\frac{\nu!^2\mathbb{E}\psi^2(\varepsilon_1)}{Nh_Nf(x)[\lambda_G'(0)]^2}\cdot\mathbf{e}_\nu^\top\mathsf{H}_N\mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}\mathsf{H}_N\mathbf{e}_\nu+o_{\mathbf{P}}\left(\frac{1}{Nh_N^{1+2\nu}}\right),$$

where $\nu\in\{0,1,\ldots,p\}$ *stands again for the order of the derivative of the regression function* $m(\cdot)$*.*

[18] In Theorem 2.2 we use the "bias" notation to refer to a difference between the parameter estimate and the true value of the parameter rather than using the original definition as the bias expression in a classical sense of a conditional expectation is not well-defined in this case (see the proof of the theorem for explanation).

Proof. See Section 2.2.3 below. ■

Applying now the classical nonparametric approach we could use Theorems 2.3 and 2.4 to express the Asymptotic Mean Square Error (AMSE) term of the estimate of the unknown regression function $m(\cdot)$ at the given point $x \in (0,1)$ for $\nu = 0$, which is commonly used in nonparametric regression to determine the optimal value of the bandwidth parameter. The AMSE quantity defined as a sum of the variance and the bias term squared would be expressed as

$$\mathsf{AMSE}_x(h_N) = \boldsymbol{\mu}_p^\top \left\{ \left(\frac{m^{(p+1)}(x)}{(p+1)!} h_N^{(p+1)} \right)^2 \cdot \boldsymbol{I}_{p+1} + \frac{\mathbb{E}\psi^2(\varepsilon_1)}{N h_N f(x) [\lambda_G'(0)]^2} \cdot \mathsf{S}_2 \right\} \boldsymbol{\mu}_p + o_{\mathbf{P}}\left(h_N^{p+1}\right),$$

where the optimal bandwidth h_{opt} is given as the one that minimizes the $\mathsf{AMSE}_x(h_N)$ quantity. This quantity can be however, equivalently approached via a minimization of the error estimates using a general loss function $\rho(\cdot)$ that the cross-validation is generally based on (see also Chapter 5). Anyhow, the value of h_{opt} is not directly expressible in an explicit form for a general loss function $\rho(\cdot)$ (its derivative $\psi(\cdot)$ respectively) unless, one considers some additional assumptions, which are not necessarily required otherwise. Iterative procedures need to be considered again.

Finally, we will formulate a result, which will state the distributional property of the M-smoothers estimate of the unknown regression function $m(\cdot)$ at some given point $x \in (0,1)$ or its derivatives evaluated at the same point respectively.

THEOREM 2.5 (Asymptotic normality for $\widehat{m}^{(\nu)}(x)$ under homoscedasticity)
For model (2.3), assumptions A1 – A7, and the notation introduced above the M-smoothers estimate follows asymptotically in law a normal distribution, more precisely it holds that

$$\sqrt{N h_N^{1+2\nu}} \cdot \left(\widehat{m}^{(\nu)}(x) - m^{(\nu)}(x) - \mathbb{B}\mathsf{ias}\left[\widehat{m}^{(\nu)}(x)\right] \right) \xrightarrow[N\to\infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{\nu!^2 \cdot \mathbb{E}\psi^2(\varepsilon_1)}{[\lambda_G'(0)]^2 f(x)} \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1} \mathsf{S}_2 \mathsf{S}_1^{-1} \mathbf{e}_\nu \right),$$

where $\nu \in \{0, 1, \ldots, p\}$ stands for the order of the derivative of the regression function $m(\cdot)$, or its estimate $\widehat{m}(\cdot)$ respectively, and $x \in (0,1)$ is the point of interest chosen in advance.

Proof. The proof of the theorem will be given in Section 2.2.3. ■

Theorem 2.5 formulates the most important result for the M-smoothers method, which is that the estimate of the unknown regression function $m(\cdot)$ at some given point $x \in (0,1)$ or its derivatives evaluated at the same point respectively, have in limit (as $N \to \infty$) a normal distributional property with a zero mean and a finite variance term. We would like to emphasise here that taking into account the asymptotic order of the bias term (see Theorem 2.3) and the convergence rate of $\sqrt{N h_N}$ there is no need to include the bias term in the normality results for $\nu = 0$ as $\sqrt{N h_N} \cdot h^{p+1} \to 0$ in this case. It is also interesting to realize what is the composition of the variance term in Theorem 2.5. As we have already mentioned at the beginning of this chapter, the local polynomial M-smoothers were proposed as a combination of two different approaches, the nonparametric local polynomial regression and the robust M-estimation methods. It is therefore quite natural to expect that the estimates produced by local polynomial M-smoothers should inherit analogous statistical properties, which could be seen as a combination of both, the statistical properties of nonparametric local polynomial estimates and the robust M-estimates as well.

Indeed, the variance term in Theorem 2.5 can be easily decomposed into two main parts where the first term would be equal to $\frac{\mathbb{E}\psi^2(\varepsilon_1)}{[\lambda_G'(0)]^2}$, which is similar to the variance expression of a simple M-estimate as given in Theorem 1.4 (Serfling's theorem) while the remaining term of the decomposition would be equal to $\frac{\nu!^2}{f(x)} \cdot \mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}$, which comes from the variance term of a nonparametric local polynomial estimate at some given point $x \in (0,1)$ (see Theorem 1.2 in Section 1.2).

Therefore, there is an obvious analogy between the M-smoothers regression approach and both methods used for its construction: the local polynomial regression and the robust M-estimation method.

2.2.3 Proofs of Theorems 2.1, 2.2, 2.3, 2.4 and 2.5

In this section we will give complete and detailed proofs of all theorems formulated above and we will justify all stated properties of the M-smoothers estimates under the homoscedastic regression model assumptions as given in A1 – A7.

We will firstly discuss a general idea and we will introduce the main concept of the whole proof. Once we derive a necessary background common for all proofs we will focus on proving the stated results individually.

Let us start with the definition of the M-smoothers estimate, where the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x = \left(\widehat{\beta}_0, \dots, \widehat{\beta}_p\right) \in \mathbb{R}p + 1$ is defined as a minimizer of the expression

$$\underset{(b_0,\dots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \quad \frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \rho\left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j\right) \cdot K\left(\frac{X_i - x}{h_N}\right), \tag{2.6}$$

where we minimize with respect to a vector of parameters $(b_0, \dots, b_p)^\top \in \mathbb{R}^{p+1}$ for some point of interest $x \in (0,1)$, which is given in advance. As we have already mentioned in case of M-estimates we can similarly consider this minimization to be an empirical version of a true but unknown theoretical functional, which defines the theoretical version of the minimization problem as

$$\underset{(b_0,\dots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \quad \lim_{h \to 0} \int_{x-h}^{x+h} \int_{\mathbb{R}} \rho\left(y - \sum_{j=0}^{p} b_j (u - x)^j\right) \mathrm{d}F_{(Y|X=u)}(y) \mathrm{d}F_X(u), \tag{2.7}$$

which is minimized for the true vector of parameters $\boldsymbol{\beta}_x = \left(\beta_0, \dots, \beta_p\right)^\top \in \mathbb{R}^{p+1}$. Function $F_{(Y|X=x)}(\cdot)$ stands here for a conditional distribution function of the random variable Y given $X = x$ and $F_X(\cdot)$ is the corresponding marginal distribution function of the random variable X. Given the random sample $\{(X_i, Y_i);\ i = 1, \dots, N \in \mathbb{N}\}$ we can define an empirical version of this functional as (2.6) where the unknown joint distribution function $F_{(X,Y)}(u,y)$ implemented in (2.7) as a product of the conditional distribution $F_{(Y|X=x)}(y)$ and the marginal distribution $F_X(u)$ is replaced in (2.6) by its empirical counterpart instead.

Now, we want to show that the empirical version (2.6) of the theoretical functional (2.7) can be effectively used in order to obtain estimates for the vector of true parameters $\boldsymbol{\beta}_x = \left(\beta_0, \dots, \beta_p\right)^\top$, which minimizes (2.7) and we will show that such vector of parameter estimates is a suitable estimate for the vector of interest, which is $\left(m(x), m'(x), \frac{m''(x)}{2!}, \dots, \frac{m^{(p)}(x)}{p!}\right) \in \mathbb{R}^{p+1}$, as $N \to \infty$.

We will use the definition of model (2.3), where $Y_i = m(X_i) + \varepsilon_i$ moreover, using a $(p+1)$-times Lipschitz property of the unknown regression function $m(\cdot)$ (see assumption A3) we can approximate the quantity $m(X_i)$ by the corresponding p-order Taylor expansion as

$$m(X_i) = m(x) + m'(x)(X_i - x) + \frac{m''(x)}{2!}(X_i - x)^2 + \cdots + \frac{m^{(p)}(x)}{p!}(X_i - x)^p + o\left(h_N^p\right), \qquad \boxed{2.8}$$

given the fact that we use the local approach defined by kernel $K(\cdot)$ where we have that $|X_i - x| \leq h_N$ as all other cases where this is not true are set to zero by the property of the kernel function.

Plugging in the Taylor expansion $\boxed{2.8}$ for $m(X_i)$ we can now replace the quantity $m(X_i)$ in $\boxed{2.6}$ with the corresponding p-order Taylor polynom $\boxed{2.8}$ and we define a new minimization problem as

$$\widetilde{\boldsymbol{\beta}}_x = \underset{(b_0,\ldots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \rho\left(\varepsilon_i - \sum_{j=0}^{p}\left(b_j - \frac{m^{(j)}(x)}{j!}\right)(X_i - x)^j\right) \cdot K\left(\frac{X_i - x}{h_N}\right), \qquad \boxed{2.9}$$

where $\widetilde{\boldsymbol{\beta}}_x = (\widetilde{\beta}_0, \ldots, \widetilde{\beta}_p)^\top \in \mathbb{R}^{p+1}$ is a vector of the corresponding parameter estimates.

We need to show now that for $N \to \infty$ the vector of parameter estimates, which solves $\boxed{2.9}$ is arbitrarily close in some sense to the vector of parameters, which was supposed to solve the original minimization problem $\boxed{2.6}$. However, this is quite obvious as the difference between two arguments of the loss function $\rho(\cdot)$ in $\boxed{2.6}$ and $\boxed{2.9}$ is of the order $O(h_N^{p+1})$ and we have that $h_N \to 0$ as $N \to \infty$ hence, using the Lipschitz property of function $\rho(\cdot)$, which comes from an existence of the derivative $\rho' = \psi$ we can consider the minimization $\boxed{2.9}$ instead of the original one as we have that the parameter estimates under both formulations can be arbitrarily close to each other, as $N \to \infty$.

For sake of simplicity we will define a new vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^\circ \in \mathbb{R}^{p+1}$ as

$$\widehat{\boldsymbol{\beta}}_x^\circ = (\widehat{\beta}_0^\circ, \ldots, \widehat{\beta}_p^\circ)^\top \overset{def.}{=} \left[\left(\widetilde{\beta}_0 - m(x)\right), \left(\widetilde{\beta}_1 - m'(x)\right) \cdot h_N, \ldots, \left(\widetilde{\beta}_p - \frac{m^{(p)}(x)}{p!}\right) \cdot h_N^p\right]^\top.$$

One can easily see that $\widehat{\boldsymbol{\beta}}_x^\circ$ is now defined as a solution of an analogous minimization problem given by

$$\widehat{\boldsymbol{\beta}}_x^\circ = \underset{(b_0,\ldots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \rho\left(\varepsilon_i - \sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot K\left(\frac{X_i - x}{h_N}\right). \qquad \boxed{2.10}$$

Unlike the previous notation where the minimizer of $\boxed{2.6}$ was supposed give us an estimate for the vector of true parameters $\boldsymbol{\beta}_x = (\beta_0, \ldots, \beta_p)^\top = \left(m(x), m'(x), \ldots, \frac{m^{(p)}(x)}{p!}\right)^\top$, using the new notation with the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^\circ \in \mathbb{R}^{p+1}$ one will get an estimate for the true vector of parameters which is $\boldsymbol{\beta}_x^\circ = (0, \ldots, 0)^\top \in \mathbb{R}^{p+1}$ instead.

Using the assumption A5 we can now apply a partial differential operator to $\boxed{2.10}$ with respect to arguments b_j for $j = 0, \ldots, p$, moreover, the position of the partial differential operator and the sum operator can be reversed, which gives us the set of equations

$$\frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \psi\left(\varepsilon_i - \sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) = 0, \qquad \boxed{2.11}$$

for $l = 0, \ldots, p,$

which are solved for the parameter estimates $\widehat{\beta}_0^\circ, \dots, \widehat{\beta}_p^\circ$. Unfortunately, this solution is expressed in an implicit form only therefore, in order to investigate statistical properties of $\widehat{\beta}_j^\circ$, for $j = 0, \dots, p$ one has to apply some further asymptotic expansions and sophisticated approximations.

The idea is to find a reasonable approximation of the solution of (2.11), which would be if effectively found, asymptotically linear in the unknown vector of parameters $\boldsymbol{\beta}_x^\circ = (\beta_0^\circ, \dots, \beta_p^\circ)$.
Given such asymptotic linear approximation one could then easily obtain a reasonable representation of the parameter estimates in explicit forms and just a straightforward computation and classical inference techniques would be required after that to obtain the final result.

We will now consider the minimization problem (2.10) and the corresponding set of equations (2.11) and we will proceed in three consecutive steps:

- firstly, we will show that the parameter estimates $\widehat{\beta}_0^\circ, \dots, \widehat{\beta}_p^\circ$ need to be all in a small neighbourhood of zero in order to satisfy the set of equations (2.11);
- next, we will prove that $\widehat{\beta}_j^\circ \to 0$ in probability as $N \to \infty$ for every $j = 0, \dots, p$, more specifically, we will show that $\sqrt{Nh_N}\ \widehat{\beta}_j^\circ \ \equiv\ \sqrt{Nh_N^{1+2j}}\ \left(\widehat{\beta}_j - \beta_j\right) = O_{\mathbf{P}}(1)$ for any $j = 0, \dots, p$.
- finally, we will prove the asymptotic normality property for the parameter estimates $\widehat{\beta}_j^\circ$, for all $j = 0, \dots, p$ namely, we will show that

$$\sqrt{Nh_N}\left(\widehat{\beta}_j^\circ - \beta_j^\circ\right) \equiv \sqrt{Nh_N^{1+2j}}\left(\widehat{\beta}_j - \frac{m^{(j)}(x)}{j!}\right) \underset{N\to\infty}{\xrightarrow{\mathscr{D}}} \mathbf{N}(\cdot,\cdot),$$

  where $\mathbf{N}(\cdot,\cdot)$ stands for a normal distribution with appropriate mean and variance quantities;

Proof of Theorem 2.1

Firstly, let us prove the first step, which we re-formulate for now as $\widehat{\beta}_j^\circ = O_{\mathbf{P}}(1)$, for all $j = 0, \dots, p$. Let $\mathbf{b} = (b_0, \dots, b_j)^\top \in \mathbb{R}^{p+1}$ be an arbitrary vector. Assuming the minimization problem (2.10) we can define the "distance function" as

$$\frac{1}{Nh_N}\sum_{i=1}^{N} h(\varepsilon_i, X_i, \mathbf{b}) \ \overset{def.}{=}\ \frac{1}{Nh_N}\sum_{i=1}^{N}\left[\rho\left(\varepsilon_i - \sum_{j=0}^{p} b_j\left(\frac{X_i - x}{h_N}\right)^j\right) - \rho\left(\varepsilon_i\right)\right]\cdot K\left(\frac{X_i - x}{h_N}\right), \tag{2.12}$$

where (2.12) is now minimized for the same vector of parameters as the minimization problem (2.10). Let $\iota_1 \in \{1, \dots, N\}$ is some index chosen arbitrarily. Then two different options are possible: either $\sum_{j=0}^{p} b_j\left(\frac{X_{\iota_1}-x}{h_N}\right)^j \geq 0$ or $\sum_{j=0}^{p} b_j\left(\frac{X_{\iota_1}-x}{h_N}\right)^j < 0$. Let us assume the first option holds. Using the absolute continuity property of function $\rho(\cdot)$ we can express $h(\varepsilon_{\iota_1}, X_{\iota_1}, \mathbf{b})$ as

$$\left[\rho\left(\varepsilon_{\iota_1} - \sum_{j=0}^{p} b_j\left(\frac{X_{\iota_1} - x}{h_N}\right)^j\right) - \rho\left(\varepsilon_{\iota_1}\right)\right]\cdot K\left(\frac{X_{\iota_1} - x}{h_N}\right) = \int_0^{\mathbf{b}^\top\mathbf{X}_{\iota_1}} \psi(\varepsilon_{\iota_1} - t)\mathrm{d}t \cdot K\left(\frac{X_{\iota_1} - x}{h_N}\right),$$

where $\mathbf{b}^\top\mathbf{X}_{\iota_1} = \sum_{j=0}^{p} b_j\left(\frac{X_{\iota_1}-x}{h_N}\right)^j$ and we will investigate the performance of $h(\varepsilon_{\iota_1}, X_{\iota_1}, \mathbf{b})$ with respect to the vector of parameters $\mathbf{b} \in \mathbb{R}^{p+1}$ using the conditional expectation conditioned on the random variable X. Given the fact that $\mathbb{E}\psi(\varepsilon_i) = 0$ for any $i \in \{1, \dots, N\}$ the following now holds

$$\begin{aligned}\mathbb{E}h(\varepsilon_{\iota_1}, X_{\iota_1}, \mathbf{b}) &= \mathbb{E}\left[\int_0^{\mathbf{b}^\top \mathbf{X}_{\iota_1}} [\psi(\varepsilon_{\iota_1} - t) - \psi(\varepsilon_{\iota_1})]\,\mathrm{d}t \cdot K\left(\frac{X_{\iota_1} - x}{h_N}\right)\right] = \\ &= -\int_0^{\mathbf{b}^\top \mathbf{X}_{\iota_1}} \lambda_G(t)\mathrm{d}t \cdot K\left(\frac{X_{\iota_1} - x}{h_N}\right) \leq 0,\end{aligned} \tag{2.13}$$

where the inequality follows easily from the non-decreasing property of function $\psi(\cdot)$ and the positiveness property of the kernel function $K(\cdot)$. Assuming now the case where $\sum_{j=0}^{p} b_j \left(\frac{X_{\iota_2} - x}{h_N}\right)^j < 0$ for some other $\iota_2 \in \{1, \dots, N\}$ we analogously obtain

$$\begin{aligned}\mathbb{E}h(\varepsilon_{\iota_2}, X_{\iota_2}, \mathbf{b}) &= \mathbb{E}\left[\int_0^{|\mathbf{b}^\top \mathbf{X}_{\iota_2}|} [\psi(\varepsilon_{\iota_2} + t) - \psi(\varepsilon_{\iota_2})]\,\mathrm{d}t \cdot K\left(\frac{X_{\iota_2} - x}{h_N}\right)\right] = \\ &= -\int_0^{|\mathbf{b}^\top \mathbf{X}_{\iota_2}|} \lambda_G(-t)\mathrm{d}t \cdot K\left(\frac{X_{\iota_2} - x}{h_N}\right) \geq 0,\end{aligned} \tag{2.14}$$

using again the non-decreasing property of function $\psi(\cdot)$ and the positiveness property of $K(\cdot)$.

Let us now assume that $\left|\sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right| \leq \daleth_N$, for some $\daleth_N > 0$ to be small enough. Then using the Taylor expansion and the fact that $\lambda_G(0) = 0$ we easily have for (2.13) that

$$0 \leq \int_0^{|\mathbf{b}^\top \mathbf{X}_{\iota_1}|} \lambda_G(t)\mathrm{d}t K\left(\frac{X_{\iota_1} - x}{h_N}\right) \leq \int_0^{\daleth_N} \lambda_G(t)\mathrm{d}t K\left(\frac{X_{\iota_1} - x}{h_N}\right) \leq \frac{\daleth_N^2 \lambda_G'(0)}{2} \cdot K\left(\frac{X_{\iota_1} - x}{h_N}\right) + o(1),$$

and analogously, we obtain the same inequalities also for (2.14). Therefore, by summing over all indexes $i \in \{1, \dots, N\}$ we easily obtain that $0 \leq \frac{1}{Nh_N} \sum_{i=1}^{N} \mathbb{E}h(\varepsilon_i, X_i, \mathbf{b}) \leq \mathcal{K}_0 \cdot \lambda_G'(0) \cdot \daleth_N^2$, for some $\mathcal{K}_0 > 0$. On the other hand, using the same arguments for $\left|\sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right| > \daleth_N$ we similarly have that $0 \leq \mathcal{K}_1 \cdot \lambda_G'(0) \cdot \daleth_N^2 < \frac{1}{Nh_N} \sum_{i=1}^{N} \mathbb{E}h(\varepsilon_i, X_i, \mathbf{b})$ for some other $\mathcal{K}_1 > 0$.

The rest already follows from assumption A5 for $\daleth_N \to 0$ as $N \to \infty$ as we have that $\lambda_G'(0) \neq 0$ and $|\lambda_G'(0)| < \infty$ therefore, we need that $|\widehat{\beta}_j^\circ| = O_{\mathbf{P}}(\daleth_N)$ for $\daleth_N \to 0$ as $N \to \infty$, which we can equivalently rewrite as $|\widehat{\beta}_j^\circ| = o_{\mathbf{P}}(1)$ for any $j = 0, \dots, p$.

In the second step, we need to show that $\sqrt{Nh_N}\, \widehat{\beta}_j^\circ = O_{\mathbf{P}}(1)$, for all $j = 0, \dots, p$, which can be thought of as an equivalent for $\sqrt{N}$-consistency property common for classical parameter estimation techniques. We will need an auxiliary lemma to finish the proof of the consistency result.

Lemma 1

For model (2.3) and assumptions A1 – A7 the following bound in probability is achieved

$$\begin{aligned}\sup_{\substack{|t_j| < T \\ j = 0, \dots, p}} \frac{1}{\sqrt{Nh_N}} \left|\sum_{i=1}^{N} \left\{\psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) - \mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right]\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| = \\ = O_{\mathbf{P}}\left((Nh_N)^{-\frac{p+1}{p+3}} \cdot \delta_N^{\frac{2(p+1)}{p+3}}\right),\end{aligned}$$

for any $l \in \{0,\ldots,p\}$ and $T > 0$. Moreover, the expectation operator $\mathbb{E}[\cdot]$ stands here for a conditional expectation conditioned on values of X and $1/\sqrt{Nh_N} \leq \delta_N \leq 1$ is chosen arbitrarily.

Proof of Lemma 1

We will use a similar idea of the proof as was used in Hušková and Marušiaková (2009) however, we will have to extend the proof into multiple dimensions. Let us define a regular $(p+1)$-dimensional grid of points in a $(p+1)$-dimensional cube $(-T,T)\times\cdots\times(-T,T)$ as $-T = \zeta_{0j} < \zeta_{1j} < \cdots < \zeta_{D_N j} < \zeta_{(D_N+1)j} = T$, for some $D_N \in \mathbb{N}$ and $j = 0,\ldots,p$, such that $\zeta_{mj} - \zeta_{(m-1)j} = \nu_N$ for $m = 1,\ldots,D_N$ and $T - \zeta_{D_N j} < \nu_N$ for some $\nu_N \to 0$, which will be specified later. Let us remind that the expectation operator stands for a conditional expectation conditioned of values of X. Using now the non-decreasing property of $\psi(\cdot)$ we have

$$\begin{aligned}
\sup_{\substack{|t_j|<T\\ j=0,\ldots,p}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) \\
-\mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right]\Bigg\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\Bigg| \leq \\
\leq \max_{\substack{1\leq m\leq D_N\\ j=0,\ldots,p}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \psi\left(\varepsilon_i - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) \\
-\mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right]\Bigg\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\Bigg| + \quad \boxed{2.15} \\
+ \max_{\substack{m=1,\ldots,D_N\\ \widetilde{\boldsymbol{\zeta}}_m \in \mathcal{V}_m^{p+1}}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right] \\
-\mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} \zeta_{\widetilde{m}_j j} \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right]\Bigg\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\Bigg|, \quad \boxed{2.16}
\end{aligned}$$

for any $l = 0,\ldots p$, where $\widetilde{\boldsymbol{\zeta}}_m = (\zeta_{\widetilde{m}_0 0},\ldots,\zeta_{\widetilde{m}_p p})^\top \in \mathcal{V}_m^{p+1}$, where $\mathcal{V}_m^{p+1}$ is a set of vectors of indexes such that $\tilde{m}_j \in \{m, m-1\}$ for $j = 0,\ldots,p$, according to the following rule: let $V'(2,p+1)$ be a set of all variations with repetitions of a class $(p+1)$ with two elements $\{0,1\}$. For each variation we have a sequence of the length $p+1$, which consists of zeros and ones only. Now, if there is a zero in this sequence on the j^{th} place for some $j \in \{0,\ldots,p\}$ we put $\widetilde{m}_j = m$ and $\zeta_{\widetilde{m}_j j} = \zeta_{mj}$ and otherwise, we put $\widetilde{m}_j = m-1$ and $\zeta_{\widetilde{m}_j j} = \zeta_{(m-1)j}$.

Let us start with the second term $\boxed{2.16}$, which is an easier one to bounds in probability. Using the definition and the Lipschitz property of function $\lambda_G(\cdot)$ we have that

$$\begin{aligned}
\left| \lambda_G\left(\sum_{j=0}^{p} \zeta_{mj}\delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) - \lambda_G\left(\sum_{j=0}^{p} \zeta_{\widetilde{m}_j j}\delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right| \leq \\
\leq \mathcal{K} \cdot \sum_{j=0}^{p} (\zeta_{mj} - \zeta_{\widetilde{m}_j j}) \cdot \delta_N \cdot \left| \left(\frac{X_i - x}{h_N}\right)^j \right| \leq \\
\leq \mathcal{K} \cdot (p+1) \cdot \nu_N \delta_N,
\end{aligned}$$

uniformly for all $m \in \{1,\dots,D_N\}$ and any $\widetilde{\zeta}_m = (\zeta_{\widetilde{m}_0 0},\dots,\zeta_{\widetilde{m}_p p})^\top \in V_m^{p+1}$, where $\mathcal{K} > 0$ and $\left|\left(\frac{X_i-x}{h_N}\right)\right| \leq 1$ given the kernel approach defined by $K(\cdot)$. Therefore, we have that the second term (2.16) is of the asymptotic order $O_{\mathbf{P}}(\sqrt{Nh_N}\cdot\nu_N\delta_N)$, as $N \to \infty$.

Let us now focus on term (2.15) and we will investigate under what conditions for ν_N it can be bounded in probability. Using Chebyshev's inequality for any $\epsilon > 0$ one easily gets that

$$\begin{aligned}
\mathbf{P}\Bigg[\max_{\substack{1\leq m\leq D_N\\ j=0,\dots,p}} \frac{1}{Nh_N\delta_N\nu_N}\Bigg|\sum_{i=1}^N\Bigg\{\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)& \\
-\mathbb{E}\left[\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| \geq \epsilon\Bigg] &\leq \\
\leq \sum_{\substack{m=1\\ j=0}}^{D_N}\sum_{\substack{m=1\\ j=1}}^{D_N}\cdots\sum_{\substack{m=1\\ j=p}}^{D_N}\mathbf{P}\Bigg[\frac{1}{Nh_N\delta_N\nu_N}\Bigg|\sum_{i=1}^N\Bigg\{\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)& \\
-\mathbb{E}\left[\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| \geq \epsilon\Bigg] &\leq \\
\leq \sum_{\substack{m=1\\ j=0}}^{D_N}\sum_{\substack{m=1\\ j=1}}^{D_N}\cdots\sum_{\substack{m=1\\ j=p}}^{D_N}(\epsilon Nh_N\delta_N\nu_N)^{-2}\cdot\mathbb{E}\Bigg[\sum_{i=1}^N\Bigg\{\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)& \\
-\mathbb{E}\left[\psi\left(\varepsilon_i-\sum_{j=0}^p\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg]^2 &\leq \tag{2.17}
\end{aligned}$$

$$\leq (\epsilon Nh_N\delta_N\nu_N)^{-2}\cdot Nh_N\cdot\nu_N^{-(p+1)}\cdot\mathcal{K}^* = \epsilon^{-2}\cdot\mathcal{K}^*\cdot(Nh_N)^{-1}\delta_N^{-2}\nu_N^{-(p+3)}, \tag{2.18}$$

for some $\mathcal{K}^* > 0$, where we have used the fact that $D_N = O(\nu_N^{-1})$ as $N \to \infty$ and assumption A5, which gives us that the second moment in (2.17) is finite. Therefore, we need that $\nu_N = o\left((Nh_N\delta_N^2)^{-\frac{1}{p+3}}\right)$, which gives us that (2.16) is of the asymptotic order $O_{\mathbf{P}}\left((Nh_N)^{-\frac{p+1}{p+3}}\cdot\delta_N^{\frac{2(p+1)}{p+3}}\right)$ therefore, we need $(\sqrt{Nh_N})^{-1} \leq \delta_N \leq 1$ in order to have terms (2.15) and (2.16) both bounded in probability. This now completes the proof of Lemma 1. Taking the least favourable choice for δ_N we get a special case of the assertion of Lemma 1 where we have that

$$\begin{aligned}
\sup_{\substack{|t_j|<T\\ j=0,\dots,p}}\frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^N\Bigg\{\psi\left(\varepsilon_i-\sum_{j=0}^p t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)& \\
-\mathbb{E}\left[\psi\left(\varepsilon_i-\sum_{j=0}^p t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| &= O_{\mathbf{P}}(1),
\end{aligned}$$

for $\delta_N = \frac{1}{\sqrt{Nh_N}}$. □

To finish the consistency proof we will use a decomposition with respect to two disjoint events and consequently we will apply the special case of Lemma 1. This gives us

$$
\begin{aligned}
\mathbf{P}\Bigg[\frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\Bigg(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)\\
+\lambda_G\Bigg(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|>\epsilon\Bigg]\leq\\
\leq\mathbf{P}\Bigg[\frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\Bigg(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)+\lambda_G\Bigg(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)\Bigg\}\times \quad (2.19)\\
\times\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|>\epsilon\ ,\ \left|\widehat{\beta}_j^{\circ}\right|\leq T,\forall_{j=0,\ldots,p}\Bigg]+\\
+\mathbf{P}\Bigg[\frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\Bigg(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)+\lambda_G\Bigg(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)\Bigg\}\times \quad (2.20)\\
\times\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|>\epsilon\ ,\ \exists_{j\in\{0,\ldots,p\}}\left|\widehat{\beta}_j^{\circ}\right|>T\Bigg],
\end{aligned}
$$

which is satisfied for all $T>0$ and $\epsilon>0$ small enough. The first term 2.19 is bounded in probability by Lemma 1 while the second term 2.20 converges to zero by the fact that $\mathbf{P}\left[\exists_{j\in\{0,\ldots,p\}}\left|\widehat{\beta}_j^{\circ}\right|>T\right]\to 0$, as we already know that $\widehat{\beta}_j^{\circ}=o_{\mathbf{P}}(1)$, for all $j=0,\ldots,p$. Therefore, we can focus on 2.19 only, which gives us

$$
\begin{aligned}
\frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\Bigg(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)+\\
+\lambda_G\Bigg(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|=O_{\mathbf{P}}(1).
\end{aligned}
$$

We can use now the Taylor expansion of the function $\lambda_G(\cdot)$ in zero as we already know that $\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j=o_{\mathbf{P}}(1)$ given the fact that $\left|\frac{X_i-x}{h_N}\right|\leq 1$, as other cases where $\left|\frac{X_i-x}{h_N}\right|>1$ are not considered by the property of the kernel function $K(\cdot)$. Hence, we obtain

$$
\lambda_G\Bigg(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg)=\lambda_G(0)+\lambda_G'(\tilde{t})\cdot\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j,
$$

where $\lambda_G(0)=0$ and $\tilde{t}$ is some value between zero and $\tilde{\tilde{t}}=\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j$. Therefore, we have

$$
\begin{aligned}
\Bigg|\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\Bigg\{\psi\Bigg(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_n}\right)^j\Bigg)+\\
+\sqrt{Nh_N}\lambda_G'(\tilde{t})\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|=O_{\mathbf{P}}(1),
\end{aligned}
$$

which is satisfied for all $l \in \{0,\dots,p\}$. The quantity $\left(\frac{X_i-x}{h_N}\right)^l \cdot K\left(\frac{X_i-x}{h_N}\right)$ is bounded in probability uniformly for all $l=0,\dots,p$ and $i=1,\dots,N$ therefore, we just have to realize that it also holds that

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi\left(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\right)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)=0, \quad \text{for} \quad l=0,\dots,p,$$

which follows from the definition of $\widehat{\beta}_j^{\circ}$ in $\boxed{2.11}$ for all $j=0,\dots,p$.
This gives us that also $\sqrt{Nh_N}\lambda_G'(\tilde{t})\sum_{j=0}^{p}\widehat{\beta}_j^{\circ} = O_{\mathbf{P}}(1)$. Using now assumption A5 we have that $|\lambda_G'(\tilde{t})| < \widetilde{\mathcal{K}}$ for some $\widetilde{\mathcal{K}}>0$, so it follows easily that $\widehat{\beta}_j^{\circ} = O_{\mathbf{P}}(1/\sqrt{Nh_N})$, for any $j=0,\dots,p$. Finally, from the definition of parameters β_j°, for $j=0,\dots,p$ we also obtain the consistency result $\sqrt{Nh_N^{1+2j}}\;\widehat{\beta}_j = O_{\mathbf{P}}(1)$, for $j=0,\dots,p$, which finishes the proof of Theorem 2.1. ■

Beside the consistency result we are also interested in additional asymptotic properties of the given estimate. However, as the parameter estimates $\widehat{\beta}_0,\dots,\widehat{\beta}_p$ are not defined in any explicit form, classical statistical techniques used for an inference will fail in this case therefore, an alternative approach needs to be found. We will now focus on the asymptotic bias and variance terms.

Proof of Theorem 2.2

Let us start with an auxiliary lemma, which is important for the proof to proceed.

Lemma 2
For model $\boxed{2.3}$ and assumptions A1 – A7 the following convergence in probability is achieved

$$\sup_{\substack{|t_j|<T\\ j=0,\dots,p}}\frac{1}{\sqrt{Nh_N}}\left|\sum_{i=1}^{N}\left\{\psi\left(\varepsilon_i-\sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)-\psi(\varepsilon_i)+\right.\right.$$
$$\left.\left.-\mathbb{E}\psi\left(\varepsilon_i-\sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right|\xrightarrow[N\to\infty]{\mathbf{P}}0,$$

uniformly for all $l=0,\dots,p$ and any $T>0$, where $\delta_N=(Nh_N)^{-1/2}$.

Proof of Lemma 2

The idea of the proof is the same as in the case of Lemma 1.
Again, we define a regular $(p+1)$-dimensional grid of points in a $(p+1)$-dimensional cube $(-T,T)\times\cdots\times(-T,T)$ such that $-T=\zeta_{0j}<\zeta_{1j}<\cdots<\zeta_{D_Nj}<\zeta_{(D_N+1)j}=T$, for $j=0,\dots,p$, where $\zeta_{mj}-\zeta_{(m-1)j}=\nu_N$ for $m=1,\dots,D_N$ and $T-\zeta_{D_Nj}<\nu_N$, for some $\nu_N\to 0$, as $N\to\infty$.

For sake of simplicity let us use the notation

$$\Xi_N^{\psi}(X_i,x,\boldsymbol{t}) \overset{def.}{=} \left[\psi\left(\varepsilon_i-\sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)-\psi(\varepsilon_i)\right], \qquad \boxed{2.21}$$

where $\boldsymbol{t} = (t_0, \dots, t_p)^\top \in \mathbb{R}^{p+1}$ is an arbitrary vector. Using now the non-decreasing property of function $\psi(\cdot)$ we can use a similar decomposition as we did for the proof of Lemma 1 hence, we obtain

$$\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \left| \sum_{i=1}^{N} \left\{ \psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) - \psi\left(\varepsilon_i\right) + \right.\right.$$

$$\left.\left. -\mathbb{E}\psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| \leq$$

$$\leq \max_{\substack{1\leq m\leq D_N\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \left| \sum_{i=1}^{N} \left\{ \Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m) - \mathbb{E}\left[\Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m)\right]\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| \tag{2.22}$$

$$+ \max_{\substack{m=1,\dots,D_N\\ \widetilde{\boldsymbol{\zeta}}_m \in \mathcal{V}_m^{p+1}}} \frac{1}{\sqrt{Nh_N}} \left| \sum_{i=1}^{N} \mathbb{E}\left[\Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m) - \Xi_N^{\psi}(X_i, x, \widetilde{\boldsymbol{\zeta}}_m)\right] \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right|, \tag{2.23}$$

for all $l = 0, \dots p$, where again we have $\boldsymbol{\zeta}_m = (\zeta_{m0}, \dots, \zeta_{mp})^\top$ and $\widetilde{\boldsymbol{\zeta}}_m = (\zeta_{\widetilde{m}_0 0}, \dots, \zeta_{\widetilde{m}_p p})^\top \in \mathcal{V}_m^{p+1}$, for the set of indexes $\mathcal{V}_m^{p+1}$ to be constructed in the same way as in the proof of Lemma 1.

The second term (2.23) can be shown to converge to zero in probability in the very same way as the second term in the proof of Lemma 1 (see p. 24), as we have that

$$\mathbb{E}\left[\Xi_N^{\psi}(X_i, x, \mathbf{t})\right] = \mathbb{E}\left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) - \psi(\varepsilon_i)\right] = -\lambda_G\left(\sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right),$$

for any $\mathbf{t} = (t_0, \dots, t_p)^\top \in \mathbb{R}^{p+1}$, given the fact that $\mathbb{E}\psi(\varepsilon_i) = 0$ from the symmetric assumptions A2 and A4. Therefore, we can point our attention to prove the corresponding convergence in probability for (2.22) only.

Using again Chebyshev's inequality we have that

$$\mathbf{P}\left[\max_{\substack{1\leq m\leq D_N\\ j=0,\dots,p}} \frac{1}{Nh_N\delta_N\nu_N} \left| \sum_{i=1}^{N} \left\{ \Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m) + \right.\right.\right.$$

$$\left.\left.\left. -\mathbb{E}\left[\Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m)\right]\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| \geq \epsilon\right] \leq$$

$$\leq \sum_{\substack{m=1\\ j=0}}^{D_N} \sum_{\substack{m=1\\ j=1}}^{D_N} \cdots \sum_{\substack{m=1\\ j=p}}^{D_N} \mathbf{P}\left[\frac{1}{Nh_N\delta_N\nu_N} \left| \sum_{i=1}^{N} \left\{ \Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m) + \right.\right.\right.$$

$$\left.\left.\left. -\mathbb{E}\left[\Xi_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m)\right]\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| \geq \epsilon\right] \leq$$

$$\leq \sum_{\substack{m=1\\j=0}}^{D_N} \sum_{\substack{m=1\\j=1}}^{D_N} \cdots \sum_{\substack{m=1\\j=p}}^{D_N} (\epsilon N h_N \delta_N \nu_N)^{-2} \times$$

$$\times \mathbb{V}\mathrm{ar} \left[\sum_{i=1}^{N} \left\{ \psi\left(\varepsilon_i - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N} \right)^j \right) - \psi(\varepsilon_i) \right\} \cdot \left(\frac{X_i - x}{h_N} \right)^l K\left(\frac{X_i - x}{h_N} \right) \right] \leq$$

$$\leq (\epsilon N h_N \delta_N \nu_N)^{-2} \cdot \nu_N^{-(p+1)} \cdot N h_N \cdot \mathbb{V}\mathrm{ar} \left[\psi\left(\varepsilon_1 - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N} \right)^j \right) - \psi(\varepsilon_1) \right] \tag{2.24}$$

for $\boldsymbol{\zeta}_m = (\zeta_{m0}, \ldots, \zeta_{mp})^\top$ and any $\epsilon > 0$, where we have again used the fact that $D_N = o(\nu_N^{-1})$ and $\left| \left(\frac{X_i - x}{h_N} \right)^l K\left(\frac{X_i - x}{h_N} \right) \right| = O_{\mathbf{P}}(1)$ uniformly for all $l = 0, \ldots, p$.

Using now the assumption A5 we easily obtain for the variance term in (2.24) that

$$\mathbb{V}\mathrm{ar} \left[\psi\left(\varepsilon_1 - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N} \right)^j \right) - \psi(\varepsilon_1) \right] \leq$$

$$\leq \mathbb{E} \left[\psi\left(\varepsilon_1 - \sum_{j=0}^{p} \zeta_{mj} \delta_N \left(\frac{X_i - x}{h_N} \right)^j \right) - \psi(\varepsilon_1) \right]^2 = O\left(\frac{1}{\sqrt{N h_N}} \right), \tag{2.25}$$

therefore, we have that (2.24) is of the asymptotic order $O_{\mathbf{P}}\left((N h_N \delta_N \nu_N)^{-2} \cdot \nu_N^{-(p+3)} \cdot \sqrt{N h_N} \right)$, which we need to tend to zero as $N \to \infty$. Just a straightforward calculation gives us that we necessarily need $\nu_N = o\left((N h_N \delta_N^2)^{-1/(p+3)} \cdot (\sqrt{N h_N})^{-1/(p+3)} \right)$, which will imply that (2.23) is of the order $O_{\mathbf{P}}\left((\sqrt{N h_N} \delta_N)^{(p+1)/(p+3)} \cdot (\sqrt{N h_N})^{-1/(p+3)} \right)$, which we need to tend to zero as well in order to have the assertion of Lemma 2 to be proved.

However, one can easily see that the choice $\delta_N = 1/\sqrt{N h_N}$ satisfies this requirement and both terms (2.22) and (2.23) now converge to zero in probability as $N \to \infty$, which was supposed to be proved. □

Using now the assertion of Lemma 2 we can proceed in a very analogous way as in the case of consistency proof (see p. 25) and we obtain

$$\frac{1}{\sqrt{N h_N}} \left| \sum_{i=1}^{N} \left[\psi\left(\varepsilon_i - \sum_{j=0}^{p} \widehat{\beta}_j^\circ \left(\frac{X_i - x}{h_N} \right)^j \right) - \psi(\varepsilon_i) + \right. \right.$$

$$\left. \left. + \lambda_G \left(\sum_{j=0}^{p} \widehat{\beta}_j^\circ \left(\frac{X_i - x}{h_N} \right)^j \right) \right] \cdot \left(\frac{X_i - x}{h_N} \right)^l K\left(\frac{X_i - x}{h_N} \right) \right| \xrightarrow[N \to \infty]{\mathbf{P}} 0,$$

for all $l = 0, \ldots, p$ and by applying the Taylor expansion of $\lambda_G(\cdot)$ in zero we get

$$\lambda_G \left(\sum_{j=0}^{p} \widehat{\beta}_j^\circ \left(\frac{X_i - x}{h_N} \right)^j \right) = \lambda_G(0) + \lambda_G'(0) \cdot \sum_{j=0}^{p} \widehat{\beta}_j^\circ \left(\frac{X_i - x}{h_N} \right)^j + o_{\mathbf{P}}\left(\frac{1}{\sqrt{N h_N}} \right),$$

where again $\lambda_G(0) = 0$ and also given the fact that

$$\left| \sum_{j=0}^{p} \widehat{\beta}_j^\circ \left(\frac{X_i - x}{h_N} \right)^j \right| \leq \sum_{j=0}^{p} \left| \widehat{\beta}_j^\circ \right| \cdot \left| \left(\frac{X_i - x}{h_N} \right)^j \right| \leq \sum_{j=0}^{p} \left| \widehat{\beta}_j^\circ \right| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{N h_N}} \right).$$

Hence, we obtain

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\left[\psi\left(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\right)-\psi\left(\varepsilon_i\right)+\lambda_G'(0)\cdot\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j+\right.$$
$$\left.+o_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right)\right]\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\xrightarrow[N\to\infty]{\mathbf{P}}0,$$

which can be easily rewritten in an asymptotically equivalent way as

$$\begin{aligned}\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi(\varepsilon_i)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right) &= \\ = \frac{\lambda_G'(0)}{\sqrt{Nh_N}}\cdot\sum_{i=1}^{N}\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)+o_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right),\end{aligned} \tag{2.26}$$

for any $l=0,\ldots,p$, given the fact that

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi\left(\varepsilon_i-\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\right)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)=0.$$

For simplicity we can rewrite (2.26) using the matrix notation to obtain

$$\frac{1}{\sqrt{Nh_N}}\cdot\mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\boldsymbol{\varepsilon}) \quad = \quad \frac{\lambda_G'(0)}{\sqrt{Nh_N}}\cdot\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\widehat{\boldsymbol{\beta}}_x^{\circ}+o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right), \tag{2.27}$$

where $\widehat{\boldsymbol{\beta}}_x^{\circ}=(\widehat{\beta}_0^{\circ},\ldots,\widehat{\beta}_p^{\circ})^{\top}$ is the vector of the parameter estimates, matrices X_N and W_N are those defined at the beginning of this section and $\boldsymbol{\psi}(\boldsymbol{\varepsilon})=(\psi(\varepsilon_1),\ldots,\psi(\varepsilon_N))^{\top}\in\mathbb{R}^N$. Once the matrix $\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N$ is regular[19] we can use an inverse matrix to get the expression

$$\widehat{\boldsymbol{\beta}}_x^{\circ} \quad = \quad \frac{1}{\lambda_G'(0)}\cdot\left(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\right)^{-1}\cdot\mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\boldsymbol{\varepsilon})+o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right), \tag{2.28}$$

which is in literature referred to as an asymptotic representation of the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}$ or Bahadur representation respectively, introduced in Bahadur (1966). For some more discussion on Bahadur representations we refer to Bose (1998) or He and Shao (1996).

To finish the proof we will firstly focus on expression (2.5), which is an easier one to prove. Taking an advantage of the given notation for $\widehat{\beta}_j^{\circ}$ for $j=0,\ldots,p$ we can express only an asymptotic conditional variance for $\widehat{\beta}_j^{\circ}$ and the conditional variance for $\widehat{\beta}_j$ for any $j=0,\ldots,p$ will follow immediately. Indeed, we have that $\widehat{\beta}_j^{\circ}-\widehat{\beta}_j h_N^j=O_{\mathbf{P}}(h_N^j)$. From (2.28), just a straightforward calculation gives us

$$\mathbb{A}\mathrm{s}.\mathbb{V}\mathrm{ar}\left[\widehat{\beta}^{\circ}\mid\mathbf{X}\right]=\frac{\mathbb{E}[\psi^2(\varepsilon_1)]}{[\lambda_G'(0)]^2}\cdot(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N)^{-1}\mathsf{X}_N^{\top}\mathsf{W}_N^2\mathsf{X}_N\cdot(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N)^{-1}+o_{\mathbf{P}}\left(\frac{1}{Nh_N}\right), \tag{2.29}$$

given the fact that random errors ε_i, for $i=1,\ldots,N$ are independent and identically distributed and moreover, from the symmetric property of the distribution function $G(\cdot)$ and the loss function $\rho(\cdot)$ it holds that $\mathbb{E}\psi(\varepsilon)=\int_{\mathbb{R}}\psi(\varepsilon_1)\mathrm{d}G(\varepsilon_1)=0$.

[19] The regularity property of matrix $\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N$ is mostly really achieved under the assumptions A1–A7.

Next, we will prove (2.4). However, one has to be aware of the fact here that the classical bias definition in sense of an expectation of the difference between the estimate and the true value of the parameter is not helpful here as a conditional expectation of the asymptotic representation (2.28) is equal to zero given the fact that $\mathbb{E}\psi(\varepsilon_1) = 0$. In order to express an imprecision in parameter estimates we will rather approach the bias term only in sense of a difference $\widehat{\boldsymbol{\beta}}_x^{\circ} - \boldsymbol{\beta}_x^{\circ}$. This difference is however, not obtained in an explicit form either and a similar asymptotic representation needs to be found.

Rather than working with the vector of parameter estimates itself we will again use function $\psi(\cdot)$ and the conditional expectation and we will compare the original approach assumed by (2.2) with the one we have actually worked with, given by (2.11). Hence, we have two sets of equations

$$\sum_{i=1}^{N} \psi\left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j\right) \cdot (X_i - x)^l \cdot K\left(\frac{X_i - x}{h_N}\right) = 0, \tag{2.30}$$

and

$$\sum_{i=1}^{N} \psi\left(\varepsilon_i - \sum_{j=0}^{p} \left(b_j - \frac{m^{(j)}(x)}{j!}\right)(X_i - x)^j\right) \cdot (X_i - x)^l \cdot K\left(\frac{X_i - x}{h_N}\right) = 0, \tag{2.31}$$

both for $l = 0, \ldots, p$ where the given vector of estimates $\widehat{\boldsymbol{\beta}}_x$ is supposed to solve the second one.

Considering now (2.30), the model definition (2.3) and the $(p+1)$-order Lipschitz property of function $m(\cdot)$ we easily obtain a set of equations

$$\sum_{i=1}^{N} \psi\left(\varepsilon_i - \sum_{j=0}^{p} b_j^{\circ} \left(\frac{X_i - x}{h_N}\right)^j + \frac{m^{(p+1)}(x)}{(p+1)!}(X_i - x)^{p+1} + o(h^{p+1})\right)(X_i - x)^l K\left(\frac{X_i - x}{h_N}\right) = 0,$$

for $l = 0, \ldots, p$, where $b_j^{\circ} = \left(b_j - \frac{m^{(j)}(x)}{j!}\right) h_N^j$. Term $\frac{m^{(p+1)}(x)}{(p+1)!}(X_i - x)^{p+1}$ can be thought of as a leading term of the bias imprecision involved by re-defining the original problem.

Using now the conditional expectation conditioned on values of the random variable X, Lemma 2 and the matrix notation already defined we easily obtain two asymptotic representations

$$\begin{aligned}
\frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N \boldsymbol{\psi}(\boldsymbol{\varepsilon}) &= \frac{\lambda_G'(0)}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\widehat{\boldsymbol{\beta}}_x^{\circ} + o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right), \\
\frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N \boldsymbol{\psi}(\boldsymbol{\varepsilon}) &= \frac{\lambda_G'(0)}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N \left(\mathsf{X}_N\widetilde{\boldsymbol{\beta}}_x^{\circ} + \frac{m^{(p+1)}(x)}{(p+1)!}(\mathbf{X} - x)\right) + o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right),
\end{aligned}$$

for $\mathbf{X} = (X_1, \ldots, X_N)^{\top}$ where $\widehat{\boldsymbol{\beta}}_x^{\circ}$ solves (2.31) and $\widetilde{\boldsymbol{\beta}}_x^{\circ}$ would be a solution to the original problem (2.30). Comparing these two quantities one obtains the asymptotic conditional bias expression in sense of the difference as

$$\widehat{\boldsymbol{\beta}}_x^{\circ} - \boldsymbol{\beta}_x^{\circ} = \left(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\right)^{-1} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!}(\mathbf{X} - x)\right) + o_{\mathbf{P}}\left(h_N^{p+1}\right). \tag{2.32}$$

It follows immediately from our notation that the difference $\widehat{\boldsymbol{\beta}}_x^{\circ} - \boldsymbol{\beta}_x^{\circ}$ is equivalent to $\mathsf{H}_N \cdot \left(\widehat{\boldsymbol{\beta}}_x - \boldsymbol{\beta}_x\right)$, which is referred to as a conditional bias term in (2.4). This completes the proof of Theorem 2.2. ■

Proof of Theorem 2.3

In order to proof Theorem 2.3 we will start with equation (2.32) and we will derive an asymptotic expression for the bias term where again we will refer to a difference between the parameter estimate and the true value rather than referring to a classical bias definition in sense of an expectation[20].

Let $s_{k,N} = \sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) \cdot \left(\frac{X_i - x}{h_N}\right)^k$. Then we have $\mathsf{S}_N = (\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N)$, where $\mathsf{S}_N = (S_{ij})_{i,j=0}^{p}$, and $S_{ij} = s_{i+j,N}$. Moreover, let function $F(\cdot)$ be the marginal distribution function of the random variables $X_1, \dots, X_N$, which corresponds to the density function $f(\cdot)$ from assumption A1. Then it is easy to see that the following holds:

$$
\begin{aligned}
s_{k,N} &= \mathbb{E}s_{k,N} + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\mathrm{ar}(s_{k,N})}\right) \qquad \text{See Fan and Gijbels (1996)}\\
&= N \cdot \int_{x-h_N}^{x+h_N} K\left(\frac{u-x}{h_N}\right) \cdot \left(\frac{u-x}{h_N}\right)^k \mathrm{d}F(u) + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\mathrm{ar}(s_{k,N})}\right)\\
&= Nh_N \cdot \int_{-1}^{1} K(y) y^k f(x + h_N y) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{Nh_N}\right)\\
&= Nh_N \cdot \left[\int_{-1}^{1} K(y) y^k (f(x) + o(1)) \mathrm{d}y + O_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right)\right]\\
&= Nh_N f(x) \cdot \int_{-1}^{1} K(y) y^k \mathrm{d}y + o_{\mathbf{P}}(Nh_N). \quad \text{for } k = 0, 1, \dots, 2p.
\end{aligned} \tag{2.33}
$$

In a quite analogous way let $s_{k,N}^{\natural} = \sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) \cdot \left(\frac{X_i - x}{h_N}\right)^k \left(\frac{m^{(p+1)}(x)}{(p+1)!}(X_i - x)^{p+1}\right)$. Then it is easy to see that we have $\mathbf{s}_N^{\natural} = \mathsf{X}_N^\top \mathsf{W}_N \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!}(\mathbf{X} - x)^{p+1}\right)$, for $\mathbf{s}_N^{\natural} = (s_{0,N}^{\natural}, \dots, s_{p,N}^{\natural})^\top$ and

$$
\begin{aligned}
s_{k,N}^{\natural} &= \mathbb{E}s_{k,N}^{\natural} + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\mathrm{ar}(s_{k,N}^{\natural})}\right) \qquad \text{See Fan and Gijbels (1996)}\\
&= N \cdot \int_{x-h_N}^{x+h_N} K\left(\frac{u-x}{h_N}\right) \cdot \left(\frac{u-x}{h_N}\right)^k \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!}(u-x)^{p+1}\right) \mathrm{d}F(u) + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\mathrm{ar}(s_{k,N}^{\natural})}\right)\\
&= Nh_N \cdot \int_{-1}^{1} K(y) y^k f(x + h_N y) \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!} h_N^{p+1} y^{p+1}\right) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\mathrm{ar}(s_{k,N}^{\natural})}\right)\\
&= Nh_N \cdot \int_{-1}^{1} K(y) y^{p+1+k} f(x + h_N y) \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!} h_N^{p+1}\right) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{Nh_N^{2p+3}}\right).
\end{aligned}
$$

Finally, once we realize that $f(x + h_N y) = f(x) + o(1)$ uniformly for $x \in (0,1)$ we will get expression

$$
s_{k,N}^{\natural} = Nh_N f(x) \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!} h_N^{p+1}\right) \cdot \int_{-1}^{1} K(y) y^{p+1+k} \mathrm{d}y + o_{\mathbf{P}}\left(Nh_N^{p+1}\right), \tag{2.34}
$$

[20] The reasoning behind the alternative bias definition remains the same as before as the expectation of the asymptotic representation used to approximate the vector of parameter estimates is zero by the fact that $\mathbb{E}\psi(\varepsilon_1) = 0$.

for $k = 0, 1, \dots, p$. Now, one needs to put together expressions (2.33) and (2.34), which gives us an asymptotic expression

$$\widehat{\beta}_j^{\circ} - \beta_j^{\circ} = \left(\frac{m^{(p+1)}(x)}{(p+1)!} h_N^{p+1} \right) \cdot \mathbf{e}_{\nu+1}^{\top} \cdot \mathsf{S}_1^{-1} \cdot \mu + o_{\mathbf{P}}\left(h_N^{p+1} \right), \tag{2.35}$$

for $j = 0, \dots, p$ where $\mu = (\int u^{p+1} K(u) \mathrm{d}u, \int u^{p+2} K(u) \mathrm{d}u, \dots, \int u^{2p+2} K(u) \mathrm{d}u)^{\top}$. Together with the definition of the vector of parameters estimates $\widehat{\boldsymbol{\beta}}_x^{\circ} = (\widehat{\beta}_0^{\circ}, \dots, \widehat{\beta}_p^{\circ})^{\top} \in \mathbb{R}^{p+1}$ we will get the bias expression as stated in Theorem (2.3).

■

Proof of Theorem 2.4

Using a sequence of similar steps we can also prove the assertion of Theorem 2.4. We will start with equation (2.29) that is

$$\mathbb{A}\text{s.}\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}_x^{\circ} \middle| \mathbf{X} \right] = \frac{\mathbb{E}\psi^2(\epsilon)}{[\lambda_G'(0)]^2} \cdot \left(\mathsf{X}_N^{\top} \mathsf{W}_N \mathsf{X}_N \right)^{-1} \cdot \mathsf{X}_N^{\top} \mathsf{W}_N^2 \mathsf{X}_N \cdot \left(\mathsf{X}_N^{\top} \mathsf{W}_N \mathsf{X}_N \right)^{-1} + o_{\mathbf{P}}\left(\frac{1}{N h_N} \right),$$

and we will derive the asymptotic variance term for $\widehat{\boldsymbol{\beta}}_x^{\circ}$. From there on just a straightforward reformulation is required to show that the expression stated in Theorem 2.4 holds true as well.

From the proof of Theorem 2.3 we already have the asymptotic expression for $\mathsf{S}_N = \mathsf{X}_N^{\top} \mathsf{W}_N \mathsf{X}_N$, where we have showed an asymptotic relationship between S_N and S_1 for $N \to \infty$. The only thing that is still left to derive is an asymptotic expression for the term $\widetilde{\mathsf{S}}_N = (\mathsf{X}_N^{\top} \mathsf{W}_N^2 \mathsf{X}_N)$. We will show there is a similar asymptotic relationship between matrices $\widetilde{\mathsf{S}}_N$ and S_2, for $N \to \infty$ as there is in the case of matrices S_N and S_1.

In an analogous way let $\widetilde{s}_{k,N} = \sum_{i=1}^{N} K^2(\frac{X_i - x}{h_N}) \cdot \left(\frac{X_i - x}{h_N} \right)^k$. Then it is quite obvious that the relation $\widetilde{\mathsf{S}}_N = (\mathsf{X}_N^{\top} \mathsf{W}_N^2 \mathsf{X}_N)$ holds for $\widetilde{\mathsf{S}}_N = (\widetilde{S}_{ij})_{i,j=0}^{p}$, where $\widetilde{S}_{ij} = \widetilde{s}_{i+j,N}$.
Hence, a just straightforward computation gives us

$$\begin{aligned}
\widetilde{s}_{k,N} &= \mathbb{E}\widetilde{s}_{k,N} + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\text{ar}(\widetilde{s}_{k,N})} \right) \qquad\qquad \text{See Fan and Gijbels (1996)} \\
&= N \cdot \int_{x-h_N}^{x+h_N} K^2\left(\frac{u-x}{h_N} \right) \cdot \left(\frac{u-x}{h_N} \right)^k \mathrm{d}F(u) + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\text{ar}(\widetilde{s}_{k,N})} \right) \\
&= N h_N \cdot \int_{-1}^{1} K^2(y) y^k f(x + h_N y) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{\mathbb{V}\text{ar}(\widetilde{s}_{k,N})} \right) \\
&= N h_N \cdot \int_{-1}^{1} K^2(y) y^k f(x + h_N y) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{N h_N} \right), \\
&= N h_N \cdot \int_{-1}^{1} K^2(y) y^k (f(x) + o(1)) \mathrm{d}y + O_{\mathbf{P}}\left(\sqrt{N h_N} \right), \\
&= N h_N f(x) \cdot \int_{-1}^{1} K^2(y) y^k \mathrm{d}y + o_{\mathbf{P}}\left(N h_N \right). \quad \text{for } k = 0, 1, \dots, 2p.
\end{aligned} \tag{2.36}$$

Now we need to get together (2.33), (2.36) and (2.29), which gives us the asymptotic variance expression

$$\mathbb{A}\text{s.}\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}_x^{\circ} \right] = \frac{\mathbb{E}\psi^2(\varepsilon_1)}{[\lambda_G'(0)]^2 f(x) N h_N} \cdot \mathsf{S}_1^{-1} \mathsf{S}_2 \mathsf{S}_1^{-1} + o_{\mathbf{P}}\left(\frac{1}{N h_N} \right). \tag{2.37}$$

It is already easy to see now, that once we use the given notation for the vector of parameters estimates $\widehat{\boldsymbol{\beta}}^{\circ}_{x}$, where $\widehat{\boldsymbol{\beta}}^{\circ}_{x} = \left((\widehat{\beta}^{\circ}_{0} - m(x)), (\widehat{\beta}^{\circ}_{1} - m'(x))h_N, \dots, \left(\widehat{\beta}^{\circ}_{p} - \frac{m^{(p)}(x)}{p!} \right) h_N^{p} \right)^{\top}$, vector $\mathbf{e}_{\nu} = (\underbrace{0, \dots, 0}_{\nu-times}, 1, 0, \dots, 0)^{\top} \in \mathbb{R}^{p+1}$ and the diagonal matrix $\mathsf{H}_N = \operatorname{diag}\{1, h_N^{-1}, \dots, h_N^{-p}\}$, then the asymptotic variance expression as stated in Theorem 2.4 follows immediately.
This completes the proof of Theorem 2.4. ■

Proof of Theorem 2.5

Finally, to show the asymptotic normality result from Theorem 2.5 we will use expression 2.27 and we will firstly focus on the left-hand side of this expression. Using the mutual independence assumption A2 we have that quantities $\psi(\varepsilon_i) \cdot \left(\frac{X_i - x}{h_N}\right) K \left(\frac{X_i - x}{h_N}\right)$ for $i = 1, \dots, N$ satisfy all necessary assumptions[21] for the Central Limit Theorem (CLT) for triangular arrays to hold therefore, we easily have that

$$\frac{1}{\sqrt{N h_N}} \sum_{i=1}^{N} \psi(\varepsilon_i) \cdot \left(\frac{X_i - x}{h_N} \right)^{l} K \left(\frac{X_i - x}{h_N} \right) \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}\left(0, \sigma^2(\psi, l)\right), \tag{2.38}$$

for any $l = 0, \dots, p$, where $\sigma^2(\psi, l)$ is a corresponding variance, which depends on the choice of loss function $\rho(\cdot)$ and $l \in \{0, \dots, p\}$ and $x \in (0,1)$. From the regularity property of matrix $\mathsf{X}_N^{\top} \mathsf{W}_N \mathsf{X}_N$ and a common property of the normal distribution we obtain that the first term on the right-hand side of 2.28 converges in distribution to the normal law with the corresponding mean and variance parameters as well. Finally, we can now apply Slutsky's theorem as we know that the remaining term on the right-hand side of 2.28 converges to zero in probability therefore, we have that the parameter estimates $\widehat{\boldsymbol{\beta}}^{\circ}_{x} = (\widehat{\beta}^{\circ}_{0}, \dots, \widehat{\beta}^{\circ}_{p})^{\top}$ converges element-wise to a normal distribution with a zero mean and the variance parameter given by 2.37.
Moreover, the bias term in sense of a difference is stated in 2.37.

Now, we just need to use the definition of the vector of parameters $\boldsymbol{\beta}^{\circ}_{x} = (\beta^{\circ}_{0}, \dots, \beta^{\circ}_{p})^{\top})$ or the parameter estimates $\widehat{\boldsymbol{\beta}}^{\circ}_{x}$ respectively, and the result from Theorem 2.5 follows immediately. ■

We have now proved all results for the homoscedastic model scenario and we have shown that the M-smoothers estimates, which have a flavour of being robust with respect to outlying observations and heavy-tailed distributions of random errors are consistent and they follow in asymptotic a normal distributional law with a zero mean and a finite variance term.
This will become an important property in Chapter 3 where we will discuss regression models with discontinuity points. Given the knowledge that the M-smoothers estimates are asymptotically normal we can easily construct confidence intervals or critical regions for statistical tests in order to decide about a hypothesis testing problem related to a question if there is a jump in the unknown regression function at some given pre-specified point (or some derivative respectively) or the function is continuous at this point (the derivative is continuous respectively).
However, before we go to regression models with change-points let us firstly discuss some popular generalization of the homoscedastic model scenario. We will prove similar statistical properties for the heteroscedastic model and a model with dependent random errors as well.

[21] These random quantities are assumed to be independent and identically distributed with a zero mean parameter and a finite variance term.

2.3 Heteroscedastic models

The homoscedastic model scenario discussed in the previous section can be considered to be the basic model for nonparametric higher order regression modelling with robust flavour and most of all further generalizations introduced in years were mostly based on this specific model assumption indeed.

There are of course many different ways, which can be taken in order to propose more general model structures: one can try to improve the flexibility with respect to the shape of the unknown regression function or the model assumptions or with respect to the random sample requirements as well. We will focus on two popular generalizations in the next two sections as we will firstly introduce a heteroscedastic model and a model with dependent random errors after that.

Unlike the homoscedastic model scenario considered in Section 2.2 where we were concerned with the unknown regression function $m(\cdot)$ only we will now introduce also a scale function $\sigma(\cdot)$, which will be together with function $m(\cdot)$ in the center of our interest. By adopting a heteroscedastic variance principle one brings more flexibility into the modelling as the local variability of random errors can freely change over the domain of interest. Moreover, this heteroscedasticity is elegantly approached and no extra additional requirements are necessary[22] while we again obtain a set of fully consistent parameters estimates with a proper statistical inference.
In general, we assume the model

$$Y_i = m(X_i) + \sigma(X_i) \cdot \varepsilon_i, \quad \varepsilon_i \sim G, \ \ i.i.d., \ \ i = 1, \dots, N, \tag{2.39}$$

where G stands again for a continuous and symmetric distribution function but moreover, we also assume that $G(1) - G(-1) = \frac{1}{2}$, which is to define a unit scale of random errors, rather than specifying a unit variance as we want to stay free of any finite moments assumptions. Unlike the homoscedastic case this is now necessary to well-define the whole model. The variability of the model is now fully described by the scale function $\sigma(\cdot)$, which is now to be estimated as well. One can easily see, that once we define the scale function $\sigma(\cdot)$ to be equal to a constant over the whole domain of interest we get back to the homoscedastic model scenario, which we have discussed in the previous section.

Considering (2.39) we are again interested in estimation of the unknown regression function $m(\cdot)$ and the scale function $\sigma(\cdot)$ while the estimates are defined at some given pre-specified point $x \in (0,1)$ from the domain of interest or over the whole interval $(0,1)$ respectively.

One can apply at least two different approaches here. Indeed, we can define a vector of parameter estimates in sense of the minimization problem (2.1) as

$$\widehat{\boldsymbol{\beta}}_x = \underset{(b_0,\dots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \sum_{i=1}^{N} \rho\left(Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j\right) \cdot K\left(\frac{X_i - x}{h_N}\right), \tag{2.40}$$

where we get the parameter estimates for the vector of true parameters $\boldsymbol{\beta}_x = (\beta_0, \dots, \beta_p)^\top$ only and the unknown scale $\sigma(x)$ is still left to be estimated using some alternative approaches.

[22] One of course has to add some additional assumptions especially with respect to the unknown scale function $\sigma(\cdot)$ however, by saying that "no extra additional requirements are necessary" we mean that the price for additional assumptions we have to incorporate into the model is balanced with respect to an additional flexibility we yield.

On the other hand, we can also implement function $\sigma(\cdot)$ directly into the minimization problem and to estimate the vector of true parameters $\boldsymbol{\beta}_x \in \mathbb{R}^{p+1}$ while taking into account the scale factor[23] at the point $x \in (0,1)$. Using this type of estimation we obtain a minimization problem defined by

$$\widehat{\boldsymbol{\beta}}_x^{(s)} = \underset{\boldsymbol{b}\,\in\,\mathbb{R}^{p+1}}{Argmin} \sum_{i=1}^{N} \rho\left(\frac{Y_i - \sum_{j=0}^{p} b_j (X_i - x)^j}{\widehat{\sigma}_N(X_i)}\right) \cdot K\left(\frac{X_i - x}{h_N}\right), \tag{2.41}$$

where $\boldsymbol{b} = (b_0, \ldots, b_p)^\top$ and $\widehat{\sigma}_N(\cdot)$ is an appropriate scale function estimate given in advance, such that $\sup_{x\in(0,1)} |\widehat{\sigma}_N(x) - \sigma(x)| = O_{\mathbf{P}}(1/\sqrt{Nh_N})$ for $N \to \infty$.

Such M-smoothers are called studentized M-smoothers (see Jurečková (2001) for some discussion on studentized M-smoothers however, for a parametric regression only) as they are capable to provide estimates for the location parameter while implicitly assume the scale. This approach however, requires slightly more technical proofs and additional assumptions therefore, we will rather focus on the first option where we firstly estimate the vector of parameter estimates $\boldsymbol{\beta}_x$ only and later on we will construct an estimate for the scale function using some additional estimation techniques. This approach is, in our opinion, more straightforward and it is an easier one to be proved.

Before we discuss in detail the main statistical properties and inference for the heteroscedastic model (2.39) let us briefly mention all additional assumptions we need for the results to hold. We will again consider assumptions A1 – A7 from Section 2.2.1, while we introduce the following modifications:

A1* The marginal density function $f(\cdot)$ of the *i.i.d.* random variables X_i, for $i = 1, \ldots, N$ is absolutely continuous, positive and bounded on interval $[0,1]$, which is the support of X.
Moreover, we assume that the scale function $\sigma(\cdot)$ is Lipschitz and positive on interval $[0,1]$;

A2* Random errors $\varepsilon_1, \ldots, \varepsilon_N$, are assumed to be *i.i.d.*, mutually independent of X_i, for $i = 1, \ldots, N$, with a symmetric distribution given by a continuous distribution function $G(\cdot)$, such that it holds that $G(1) - G(-1) = \frac{1}{2}$;

A5* We assume that function $\lambda_G(t,v) = -\int \psi(ve - t)\mathrm{d}G(e)$ is Hölder of the order $\alpha > \frac{\iota-1}{2\iota}$ in argument $v > 0$ where ι is defined in assumption A7, the partial derivative $\lambda_G'(t,v) = \frac{\partial}{\partial t}\lambda_G(t,v)$ exists and it is continuous in t and $\int_{\mathbb{R}} \left[\psi(ve - \epsilon_N) - \psi(e)\right]^2 \mathrm{d}G(e) < \mathcal{K} \cdot |\epsilon_N|$ both for some neighbourhoods of $t = 0$ and $v = \sigma(x)$, for the point $x \in (0,1)$, any sequence $\epsilon_N \to 0$ and some $\mathcal{K} > 0$. Moreover, $\int \psi^2(\sigma(x)e)\mathrm{d}G(e) < \infty$ and $\lambda_G'(0,\sigma(x)) = \frac{\partial}{\partial t}\, \lambda_G(t,\sigma(x))|_{t=0} \neq 0$, for the given point $x \in (0,1)$;

The modified assumptions are proposed in order to properly incorporate the scale function $\sigma(\cdot)$ into the model. In assumption A1* we have the same conditions posed on the density function $f(\cdot)$ however, additional conditions are required to define all necessary properties for the scale function $\sigma(\cdot)$. A positivity assumption is quite obvious while the Lipschitz property is natural as well[24].

Assumption A2 is enhanced for a requirement on a unit scale of random errors, which is necessary because the scale is now taken over by the scale function $\sigma(\cdot)$ otherwise, the estimation problem would

[23] It is important to realize here that once we are about to estimate the unknown regression function $m(\cdot)$ or the scale function $\sigma(\cdot)$ at some given point $x \in (0,1)$ only then both functions defined over interval $(0,1)$ reduce to single values only therefore $m(x)$ and $\sigma(x)$ are treated by the minimization problem as single parameter values rather than real-valued functions on interval $(0,1)$.

[24] One can also propose a model with a heteroscedastic variance structure while assuming some sudden changes - jumps in the scale function $\sigma(\cdot)$. However, this approach is not the subject of this thesis therefore, we will omit discussions on such scenarios here.

be not well-defined. Finally, assumption A5 has to be also modified in order to take into account the scale function and the fact that it can flexibly change now for different values of $x \in (0,1)$.

Once we are interested in estimation of the unknown regression function $m(\cdot)$ or its derivatives respectively in a local sense only – at one pre-specified point $x \in (0,1)$, it is sufficient to consider assumptions A1**, A3 and A5** to hold at some small neighbourhood of $x \in (0,1)$ only.

2.3.1 The main asymptotic results

Let us firstly recall the notation, which we have already established for the homoscedastic model (2.3) in Section 2.2.2. Basically, we will show the same results however, unlike the homoscedastic model we will now omit both expressions for the asymptotic conditional bias and variance terms as they can be derived immediately from the asymptotic Bahadur representation later on.

THEOREM 2.6 (Consistency for heteroscedastic M-smoothers)
For model (2.39) and assumptions A1 – A7 (for A1, A2 and A5 being modified by A1, A2* and A5*) the M-smoothers estimates of the regression function $m(\cdot)$ and its derivatives respectively are consistent. In other words it holds that*

$$\sqrt{Nh_N^{1+2\nu}} \cdot \left(\widehat{\beta}_\nu - \frac{m^{(\nu)}(x)}{\nu!}\right) = O_{\mathbf{P}}(1),$$

for $N \to \infty$ and any $\nu \in \{0,\dots,p\}$ given at the chosen point $x \in (0,1)$.

Proof. See Section 2.3.2 below. ■

THEOREM 2.7 (Asymptotic bias term for $\widehat{m}^{(\nu)}(x)$ under heteroscedasticity)
For model (2.39), assumptions A1 – A7 (for A1, A2 and A5 being modified by A1, A2* and A5*) and the given notation the asymptotic bias term for the M-smoothers estimates is equal to*

$$\mathbb{A}\text{s.}\mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right] \overset{def.}{=} \widehat{\beta}_\nu - \beta_\nu = \nu! h_N^{p+1-\nu} \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!}\right) \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1} \boldsymbol{\mu} + o_{\mathbf{P}}(h_N^{p+1-\nu}),$$

for $\nu \in \{0,\dots,p\}$ and $\boldsymbol{\mu} = \left(\int_{-1}^{1} u^{p+1}K(u)du, \int_{-1}^{1} u^{p+2}K(u)du, \dots, \int_{-1}^{1} u^{2p+2}K(u)du\right)^\top \in \mathbb{R}^{p+1}$.

Proof. See Section 2.3.2 below. ■

The bias term in Theorem 2.7 is again derived in sense of a difference rather than expectation while using the same argumentation as before. It is also not surprising that the bias expressions for homoscedastic as well as heteroscedastic models are the same. Indeed, using the local approach defined by the given kernel function $K(\cdot)$ the heteroscedasticity in random errors does not play any role with respect to the bias term as the scale (variability respectively) in some small neighbourhood of the given point $x \in (0,1)$ can be assumed to be constant.

Finally, we can formulate the asymptotic distributional property for the heteroscedastic M-smoothers estimates, which can be again seen in an analogy with the homoscedastic normality result.

THEOREM 2.8 (Asymptotic normality for $\widehat{m}^{(\nu)}(x)$ under heteroscedasticity)

For model 2.39, assumptions A1 – A7 (A1, A2 and A5 replaced by A1, A2* and A5*) and the given notation the M-smoothers estimates follow in asymptotic in law a normal distribution given by*

$$\sqrt{Nh_N^{1+2\nu}}\cdot\left(\widehat{m}^{(\nu)}(x)-m^{(\nu)}(x)-\mathbb{B}\mathrm{ias}\left[\widehat{m}^{(\nu)}(x)\right]\right)\xrightarrow[N\to\infty]{\mathscr{D}}\mathbb{N}\left(0,\frac{\nu!^2\cdot\mathbb{E}\psi^2(\sigma(x)\varepsilon_1)}{[\lambda_G'(0,\sigma(x))]^2f(x)}\cdot\mathbf{e}_\nu^\top\mathsf{V}\mathbf{e}_\nu\right),$$

where $\nu\in\{0,1,\ldots,p\}$ stands for the order of the derivative of function $m(\cdot)$ or its estimate $\widehat{m}(\cdot)$ respectively, and $\mathsf{V}=\mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}$.

Proof. See Section 2.3.2 below. ■

In the same way as before we can again express the Asymptotic Mean Square Error (AMSE) quantity, which is used to determine the optimal value of the bandwidth parameter h_N but a common problem related to no explicit definition of this quantity remains for the heteroscedastic scenario as well. Iterative approaches similar to those used for the homoscedastic model are proposed to solve this problem. Mostly, they all can be thought of as straightforward generalizations of those already designed for homoscedastic cases. Some of them are also discussed in more detail in Section 5.1.

The only clear difference in the results derived for homoscedastic and heteroscedastic M-smoothers is the scale factor $\sigma(x)$ used in addition for the heteroscedastic scenario where it multiplies the random error quantities $\{\varepsilon_i\}_{i=1}^N$ and it brings them to a proper local scale. Unlike the homoscedastic model where we could assume an arbitrary scale of random errors, in heteroscedastic models the scale of random errors is by the assumption A2* restricted to one only and the scale factor $\sigma(x)$ is used instead to take into account scales different from one.

Using the local estimation approach and considering a small region around some specific point $x\in(0,1)$ the whole heteroscedastic model scenario reduces to a homoscedastic structure as the scale function $\sigma(\cdot)$ can be considered to be constant within a small neighbourhood of $x\in(0,1)$. Moreover, one can easily seen that all results and expressions derived under the heteroscedastic model simplify in a straightforward way to a homoscedastic model once the scale function is set to be constant.

2.3.2 Proofs of Theorems 2.6, 2.7 and 2.8

In the proofs we will consider a vector of parameter estimates defined by the minimization problem 2.40. Moreover, as the lines of the proofs will closely follow the lines of the proofs of Theorems 2.3, 2.4 and 2.5 we will state in detail the main differences only and otherwise, we will refer to the corresponding lines of the proofs in Section 2.2.3.

Let us start with the minimization problem 2.40. Using the model definition 2.39, the Taylor expansion of function $m(\cdot)$ and the convex property of the loss function $\rho(\cdot)$ and the corresponding partial differential operators we obtain a set of equations

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^N\psi\left(\sigma\left(X_i\right)\varepsilon_i-\sum_{j=0}^p b_j\left(\frac{X_i-x}{h_N}\right)^j\right)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)=0,\tag{2.42}$$

for $l=0,\ldots,p$,

which is solved for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^\circ=(\widehat{\beta}_0^\circ,\ldots,\widehat{\beta}_p^\circ)^\top\in\mathbb{R}^{p+1}$ where the definition of the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^\circ$ is the same as in 2.10 and the vector of the true parameters is again $\boldsymbol{\beta}_x^\circ=(0,\ldots,0)^\top\in\mathbb{R}^{p+1}$.

Analogously as before we will proceed in three consecutive steps:

❑ firstly, we need show that all parameter estimates $\widehat{\beta}_0^\circ, \dots, \widehat{\beta}_p^\circ$ are close enough to zero in order to satisfy the set of equations $\boxed{2.42}$ however, this can be done in very the same way as in the proof for the homoscedastic model therefore, we will omit it here;

❑ in the second step we will show the consistency property of the parameter estimates, which means that $\sqrt{Nh_N}\ \widehat{\beta}_j^\circ \ \equiv\ \sqrt{Nh_N^{1+2j}}\ \left(\widehat{\beta}_j - \beta_j\right) = O_{\mathbf{P}}(1)$ for any $j = 0, \dots, p$.

❑ finally, we will prove the asymptotic normality property and we will show that

$$\sqrt{Nh_N}\left(\widehat{\beta}_j^\circ - \beta_j^\circ\right) \equiv \sqrt{Nh_N^{1+2j}}\left(\widehat{\beta}_j - \frac{m^{(j)}(x)}{j!}\right) \underset{N\to\infty}{\xrightarrow{\mathscr{D}}} \mathbf{N}(\cdot,\cdot), \quad \text{for } j = 0, \dots, p,$$

where $\mathbf{N}(\cdot,\cdot)$ stands for a normal distribution with appropriate mean and variance quantities;

Once we show the validity of all three steps just a straightforward calculations are required after that to finish all proofs of the theorems stated above.

Proof of Theorem 2.6

The basic idea of the proof remains the same: we will introduce two auxiliary lemmas where the assertions of both lemmas will be crucial for the rest of the proofs under the heteroscedastic model scenario.

Lemma 3

For model $\boxed{2.39}$ and the same assumptions as in Theorem 2.6 the following bound in probability is achieved

$$\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}}\left|\sum_{i=1}^{N}\left\{\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N\left(\tau_{iN}(x)\right)^j\right) - \mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N\left(\tau_{iN}(x)\right)^j\right)\right]\right\}\left(\tau_{iN}(x)\right)^l K\left(\tau_{iN}(x)\right)\right| =$$

$$= O_{\mathbf{P}}\left((Nh_N)^{-\frac{p+1}{p+3}} \cdot \delta_N^{\frac{2(p+1)}{p+3}}\right),$$

for any $T > 0$ and $l \in \{0, \dots, p\}$ where $\tau_{iN}(x) = \left(\frac{X_i - x}{h_N}\right)$. Moreover, the expectation operator $\mathbb{E}[\cdot]$ stands here for a conditional expectation conditioned on values of the random variable X and $1/\sqrt{Nh_N} \le \delta_N \le 1$ is chosen arbitrarily.

Proof of Lemma 3

The proof of Lemma 3 follows the same idea as the proof of Lemma 1 even with the scale factor $\sigma(X_i)$ used to multiply random error terms ε_i, for $i = 1, \dots, N$, therefore, the detailed proof is omitted. □

Using now the same argumentation as in the proof of the consistency for the homoscedastic model and applying the assertion of Lemma 3 we obtain that

$$\left|\sum_{i=1}^{N}\left\{\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}\widehat{\beta}_j^\circ\left(\frac{X_i - x}{h_N}\right)^j\right) + \right.\right.$$

$$\left.\left. +\lambda_G\left(\sum_{j=0}^{p}\widehat{\beta}_j^\circ\left(\frac{X_i - x}{h_N}\right)^j, \sigma(X_i)\right)\right\} \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| = O_{\mathbf{P}}\left(\sqrt{Nh_N}\right),$$

and we can again use the Taylor expansion of function $\lambda_G(\cdot,\sigma(X_i))$, for each $i = 1,\ldots,N$, as we have the Lipschitz property assured by assumption A5*. This gives us

$$\lambda_G\left(\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j,\sigma(X_i)\right) = \lambda_G(0,\sigma(X_i)) + \lambda_G'(\tilde{t},\sigma(X_i))\cdot\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j,$$

for $i = 1,\ldots,N$ and $l = 0,\ldots,p$, where $\lambda_G'(\cdot,\cdot)$ stands for a partial derivative of function $\lambda_G(\cdot,\cdot)$ with respect to its first argument.
In addition, $\tilde{t} \in \mathbb{R}$ is some value between zero and $\tilde{\tilde{t}} = \sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j$.

It is important to realize here that rescaling of random quantities ε_i's, for $i = 1,\ldots,N$ does not have any effect on their symmetric distributional property therefore, we also have that $\lambda_G(0,\sigma(X_i)) = 0$, for any $i = 1,\ldots,N$. Hence, we obtain

$$\left|\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\left\{\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_n}\right)^j\right) + \right.\right.$$
$$\left.\left. + \sqrt{Nh_N}\lambda_G'(\tilde{t},\sigma(X_i))\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^j\right\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right| = O_{\mathbf{P}}(1).$$

uniformly in $i = 1,\ldots,N$. One can easily bound $\lambda_G'(\tilde{t},\sigma(X_i))$ by some $0 < \mathcal{K} < \infty$, which follows from the Lipschitz property of function $\lambda_G(\cdot,\cdot)$. Using now the fact that parameter estimates $\widehat{\beta}_0^{\circ},\ldots,\widehat{\beta}_p^{\circ}$ solve the equations (2.42) and applying the same argumentation as in the consistency proof for the homoscedastic case one will easily obtain the consistency property for parameter estimates $\widehat{\beta}_0^{\circ},\ldots,\widehat{\beta}_p^{\circ}$. The consistency property for the M-smoothers estimates as stated in Theorem 2.6 follows immediately once we recall the definition of the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ} \in \mathbb{R}^{p+1}$. ■

Proof of Theorem 2.7

Using now the consistency result for the M-smoothers estimates we can proceed to derive the asymptotic bias term. We will firstly derive the asymptotic Bahadur representation for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}$ and we will use this representation to derive the asymptotic bias term and the asymptotic normality result for. We will need the assertion of the following lemma to continue.

Lemma 4
Let us assume model (2.39) and the same assumptions as in Theorem 2.7. Then the following convergence in probability is achieved

$$\sup_{\substack{|t_j|<T\\ j=0,\ldots,p}}\frac{1}{\sqrt{Nh_N}}\left|\sum_{i=1}^{N}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) + \right.\right.$$
$$\left.\left. - \mathbb{E}\psi\left(\sigma(x)\varepsilon_i - \sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right| \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

uniformly for all $l = 0,\ldots,p$ any $T > 0$ and the given point of interest $x \in (0,1)$, where it holds that $\delta_N = (Nh_N)^{-1/2}$.

Proof of Lemma 4

The proof of Lemma 4 is slightly more complicated than the proof of Lemma 2 as one needs to correctly deal with a relationship between quantities $\sigma(X_i)$ and $\sigma(x)$ respectively. Using a triangle inequality we can however rewrite the expression in Lemma 4 as

$$\begin{aligned}
\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) \\
- \mathbb{E}\psi\left(\sigma(x)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right)\Bigg\} \cdot \left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| \le
\end{aligned}$$

$$\begin{aligned}
\le \sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) \\
- \mathbb{E}\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right)\Bigg\} \cdot \left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| +
\end{aligned} \tag{2.43}$$

$$\begin{aligned}
+ \sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \Bigg| \sum_{i=1}^{N} \Bigg\{ \lambda_G\left(\sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j, \sigma(X_i)\right) \\
- \lambda_G\left(\sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j, \sigma(x)\right)\Bigg\} \cdot \left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg|,
\end{aligned} \tag{2.44}$$

and we will prove that both terms converge to zero in probability once we have $\delta_N = (Nh_N)^{-1/2}$. Using an analogous notation

$$\widetilde{\Xi}_N^{\psi}(X_i, x, \boldsymbol{t}) \overset{def.}{=} \left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right)\right],$$

where $\boldsymbol{t} = (t_0, \dots, t_p)^\top \in \mathbb{R}^{p+1}$ we can precisely go along the lines of the proof of Lemma 2 and to decompose (2.43) into two terms, which are both analogous to (2.22) and (2.23), while the second term of this decomposition can be easily dealt with using the Lipschitz property of function $\lambda_G(\cdot, \sigma(X_i))$, for any $i = 1, \dots, N$ and the same argumentation as in Lemma 2.

The first term of the decomposition (the one analogous to (2.22)) can be taken care of using Chebyshev's inequality in the a similar way as in the proof of Lemma 2 however, this time we obtain

$$\begin{aligned}
\mathbf{P}\left[\max_{\substack{1\le m\le D_N\\ j=0,\dots,p}} \frac{1}{Nh_N\delta_N\nu_N} \left| \sum_{i=1}^{N} \left\{\widetilde{\Xi}_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m) - \mathbb{E}\left[\widetilde{\Xi}_N^{\psi}(X_i, x, \boldsymbol{\zeta}_m)\right]\right\} \left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right| \ge \epsilon\right] \\
\le (\epsilon N h_N \delta_N \nu_N)^{-2} \cdot \nu_N^{-(p+1)} \cdot Nh_N \cdot \mathsf{Var}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} \zeta_{mj}\delta_N \left(\frac{X_i-x}{h_N}\right)^j\right) - \psi(\sigma(x)\varepsilon_i)\right],
\end{aligned}$$

for $i \in \{1, \dots, N\}$ and any $\epsilon > 0$, where D_N, ν_N and $\boldsymbol{\zeta}_m = (\zeta_{m0}, \dots, \zeta_{mp})^\top$ are defined in the proof of Lemma 2. We need to show now that the variance term is of the order $O\left(\frac{1}{\sqrt{Nh_N}}\right)$.

However, analogously as in (2.25) we easily obtain from assumption A5* that

$$\mathbb{V}\text{ar}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) - \psi(\sigma(x)\varepsilon_i)\right] \leq$$

$$\leq \mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) - \psi(\sigma(x)\varepsilon_i)\right]^2 = O\left(\frac{1}{\sqrt{Nh_N}}\right),$$

where in addition we have also used the Hölder property of the order $\alpha > \frac{\iota-1}{2\iota}$ for function $\lambda_G(0,\cdot)$. The required rate of convergence follows now immediately given the same argumentation as in the proof of Lemma 2.

It is still left to prove the asymptotic convergence in probability for the second term (2.44) but using the partial derivative of function $\lambda_G(\cdot,\cdot)$ in its first argument we obtain

$$\lambda_G\left(\sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j, \sigma(x)\right) = \lambda_G\left(0,\sigma(x)\right) + \lambda'_G(0,\sigma(x)) \cdot \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j + o_{\mathbf{P}}\left(\delta_N\right),$$

and analogously also

$$\lambda_G\left(\sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j, \sigma(X_i)\right) = \lambda_G\left(0,\sigma(X_i)\right) + \lambda'_G(0,\sigma(X_i)) \cdot \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j + o_{\mathbf{P}}\left(\delta_N\right),$$

both for $|X_i - x| \leq h_N$ and $i = 1,\ldots,N$. To finish the proof one just needs to realize that $\lambda_G\left(0,\sigma(x)\right) = \lambda_G\left(0,\sigma(X_i)\right) = 0$ for $i = 1,\ldots,N$ and $|\lambda'_G(0,\sigma(x))| \leq \mathcal{K}_1$ as well as $|\lambda'_G(0,\sigma(X_i))| \leq \mathcal{K}_2$ uniformly over $i \in \{1,\ldots,N\}$, for some $\mathcal{K}_1,\mathcal{K}_2 > 0$, both under the Lipschitz property of function $\lambda_G(\cdot,\cdot)$ in its first argument, which follows directly from the existence of the partial derivative. Moreover, we have that $\left|\sum_{j=1}^{p} t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right)$ therefore, the convergence in probability for the second term (2.44) follows immediately as well. This now finishes the proof of Lemma 4. □

Using the assertion of Lemma 4 and a similar argumentations as for the proof in the homoscedastic model we obtain that

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i - x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) + \right.$$

$$\left. + \lambda'_G(0,\sigma(x)) \cdot \sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i - x}{h_N}\right)^j + o_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right)\right] \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

uniformly for $l = 0,\ldots,p$ and hence, applying the matrix notation we have

$$\frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) \;=\; \frac{\lambda'_G(0,\sigma(x))}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\widehat{\boldsymbol{\beta}}_x^{\circ} + o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right), \tag{2.45}$$

where $\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) = (\sigma(x)\varepsilon_1,\ldots,\sigma(x)\varepsilon_N)^{\top}$. Under the regularity property of matrix $\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N$ we finally obtain the Bahadur representation for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}$ given by

$$\widehat{\boldsymbol{\beta}}_x^{\circ} \;=\; \frac{1}{\lambda'_G(0,\sigma(x))} \cdot \left(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\right)^{-1} \cdot \mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) + o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right). \tag{2.46}$$

This asymptotic Bahadur representation is just a heteroscedastic analogy of the Bahadur representation (2.28) derived under the homoscedastic model scenario where the only difference involved in is the scale factor $\sigma(x)$, which needs to be considered for the given point $x \in (0,1)$.

To finish the proof just a straightforward computation is needed similarly as in the case of homoscedastic model. One just has to use function $\lambda'_G(0,\sigma(x))$ instead of $\lambda'_G(0)$ and also $\psi(\sigma(x)\varepsilon_1)$ instead of $\psi(\varepsilon_1)$. The rest of the proof is therefore omitted. ■

Proof of Theorem 2.8

To show the asymptotic normality results we will start with expression (2.45). The left-hand side of this expression converges again to a normal distribution using the same argumentation as in the proof of the asymptotic normality for the homoscedastic model. From the regularity property of $\mathsf{X}_N\mathsf{W}_N\mathsf{X}_N$ we obtain that also the first term on the right-hand side of (2.46) converges element-wise to a normal distribution with corresponding mean and variance parameters.

Applying now the Slutsky theorem the result follows immediately. Moreover, the corresponding variance quantity can be easily expressed using the asymptotic Bahadur representation (2.45) similarly as in the case of homoscedastic scenario before. ■

We have shown in this section that the M-smoothers estimation approach under the heteroscedastic assumption yields the same statistical properties as the M-smoothers approach under the homoscedastic regression model. Indeed, we have proved consistency of the M-smoothers estimates as well as their asymptotic normality property. Moreover, one can easily see a nice correspondence between the results derived for homoscedastic and heteroscedastic models where the scale factor $\sigma(x)$ is introduced in addition in the heteroscedastic case in order to take care of an additional flexibility within slightly less strict variance assumptions. Defining this scale factor $\sigma(x)$ to be constant for all $x \in (0,1)$ the results derived for the heteroscedastic model easily reduce back to the results derived for the homoscedastic scenario. This is a nice additional property, which will simplify our next theory development for the change-point models and a discontinuous robust regression approach.

The local polynomial M-smoothers introduced with independent random errors give us a flexible and powerful tool in the regression estimation approach and the results derived under homoscedastic as well as heteroscedastic scenarios are in a reasonable balance with assumptions required for their proofs. The consistency property is important in order to have an estimation approach, which can provide us with a good estimate for the unknown but true regression function of interest (or its derivatives respectively) while the asymptotic normality property makes it easier to work with such estimates in real data cases, as one can easily construct confidence bounds and acceptance/rejection regions for hypothesis testing problems, which we will also focus on in the next chapter.

We will however, discuss one more extension of the M-smoothers regression approach and we will propose the M-smoothers estimation for a model with dependent random errors, which is mostly the case indeed in modern real data situations. Providing an analogous statistical inference for the model with some dependent data concept we will obtain a quite complete set of statistical modelling tools for the robust M-estimation techniques under the most common situations, which can be possibly considered for different regression model scenarios.

2.4 Models with dependent observations

We have already discussed a standard model based on the homoscedastic variance assumption and we have also extended such model into the heteroscedastic variance scenario later on. In order to develop even more flexible regression techniques one can still think of some further generalizations with respect to less strict assumptions posed on the variance (scale respectively) function as we were only restricted to Lipschitz functions so far. However, rather than discussing further variance/scale generalizations we would like to focus on a generalization with respect to another important assumption, which was required to be strictly satisfied until now.

It is always nice and convenient to start with an condition posed on a random sample from some unknown distribution function however, an independence assumption, which is implicitly assumed for every random sample may not be always the real case. However, this assumption is rather crucial and even small disturbances from an independently generating random system may have a serious impact on the final performance of the given estimator therefore, one needs to always consider a set of proper statistical tools in order to deal with data samples, which do not necessarily satisfy the independence assumption.

Before we discuss the M-smoothers estimation techniques together with some dependence structures of random error sequences we will shortly introduce the most common dependence forms, which will be considered later on.

2.4.1 Weak dependence

The concept of dependent statistical data is well known for many years in statistics however, the most often used generalization of a sequence of independent, identically distributed random variables are martingales and mixing random processes in discrete time. The notation "mixing" is used here in term that random variables temporally far apart from one another have a tendency to behave like independent ones. This is also in literature referred to as a weak dependence structure.

As far as we do not assume independent observations any more it is necessary to specify an exact form of dependence using a statistical background to be able to proceed with further calculations. Let us assume a sequence of random variables $\{\xi_n\}_{n=1}^{\infty}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$ are two sub-σ-fields in $\mathcal{F}$ and let us introduce two measures on $\mathcal{A} \times \mathcal{B}$ defined by

$$\begin{aligned} \alpha(\mathcal{A},\mathcal{B}) &\overset{def.}{=} \sup_{A\in\mathcal{A},\ B\in\mathcal{B}} \left|\mathbf{P}\left(A\cap B\right) - \mathbf{P}\left(A\right)\mathbf{P}\left(B\right)\right|, \\ \varphi(\mathcal{A},\mathcal{B}) &\overset{def.}{=} \sup_{A\in\mathcal{A},\ B\in\mathcal{B},\ \mathbf{P}(A)>0} \left|\mathbf{P}\left(B|A\right) - \mathbf{P}\left(B\right)\right|. \end{aligned}$$

Both measures can be used to measure the dependence of random events from the σ-field $\mathcal{A}$ on random events from the σ-field $\mathcal{B}$. It is also important to be aware here of the form of symmetry where the first measure is obviously symmetric as we easily have that $\alpha(\mathcal{A},\mathcal{B}) = \alpha(\mathcal{B},\mathcal{A})$ while this is in general not true for the second measure based on the conditional probability.

DEFINITION 1 (α-mixing and φ-mixing dependent variables)
Let $\mathcal{F}_l^k$ be a σ-field generated by the sequence of random variables $\{\xi_i;\ l \leq i \leq k\}$. Then the sequence of random variables $\{\xi_n\}_{n=1}^{\infty}$ is said to form a strong mixing (α-mixing) process if

$$\alpha(n) \overset{def.}{=} \sup_{k\in\mathbb{N}} \ \alpha(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \to 0,$$

for $n \to \infty$. Moreover, the sequence $\{\xi_n\}_{n=1}^{\infty}$ is said to be uniformly strong mixing (φ-mixing) if

$$\varphi(n) \overset{def.}{=} \sup_{k \in \mathbb{N}} \varphi(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \to 0,$$

for $n \to \infty$ again.

The concept of the uniformly strong mixing dependence was introduced by Rosenblatt (1956) and it holds that the uniformly strong mixing dependence implies the strong mixing dependence (see Lin and Lu (1997) for the proof), which was introduced later on by Ibragimov (1959). Quantities $\alpha(n)$ and $\varphi(n)$ are called the coefficients of dependence and they express how much dependence there exists between two events separated in time by at least $n \in \mathbb{N}$ other events.

The concept of the weak dependence is very useful in statistics as it effectively deals with most common dependence structures, random processes and time series. Indeed, it was shown in Anderson (1958) that m-dependent processes as well as finite order ARMA processes with innovations satisfying Doeblin's condition are φ-mixing. On the other hand, finite order processes, which do not satisfy Doeblin's condition were showed to be α-mixing instead (see Ibragimov and Linik (1971)). Additionally, Rosenblatt (1971) provides general conditions for stationary Markov Processes to by α-mixing as well.

We will consider the α-mixing dependence for random error terms and we will discuss an extension of the M-smoothers estimation approach for heteroscedastic[25] model. This will also cover the most common dependence structures of random error terms used in different regression settings including auto-regressive (AR) processes of finite orders.

2.4.2 α- mixing model

Let us consider a bivariate sequence of random variables $\{(X_i, Y_i);\ i = 1, \dots, N\}$, which comes from some population (X, Y) taking values in $[0,1] \times \mathbb{R}$ and let this sequence forms a strictly stationary α-mixing process with α-mixing dependence coefficients $\alpha(N)$, such that $\alpha(N) \to 0$ for $N \to \infty$.
As far as we still assume the model

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \tag{2.47}$$

the random error terms can be easily expressed as $\varepsilon_i = (Y_i - m(X_i))/\sigma(X_i)$ for $i = 1, \dots, N$. This also means that the random sequence $\{\varepsilon_i\}_{i=1}^{N}$ forms a strong mixing process with the same α-mixing dependence coefficients as well given the fact that any measurable mapping of an α-mixing process forms a strong mixing process with the same dependence coefficients again.
We are again primarily interested in estimation of the unknown regression function $m(\cdot)$ in sense of the local polynomial M-smoothers approach defined by the minimization problem (2.1). We will show that under some appropriate assumptions posed on the convergence rates of the mixing coefficients the performance of the M-smoothers estimator and its asymptotic behaviour is essentially the same as in the case of *i.i.d.* observations.

We will again start with the formulation of all necessary assumptions, which are required to deal with the weak dependence concept considered in this section. They are somewhat analogous to those considered in the previous sections however, some modifications are introduced.

[25] By introducing a generalization of the heteroscedastic model scenario together with the concept of weak dependent random errors we also account for any homoscedastic model scenario as this follows easily as a special case of the heteroscedastic model.

A1** The marginal density function $f(\cdot)$ of the random variables X_i, for $i = 1, \dots, N$ is absolutely continuous, positive and bounded on $[0,1]$, which is the support of X. Moreover, we assume that the scale function $\sigma(\cdot)$ is Lipschitz and positive on interval $[0,1]$. Additionally, we need the joint density function $f_\ell(\cdot,\cdot)$ of random variables X_1 and $X_{1+\ell}$ to be bounded for all $\ell \geq 1$;

A2** The sequence $\{(X_i, Y_i)\}_{i=1}^N$ forms a strongly stationary α-mixing process with the α-mixing dependence coefficients $\{\alpha(n)\}_{n=1}^\infty$ such that

$$\sum_{n=1}^{\infty} [\alpha(n)]^{\delta/(2+\delta)} < \infty, \tag{2.48}$$

for some $\delta > 0$. Let X_i and ε_i are mutually independent for $i = 1, \dots, N$ and moreover, the distribution function $G(\cdot)$ of $\{\varepsilon_i\}_{i=1}^N$ is assumed to be continuous and symmetric with a unit scale, $G(1) - G(-1) = \frac{1}{2}$. It also holds that $\sum_{\ell=1}^\infty \mathbb{E}[\psi(\sigma(x)\varepsilon_1)\psi(\sigma(x)\varepsilon_\ell)] \leq \mathcal{K}_2 < \infty$ for the chosen $x \in (0,1)$ and some $\mathcal{K}_2 > 0$;

A5** We assume that function $\lambda_G(t,v) = -\int \psi(ve - t) \mathrm{d}G(e)$ is Hölder of the order $\alpha > \frac{\iota - 1}{2\iota}$ in argument $v > 0$ for ι being defined in assumption A7. The partial derivative $\lambda_G'(t,v) = \frac{\partial}{\partial t}\lambda_G(t,v)$ exists and it is continuous in t and $\int_{\mathbb{R}} (\psi(\sigma(x)e - \epsilon_N) - \psi(\sigma(x)e))^2 \mathrm{d}G(e) < \mathcal{K} \cdot |\epsilon_N|$ both for some neighbourhoods of $t = 0$ and $v = \sigma(x)$, for the given point $x \in (0,1)$, any sequence $\epsilon_N \to 0$ and some $\mathcal{K} > 0$. Moreover, it holds that $\int |\psi(\sigma(x)e)|^{2+\delta} \mathrm{d}G(e) < \infty$ and $\lambda_G'(0, \sigma(x)) = \frac{\partial}{\partial t}\, \lambda_G(t, \sigma(x))|_{t=0} \neq 0$ for $\delta > 0$ small enough and the given point $x \in (0,1)$;

In assumption A1** the requirements posed on the marginal density function and the scale function remain the same as in the heteroscedastic case however, we have to account for the fact that random variables $\{X_i\}_i$ are not independent anymore therefore, strict stationarity is assumed instead. Moreover, we also need the joint density function $f_\ell(\cdot,\cdot)$ to be bounded for all $\ell > 1$ to be able to derive the asymptotic normality result.
Assumption A2** defines a set of classical regularity conditions for the α-mixing coefficients in order to obtain the same asymptotic performance of the M-smoothers estimator as in the *i.i.d.* case.
Finally, assumption A5** needs to be slightly alter in order to prepare a sufficient background for the weak dependence inference however, the main idea remains the same.

We will now introduce the main results and will discuss the main statistical properties of the local polynomial M-smoothers approach when considering weakly dependent strongly mixing sequences of data points.

2.4.3 The main asymptotic results

We have already mentioned that under some appropriate conditions posed on the mixing coefficients one mostly obtains estimates with the same asymptotic performance as those derived for the *i.i.d.* cases. Indeed, it was proved in Baek and Wehrly (1993) or Boente and Fraiman (1995) that the performance of the local constant kernel regression estimator for weakly dependent data (α-mixing and φ-mixing dependence) is under the right regularity conditions essentially the same as for the independent and identically distributed data.

We will extend this idea onto the local polynomial M-smoothers as well. We will firstly show the consistency property for the M-smoothers estimates under the α-mixing dependence and after that we will also state the asymptotic normality result and we will provide complete proofs for all stated results.

THEOREM 2.9 (Consistency for α-mixing dependence)
For model (2.47) and assumptions A1 – A7 (for A1, A2 and A5 being replaced by A1**, A2** and A5**) the M-smoothers estimates of the regression function $m(\cdot)$ and its derivatives respectively are consistent. Equivalently, we can state that

$$\sqrt{Nh_N^{1+2\nu}} \cdot \left(\widehat{\beta}_\nu - \frac{m^{(\nu)}(x)}{\nu!} \right) = O_{\mathbf{P}}(1),$$

for $N \to \infty$ any $\nu \in \{0, \ldots, p\}$ and the given point $x \in (0,1)$.

Proof. See Section 2.4.4 below. ■

THEOREM 2.10 (Asymptotic bias term for $m^{(\nu)}(x)$ under the α-mixing dependence)
Let model (2.47) holds. Then for assumptions A1 – A7 with A1, A2 and A5 to be replaced by A1**, A2** and A5** and the given notation the asymptotic bias term for the M-smoothers estimates of the unknown regression function and its derivatives respectively can be expressed as

$$\mathbb{A}\text{s.}\mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right] \overset{def.}{=} \nu! \left(\widehat{\beta}_\nu - \beta_\nu\right) = \nu! h_N^{p+1-\nu} \cdot \left(\frac{m^{(p+1)}(x)}{(p+1)!} \right) \cdot \mathbf{e}_\nu^\top \mathsf{S}_1^{-1} \boldsymbol{\mu} + o_{\mathbf{P}}(h_N^{p+1-\nu}),$$

where $\boldsymbol{\mu} = \left(\int_{-1}^{1} u^{p+1} K(u) du, \int_{-1}^{1} u^{p+2} K(u) du, \ldots, \int_{-1}^{1} u^{2p+2} K(u) du \right)^\top \in \mathbb{R}^{p+1}$ and $\nu \in \{0, \ldots, p\}$ stands for the order of the corresponding derivative of the regression function $m(\cdot)$.

Proof. See Section 2.4.4 below. ■

THEOREM 2.11 (Asymptotical normality for $m^{(\nu)}(x)$ under the α-mixing dependence)
Let model (2.47) holds. Then for assumptions A1 – A7 with A1, A2 and A5 to be replaced by A1**, A2** and A5** and the given notation the M-smoothers estimate follows in asymptotic in law a normal distribution given by

$$\sqrt{Nh_N^{1+2\nu}} \cdot \left(\widehat{m}^{(\nu)}(x) - m^{(\nu)}(x) - \mathbb{B}\text{ias}\left[\widehat{m}^{(\nu)}(x)\right] \right) \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{\nu!^2 \cdot \mathbb{E}\left[\psi(\sigma(x)\varepsilon_1)\right]^2}{[\lambda_G'(0, \sigma(x))]^2 f(x)} \cdot \mathbf{e}_\nu^\top \mathsf{V} \mathbf{e}_\nu \right),$$

where $\nu \in \{0, 1, \ldots, p\}$ stands for the order of the derivative of the regression function $m(\cdot)$, or its estimate $\widehat{m}(\cdot)$ respectively. Moreover, $\mathsf{V} = \mathsf{S}_1^{-1}\mathsf{S}_2\mathsf{S}_1^{-1}$ as in Theorem 2.8.

Proof. See Section 2.4.4 below. ■

We have omitted stating the theorem for asymptotic variance term of the M-smoothers estimates as it can be directly obtained from the variance term in the asymptotic normality expression. For the bias term however, we have rather formulated the corresponding theorem above.

We can now compare all results derived under three different scenarios under consideration. We have already shortly discussed the main differences between the homoscedastic and heteroscedastic model, both derived for *i.i.d.* random errors. Now, we can compare those results with the results derived under the weak dependence assumption namely, the α-mixing dependence. One can easily see a full correspondence of the results, which we have expected as we had said that under the right regularity conditions the performance of the estimator should be the same for *i.i.d.* cases as well as for the weak dependence concept. We will now prove the stated results in order to justify such conclusions.

2.4.4 Proofs of Theorems 2.9, 2.10 and 2.11

The main idea of all proofs remains the same as in the previous sections however, additional caution needs to be paid to deal with the rates of convergence for the α-mixing dependence coefficients.

We will start with the minimization problem

$$\widehat{\boldsymbol{\beta}}_x^{\circ} = \underset{(b_0,\dots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \rho\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot K\left(\frac{X_i - x}{h_N}\right), \qquad \boxed{2.49}$$

which we can easily get using the original minimization problem $\boxed{2.1}$, the definition of the model and the $(p+1)$ order Lipschitz assumption posed on the unknown regression function $m(\cdot)$. The vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}$ is supposed to give us a reasonable estimate for the vector of true parameters, which is the vector $\boldsymbol{\beta}_x^{\circ} = (0,\dots,0)^\top \in \mathbb{R}^{p+1}$ where the notation used here is the same as in the case of minimization $\boxed{2.10}$.

Using now the convex property of the loss function $\rho(\cdot)$ we can express the minimization problem $\boxed{2.49}$ in terms of a set of "normal" equations as

$$\frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) = 0, \qquad \boxed{2.50}$$
$$\text{for} \quad l = 0,\dots,p,$$

which is solved for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ} = (\widehat{\beta}_0^{\circ},\dots,\widehat{\beta}_p^{\circ})^\top \in \mathbb{R}^{p+1}$. Similarly, as in the homoscedastic or heteroscedastic cases there is no explicit solution given and iterative numerical approaches need to be used to express the parameter estimates in real data examples.
On the other hand, the asymptotic approximations and representations need to be found as well in order to investigate the corresponding statistical inference of the parameter estimates.

Let us again proceed in three separate but consecutive steps:

- ❑ firstly, we will show that the parameter estimates $\widehat{\beta}_0^{\circ},\dots,\widehat{\beta}_p^{\circ}$ need to be all in a small neighbourhood of zero in order to satisfy the set of equations $\boxed{2.50}$;
- ❑ next, we will prove that $\widehat{\beta}_j^{\circ} \to 0$ as $N \to \infty$ for every $j = 0,\dots,p$, more specifically, we will show that $\sqrt{Nh_N}\, \widehat{\beta}_j^{\circ} \;\equiv\; \sqrt{Nh_N^{1+2j}}\, \left(\widehat{\beta}_j - \beta_j\right) = O_{\mathbf{P}}(1)$ for any $j = 0,\dots,p$.
- ❑ finally, we will prove the asymptotic normality property and we will show that

$$\sqrt{Nh_N}\left(\widehat{\beta}_j^{\circ} - \beta_j^{\circ}\right) \equiv \sqrt{Nh_N^{1+2j}}\left(\widehat{\beta}_j - \frac{m^{(j)}(x)}{j!}\right) \xrightarrow[N\to\infty]{\mathscr{D}} \mathbf{N}(\cdot,\cdot),$$

for some appropriate mean and variance parameters;

The first step of the proof can be easily shown using the same argumentation as the one we have used for the proof of the results derived under the homoscedastic model scenario therefore, we will omit copying it here again and we will proceed with the second step now.

Proof of Theorem 2.9

We want to prove the consistency result for the parameter estimates $\widehat{\beta}_0^{\circ}, \ldots, \widehat{\beta}_p^{\circ}$ derived under the weak dependence assumption. This is slightly more complicated to prove as one needs to properly take into account the right form of the dependence structure within given data points. We will therefore formulate two necessary lemmas at first.

Lemma 5

Let $\{\xi_i\}_{i=1}^N$ be some strictly stationary α-mixing process with the dependence coefficients $\alpha(i)$. Let moreover, $g_N(\cdot)$ be some measurable mappings such that $\mathbb{E} g_N(\xi_i) = 0$. Then the following holds:

$$\mathbb{E}\left|\sum_{i=1}^{N} g_N(\xi_i)\right|^2 \leq \mathcal{K}N \cdot \left(\mathbb{E}\left|g_N(\xi_1)\right|^{2+\epsilon}\right)^{\frac{2}{2+\epsilon}} \cdot \sum_{j=1}^{\infty}(\alpha(j))^{\frac{\epsilon}{2+\epsilon}},$$

for some $\mathcal{K} > 0$ and any $\epsilon > 0$.

Proof of Lemma 5

The proof of the lemma follows easily from the proof of Theorem 1 in Yokoyama (1980). □

Lemma 6

Let us assume model 2.47 and the set of the same assumptions as in Theorem 2.9. Then the following bound in probability is achieved

$$\sup_{\substack{|t_j|<T\\ j=0,\ldots,p}}\left|\sum_{i=1}^{N}\left\{\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) - \mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right)\right]\right\}\left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right| =$$

$$= O_{\mathbf{P}}\left((Nh_N)^{-\frac{p+1}{p+3}} \cdot \delta_N^{\frac{2(p+1)}{p+3}}\right),$$

for any $T > 0$ and $l \in \{0, \ldots, p\}$ and some arbitrary $1/\sqrt{Nh_N} \leq \delta_N \leq 1$. The expectation operator $\mathbb{E}[\cdot]$ stands here for a conditional expectation conditioned on values of the random variable X.

Proof of Lemma 6

Using again a $(p+1)$-dimensional grid of points in a $(p+1)$-dimensional cube $(-T,T) \times \cdots \times (-T,T)$ defined as in the proof of Lemma 1 we can again expand the expression in Lemma 6 into a sum of two terms similarly as in the proof of Lemma 1. Now, the term which corresponds with 2.16 can be easily bounded in probability using the Lipschitz property of function $\lambda_G(\cdot,\cdot)$ in its first argument and the definition of the mesh of grid points namely, the asymptotic rate of its expansion. We will therefore omit this part of the proof here and we will pay attention to prove the rest.

For the second term, which corresponds with (2.15) (see the proof of Lemma 1 – p.23) we have

$$
\begin{aligned}
&\mathbf{P}\Bigg[\max_{\substack{1\le m\le D_N\\ j=0,\dots,p}} \frac{1}{Nh_N\delta_N\nu_N}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\\
&\qquad -\mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)K\left(\frac{X_i-x}{h_N}\right)\Bigg|\ge\epsilon\Bigg]\le\\
&\le\sum_{\substack{m=1\\ j=0}}^{D_N}\sum_{\substack{m=1\\ j=1}}^{D_N}\cdots\sum_{\substack{m=1\\ j=p}}^{D_N}\mathbf{P}\Bigg[\frac{1}{Nh_N\delta_N\nu_N}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\\
&\qquad -\mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)K\left(\frac{X_i-x}{h_N}\right)\Bigg|\ge\epsilon\Bigg]\le\\
&\le\sum_{\substack{m=1\\ j=0}}^{D_N}\sum_{\substack{m=1\\ j=1}}^{D_N}\cdots\sum_{\substack{m=1\\ j=p}}^{D_N}(\epsilon Nh_N\delta_N\nu_N)^{-2}\cdot\mathbb{E}\Bigg[\sum_{i=1}^{N}\Bigg\{\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\\
&\qquad -\mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg]^2\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg\},
\end{aligned}
\tag{2.51}
$$

uniformly for $l=0,\dots,p$, and $\epsilon>0$ given Chebyshev's inequality and the fact that $D_N=O(\nu_N^{-1})$, where ν_N and ζ_{mj} for $m=1,\dots,D_N$ and $j=0,\dots,p$ are defined in the proof of Lemma 1.

Now we can apply Lemma 5 to function $g_N(\cdot)$, which we defined as

$$
g_N(\varepsilon_i)\overset{def.}{=}\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)-\mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right],
$$

which clearly satisfies the condition $\mathbb{E}g_N(\varepsilon_i)=0$. Therefore, we finally have

$$
\begin{aligned}
&\mathbf{P}\Bigg[\max_{\substack{1\le m\le D_N\\ j=0,\dots,p}} \frac{1}{Nh_N\delta_N\nu_N}\Bigg|\sum_{i=1}^{N}\Bigg\{\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\\
&\qquad -\mathbb{E}\left[\psi\left(\sigma(X_i)\varepsilon_i-\sum_{j=0}^{p}\zeta_{mj}\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)K\left(\frac{X_i-x}{h_N}\right)\Bigg|\ge\epsilon\Bigg]\le\\
&\le(\epsilon Nh_N\delta_N\nu_N)^{-2}\cdot Nh_N\cdot\nu_N^{-(p+1)}\cdot\mathcal{K}^*=\epsilon^{-2}\cdot\mathcal{K}^*\cdot(Nh_N)^{-1}\delta_N^{-2}\nu_N^{-(p+3)},
\end{aligned}
\tag{2.52}
$$

for some $\mathcal{K}^*>0$ and given the property that $\mathbb{E}\left|\psi(\varepsilon_1)\right|^{2+\delta}<\infty$ for some $\delta>0$ and the fact that for the α-mixing dependence coefficients $\alpha(i)$ we have $\sum_{n=1}^{\infty}n^a[\alpha(n)]^b<\infty$, for some $a>b>0$.

To conclude, we can use the same argumentation as in the proof of Lemma 1 and we obtain that $\nu_N=o\left((Nh_N\delta_N^2)^{-\frac{1}{p+3}}\right)$. The first term (the one analogous to (2.16)) is therefore of the asymptotic order $O_{\mathbf{P}}\left((Nh_N)^{-\frac{p+1}{p+3}}\cdot\delta_N^{\frac{2(p+1)}{p+3}}\right)$ and we need $(\sqrt{Nh_N})^{-1}\le\delta_N\le 1$ in order to have the assertion of Lemma 6 to be proved. □

The proof of the consistency for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ} = (\widehat{\beta}_0^{\circ}, \dots, \widehat{\beta}_p^{\circ})^\top \in \mathbb{R}^{p+1}$ follows now easily using an analogous argumentation as in the proof for the homoscedastic or heteroscedastic model respectively and the assertion of Lemma 6. Therefore, we easily obtain that

$$\sqrt{Nh_N}\,\widehat{\beta}_j^{\circ} \equiv \sqrt{Nh_N^{1+2j}}\left(\widehat{\beta}_j - \beta_j\right) = O_{\mathbf{P}}(1),$$

for $\forall j = 0, \dots, p$, which follows directly from the definition of the vector $\boldsymbol{\beta}_x^{\circ} \in \mathbb{R}^{p+1}$. ■

Proof of Theorem 2.10

Using the consistency property for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}$ derived under the α-mixing dependence we can derive the asymptotic Bahadur representation for $\widehat{\boldsymbol{\beta}}_x^{\circ}$, which can be directly used to get the bias expression as stated in Theorem 2.10. We again need an auxiliary lemma to finish the proof.

Lemma 7

For model (2.47) *and the same assumption as in Theorem 2.10 the following convergence in probability is achieved*

$$\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}} \left| \sum_{i=1}^{N} \left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) + \right.\right.$$
$$\left.\left. - \mathbb{E}\psi\left(\sigma(x)\varepsilon_i - \sum_{j=0}^{p} t_j \delta_N \left(\frac{X_i - x}{h_N}\right)^j\right)\right] \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \right| \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

for any $T > 0$ and $l = 0, \dots, p$ where $x \in (0,1)$ is the given point of interest and $\delta_N = (Nh_N)^{-1/2}$.

Proof of Lemma 7

We will use the idea of the proof of Lemma 4 where the assertion of Lemma 5 needs to be used again in addition to give bounds in probability for a sum of dependent random variables similarly, as we have done it in the proof of Lemma 6. □

Adopting now the same argumentation as in the case of the proofs for the homoscedastic and heteroscedastic models we can use the assertion of Lemma 7 and the definition of function $\lambda_G(\cdot, \sigma(x))$ and we obtain that

$$\frac{1}{\sqrt{Nh_N}} \left| \sum_{i=1}^{N} \left[\psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^{p} \widehat{\beta}_j^{\circ} \left(\frac{X_i - x}{h_N}\right)^j\right) - \psi\left(\sigma(x)\varepsilon_i\right) + \right.\right.$$
$$\left.\left. + \lambda_G\left(\sum_{j=0}^{p} \widehat{\beta}_j^{\circ} \left(\frac{X_i - x}{h_N}\right)^j, \sigma(x)\right)\right] \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \right| \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

uniformly for $l = 0, \dots, p$. Applying the Taylor expansion of function $\lambda_G(\cdot, \sigma(x))$ in zero, using the fact that $\left|\sum_{j=0}^{p} \widehat{\beta}_j^{\circ} \left(\frac{X_i - x}{h_N}\right)\right| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right)$ and the fact that parameter estimates $\widehat{\beta}_0^{\circ}, \dots, \widehat{\beta}_p^{\circ}$ solves the set of equations (2.42) we obtain the expression

$$\begin{aligned}\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi(\sigma(x)\varepsilon_i)\cdot\left(\frac{X_i-x}{h_N}\right)^{l}K\left(\frac{X_i-x}{h_N}\right) &= \qquad (2.53)\\ &= \frac{\lambda_G'(0,\sigma(x))}{\sqrt{Nh_N}}\cdot\sum_{i=1}^{N}\sum_{j=0}^{p}\widehat{\beta}_j^{\circ}\left(\frac{X_i-x}{h_N}\right)^{j}\cdot\left(\frac{X_i-x}{h_N}\right)^{l}K_{h_N}\left(\frac{X_i-x}{h_N}\right)+o_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}}\right),\end{aligned}$$

which holds again for any $l=0,\ldots,p$. Rewriting the expression (2.53) using the matrix notation from before we finally obtain the asymptotic Bahadur representation for the vector of parameter estimates under the α-mixing dependence where

$$\widehat{\boldsymbol{\beta}}_x^{\circ} \;=\; \frac{1}{\lambda_G'(0,\sigma(x))}\cdot\left(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\right)^{-1}\cdot\mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon})+o_{\mathbf{P}}\left(1/\sqrt{Nh_N}\right), \qquad (2.54)$$

where we have assumed a regularity property of matrix $\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N$. In addition to the given notation we have $\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon})=(\psi(\sigma(x)\varepsilon_1),\ldots,\psi(\sigma(x)\varepsilon_N))^{\top}\in\mathbb{R}^N$.

In order to express the asymptotic bias term one can proceed in a very analogous way as we have already done in proving Theorem 2.3 one just needs to assume function $\lambda_G'(0,\sigma(x))$ at the pre-specified point $x\in(0,1)$ instead of function $\lambda_G'(0)$ that we were dealing with in the homoscedastic model scenario. This finishes the proof of Theorem 2.10. ■

Proof of Theorem 2.11

We will show, that the left-hand side of expression (2.53) converges for any $l=0,\ldots,p$ element-wise in law to a normal distribution with a zero mean and the corresponding variance-covariance parameter. Unlike the homoscedastic and heteroscedastic models the sum on the left-hand side of (2.53) is not over independent random variables any more therefore, we need to introduce an appropriate statistical machinery in order to deal with a sum of dependent variables.

THEOREM 2.12 (Central Limit Theorem for strongly mixing variables)

Let $\{\xi_{Ni}\}_{i=1}^{k_N}$ be a triangular scheme of random variables, which are strongly mixing with finite second moments, such that $\mathbb{E}\xi_{Ni}=0$, for all $i\in\mathbb{N}$. Let moreover, the following conditions hold

(i) $\sup_{N\in\mathbb{N}}\frac{1}{\sigma_N^2}\sum_{i=1}^{k_N}\mathbb{E}\xi_{Ni}^2<\infty$,

(ii) $\frac{1}{\sigma_N^2}\sum_{i=1}^{k_N}\mathbb{E}\left[\xi_{Ni}^2\mathbb{I}_{\{|\xi_{Ni}|>\epsilon\cdot\sigma_N\}}\right]\overset{N\to\infty}{\longrightarrow}0$,

for $\sigma_N^2=\mathrm{Var}\left[\sum_{i=1}^{k_N}\xi_{Ni}\right]$ *and any* $\epsilon>0$.
Then the following convergence in distribution holds true

$$\frac{\sum_{i=1}^{k_N}\xi_{Ni}}{\sigma_N}\;\underset{N\to\infty}{\overset{\mathscr{D}}{\longrightarrow}}\;\mathbb{N}(0,1),$$

as $N\to\infty$.

Proof of Theorem 2.12

The proof of this theorem can be found in Peligrad (1996). Some more general version of the proof can be also found in Lin and Lu (1997).

■

We want to show now that

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi(\sigma(x)\varepsilon_i)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\underset{N\to\infty}{\xrightarrow{\mathscr{D}}}\mathbb{N}(0,\sigma^2(\psi,l,x)),\tag{2.55}$$

for any $l\in\{0,\ldots,p\}$ and some appropriate variance-covariance parameter $\sigma^2(\psi,l,x)$.

We will apply Theorem 2.12 for $\xi_{iN}=\frac{1}{\sqrt{Nh_N}}\psi(\sigma(x)\varepsilon_i)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)$ and $k_N=Nh_N$, for $i=1,\ldots,N$. We need to verify the conditions of Theorem 2.12 for the sequence $\{\xi_{Ni}\}_{i=1}^N$.

Firstly, we need to realize that the sequence $\left\{\frac{1}{\sqrt{Nh_N}}\psi(\sigma(x)\varepsilon_i)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right\}_{i=1}^N$ forms a weakly dependent strictly stationary random process, which is moreover, α-mixing as well with the same α-mixing dependence coefficients as those for the original process of random errors $\{\varepsilon_i\}_{i=}^N$. This is indeed a very convenient property common for α-mixing processes. Next, the process $\{\xi_{Ni}\}_i$ is conditionally on X centered around its zero mean, which is easily implied from the symmetric property of the distribution function $G(\cdot)$ and the symmetric property of the loss function $\rho(\cdot)$. We easily obtain the finite second order property as well, which follows directly from assumption A5**. What remains is to show that conditions *(i)* and *(ii)* are also satisfied. Let us introduce the following lemma.

Lemma 8 (Davydov's inequality for α-mixing processes)
Let $\{\xi_i\}_i$ be a strictly stationary α-mixing process with mixing coefficients $\{\alpha(n)\}_n$. Let $\mathcal{F}_j^k$ denotes a σ-field generated by $\xi_j,\ldots,\xi_k$, for $j\leq k$ and let Z_1 and Z_2 be some measurable random variables with respect to $\mathcal{F}_1^k$ and $\mathcal{F}_{k+n}^\infty$ respectively. Then the following inequality holds

$$|\mathbb{C}\mathrm{ov}(Z_1,Z_2)|\leq 12\,(\alpha(n))^{\frac{\delta}{2+\delta}}\cdot\left(\mathbb{E}|Z_1|^{2+\delta}\right)^{1/(2+\delta)}\cdot\left(\mathbb{E}|Z_2|^{2+\delta}\right)^{1/(2+\delta)},\tag{2.56}$$

for some $\delta>0$. Moreover, for a measurable and bounded function $f:\mathbb{R}^k\to\mathbb{R}$ it holds that

$$|\mathbb{C}\mathrm{ov}(f(\xi_1,\ldots,\xi_k),Z_2)|\leq 12\,(\alpha(n))^{1/2}\cdot\sqrt{\mathbb{E}(Z_2)^2}\cdot\|f\|_\infty,\tag{2.57}$$

where $\|f\|_\infty=\sup_{\mathbf{x}\in\mathbb{R}^k}|f(\mathbf{x})|$.

Proof of Lemma 8

The proof of (2.56) can be found in Davydov (1970) or Ibragimov (1962). Assertion (2.57) follows easily from Davydov's inequality, using the boundedness of the density function $f(\cdot)$. □

Due to stacionarity we can easily express the variance-covariance parameter as

$$\begin{aligned}\sigma_N^2(\psi, l, x) = \mathbb{V}\mathrm{ar}\left[\xi_{N1} + \ldots, \xi_{NN}\right] = \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \psi(\sigma(x)\varepsilon_i) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right] = \\ = \mathbb{E}\left[\psi(\sigma(x)\varepsilon_1)\right]^2 \cdot \frac{1}{Nh_N} \sum_{1=1}^{N} \left(\frac{X_i - x}{h_N}\right)^{2l} K^2\left(\frac{X_i - x}{h_N}\right) + \\ + \frac{2}{Nh_N} \sum_{i=1}^{N} \sum_{\ell=1}^{N-i} \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i) \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \times \right. \\ \left. \times \; \psi(\sigma(x)\varepsilon_{i+\ell}) \left(\frac{X_{i+\ell} - x}{h_N}\right)^l K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right],\end{aligned} \tag{2.58}$$

for any $l = 0, \ldots, p$ and we will use Davydov's covariance inequality to show that the covariance expression (2.58) converges to zero for $N \to \infty$. We will firstly use the mutual independence assumption between $\{\varepsilon_i\}_i$ and $\{X_i\}_i$, for $i = 1, \ldots, N$, which allows us to express the joint expectation as

$$\begin{aligned}\mathbb{E}\left[\psi(\sigma(x)\varepsilon_i) \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \cdot \psi(\sigma(x)\varepsilon_{i+\ell}) \left(\frac{X_{i+\ell} - x}{h_N}\right)^l K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right] = \\ = \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i)\psi(\sigma(x)\varepsilon_{i+\ell})\right] \cdot \mathbb{E}\left[\left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) \left(\frac{X_{i+\ell} - x}{h_N}\right)^l K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right],\end{aligned}$$

where $\mathbb{E}\left[\psi(\sigma(x)\varepsilon_i)\psi(\sigma(x)\varepsilon_{i+\ell})\right]$ can be taken care of directly using Davydov's inequality as we have that $\mathbb{E}\psi(\sigma(x)\varepsilon_i) = 0$ for all $i = 1, \ldots, N$ therefore, it also holds that

$$\mathbb{C}\mathrm{ov}\left[\psi(\sigma(x)\varepsilon_i)\psi(\sigma(x)\varepsilon_{i+\ell})\right] = \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i)\psi(\sigma(x)\varepsilon_{i+\ell})\right] \leq 12\left(\alpha(\ell)\right)^{\frac{\delta}{2+\delta}} \cdot \mathcal{K}_1,$$

uniformly for $i = 1, \ldots, N$, for some $0 < \mathcal{K}_1 < \infty$, which follows from the stacionarity assumption and the fact that $\int_{\mathbb{R}} |\psi(\sigma(x)e)|^{2+\delta} \mathrm{d}G(e) < \infty$ for some $\delta > 0$.

Therefore, we have

$$\sum_{\ell=1}^{N-i} \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i) \cdot \psi(\sigma(x)\varepsilon_{i+\ell})\right] \leq \sum_{\ell=1}^{N} \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i) \cdot \psi(\sigma(x)\varepsilon_{i+\ell})\right] \leq 12 \cdot \mathcal{K}_1 \cdot \sum_{\ell=1}^{N} \left(\alpha(\ell)\right)^{\frac{\delta}{2+\delta}},$$

and hence, by reversing the order of the sum operators and using an appropriate re-indexing we obtain for the variance-covariance term the expression

$$\begin{aligned}\sigma_N^2(\psi, l, x) \leq \mathbb{E}\left[\psi(\sigma(x)\varepsilon_1)\right]^2 \cdot \frac{1}{Nh_N} \cdot \sum_{i=1}^{N} \left(\frac{X_i - x}{h_N}\right)^{2l} K^2\left(\frac{X_i - x}{h_N}\right) + \\ + \frac{24 \cdot \mathcal{K}_1}{Nh_N} \cdot \sum_{\ell=1}^{N-1} \left(\alpha(\ell)\right)^{\frac{\delta}{2+\delta}} \cdot \sum_{i=1}^{N-\ell} \mathbb{E}\left[\left|\frac{X_i - x}{h_N}\right|^l \left|\frac{X_{i+\ell} - x}{h_N}\right|^l K\left(\frac{X_i - x}{h_N}\right) K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right],\end{aligned} \tag{2.59}$$

which clearly holds for any $l = 0, \ldots, p$.

Let now $f_\ell(\cdot,\cdot)$ be a joint density function of random variables X_1 and $X_{1+\ell}$. Given assumption A1** we have that its bounded for all $\ell > 1$ therefore, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{X_i - x}{h_N}\right|^l \left|\frac{X_{i+\ell} - x}{h_N}\right|^l K\left(\frac{X_i - x}{h_N}\right) K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right] = \\
&\quad = \int_{x-h_N}^{x+h_N}\int_{x-h_N}^{x+h_N} \left|\frac{u_1 - x}{h_N}\right|^l \left|\frac{u_2 - x}{h_N}\right|^l K\left(\frac{u_1 - x}{h_N}\right) K\left(\frac{u_2 - x}{h_N}\right) \cdot f_\ell(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2 = \\
&\quad = h_N^2 \int_{-1}^{1}\int_{-1}^{1} |v_1|^l |v_2|^l K(v_1)K(v_2) f_\ell(x + h_N v_1, x + h_N v_2)\mathrm{d}v_2\mathrm{d}v_2 \leq \\
&\quad \leq \mathcal{K}_2 \cdot h_N^2,
\end{aligned}$$

uniformly for $i = 1, \ldots, N$ and any $l \in \{0, \ldots, p\}$ where $0 < \mathcal{K}_2 < \infty$.

Therefore, we finally obtain for 2.59 that it holds that

$$\begin{aligned}
&\frac{24 \cdot \mathcal{K}_1}{N h_N} \cdot \sum_{\ell=1}^{N-1} (\alpha(\ell))^{\frac{\delta}{2+\delta}} \cdot \sum_{i=1}^{N-\ell} \mathbb{E}\left[\left|\frac{X_i - x}{h_N}\right|^l \left|\frac{X_{i+\ell} - x}{h_N}\right|^l K\left(\frac{X_i - x}{h_N}\right) K\left(\frac{X_{i+\ell} - x}{h_N}\right)\right] \leq \\
&\qquad\qquad \leq \frac{24 \cdot \mathcal{K}_1}{N h_N} \cdot \sum_{\ell=1}^{N} (\alpha(\ell))^{\frac{\delta}{2+\delta}} \cdot \sum_{i=1}^{N} \mathcal{K}_2 h_N^2 \underset{N\to\infty}{\longrightarrow} 0,
\end{aligned}$$

due to the fact that $h_N \cdot \sum_i (\alpha(i))^{\delta/(2+\delta)} \to 0$ for $N \to \infty$.

Moreover, for the sum of the second moments of ξ_{Ni} we obtain

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}\xi_{Ni}^2 &= \frac{1}{N h_N} \sum_{i=1}^{N} \mathbb{E}\left[\psi(\sigma(x)\varepsilon_i) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right]^2 \leq \\
&\leq \frac{\mathcal{K}_3}{N h_N} \sum_{i=1}^{N} \mathbb{E}\left[\left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)\right]^2 \leq \mathcal{K}_3^* h_N \underset{N\to\infty}{\longrightarrow} 0,
\end{aligned} \tag{2.60}$$

for some constants $\mathcal{K}_3 > 0$ and $\mathcal{K}_3^* > 0$ and any $N \in \mathbb{N}$ therefore, conditions *(i)* and *(ii)* are also both easily satisfied.

Using now the matrix notation it is easy to see that

$$\sigma_N^2(\psi, l, x) \underset{N\to\infty}{\longrightarrow} \sigma^2(\psi, l, x) \overset{def.}{=} \mathbb{E}\left[\sigma(x)\psi(\varepsilon_1)\right]^2 \cdot f(x) \cdot \mathsf{S}_2,$$

where matrix S_2 is the same as in Theorem 2.5 or 2.8 respectively.

To finish the proof of Theorem 2.11 one just needs to use the same arguments and an analogous computations as in the proof of Theorem 2.4 and the result follows immediately therefore, we can omit the rest of the proof. ■

We would like to note that the Central Limit Theorem for triangular arrays was required here mainly because of an adaptive bandwidth parameter h_N used for the minimization. Indeed, once we adopt a constant bandwidth approach ($h_N \equiv h > 0$ for all $N \in \mathbb{N}$) it is sufficient to use a standard Central Limit Theorem for functional data more specifically, its special version related to a weak invariance principle (WIP) and a convergence to a standard Wiener process $\mathcal{W}(t)$ for a fixed time $t = 1$.

Using an asymptotic bandwidth parameter h_N, which depends on the sample size $N \in \mathbb{N}$ one introduces more or less the same amount of flexibility and smoothness even when the sample size increases as the bandwidth parameter flexibly adapts to correspond with the given data sample. On the other hand one needs to pay more attention to derive the proper statistical inference using such asymptotic bandwidth parameter h_N.

However, given the fixed bandwidth parameter $h > 0$ the covariance quantities would not reduce so easily and the variance-covariance quantity would be (asymptotically) equal to

$$\sigma_N^2(\psi, l, x) = \mathbb{E}\left[\psi(\sigma(x)\varepsilon_1)\right]^2 + 2\sum_{\ell=1}^{\infty} \mathbb{E}\left[\psi(\sigma(x)\varepsilon_1)\psi(\sigma(x)\varepsilon_{\ell+1})\right],$$

where the sum of covariances would not tend to zero any more for $N \to \infty$. This is caused by the fact that the fixed bandwidth parameter do not reduce the dependence effect as the number of observations increases therefore, this also needs to be reflected in the variance-covariance expression.

Referring to the bandwidth parameter, let us just mention that there are also statistical approaches proposed for nonparametric regression techniques where so called adaptive bandwidth parameter is used instead of h_N, or h respectively, which are both the same along the whole domain of interest.
The adaptive bandwidth parameter is more suitable especially in situations where the presence (the amount) of measurements (observations respectively) markedly differs along the domain of interest. We have however, not considered such cases in this thesis.

It would be also interesting to show analogous results for the φ-mixing dependence concept. We have already said that the uniformly strong mixing condition implies the strong mixing therefore, by proving the results for φ-mixing dependence one would obtain slightly stronger results, which could be easily adopted for α-mixing concept as well. The proofs of analogous results for the φ-mixing dependence follow very the same idea as we have presented in this chapter for the case of α-mixing dependence however, some additional assumptions need to be considered for such proofs to hold.

Proving all theorems stated under the weak dependence concept we have completed our study of the local polynomial M-smoothers estimates under the variety of different situations. We have also proved that using an appropriate regularization assumptions one can get the same asymptotic performance of the M-smoothers estimators for independent as well as dependent data specifically, the strong mixing dependence. Basically, we have covered most common situations, which can be commonly considered for nonparametric regression scenarios: homoscedasticity as well as heteroscedasticity and independent random errors as well as dependent random errors too. We should now have sufficient robust modelling tools in hand to handle most of all practical situations.

On the final note, there is also a way to extend the results to account for another concept of dependent data, which we have not discussed yet, the ϕ-mixing dependence. The results can be however, derived in a quite analogous way as well.

2.5 Discussion on M-smoothers

We have proposed and discussed a set of very flexible statistical tools for a nonparametric regression modelling where the flexibility in the model is approached from different angles at the same time. Firstly, applying the local polynomial estimation techniques we account for a huge class of various regression functions, which can be possibly allowed to be considered by the estimation approach while yielding a consistent estimate. Moreover, asymptotic normality property is achieved as well.

From another point of view, using the M-estimation techniques (or so called robust approaches respectively) we extend the flexibility of the model into another direction as we allow for some presence of outlying observations or even heavy tailed distributions of random errors. This is not the case for the classical L_2-norm based estimation methods, which would fail once some outliers are present. Moreover, the performance of the L_2-norm based approaches in case of heavy-tailed random error distributions is even roughly inconsistent and such approaches are definitely not suitable unless the normality condition for the random error terms is assumed in advance.

Finally, the flexibility in modelling is approached from one more direction here – an independence assumption. By introducing M-smoothers together with the weak dependence concept we extent the flexibility of the model with respect to the set of assumptions even further more.

Local polynomial M-smoothers have a great potential to be applied in real situation mainly because of two nice properties they inherit: a huge amount of flexibility while staying at a reasonable level with a set of required conditions and on the other hand, assuring nice statistical properties (consistency and asymptotic normality), which makes them usable in real data situations.

There is however, one more aspect of the flexibility in the final model, which have been not addressed yet: a flexibility with respect to the smoothness assumption posed on the unknown regression function and its derivatives respectively. We will discuss an extension with respect to such smoothness flexibility in the next chapter, and we will generalize the class of admissible regression functions as we slightly relax the assumption related to the Lipschitz property considered so far.

*"To improve is to change.
To be perfect is to change often."*

Winston Churchill
(1874 – 1965)

3 Change-point models

As we already have all necessary background for the M-smoothers estimation we can now describe a model where the smoothness assumption posed in the assumption A3 is relaxed a little as we will allow also a model with some discontinuity points – change-point occurrences respectively. To be more specific, one should rather refer to change-points as to structural breaks instead as we allow for discontinuity points not only in the unknown regression function itself but also in all order derivatives (up to the order of the local polynomial approximation given by $p \in \mathbb{N}$). This implicitly means that we take into account sudden changes not only in a location parameter but also in direction, curvature and all higher order levels of topology as well. One has to realize here that except the location and direction changes all higher order discontinuity points are barely visible with a naked eye and any further interpretation of such changes requires a huge amount of imagination at first.

Unlike some other authors and papers related to this topic, in our approach we want to stay fully free of any assumptions given with respect to some ties among discontinuity points. There are many methods proposed where there is a kind of hierarchy considered between jump occurrences. Specifically, if there is a structural break (jump) considered in the first derivative – a change in direction – there is automatically a break (jump) assumed also in all lower order derivatives – in this case a change in location. However, this will not be the case in our approach as our methods can perfectly handle a situation if there is (e.g.) a sudden change in curvature only (jump in the second derivative) while all lower and higher order derivatives can stay nicely continuous. This is, in our opinion a more realistic scenario therefore, there are no ties considered between change-point occurrences in our approach at all.

We will distinguish two different options in this chanper. Firstly we will discuss a model where all considered change-points (structural breaks respectively) are given in advance hence, we have an a priori knowledge where the jump locations could occur. On the other hand, in the second scenario we assume no a priori knowledge about the change-point occurrences at all therefore, jump locations are treated as unknown parameters and they are subject of the whole estimation process as well.

We will mostly focus on change-point models under the heteroscedastic model assumption. The homoscedastic scenario follows easily as a special case of the heteroscedastic model and change-point models under the weak dependence assumption can be derived in an analogous way as well however, slightly more technical proofs are required to handle the given dependence concept.

3.1 Inconsistency with classical M-smoothers approach

Let us firstly propose a situation where we will consider a model with some change-point occurrences but the method which is used to estimate the unknown regression function (possibly discontinuous) or its derivatives respectively (possibly discontinuous as well) will remain as before – with no additional adaptation. We will show that using the proposed M-smoothers approach together with discontinuity assumptions however, with no additional modification will cause a serious inconsistency and it will introduce a systematic bias term, which can not diminish even if the sample size tends to infinity.

Indeed, just a common sense dictates that once the regression function $m(\cdot)$ is discontinuous at some point $x_0 \in (0,1)$ while we try to impose a continuous estimation at this point even though then there necessarily has to occur some imprecision involved at least within a small neighbourhood $(x_0 - h_N, x_0 + h_N)$ of $x_0 \in (0,1)$ for $h_N > 0$ to be the given bandwidth parameter. To be more specific, this imprecision can be expressed explicitly and it is given by the theorem below.

THEOREM 3.1 (Inconsistency of M-smoothers in change-point locations)
Let the assumption of Theorem 2.8 are satisfied and moreover, we assume that the regression function $m(\cdot)$ has a discontinuity point at $x_0 \in (0,1)$. Then the proposed M-smoothers estimator introduces a systematic bias within the point x_0 and it holds that

$$\sqrt{Nh_N} \cdot (\widehat{m}(x_0) - m(x_0)) \underset{N \to \infty}{\xrightarrow{\mathscr{D}}} \mathbb{N}\left(\frac{\Delta \cdot \mathbf{e}_0^\top \boldsymbol{\mu}_0}{\mathbf{e}_0^\top \mathsf{S}_1 \mathbf{e}_0}, \frac{\mathbb{E}\psi^2(\sigma(x_0)\varepsilon_1)}{[\lambda'_G(0, \sigma(x_0))]^2 f(x_0)} \cdot \mathbf{e}_0^\top \mathsf{S}_1^{-1} \mathsf{S}_2 \mathsf{S}_1^{-1} \mathbf{e}_0 \right),$$

where $\boldsymbol{\mu}_0 = \left(\int_0^1 u^0 K(u) du, \ldots, \int_0^1 u^p K(u) du \right)^\top \in \mathbb{R}^{p+1}$ and $\Delta \neq 0$ stands for the corresponding size of a jump at the given point $x_0 \in (0,1)$.

Proof. We will formulate some important issues of the proof only. The rest of the proof follows easily according to the proofs for the heteroscedastic model (see Section 2.3.2).

Assuming the model with a discontinuity point at $x_0 \in (0,1)$ we can express the model relationship as $Y_i = m(X_i) + \Delta \cdot \mathbb{I}_{\{X_i > x_0\}} + \sigma(X_i)\varepsilon_i$, for $i = 1, \ldots, N$ and $\Delta \neq 0$ to be the size of the jump.
Let us consider the minimization problem (2.1) for $x = x_0$. Then the analogy of (2.42) leads us in this case to the set of equations

$$\frac{1}{\sqrt{Nh_N}} \sum_{i=1}^{N} \psi \left(\sigma(X_i)\, \varepsilon_i + \mathbb{I}_{\{X_i > x_0\}} - \sum_{j=0}^{p} b_j \left(\frac{X_i - x_0}{h_N} \right)^j \right) \cdot \left(\frac{X_i - x_0}{h_N} \right)^l K\left(\frac{X_i - x_0}{h_N} \right) = 0,$$
for $l = 0, \ldots, p,$

which is solved for the vector of parameter estimates $\widetilde{\boldsymbol{\beta}}_{x_0}^{\circ} = (\widetilde{\beta}_0^{\circ}, \ldots, \widetilde{\beta}_p^{\circ})^\top \in \mathbb{R}^{p+1}$. Using the same reasoning as in the proof of Theorem 2.8 and the sequence of analogous steps one can easily obtain an asymptotic conditional representation of the vector of parameter estimates as

$$\frac{1}{\sqrt{Nh_N}\, \lambda'_G(\sigma(x_0)\varepsilon_1)} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(x_0)\boldsymbol{\varepsilon}) = \frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N \widetilde{\boldsymbol{\beta}}_{x_0}^{\circ} + o_{\mathbf{P}}\left(\frac{1}{\sqrt{Nh_N}} \right) + \\ + \frac{\Delta}{\sqrt{Nh_N}} \sum_{\substack{i=1 \\ X_i > x_0}}^{N} \left(\frac{X_i - x_0}{h_N} \right)^l K\left(\frac{X_i - x_0}{h_N} \right), \qquad (3.1)$$

for any $l = 0, \ldots, p$, where the notation is the same as in Chapter 2. It is easy to verify now that

$$\mu_l^{(N)} \overset{def.}{=} \frac{1}{Nh_N} \sum_{\substack{i=1 \\ X_i > x_0}}^{N} \mathbb{E}\left[\left(\frac{X_i - x_0}{h_N} \right)^l K\left(\frac{X_i - x_0}{h_N} \right) \right] \underset{N \to \infty}{\longrightarrow} f(x_0) \cdot \int_0^1 y^l K(y) \mathrm{d}y,$$

for any $l = 0, \ldots, p$ therefore, $\mu_l^{(N)} \longrightarrow f(x_0)\widetilde{\mu}_l$ as $N \to \infty$ where $\boldsymbol{\mu}_0 = (\widetilde{\mu}_0, \ldots, \widetilde{\mu}_p)^\top \in \mathbb{R}^{p+1}$. The results now follows from the definition of matrix S_1 (see the proof of Theorem 2.3). ■

Corollary 1

An analogous theorem to Theorem 3.1 can be formulated in case of higher order discontinuities as well. Once the quantity $\Delta \neq 0$ reflects the size of a jump in some derivative of the unknown regression function $m(\cdot)$ at the point $x_0 \in (0,1)$, the systematic bias result for $\sqrt{Nh_N^{1+2\nu}}\left(\widehat{m}^{(\nu)}(x_0) - m^{(\nu)}(x_0)\right)$ is obtained once we replace vector $\mathbf{e}_0$ in the limit distribution by the corresponding vector $\mathbf{e}_\nu$.

Theorem 3.1 states the asymptotic bias imprecision only therefore, it is only defined for the point $x_0 \in (0,1)$ where the jump occurs. However, using a finite sample approximation one will experience a systematic bias not only at the point $x_0 \in (0,1)$ but in some small neighbourhood $(x_0 - h_N, x_0 + h_N)$ around this point as well for some fixed value of the bandwidth parameter $h_N > 0$. In this case the systematic bias imprecision can be approximated by an appropriately standardized sum in (3.1).

3.2 Change-points known in advance

Let us start with the simplest case where all possible change-point occurrences (jump locations) in the model are known in advance, which means we do have an a priori knowledge about all possible candidates where the structural breaks could occur. This can be given by some previous exploration or the nature of an experiment or just by a subjective opinion of a statistician.

3.2.1 Single change-point model

The simplest scenario to consider is a model where only one change-point is present in the model. We will assume it is a jump located in the unknown regression function itself while its position is well-known in advance. Considering such model we would like to investigate the main statistical properties for such generalization. Indeed, change-point models follow as an obvious generalization of models discussed in Chapter 2 as the Lipschitz assumption is slightly released and some discontinuity points are allowed to be a part of the model. Some further generalizations into models with structural breaks (higher order discontinuities) or models with multiple change-points follow easily and we will mention them later on.

Let again $\{(X_i, Y_i);\ i = 1, \ldots, N \in \mathbb{N}\}$ be a random sample given from some unknown joint distribution function $F_{(X,Y)}(x,y)$ where the decomposition

$$Y_i = m(X_i) + \sigma(X_i) \cdot \varepsilon_i, \qquad \varepsilon_i \sim G, \ \ i.i.d.,\ i = 1, \ldots, N, \tag{3.2}$$

is assumed to hold, where for function $m(\cdot)$ we assume that

$$m(X_i) = m_0(X_i) + \Delta \cdot \mathbb{I}_{\{X_i > x_0\}}, \ \text{ for some } \ \Delta \in \mathbb{R} \setminus \{0\}.$$

The point $x_0 \in (0,1)$ is chosen in advance and function $m_0(\cdot)$ satisfies assumption A3. Moreover, function $\mathbb{I}_{\{\cdot\}}$ stands for an identifier function of the given event of interest.

For model (3.2) we will assume the same set of assumptions as those required for the heteroscedastic model[26] (see p.35) except assumption A3 related to the $(p+1)$-order Lipschitz property, which is now assumed to hold for function $m_0(\cdot)$ rather than $m(\cdot)$.

[26] In case of chance-point models considered under the homoscedastic scenario one only needs to consider assumptions derived for the homoscedastic model. Similarly, if we adopt a change-point model under the weak dependence concept the corresponding weak dependence assumptions will be required for the results to hold.

Given such model the point $x_0 \in (0,1)$ is called the change-point (jump location respectively) and $\Delta \neq 0$ is some unknown size of a jump, which is a subject of the estimation process too. Under this model scenario one can be interested in many different aspects of the model e.g. the estimation of the unknown regression function, the estimation of the smooth function $m_0(\cdot)$ or the estimation of the size of the jump. One has to be aware of the fact that the jump location $x_0 \in (0,1)$ is assumed to be known in advance in this case therefore, there is no need to estimate the location of the change-point. In this chapter we will however, rather focus on statistical procedures, which are used to "confirm" (or reject) an existence of such change-point occurrence and we will discuss some important issues related to the statistical hypothesis testing problems.

The idea is to propose a statistical test, which will be suitable to decide whether the jump at the point $x_0 \in (0,1)$ is statistically significant or not. We propose the following pair of hypothesis

$$\left.\begin{array}{ll} \mathsf{H}_0: & \Delta = 0 \\ \mathsf{H}_1: & \Delta \neq 0 \end{array}\right\} \text{ for given } x_0 \in (0,1) \tag{3.3}$$

and we want to find an appropriate test statistic and the respective limit distribution of this test statistic under the null hypothesis in order to draw critical values required for the test to decide. We want the test statistic to be sensitive at possible jump locations therefore, we propose to use the following quantity

$$T_N(x_0) = \sqrt{N h_N} \cdot |\widehat{m}_+(x_0) - \widehat{m}_-(x_0)|, \tag{3.4}$$

where $\widehat{m}_+(x_0)$ and $\widehat{m}_-(x_0)$ respectively are just one-sided kernel estimates given by the minimization problem (2.1) where the original kernel function $K(\cdot)$ is replaced here by the corresponding one-sided counterparts $K^+(\cdot) = K(\cdot)\mathbb{I}_{\{\cdot\ \geq 0\}}$ and $K^-(\cdot) = K(\cdot)\mathbb{I}_{\{\cdot\ \leq 0\}}$ respectively.

The idea behind introducing one-sided kernels is to produce one-sided M-smoothers estimates (the left-hand side estimate and the right-hand side estimate) of the unknown regression function $m(\cdot)$ or its derivatives respectively. Indeed, once we construct one-sided M-smoothers estimates and once we prove their consistency we can easily use the difference between the left-hand side and the right-hand side estimate (appropriately standardized) as the corresponding test statistic as it consistently estimates the size of the jump, which is of interest.

However, unlike the M-smoothers estimator and the corresponding set of assumptions described in Chapter 2, one sided kernels, which are introduced here are not symmetric any more. We will firstly show that the asymptotic performance of the M-smoothers estimators based on one-sided kernel functions $K^+(\cdot)$ and $K^-(\cdot)$ is actually the same as the performance based on the symmetric kernel $K(\cdot)$.

THEOREM 3.2 (Asymptotic normality for one sided estimates)
For model (3.2), the given point $x \in (0,1)$ such that $x \neq x_0$ and assumptions A1 – A8, with A3 satisfied for $m_0(\cdot)$ and A1, A2 and A5 replaced by A1, A2* and A5* it holds that*

$$\sqrt{N h_N} \cdot \left(\widehat{m}_+^{(\nu)}(x) - m_+^{(\nu)}(x) - \mathbb{B}\text{ias}\left[\widehat{m}_+^{(\nu)}(x)\right]\right) \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{\nu! \mathbb{E}\left[\psi^2(\sigma(x)\varepsilon_1)\right]}{[\lambda'_G(0,\sigma(x))]^2 f(x)} \mathbf{e}_\nu^\top \underline{\mathsf{S}}_1^{-1} \underline{\mathsf{S}}_2 \underline{\mathsf{S}}_1^{-1} \mathbf{e}_\nu\right),$$

where matrices $\underline{\mathsf{S}}_1$ and $\underline{\mathsf{S}}_2$ are the same as those in Theorem 2.8 just the kernel function $K(\cdot)$ is replaced by the corresponding one-sided counterpart $K^+(\cdot)$ and $\nu \in \{0, \ldots p\}$ stands for the order of a derivative of the regression function $m(\cdot)$ or its estimate respectively.

Proof. The idea of the proof is the same as in the proof of Theorem 2.8 some minor differences however, needs to be taken care of with respect to different, non-symmetric kernel function $K^{+}(\cdot)$.

Firstly, one has to be aware of a different structure of the kernel function used here. Indeed, unlike symmetric kernels where we had $\int u^{k}K(u)\mathrm{d}u = 0$ for any odd $k \in \{1,\ldots,p\}$, this is not the case for one-sided kernels any more. Additionally, for a general kernel function $K(\cdot)$ we can not specify the mean value $\int uK^{+}(u)\mathrm{d}u$, which now also depends on the given kernel however, we would like to work with standardized kernels in order to simplify further formulations. This can be nicely solved by introducing a slightly modified assumption A6 where

A6† Let $K^{+}(\cdot) = K(\cdot)\mathbb{I}_{\{\cdot\ \geq 0\}}$, where $K(\cdot)$ is a symmetric kernel from assumption A6. Moreover, we assume that $\int_{\mathbb{R}} K^{+}(u)\mathrm{d}u \int_{\mathbb{R}} u^{2}K^{+}(u)\mathrm{d}u - \left(\int_{\mathbb{R}} uK^{+}(u)\mathrm{d}u\right)^{2} = 1$ and it holds that $K^{+}(0) \neq 0$;

An analogous condition needs to be satisfied for the left-hand side kernel function $K^{-}(\cdot)$ as well and we also assume a mutual symmetric property where we have $K^{+}(u) = -K^{-}(-u)$ for all $u \in \mathbb{R}$.

Let us shortly motivate the assumption $K^{+}(0) \neq 0 \neq K^{-}(0)$: as we are primarily interested in estimation of the unknown regression function $m(\cdot)$ or its derivatives respectively form the right-hand side or the left-hand side respectively, it is important to assign some positive weights to points, which are close from the right-hand side the left-hand side of the point of interest $x \in (0,1)$ where the kernel function is seated on.

Under the modified assumption A6† the proof of Theorem 3.2 goes now along the lines of the proof of Theorem 2.8 therefore, we will omit the rest. ■

Corollary 2
A quite analogous theorem can be also formulated and proved to state the asymptotic normality result for the left-hand side M-smoothers estimates. The proof of such theorem is analogous to the proof of Theorem 3.2.

We have shown that using one-sided kernels is not an issue here and that the results hold quite analogously for one-sided estimates as well. This will be the key point in the next proofs related to the testing problem defined by (3.3) and based on the test statistic (3.4).

One can expect quantities $\widehat{m}_{+}(x_0)$ and $\widehat{m}_{-}(x_0)$ in (3.4) to be nonparametric estimates of the corresponding theoretical values $m_{+}(x_0) = \lim_{y \searrow x_0} m(y)$ and $m_{-}(x_0) = \lim_{y \nearrow x_0} m(y)$ therefore, such test statistics should be sensitive at a possible jump location and one would expect it to be large if there is a change-point present at the given point $x_0 \in (0,1)$ and it should be negligible if there is not. Given this behaviour, large values of $T_N(x_0)$ should cause rejection of the null hypothesis against the alternative one and small values of the test statistic $T_N(x_0)$ should not reject the null hypothesis.

A proper statistical criterion is given by the corresponding critical value $t_{x_0}(\alpha)$ where

$$\textit{iff} \qquad |T_N(x_0)| > t_{x_0}(\alpha) \qquad \Longrightarrow \qquad \text{reject } \mathrm{H}_0,$$

where $\alpha \in (0,1)$ is a given level of the test, which restricts the probability of the second type error in the test. In order to know what distribution to use for drawing critical values let us now state the main distributional results for this test based on the test statistic $T_N(x_0)$.

THEOREM 3.3 (Asymptotic distribution of the test statistics under the null hypothesis)
Let the assumptions of Theorem 3.2 are fulfilled. Then the asymptotic distribution of the test statistic $T_N(x_0)$ *under the null hypothesis* H_0 *from* $\boxed{3.3}$ *is given by*

$$\sqrt{Nh_N} \cdot [\widehat{m}_+(x_0) - \widehat{m}_-(x_0)] \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{2\mathbb{E}\left[\psi^2(\sigma(x_0)\varepsilon_1)\right]}{[\lambda_G'(0,\sigma(x_0))]^2 f(x_0)} \mathbf{e}_0^\top \underline{\mathsf{S}}_1^{-1} \underline{\mathsf{S}}_2 \underline{\mathsf{S}}_1^{-1} \mathbf{e}_0\right),$$

where matrices $\underline{\mathsf{S}}_1$ *and* $\underline{\mathsf{S}}_2$ *are the same as in Theorem 3.2 and* $\mathbf{e}_0 = (1,0,\ldots,0)^\top \in \mathbb{R}^{p+1}$.

Proof. Given Theorem 3.2 one just needs to realize that the same result holds also for an asymptotic behaviour of the quantity $\sqrt{Nh_N} \cdot [\widehat{m}_-(x) - m_-(x)]$, where in turn the one-sided kernel function $K^+(\cdot)$ is replaced by its mirror version $K^-(\cdot)$. Moreover, both estimates $\widehat{m}_+(x_0)$ and $\widehat{m}_-(x_0)$ are independent of each other as they are computed from two separate parts of data therefore, one just needs to use the generic property of the normal distribution and the result already follows directly. ■

THEOREM 3.4 (Consistency of the test)
For model $\boxed{3.2}$*, assumptions as in Theorem 3.3 and the alternative hypothesis* H_1 *from* $\boxed{3.3}$ *the following convergence in probability holds true*

$$\sqrt{Nh_N} \cdot |\widehat{m}_+(x_0) - \widehat{m}_-(x_0)| \xrightarrow[N \to \infty]{\mathbf{P}} \infty.$$

Proof. Under the alternative hypothesis H_1 it holds that $[\widehat{m}_+(x_0) - \widehat{m}_-(x_0)] \xrightarrow{\mathbf{P}} \Delta \neq 0$ so the result holds under the assumption A8 as $\sqrt{Nh_N} \to \infty$ for $N \to \infty$. ■

Given these two theorems we are now fully competent to use the proposed statistical test in order to decide if some expected change (jump) at some given point $x_0 \in (0,1)$ in the unknown regression function $m(\cdot)$ really occurs or if it does not.

Another question that could arise here is a capability of computing the limit distribution under the null hypothesis given by Theorem 3.3, which should be use to obtain critical values, which are of the key importance in making the test decisions. We already know that this distribution is normal however, the variance term of this distribution depends on some unknown quantities. One will never know in real situations how does the scale function $\sigma(\cdot)$ look like, or what is the true distribution function $G(\cdot)$ of random errors or what is the real shape of the design density function $f(\cdot)$ either. One way to deal with this problem is to use appropriate empirical estimates for these quantities based on the random sample $\{(X_i, Y_i);\ i = 1, \ldots, N\}$ and plugging them in to the variance expression in Theorem 3.3 we should obtain a reasonable approximation of the true but unknown asymptotic distribution under the null hypothesis H_0. However, such plug-in approximations give rather slow convergence and mostly they are also too much time expensive therefore, we will consider a bootstrap based simulations in Chapter 4, which can handle this situation in a much more convenient way.

3.2.2 Multiple change-point models

Let us now discuss a more complex generalizations of the model defined in $\boxed{3.2}$ especially, we will focus on a model with multiple change-point occurrences and a model with jumps of different hierarchy levels.

In order to implement multiple change-points into the model one just needs to understand the basic structure of the model (3.2) where $x_0 \in (0,1)$ stands for a change-point location and $\Delta \neq 0$ stands for the size of a jump. Let us define a sequence of change-point locations $0 < x_1 < x_2 < \cdots < x_n < 1$, for some $n \in \mathbb{N}$ such that $|x_i - x_j| > 2h_N$ for $\forall i \neq j$. This is assumed to assure that two neighbouring change-points are not too much close to each other and to make sure that the unknown regression function $m(\cdot)$, which is continuous between these jumps locations is still estimable using a local approach given by the bandwidth parameter $h_N > 0$. This will also allow us to consider a set of tests to be mutually independent as each test statistic is actually computed from some other (unique) part of data. We define a model with multiple change-points (in location) as follows:

$$Y_i = m(X_i) + \sigma(X_i) \cdot \varepsilon_i, \quad \varepsilon_i \sim G, \ \textit{i.i.d.}, \ \text{for } i = 1, \ldots, N, \tag{3.5}$$

where for the unknown regression function $m(\cdot)$ it holds that

$$m(X_i) = m_0(X_i) + \sum_{k=1}^{n} \Delta_k \cdot \mathbb{I}_{\{X_i > x_k\}}, \ \text{where } \Delta_k \in \mathbb{R} \setminus \{0\}, \quad \forall k = 1, \ldots, n, \ \text{and } n \in \mathbb{N},$$

where again function $m_0(\cdot)$ is assumed to satisfy assumption A3. One can define even more general models where there are jumps assumed in the regression function itself and also in its derivatives.
For simplicity, let us define a set of polynomial basis functions $\mathscr{R}^p = \{\varphi_0, \ldots, \varphi_p\}$, where $\varphi_j(x) = x^j$, for $j = 0, \ldots, p$ and $p \in \mathbb{N}$ being the order of the local polynomial approximation. Under this notation we can formulate a very general change-point model using the following three steps:

① $Y_i = m(X_i) + \sigma(X_i) \cdot \varepsilon_i$, where $\varepsilon_i \sim G$, are *i.i.d.*, for $i = 1, \ldots, N$;

② $m(X_i) = m_0(X_i) + \sum_{k=1}^{n} \Delta_k \cdot \varphi_k(X_i) \cdot \mathbb{I}_{\{X_i > x_k\}}$, where $m_0 \in \mathscr{L}_{p+1}(0,1)$;

③ $\varphi_k \in \mathscr{R}^p$, and $\Delta_k \in \mathbb{R} \setminus \{0\}$ for $\forall k = 1, \ldots, n$;

Under this model definition we implicitly allow all possible structural breaks – sudden changes to occur in the model (jumps of different hierarchy level in the unknown regression function itself and all order derivatives, up to the given order of the local polynomial approximation given by the value of $p \in \mathbb{N}$).

However, in all these models the jump locations (structural breaks) are given in advance and the only common hypothesis testing problem for such model scenarios is to test if some jump at the given location $x_k \in (0,1)$ for some $k \in \{1, \ldots, n\}$ is significant or if it is not. The global hypothesis problem can be introduced via a set of n pairs of hypothesis where

$$\left.\begin{array}{ll} \mathsf{H}_0^{(k)}: & \Delta_k = 0 \\ \mathsf{H}_1^{(k)}: & \Delta_k \neq 0 \end{array}\right\} \ \text{for } k = 1, \ldots, n, \tag{3.6}$$

where one needs to use an appropriate test statistic and its corresponding asymptotic limit distribution derived for the particular level of the discontinuity level he wants to test for.

As far as we have derived only a test for testing change-point occurrences in the regression function itself it is still left to be shown that a similar approach also applies for statistical tests for testing the significance of higher order structural breaks – jump occurrences in the corresponding derivatives of the unknown regression function $m(\cdot)$. However, this can be easily derived using already known facts.

Indeed, from Theorems 2.8 and 3.3 one can easily obtain the asymptotic distributional property for a general test statistic

$$T_N^{(\nu)}(x_k) = \sqrt{N h_N^{1+2\nu}} \cdot \left| \widehat{m}_+^{(\nu)}(x_k) - \widehat{m}_-^{(\nu)}(x_k) \right|, \ \text{for any } k = 1, \ldots, n \text{ and } \nu = 0, \ldots, p, \tag{3.7}$$

where quantities $\widehat{m}_{+}^{(\nu)}(x_k)$ and $\widehat{m}_{-}^{(\nu)}(x_k)$ are just the corresponding one-sided estimates of one-sided derivatives of the unknown regression function $m(\cdot)$, defined by an analogous minimization problem as $\boxed{2.1}$ where the kernel function $K(\cdot)$ is replaced by its one-sided counterparts.

One naturally assumes that the quantity in $\boxed{3.7}$ will be sensitive at those points where some change-points in the ν^{th} derivative of the unknown regression function really occur. The quantity in $\boxed{3.7}$ is therefore e used as a suitable candidate for the test statistic. The next theorem follows just as a straightforward extension of the previous assertions and conclusions.

THEOREM 3.5 (Distribution of the general test statistics under the null hypothesis)

Let the assumptions of Theorem 3.3 are fulfilled. Then the general test statistic $T_N^{(\nu)}(x_k)$ follows under the null hypothesis in limit in law a normal distribution given by

$$\sqrt{Nh_N^{1+2\nu}} \cdot \left(\widehat{m}_{+}^{(\nu)}(x_k) - \widehat{m}_{-}^{(\nu)}(x_k)\right) \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}\left(0, \frac{2\nu!\mathbb{E}\left[\psi^2(\sigma(x_k)\varepsilon_1)\right]}{[\lambda_G'(0,\sigma(x_k))]^2 f(x_k)} \mathbf{e}_\nu^\top \underline{\mathsf{S}}_1^{-1} \underline{\mathsf{S}}_2 \underline{\mathsf{S}}_1^{-1} \mathbf{e}_\nu\right),$$

for any $\nu = 0, \ldots, p$ and some $k \in \{1, \ldots, n\}$, where $\nu \in \{0, \ldots, p\}$ stands for the corresponding hierarchy level of the discontinuity point to be considered.

Proof of Theorem 3.5

The proof of this theorem follows easily as a straightforward generalization of the proof of Theorem 3.3 while the same argumentation as above holds here as well.

■

Considering now the set of $n \in \mathbb{N}$ pairs of hypothesis $\boxed{3.6}$ we need to be aware of the fact that by providing a separate statistical test for each pair of the hypothesis for $k = 1, \ldots, n$ we introduce during the processing a common problem related to the multiple testing procedures where the probability of the second type error is not managed by the level of the test any more as it now also depends on the total number of independent pairs of hypothesis to be tested.

We will however, not discuss this matter in this thesis in detail let us just mention that appropriate statistical correction procedures have been proposed in order to deal with the multiple testing problems, under the variety of different settings and scenarios. Probably the most common one is the Bonferroni correction[27] introduced by Dunn (1961).

[27] Perhaps, the Bonferroni correction is more accurately described as Dunn-Bonferroni correction as it was originally proposed by Olive Jean Dunn as a generalization of the Boole inequality. Many alternatives to the Bonferroni correction can be easily find in literature.

3.3 Unknown change-points

Another, somehow more complex approach will arise when we have no a priori knowledge about any change-point locations, which may or may not occur in a considered regression model. Under such situations one is mostly interested in a statistical testing problem that the unknown regression function $m(\cdot)$, which is of interest (or its derivatives up to the order p respectively) is continuous (or smooth of some specific order) or if there is a jump (at least one) present along its domain[28].

We will firstly discuss a situation where we want to test the continuity of the regression function $m(\cdot)$ against an alternative that there is a jump present in $m(\cdot)$. Let us introduce a pair of statistical hypothesis where

$$\begin{aligned} \mathsf{H}_0: &\quad m(\cdot) \in \mathcal{C}_1(0,1); \\ \mathsf{H}_1: &\quad m(\cdot) \notin \mathcal{C}_1(0,1); \end{aligned} \tag{3.8}$$

where $\mathcal{C}_1(0,1)$ stands for a set of continuous function of the first order defined on $(0,1)$ or analogously,

$$\begin{aligned} \mathsf{H}_0: &\quad \forall x \in (0,1): \quad m_+(x) - m_-(x) = 0; \\ \mathsf{H}_1: &\quad \exists_{\tau \in (0,1)}: \qquad \int_{\tau-\epsilon}^{\tau+\epsilon} (m_+(x) - m_-(x))^2 \, \mathsf{d}F_X(x) > 0; \end{aligned} \tag{3.9}$$

for some arbitrarily small $\epsilon > 0$ and $F_X(\cdot)$ to be the marginal distribution function of the random variable X. Instead of the L_2 norm in (3.9) we could also adopt a general loss function $\rho(\cdot)$ however, we will not discuss such options in this thesis.

The idea of the test is to decide if there exists a statistically significant evidence for a jump to occur in the unknown regression function and if the test is rejected one is then interested in specifying an exact position where the discontinuity point should be located. Two steps of the same procedure are involved: providing a statistically consistent decision and estimating the position of the jump.

There are actually two different approaches, which can be used to deal with such statistical tests. The first one is based on cumulative sums statistic (CUSUM statistic) proposed by Page (1954) and the second one, technically more advanced is based on maximum type of test statistics, which lead to Extreme Value Theory. We will however, discuss only the first option here.

Let us partially base our approach on the one proposed by Gao et al. (2008) and all further details can be found there as well. Let us propose an auxiliary statistic

$$\widetilde{T}_{h_N}(x) \overset{def.}{=} (\widehat{m}_+(x) - \widehat{m}_-(x))^2, \tag{3.10}$$

which is again assumed to be sensitive with respect to a possible jump occurrence for any $x \in (0,1)$. We already know from Theorem 3.3 that under the null hypothesis (if there is no jump located at the given point $x \in (0,1)$) the quantity $(\widehat{m}_+(x) - \widehat{m}_-(x))$ follows in limit in law a normal distribution with a zero mean parameter and an appropriate variance term. As far as we are interested in the whole domain of interest rather than one specific point only we need to account for this requirement, which can be done by introducing the test statistic in the form

$$T_{h_N} = \frac{\int_0^1 \widetilde{T}_{h_N}(x) \mathsf{d}F_X(x) - \int_0^1 \mathbb{E}\widetilde{T}_{h_N}(x) \mathsf{d}F_X(x)}{\int_0^1 \sigma\left(\widetilde{T}_{h_N}(x)\right) \mathsf{d}F_X(x)}, \tag{3.11}$$

[28] We are still staying focused on interval $(0,1)$ only to be the domain of interest for the random variable X. Adaptations onto more general intervals (a,b), for $-\infty < a < b < \infty$ are quite straightforward.

where $\sigma\left(\widetilde{T}_{h_N}(x)\right) = \sqrt{\mathbb{V}\mathrm{ar}[\widetilde{T}_{h_N}(x)]}$, for $x \in (0,1)$.

THEOREM 3.6 (Asymptotic distribution for the global test statistic)
Let us assume the hypothesis testing problem $\boxed{3.9}$ *and the set of assumptions as in Theorem 3.2. Then under the null hypothesis* H_0 *the following convergence in distribution is achieved*

$$T_{h_N} \xrightarrow[N \to \infty]{\mathscr{D}} \mathbb{N}(0,1).$$

Proof. The proof of this theorem can be shown as a generalization of the proof in Gao et al. (2008). ■

The decision about the hypothesis testing problem $\boxed{3.9}$ is again based on the corresponding critical value of the limit distribution and it holds that

$$\textit{iff} \qquad |T_{h_N}| > u_\alpha \qquad \Longrightarrow \qquad \text{reject } \mathrm{H}_0,$$

where $u_\alpha \in \mathbb{R}$ is the corresponding critical value of the standard normal distribution $\mathbb{N}(0,1)$. In real situations however, it is more common to consider a test statistic given by

$$T^{\dagger}_{h_N} = \int_0^1 \widetilde{T}_{h_N}(x)\mathrm{d}F_X(x),$$

which also follows in asymptotic a normal distributional law where for the variance parameter we have that $\sigma^2_{h_N}(T^{\dagger}_{h_N}(x)) = \int_0^1 \widetilde{T}_{h_N}(x)\mathrm{d}F_X(x)$. The final decision can be based on bootstrap simulations where we would mimic the unknown distribution of interest or on some plug-in techniques which, however, mostly perform much more poorly.

Corollary 3
The result of Theorem 3.6 can be generalized into the α*-mixing dependence concept as well. Moreover, just a straightforward extension is needed to formulate the asymptotic distribution of the test statistic* $T^{(\nu)}_{h_N}$ *used for an analogous test however, for testing continuity vs. discontinuity of some higher levels of hierarchy, where* $T^{(\nu)}_{h_N}$ *is based on quantities* $\left(\widehat{m}^{(\nu)}_{+}(x) - \widehat{m}^{(\nu)}_{-}(x)\right)^2$, *for* $x \in (0,1)$.

3.4 Discussion on Change-points

Change-point models represent a very popular area in a statistical research today unfortunately, just a very small portion of the whole concept of models with discontinuities can be mentioned in this thesis. We have only discussed some useful scenarios together with some common statistical approaches but many other scenarios as well as statistical methods on this topic can be found in literature.

One interesting idea for a further research would be considering the M-smoothers estimators together with maximum type test statistics and to test the continuity property of the whole regression function. On the other hand, one could also possibly think of another generalizations of the proposed concept and to introduce a model where discontinuities are allowed to be present in the scale function $\sigma(\cdot)$ or its derivatives as well. Some nonparametric simultaneous testing procedures based on the L_2 approach were already introduced to handle such situations in practical cases.

"Only a good detective can tell you who is the murderer if nobody saw anything, nobody heard and nobody knows anything! That is what the bootstrap is all about!"

Martin Marušic
(1986 – . . .)

4

Bootstrapping of M-smoothers

We already have a sufficient knowledge on the asymptotic performance of the proposed M-smoothers estimates under the all three considered scenarios. However, focussing now on the limit distributions stated in the theorems in Chapter 2 as well as Chapter 3 one can easily see that they all heavily depends on some unknown quantities, which are in real applications not given explicitly and they rather remain unknown.

One plausible way to deal with such unknown quantities is to use some additional estimation approaches to estimate them firstly and to plug the estimates into the asymptotic normality expression to get the final distribution of interest. Unfortunately, as we have already mentioned, using the plug-in techniques could be considered to be quite straightforward in practice but their asymptotic performance is rather poor no mention that some quantities may not be even so easily obtainable.

An alternative approach was introduced within a bootstrap idea. Since the bootstrap methods were firstly introduced in statistic by Efron and Tibshirani (1993) (an extensive study on bootstrap methods can be found in Davison and Hinkley (1997) and Bickel and Freedman (1981) where the second one refers to the nonparametric regression especially) they have grown a huge attention and importance in almost all areas of the modern statistic. In principle, the bootstrap approach works with the empirical distribution function of some random sample to obtain a reasonable approximation of the unknown distribution for some statistic, which is of interest. The main advantage of the bootstrap methods is their simplicity and once a corresponding set of regularity conditions is derived the bootstrap methods can be consistently used even for very complex problems and sophisticated statistical questions as well.

In order to apply the bootstrap idea in real data problems the bootstrap simulations are used instead. They can be thought of as computer-based re-sampling approaches for assigning measures of accuracy to some sample estimates. The idea of bootstrap simulations is to impose a well defined re-sampling in order to replicate the quantity of interest (mostly some statistic) where the set of obtained replicates can be effectively and consistently used to mimic the unknown distribution of the quantity of interest (as well as all further properties coming from this distribution e.g. critical values).

On the other hand, the bootstrap approximation usually does not provide a general finite sample guarantees and additionally, there is a tendency for slightly optimistic results to be expected however, the bootstrap methods can provide very effectively in general. Moreover, using simulation methods instead of plug-in techniques one can achieve both a reasonable asymptotic performance as well as a quite straightforward application in real data cases. As far as the proper conditions required for the bootstrap simulations to work are crucial we will separately discuss an appropriate bootstrap algorithm for each of the three scenarios considered in Chapter 2 and a proper theoretical justification of the bootstrapping idea will be given for each algorithm afterwards.

4.1 Smooth residual bootstrap

Focusing now on different bootstrap adaptations proposed for regression models and nonparametric ones especially, one has in hand a variety of different approaches each of them based on somehow another principle. Three of them have however, mainly attached the most popularity in last years:

- ❑ **Model based bootstrap** - given the original random data sample $\{(X_i, Y_i;\ i = 1, \dots, N\}$ we assume the data pairs (X_i, Y_i) for $i = 1, \dots, N$ to be fixed and bootstrap replicates come directly when re-sampling from these data pairs.
- ❑ **Residual based bootstrap** - given the random sample $\{(X_i, Y_i;\ i = 1, \dots, N\}$ one firstly needs to obtain a corresponding estimate for a functional regression relationship, which can be consequently used to estimate the vector of residuals. The residual based bootstrap relies on re-sampling from the set of these residuals.
- ❑ **Blocks bootstrap** - is technically based on re-sampling from the set of pre-specified blocks of consecutive data pairs (X_i, Y_i), for $i = 1, \dots, N$. This approach is especially suitable for most cases under some data dependence concept.

We will discuss two principles here: a generalization of the residual based bootstrap, which will be proposed for the homoscedastic and heteroscedastic model scenario with independent random error terms and also a generalization of the blocks bootstrap approach, which we will used to handle the model with the α-mixing dependence concept.

4.1.1 Homoscedastic model with independent random errors

Let us start with the simplest model scenario for the M-smoothers estimator and let us recall the definition of model 2.3 together with the set of assumptions A1 – A7. We will propose a residual based bootstrap algorithm, which can be used to mimic the limit distribution stated in Theorem 2.5 and Theorem 3.2 both for $\nu = 0$ and Theorem 3.3 respectively.

Algorithm: SMOOTH RESIDUAL BOOTSTRAP (homoscedasticity)

start

B1 Compute residuals $\{\widehat{\varepsilon}_i;\ i = 1, \dots, N\}$, where $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i)$, where $\widehat{m}(X_i)$ is a corresponding M-smoothers estimate of $m(X_i)$ at X_i defined by minimization 2.1;

B2 Resample with replacement from the set of residuals $\{\widehat{\varepsilon}_i;\ i = 1, \dots, N\}$ in order to obtain new residuals $\tilde{\varepsilon}_i$, for $i = 1, \dots, N$;

B3 Define new bootstrap residuals $\varepsilon_i^\star = V_i \cdot \tilde{\varepsilon}_i + a_N \cdot Z_i$, where $\mathbf{P}[V_i = -1] = \mathbf{P}[V_i = 1] = \frac{1}{2}$, and $Z_i \sim N(0, 1)$ are i.i.d. standard normal random variables and $a_N = o(1)$ is a bootstrap bandwidth parameter, such that $N h_N a_N^2 / \log h_N^{-1} \to \infty$ and $a_N^2 / h_N^{1+\delta} = o(1)$ as $N \to \infty$, for $\delta > 0$ small enough;

B4 Define a new bootstrap data sample $\{(X_i, Y_i^\star);\ i = 1, \dots, N\}$, where $Y_i^\star = \widehat{m}(X_i) + \varepsilon_i^\star$;

B5 Re-estimate the unknown function of interest $m(x_0)$ (or $m_+(x_0)$ and $m_-(x_0)$ respectively) based on the new data sample $\{(X_i, Y_i^\star);\ i = 1, \dots, N\} \to$ obtain $\widehat{m}^\star(x_0), \widehat{m}_+^\star(x_0)$ and $\widehat{m}_-^\star(x_0)$;

B6 Repeat steps B2 → B3 → B4 → B5 to get the estimates $\widehat{m}_b^\star(x_0), \widehat{m}_{b+}^\star(x_0)$ and $\widehat{m}_{b-}^\star(x_0)$, for $b = 1, \dots, B$, where $B \in \mathbb{N}$ is sufficiently large;

B7 Use the quantities produced in step B6 to mimic the unknown distribution of interest;

end of SRB

The algorithm described above follows the idea of the smooth residual based bootstrap introduced in Neumeyer (2006). The notion of the smooth bootstrap comes from the step B3, which is introduced in the algorithm to ensure the right centering of bootstrapped residuals while the second part of this step – the smoothing part $a_N Z_i$ – ensures a proper convergence of the bootstrapped distribution to an unknown distribution of the true random errors $\{\varepsilon_1\}_{i=1}^N$.

Using the smooth version of the bootstrap algorithm one can conveniently handle both, a proper centering of bootstrapped residuals in order to eliminate the systematic bias and also preserving the robust flavour of the whole procedure.

The basic principle is to withdraw a new set of residuals $\{\widetilde{\varepsilon}_i;\ i = 1,\ldots,N\}$ from the set of original residuals $\{\widehat{\varepsilon}_i;\ i = 1,\ldots,N\}$. In order to correct for a systematic bias, which would be implicitly involved in all further calculations one needs to make sure that the bootstrapped residuals $\{\widetilde{\varepsilon}_i\}$ are, conditionally on $\{(X_i, Y_i);\ i = 1,\ldots,N\}$ centered, which is mostly done by subtracting the average (e.g. $\widetilde{\varepsilon}_i^* = \widetilde{\varepsilon}_i - \frac{1}{N}\sum_j \widetilde{\varepsilon}_j$). However, bearing in mind the robust flavour of the M-smoothers approach where we allow for outlying observations and heavy-tailed distributions of random errors such centering would be insane as it is not stable with respect to outliers moreover, it may not be even defined for some heavy-tailed distributions.

On the other hand, when using the bootstrapping idea to mimic the unknown distribution of a test statistic one has to be aware of the fact that such limit distribution is given under the null hypothesis only. For the M-smoothers approach and change-point tests this would mean that the set of residuals should be appropriately obtained using the data sample and the M-smoothers estimate both in a full correspondence with the null hypothesis however, if there is a change-point really present in the model, which we do not know yet this would not be the case.
Indeed, if the alternative hypothesis holds true we use a continuous M-smoothers estimate to compute the residuals while the data points actually come from a discontinuous function instead.

The effect of such "imprecision" will be however, only reflected in a small artificial increase of the variability (scale respectively) within the set of bootstrapped residuals $\{\varepsilon_i^\star\}_{i=1}^N$ conditioned on the random sample $(\mathcal{X}, \mathcal{Y})$, which will end up with a slightly more optimistic result than expected (which is common for the bootstrap approaches in general).

However, given the fact that inappropriate residuals are produced in a small neighbourhood of the jump location only while over the rest of the domain we assume to have consistent estimates of the error terms this does not seriously effect the final result. Moreover, an artificial increase of variability in the bootstrapped residuals is also cased by the centering part as well as the smoothness part in B3. Asymptotically, all three variance increase factors are negligible therefore, we will not investigate these issues in more details and we will focus on justifying the bootstrapping idea instead.

THEOREM 4.1 (Bootstrap consistency for homoscedastic model)
Let us assume model (2.3) and assumptions A1 – A7 and let the bootstrap bandwidth parameter a_N satisfies all conditions in B3. Then the following convergence in probability is achieved

$$\sup_{z\in\mathbb{R}} \left\{ \mathbf{P}^\star\left[\sqrt{Nh_N}\left(\widehat{m}^\star(x) - \widehat{m}(x)\right) \leq z\right] - \mathbf{P}\left[\sqrt{Nh_N}\left(\widehat{m}(x) - m(x)\right) \leq z\right]\right\} \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

where $\mathbf{P}^\star[\ \cdot\ ]$ stands for a conditional probability conditioned on the given random sample $(\mathcal{X}, \mathcal{Y}) = \{(X_i, Y_i);\ i = 1,\ldots,N \in \mathbb{N}\}$ and $x \in (0,1)$ is a given point of interest.

Proof. The proof of the bootstrap consistency result for the homoscedastic model scenario is given in Section 4.2 below. ■

Corollary 4
An analogous theorem can be formulated and proved to justify the bootstrap consistency with respect to the asymptotic distribution of one-sided M-smoothers as well as the asymptotic distributions of the corresponding test statistics discussed in Chapter 3 (under the homoscedastic assumption of course). Moreover, the assertion in Theorem 4.1 can be extended to cover the asymptotic distributions of the corresponding M-smoothers estimates of all order derivatives for any $\nu \in \{1, \dots, p\}$.

Given Theorem 4.1 one obtains a very powerful tool for a practical application of the M-smoothers estimators under the homoscedastic model scenario as it nicely avoids estimating the unknown scale, the distribution of random errors or the original density function of the design variable X. Instead, it is sufficient to employ a quite straightforward simulation exercise and all required conclusions can be consistently derived using the bootstrap distribution. This is useful especially for making decisions on hypothesis testing problems or constructing confidence intervals for some quantity of interest.

Introducing the bootstrap algorithm is however, just one part of the whole bootstrapping concept, which is used to avoid the plug-in techniques. Another important part relies on justifying the proposed bootstrapping idea and showing that the algorithm as described above can be indeed effectively used to mimic the unknown distribution of interest.

We will prove the consistency property for the proposed bootstrap algorithm and we will also compare the asymptotic and the finite sample performance for this method however, let us firstly discuss another model scenario, which can be also considered together with the smooth residual bootstrap – the heteroscedastic model introduced in Section 2.3.

4.1.2 Heteroscedastic model with independent random errors

The idea of bootstrapping in heteroscedastic models remains very the same as in the homoscedastic scenario already discussed. Indeed, the asymptotic distribution derived for the heteroscedastic model (e.g Theorem 2.8) depends again on some unknown quantities. Beside the distribution function $G(\cdot)$ and the density function $f(\cdot)$ there is in addition the scale function $\sigma(\cdot)$ incorporated into the variance expression in Theorem 2.8 (analogously in Theorem 3.2 and Theorem 3.3 as well), which needs to be properly taken care of by the heteroscedastic version of the bootstrap algorithm.

The heteroscedastic version of the bootstrap algorithm will mostly correspond with the algorithm introduced for the homoscedastic model however, one modification is implemented in the first step B1, where the set of residuals $\{\widehat{\varepsilon}_i;\ i = 1, \dots, N\}$ needs to be standardized with respect to the scale function $\sigma(\cdot)$ in order to fulfil the assumption A2* where the unit scale is assumed now.

It is obvious, that the special case of the heteroscedastic bootstrap algorithm where we have the scale function to be constant over the whole domain of interest will lead back to the homoscedastic version of the proposed algorithm.
Given this fact we will only assume one version of the bootstrap algorithm in the proofs and we will justify the bootstrapping idea under the heteroscedastic model scenario only.
Let us however introduce the heteroscedastic version of the bootstrap algorithm firstly.

Algorithm: SMOOTH RESIDUAL BOOTSTRAP (heteroscedasticity)

start

B1 Compute residuals $\{\widehat{\varepsilon}_i;\ i = 1, \dots, N\}$, where $\widehat{\varepsilon}_i = \frac{Y_i - \widehat{m}(X_i)}{\widehat{\sigma}(X_i)}$, where $\widehat{m}(X_i)$ is an estimate of $m(X_i)$ at X_i defined by $\boxed{2.1}$ and $\widehat{\sigma}(X_i)$ is the corresponding estimate[29] of the scale function $\sigma(\cdot)$ given at the point X_i as well;

B2 Resample with replacement from the set of residuals $\{\widehat{\varepsilon}_i;\ i = 1, \dots, N\}$ in order to obtain new residuals $\tilde{\varepsilon}_i$, for $i = 1, \dots, N$;

B3 Define new bootstrap residuals $\varepsilon_i^\star = V_i \cdot \tilde{\varepsilon}_i + a_N \cdot Z_i$, where $\mathbf{P}[V_i = -1] = \mathbf{P}[V_i = 1] = \frac{1}{2}$, $Z_i \sim N(0,1)$ are $i.i.d.$ standard normal random variables and $a_N = o(1)$ is an appropriate bootstrap bandwidth parameter, such that $N h_N a_N^2 / \log h_N^{-1} \to \infty$, $N a_N^{2(p+1)} \to 0$ and $a_N^2 / h_N^{1+\delta} = o(1)$ as $N \to \infty$, for some $\delta > 0$ small enough;

B4 Define a new bootstrap data sample $\{(X_i, Y_i^\star);\ i = 1, \dots, N\}$, where $Y_i^\star = \widehat{m}(X_i) + \widehat{\sigma}(X_i) \cdot \varepsilon_i^\star$;

B5 Re-estimate the unknown functions $m(x_0)$ (or $m_+(x_0)$ and $m_-(x_0)$ respectively) based on new data sample $\{(X_i, Y_i^\star);\ i = 1, \dots, N\} \to$ obtain $\widehat{m}^\star(x_0)$ ($\widehat{m}_+^\star(x_0)$ and $\widehat{m}_-^\star(x_0)$ respectively);

B6 Repeat steps B2 $\to$ B3 $\to$ B4 $\to$ B5 to get estimates $\widehat{m}_b^\star(x_0), \widehat{m}_{b+}^\star(x_0)$ and $\widehat{m}_{b-}^\star(x_0)$, for $b = 1, \dots, B$, for $B \in \mathbb{N}$ to be sufficiently large;

B7 Use the quantities produced in step B6 to mimic the unknown distribution of interest;

end of SRB

THEOREM 4.2 (Bootstrap consistency for heteroscedastic model)

Let us assume model $\boxed{2.39}$ *and the same assumptions as in Theorem 2.8. Moreover, let the additional assumption posed on the bootstrap bandwidth parameter* a_N *in B3 is satisfied as well. Then the following convergence in probability is achieved*

$$\sup_{z \in \mathbb{R}} \left\{ \mathbf{P}^\star \left[\sqrt{N h_N} \left(\widehat{m}^\star(x) - \widehat{m}(x) \right) \leq z \right] - \mathbf{P} \left[\sqrt{N h_N} \left(\widehat{m}(x) - m(x) \right) \leq z \right] \right\} \xrightarrow[N \to \infty]{\mathbf{P}} 0,$$

where $\mathbf{P}^\star[\ \cdot\ ]$ *stands for a conditional probability given data* $(\mathcal{X}, \mathcal{Y}) = \{(X_i, Y_i);\ i = 1, \dots, N\}$.

Proof. The proof of this theorem is given in Section 4.2 below. ■

Corollary 5

An analogous theorem can be formulated to state the consistency with respect to the asymptotic distributions of one-sided M-smoothers as well as the asymptotic distributions stated in Theorem 3.2 and Theorem 3.3. Similarly, the result in Theorem 4.2 can be easily extended to cover the asymptotic distributions of the corresponding M-smoothers estimates of all order derivatives for any $\nu \in \{1, \dots, \nu\}$.

Going now along the steps of both proposed versions of the smooth residual bootstrap algorithm (homoscedastic and heteroscedastic) one can acknowledge their simplicity and intuitiveness. Using the bootstrap techniques to investigate the asymptotic properties and providing statistical decisions is far more convenient than using plug-in techniques and expressing the unknown distribution of interest throughout its some additional finite sample estimates.

[29] One possible way to introduce an estimate of the scale function $\sigma(\cdot)$ is to run a minimization problem similar to $\boxed{2.1}$ where in turn one uses quantities $\rho(Y_i - \widehat{m}(X_i))$ for $i = 1, \dots, N$ instead of Y_i's and function $\sigma(\cdot)$ instead of $m(\cdot)$ while using the same loss function $\rho(\cdot)$ and the local approach defined by the kernel function $K(\cdot)$. Some alternative estimation approaches however, for the fixed design only are also shortly discussed in Chapter 5.

4.2 Justification of the smooth residual bootstrap

In order to complete the discussion on bootstrapping of M-smoothers with independent random error terms we just need to justify the proposed simulation algorithms and we should give theoretical proofs for Theorems 4.1 and 4.2. However, as far as the homoscedastic scenario can be easily derived as a special case of the heteroscedastic bootstrap principle where the scale function $\sigma(\cdot)$ is set to be constant over the whole domain of interest we will prove Theorem 4.2 only and the proof of Theorem 4.1 will follow analogously.

Firstly, we recall that $\mathbf{P}^{\star}[\cdot]$ stands for a conditional probability conditioned on the given random sample $(\mathcal{X},\mathcal{Y}) = \{(X_i,Y_i);\ i=1,\ldots,N\}$, which we will bear in mind along the whole proof below.

Let $\{\varepsilon_i^{\star}\}_{i=1}^{N}$ be a sequence of the bootstrapped residuals obtained in step B3 of the smooth residual bootstrap algorithm for the heteroscedastic model. Then the corresponding vector of the parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\star} = \left(\widehat{\beta}_0^{\star},\ldots,\widehat{\beta}_p^{\star}\right)^{\top} = (\widehat{m}^{\star}(x),\widehat{m}^{\star\prime}(x),\ldots,\frac{\widehat{m}^{\star(p)}(x)}{p!})^{\top} \in \mathbb{R}^{p+1}$ is given as a solution of the /mboxminimization problem

$$\widehat{\boldsymbol{\beta}}_x^{\star} = \underset{(b_0,\ldots,b_p)^{\top}\in\mathbb{R}^{p+1}}{Argmin} \sum_{i=1}^{N}\rho\left(Y_i^{\star} - \sum_{j=0}^{p} b_j(X_i-x)^j\right)\cdot K\left(\frac{X_i-x}{h_N}\right), \qquad \boxed{4.1}$$

where $Y_i^{\star}$, for $i=1,\ldots,N$ are defined in step B4 as

$$Y_i^{\star} \overset{def.}{=} \widehat{m}_N(X_i) + \widehat{\sigma}_N(X_i)\varepsilon_i^{\star},$$

where $\widehat{m}_N(\cdot)$ is the M-smoothers estimate of the unknown regression function $m(\cdot)$ defined by $\boxed{2.41}$ given the finite sample data $(\mathcal{X},\mathcal{Y})$ and $\widehat{\sigma}_N(\cdot)$ is a corresponding estimate of the scale function defined in step B1 based on the finite data sample $(\mathcal{X},\mathcal{Y})$ again.

The idea of the proof is now the same as in the case of Theorem 2.8. Using a sequence of similar steps we obtain an equivalent problem given by the set of equations

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}\psi\left(\widehat{\sigma}_N(X_i)\,\varepsilon_i^{\star} - \sum_{j=0}^{p} b_j\left(\frac{X_i-x}{h_N}\right)^j\right)\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right) = 0, \qquad \boxed{4.2}$$
$$\text{for} \quad l = 0,\ldots,p,$$

which is solved for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\bullet} = (\widehat{\beta}_0^{\bullet},\ldots,\widehat{\beta}_p^{\bullet})^{\top}$, where the definition of the vector $\widehat{\boldsymbol{\beta}}_x^{\bullet} \in \mathbb{R}^{p+1}$ is analogous with the definition of $\widehat{\boldsymbol{\beta}}_x^{\circ} \in \mathbb{R}^{p+1}$ in Chapter 2.

We want to find now a reasonable asymptotic representation for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\bullet}$ similarly, as we did for $\widehat{\boldsymbol{\beta}}_x^{\circ}$ in $\boxed{2.46}$. The idea of the three subsequent steps in Chapter 2 will be therefore, applied again.

The first step of the proof, which in this case refers to the fact that $\mathbf{P}^{\star}\left[|\widehat{\beta}_j^{\star}| \leq \daleth_N^{\star}\right] \xrightarrow{\mathbf{P}} 0$, for any $j=0,\ldots,p$ and some sequence $\daleth_N^{\star} \geq 0$ such that $\daleth_N^{\star} \to 0$ as $N\to\infty$ can be proved using the same arguments as before and hence we will omit it here. We will rather focus on the rest of the proof and we will formulate two important lemmas which will be used afterwards.

Lemma 9

Let model 2.39 holds and let the assumptions from Theorem 2.8 together with the assumption posed on the bootstrap bandwidth parameter a_N in step B3 are satisfied. Then it holds that

(i) $$\sup_{e\in\mathbb{R}} |G^\star(e) - G(e)| = o(1), \qquad [\mathbf{P}] - a.s.,$$

where $G^\star(\cdot)$ stands for a distribution function of the bootstrapped random errors $\{\varepsilon_i\}_{i=1}^N$ defined in B3 conditioned on the random sample $(\mathcal{X},\mathcal{Y})$. Additionally, if the corresponding density functions exist it also holds that

(ii) $$\sup_{e\in\mathbb{R}} |g^\star(e) - g(e)| = o(1), \qquad [\mathbf{P}] - a.s..$$

Proof of Lemma 9

The proof of the lemma can be found in Neumeyer (2006) as a proof of Lemma 2.19. For the homoscedastic scenario the proof follows from the proofs of Lemma 2.1 and 2.4 (Neumeyer (2006)). □

Lemma 10

For model 2.39 and the same assumptions as in Lemma 9 the following convergence in probability is achieved

$$\begin{aligned}\mathbf{P}^\star\Bigg[&\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}}\Bigg|\sum_{i=1}^N \Bigg\{\psi\left(\widehat{\sigma}_N(X_i)\varepsilon_i^\star - \sum_{j=0}^p t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right) + \\ &-\mathbb{E}^\star\left[\psi\left(\widehat{\sigma}_N(X_i)\varepsilon_i^\star - \sum_{j=0}^p t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\Bigg\}\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\Bigg| \leq \mathcal{K}^\star\Bigg] \xrightarrow[N\to\infty]{\mathbf{P}} 0,\end{aligned}$$

for any $T > 0$ and $l \in \{0,\dots,p\}$ and some $\mathcal{K}^\star > 0$ where the expectation operator $\mathbb{E}^\star[\cdot]$ stands here for a conditional expectation conditioned on the random sample $(\mathcal{X},\mathcal{Y})$ given with respect to the distribution function $G^\star(\cdot)$.

The distribution function $G^\star(\cdot)$ of the bootstrapped random errors $\{\varepsilon_i\}_{i=1}^N$ should be continuous given the smoothness correction involved in the heteroscedastic algorithm in step B3. Moreover, it is given conditionally on the random sample $\{(X_i,Y_i);\ i = 1,\dots,N\}$ therefore, it also depends on $N \in \mathbb{N}$, which is however, not explicitly expressed in the notation for the function $G^\star(\cdot)$.

Proof of Lemma 10

Firstly, we will show that the distribution function $G^\star(\cdot)$ conditioned on $(\mathcal{X},\mathcal{Y})$ and given for the fixed value of $N \in \mathbb{N}$ is indeed (even absolutely) continuous and symmetric. This will allow us to assume function $\lambda_{G^\star}(t)$ in sense of definition 1.8 to be a bootstrap counterpart of $\lambda_G(t)$ for $t \in \mathbb{R}$ (or at least at some small neighbourhood of $t = 0$). We easily have that

$$\begin{aligned} G^\star(x) &= \mathbf{P}^\star[\varepsilon_i^\star \le x] = \mathbf{P}^\star[V_i\cdot\tilde{\varepsilon}_i + a_N\cdot Z_i \le x] = \\ &= \int_{\mathbb{R}} \mathbf{P}^\star[V_i\cdot\tilde{\varepsilon}_i \le x - a_N y]\,\mathrm{d}\Phi(y) = \frac{1}{2}\int_{\mathbb{R}}[G_N(x-a_N y) - G_N(-x+a_N y) + 1]\,\mathrm{d}\Phi(y) = \\ &= \frac{1}{2N}\sum_{i=1}^N\left[\Phi\left(\frac{x-\widehat{\varepsilon}_i}{a_N}\right) + \Phi\left(\frac{x+\widehat{\varepsilon}_i}{a_N}\right)\right], \end{aligned}$$

where $\Phi(\cdot)$ stands for a distribution function of a standardized and zero mean normal distribution. The distribution function $G^{\star}(\cdot)$ is clearly given as a finite sum of absolutely continuous functions therefore, it is continuous as well. To show the symmetric property we just need to use the same calculations again and using the fact that $\Phi(x) = 1 - \Phi(-x)$ for any $x \in \mathbb{R}$ we can proceed backwards in the expressions above to finally obtain that $G^{\star}(x) = 1 - G^{\star}(-x)$. Details are omitted here.

In order to proceed with the proof of Lemma 10 we can use the continuity property of the bootstrap distribution function $G^{\star}(\cdot)$ and hence, we obtain that

$$\begin{aligned}
&\mathbb{E}^{\star}\psi\left(\widehat{\sigma}_N(X_i)\varepsilon_i^{\star} - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) = \\
&= \int_{\mathbb{R}} \psi\left(\widehat{\sigma}_N(X_i)e - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) \mathsf{d}\left(G(e) + G^{\star}(e) - G(e)\right) = \\
&= \lambda_G\left(\sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j, \widehat{\sigma}(X_i)\right) - \int_{\mathbb{R}} \psi\left(\widehat{\sigma}_N(X_i)e - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) \mathsf{d}(G - G^{\star})(e),
\end{aligned}$$

and we will show that the last integral in the expression above is asymptotically negligible.

Given the assertion of Lemma 9, which holds true $[\mathbf{P}]$-almost surely we can obviously assume that we have such data sample $(\mathcal{X}, \mathcal{Y})$ that it holds that $\lim_{N\to\infty}\sup_{e\in\mathbb{R}}|G^{\star}(e) - G(e)| = 0$ and similarly also $\lim_{N\to\infty}\sup_{e\in\mathbb{R}}|g^{\star}(e) - g(e)| = 0$.

Moreover, it is quite obvious that the following inequality holds as well

$$(g - g^{\star}) \leq |g - g^{\star}| = (g \vee g^{\star}) - (g \wedge g^{\star}),$$

where also $(g \wedge g^{\star}) \leq g$ therefore, for an arbitrary function $\widetilde{\psi}(\cdot)$ such that $\int_{\mathbb{R}}|\widetilde{\psi}(e)|\mathsf{d}G(e) \leq \infty$ we have from the Dominated Convergence Theorem that

$$\lim_{N\to\infty}\int_{\mathbb{R}} \widetilde{\psi}(e)\mathsf{d}(G \wedge G^{\star})(e) = \int_{\mathbb{R}} \widetilde{\psi}(e)\mathsf{d}G(e). \tag{4.3}$$

On the other hand, we also have that

$$\lim_{N\to\infty}\int_{\mathbb{R}} \widetilde{\psi}(e)\mathsf{d}(G \vee G^{\star})(e) = \lim_{N\to\infty}\int_{\mathbb{R}} \widetilde{\psi}(e)\mathsf{d}(G + G^{\star} - (G \wedge G^{\star}))(e) = \int_{\mathbb{R}} \widetilde{\psi}(e)\mathsf{d}G(e), \tag{4.4}$$

therefore, combining (4.3) and (4.4) we obtain that

$$\int_{\mathbb{R}} \psi\left(\widehat{\sigma}_N(X_i)e - \sum_{j=0}^{p} t_j\delta_N\left(\frac{X_i - x}{h_N}\right)^j\right) \mathsf{d}(G - G^{\star})(e) = o_{\mathbf{P}}(1).$$

Repeating the same argumentation over again one can also show that $\mathbb{E}^{\star}\left[\psi(\cdot)\right]^2 \to 0$ in probability as $N \to \infty$, which is required to complete the proof of the lemma while the rest goes precisely along the lines of the proof of Lemma 3 and therefore, it is omitted. □

Lemma 11
For model $\boxed{2.39}$ *and the same assumptions as in Lemma 9 the following convergence in probability is achieved*

$$\mathbf{P}^{\star}\left[\sup_{\substack{|t_j|<T\\ j=0,\dots,p}} \frac{1}{\sqrt{Nh_N}}\left|\sum_{i=1}^{N}\left[\psi\left(\widehat{\sigma}_N(X_i)\varepsilon_i^{\star}-\sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)-\psi\left(\widehat{\sigma}_N(x)\varepsilon_i^{\star}\right)+\right.\right.\right.$$
$$\left.\left.\left.-\mathbb{E}^{\star}\psi\left(\widehat{\sigma}_N(x)\varepsilon_i^{\star}-\sum_{j=0}^{p}t_j\delta_N\left(\frac{X_i-x}{h_N}\right)^j\right)\right]\cdot\left(\frac{X_i-x}{h_N}\right)^l K\left(\frac{X_i-x}{h_N}\right)\right|\geq\epsilon\right]\xrightarrow[N\to\infty]{\mathbf{P}}0,$$

for any $\epsilon>0$ *and* $T>0$ *and* $l\in\{0,\dots,p\}$. *Additionally, the expectation operator* $\mathbb{E}^{\star}[\cdot]$ *stands here for a conditional expectation conditioned on the random sample* $(\mathcal{X},\mathcal{Y})$ *with respect to the bootstrap distribution function* $G^{\star}(\cdot)$ *where* $\delta_N=(Nh_N)^{-1/2}$.

Proof of Lemma 11

The proof follows as an extension of the proof of Lemma 4, using the assertion of Lemma 9 and the idea of the proof of Lemma 10 together with the fact that

$$\begin{aligned}|\widehat{\sigma}_N(X_i)-\widehat{\sigma}_N(x)| &\leq |\widehat{\sigma}_N(X_i)-\sigma(X_i)|+|\sigma(X_i)-\sigma(x)|+|\sigma(x)-\widehat{\sigma}_N(x)|\leq\\ &\leq K|X_i-x|^{\alpha}+o_{\mathbf{P}}(1),\end{aligned}$$

which follows from the assumptions posed on the scale function $\sigma(\cdot)$ and its finite sample estimate $\widehat{\sigma}_N(\cdot)$ respectively, where $\alpha>0$ is defined in assumption A5* (see p.35). □

Using now the assertions of Lemma 11 similarly as in the case of Lemma 4 we obtain the asymptotic Bahadur representation for the bootstrapped estimates $\widehat{\boldsymbol{\beta}}_x^{\bullet}\in\mathbb{R}^{p+1}$, which can be expressed as

$$\mathbf{P}^{\star}\left[\sqrt{Nh_N}\cdot\left|\widehat{\boldsymbol{\beta}}_x^{\bullet}-\frac{1}{\lambda'_{G^{\star}}(0,\widehat{\sigma}_N(x))}\cdot\left(\mathsf{X}_N^{\top}\mathsf{W}_N\mathsf{X}_N\right)^{-1}\cdot\mathsf{X}_N^{\top}\mathsf{W}_N\boldsymbol{\psi}(\widehat{\sigma}_N(x)\boldsymbol{\varepsilon}^{\star})\right|>\epsilon\right]\xrightarrow[N\to\infty]{\mathbf{P}}0, \qquad \boxed{4.5}$$

for any $\epsilon>0$ where in addition to the previous notation we have $\boldsymbol{\varepsilon}^{\star}=(\varepsilon_1^{\star},\dots,\varepsilon_N^{\star})^{\top}$.

We need to show now that the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\bullet}\in\mathbb{R}^{p+1}$ converges conditionally on the random sample $(\mathcal{X},\mathcal{Y})$ to the same limit distribution as the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\circ}=(\widehat{\beta}_0^{\circ},\dots,\widehat{\beta}_p^{\circ})$, which converges in law to a limit distribution stated in Theorem 2.8.

Once we prove that we will be done with the whole bootstrap justification. Let therefore us introduce a concept of a conditional weak convergence in probability and the Mallow metric idea[30], which metricates such convergence in probability.
We will need both statistical tools to complete the rest of the proof.

[30] The Mallow metric is also known in literature as the Wasserstein (Vasershtein respectively) metric where the name was coined by a Soviet mathematician and probability theorist Roland Lvovich Dobrushin (see Dobrushin (1970)) in 1970 to honour another Soviet mathematician Leonid Nasonovich Vasershtein who introduced the concept of this metric in 1969. For more details on the Mallow metric we refer to Belyaev (1995) or Belyaev and Sjöstedt-de Luna (2000) or Bickel and Freedman (1981).

DEFINITION 2 (Conditional weak convergence in probability)
Let $\{\mathbf{T}_N, \mathbf{T}_N^{'}\}_{N=1}^{\infty}$ be some random vectors. If for every real-valued and bounded continuous function $f(\cdot)$ holds that

$$\mathbb{E}[f(\mathbf{T}_N^{'})|\mathbf{S}_N] - \mathbb{E}[f(\mathbf{T}_N)] \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

then $\mathbf{T}_N^{'}$ condition on $\mathbf{S}_N$ and $\mathbf{T}_N$ are said to be approaching (each other) in distribution in probability along all sequences $\mathbf{S}_N$. In short we use the notation

$$\mathbf{T}_N^{'}|\mathbf{S}_N \overset{\mathscr{D}(\mathbf{P})}{\underset{N\to\infty}{\longleftrightarrow}} \mathbf{T}_N.$$

DEFINITION 3 (Mallow's metric)
For two random distributions $\mathscr{D}_{\mathcal{A}}$ and $\mathscr{D}_{\mathcal{B}}$ we define the Mallow metric as

$$d_{M,2}\left(\mathscr{D}_{\mathcal{A}}, \mathscr{D}_{\mathcal{B}}\right) \overset{def.}{=} \inf_{\mathscr{D}_{\mathcal{AB}} \in \mathcal{P}_{AB}} \left(\mathbb{E}_{\mathscr{D}_{\mathcal{AB}}} |X_A - X_B|^2\right)^{\frac{1}{2}},$$

where the infimum is taken over all joint distributions of some random variables X_A and X_B such that their marginal distributions are $\mathscr{D}_{\mathcal{A}}$ and $\mathscr{D}_{\mathcal{B}}$ respectively. For short we will also use the notation

$$d_{M,2}\left(X_A, X_B\right) \equiv d_{M,2}\left(\mathscr{D}_{\mathcal{A}}, \mathscr{D}_{\mathcal{B}}\right),$$

where $X_A \sim \mathscr{D}_{\mathcal{A}}$ and $X_B \sim \mathscr{D}_{\mathcal{B}}$.

We will use the concept of the weak conditional convergence in distribution in probability and some theory developed for the Mallow metric and will now show that

$$\frac{\left(\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N\right)^{-1}}{\lambda_{G^\star}^{\prime}(0, \widehat{\sigma}_N(x))} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\widehat{\sigma}_N(x)\boldsymbol{\varepsilon}^\star)\bigg| (\mathcal{X}, \mathcal{Y}) \overset{\mathscr{D}(\mathbf{P})}{\underset{N\to\infty}{\longleftrightarrow}} \frac{\left(\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N\right)^{-1}}{\lambda_{G}^{\prime}(0, \sigma(x))} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}), \qquad \boxed{4.6}$$

which can be equivalently re-written using the second degree Mallow's metric as

$$d_{M,2}\left(\frac{\left(\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N\right)^{-1}}{\lambda_{G^\star}^{\prime}(0, \widehat{\sigma}_N(x))} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\widehat{\sigma}_N(x)\boldsymbol{\varepsilon}^\star)\bigg| (\mathcal{X}, \mathcal{Y}),\ \frac{\left(\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N\right)^{-1}}{\lambda_{G}^{\prime}(0, \sigma(x))} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) \right) \xrightarrow[N\to\infty]{\mathbf{P}} 0.$$

Let us however, firstly show that

$$d_{M,2}\left(\frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\widehat{\sigma}_N(x)\boldsymbol{\varepsilon}^\star)\bigg| (\mathcal{X}, \mathcal{Y}),\ \frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) \right) \xrightarrow[N\to\infty]{\mathbf{P}} 0. \qquad \boxed{4.7}$$

The convergence in $\boxed{4.7}$ can be easily shown using the Central Limit Theorem for triangular arrays where both terms converge in law to a normal distribution with the appropriate mean and variance parameters. We only need to verify that $\mathbb{E}^\star \psi(\widehat{\sigma}_N(x)\varepsilon_i^\star)$ and $\mathbb{E}\psi(\sigma(x)\varepsilon_i)$ as well as $\mathbb{E}^\star \psi^2(\widehat{\sigma}_N(x)\varepsilon_i^\star)$ and $\mathbb{E}^\star \psi^2(\sigma(x)\varepsilon_i)$ are asymptotically equal.
However, both expressions follow from the assumption A5*, the property of the scale function estimate $\widehat{\sigma}_N(x)$ and the arguments used in the proof of Lemma 10.

Hence, we can use a common property of the Mallow metric to obtain

$$\underbrace{d_{M,2}\left(\sqrt{Nh_N}\ \mathsf{S}_N^{-1} \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\widehat{\sigma}_N(\mathbf{x})\boldsymbol{\varepsilon}^\star)\Big| (\mathcal{X}, \mathcal{Y}),\ \sqrt{Nh_N} \mathsf{S}_N^{-1} \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(\mathbf{x})\boldsymbol{\varepsilon})\right)}_{\hookrightarrow d_{M,2}(\bullet, \circ)} \leq$$

$$\leq Nh_N \cdot \left\|\mathsf{S}_N^{-1}\right\|_2 \cdot d_{M,2}\left(\frac{1}{\sqrt{Nh_N}} \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\widehat{\sigma}_N(x)\boldsymbol{\varepsilon}^\star)\bigg| (\mathcal{X}, \mathcal{Y}),\ \frac{1}{\sqrt{Nh_N}} \cdot \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(\mathbf{x})\boldsymbol{\varepsilon}) \right) \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

where $\mathsf{S}_N^{-1} = \left(\mathsf{X}_N^\top \mathsf{W}_N \mathsf{X}_N\right)^{-1}$ can be thought of as a random linear operator on the Euclidean $\mathbb{R}^{p+1}$ space while the norm of this operator "converges" in probability to $\frac{\|\mathsf{S}_1^{-1}\|_2}{Nh_N f(x)}$ for $\mathsf{S}_1^{-1} = (\mathsf{S}_1)^{-1}$ where S_1 is the matrix defined in Section 2.2.2. Moreover, its norm is finite and we also know that $f(x) \neq 0$ due to the assumption A1* (or A1 respectively).

Finally, we need to prove that $\lambda'_{G^\star}(0, \widehat{\sigma}_N(x)) \xrightarrow{\mathbf{P}} \lambda'_G(0, \sigma(x))$, for $N \to \infty$. Let us firstly show that following convergence holds true

$$\sup_{|t| \leq T} |\lambda_{G^\star}(t) - \lambda_G(t)| = \sup_{|t| \leq T} \left| \int_{\mathbb{R}} \psi(e-t) \mathrm{d}(G^\star - G)(e) \right| = o_{\mathbf{P}}(1), \tag{4.8}$$

for some $T > 0$ to be small enough. However, the limit term $(G^\star - G)$ where $G^\star$ depends on $N \in \mathbb{N}$ and $\psi(e-t)$ for $e \in \mathbb{R}$ and $|t| \leq T$ that the supremum in taken over are both independent of each other and therefore, the expression (4.8) holds due to *(i)* in Lemma 9 using a similar argumentation as at the begining of the proof of Lemma 9.

Using now the triangular inequality we obtain

$$\begin{aligned} &\left|\lambda'_{G^\star}(0, \widehat{\sigma}_N(x)) - \lambda'_G(0, \sigma(x))\right| \leq \\ &\qquad \leq \left|\lambda'_{G^\star}(0, \widehat{\sigma}_N(x)) - \lambda'_{G^\star}(0, \sigma(x))\right| + \left|\lambda'_{G^\star}(0, \sigma(x)) - \lambda'_G(0, \sigma(x))\right| \xrightarrow[N \to \infty]{\mathbf{P}} 0, \end{aligned}$$

where the first term converges to zero in probability due to the continuity property of $\lambda_{G^\star}(0, \cdot)$ in some neighbourhood of $\sigma(x)$ for $x \in (0,1)$ and the convergence of the second term is achieved due to (4.8).

In order to complete the whole proof we just need to put together both parts of the proof that we have shown so far hence, we obtain

$$\lambda'_{G^\star}(0, \widehat{\sigma}_N(x)) \xrightarrow[N \to \infty]{\mathbf{P}} \lambda'_G(0, \sigma(x)), \qquad \text{as well as} \qquad d_{M,2}(\bullet, \circ) \xrightarrow[N \to \infty]{\mathbf{P}} 0,$$

and applying the Slutzsky theorem finishes now the whole bootstrap consistency proof. ■

We have shown the bootstrap consistency for the heteroscedastic model scenario and we have shown that the bootstrap simulation can be consistently used to mimic the unknown distribution of interest. We have also mentioned that the assertion of Theorem 4.2 can be extended to provide a complete justification of the proposed bootstrap algorithm even when it is used to mimic the unknown limit distributions of the M-smoothers estimates of any order derivatives of $m(\cdot)$ up to the order $p \in \mathbb{N}$. The arguments above hold as well for one-sided M-smoothers and the corresponding limit distribution of the test statistic under the given null hypothesis.

However, an additional assumption needs to be posed on the kernel function $K(\cdot)$ when we want to use the bootstrap approach to mimic the limit distribution for the estimates of derivatives rather than the regression function itself. Indeed, in this case one also needs to consider the Lipschitz property to be satisfied for the corresponding derivatives of the kernel function $K(\cdot)$.

Finally, the same conclusions can be made with respect to the homoscedastic scenario and the corresponding bootstrap algorithm. A complete proof for the bootstrap consistency results under the homoscedasticity follows as a special case of the heteroscedastic scenario discussed above.

4.3 Block-bootstrap

Let us finally deal with the case where the M-smoothers estimator is considered under the α-mixing dependence structure of the random error terms. Given the fact that the residual based bootstrap draws new residuals independently, it is not suitable to be used for the α-mixing concept as it would definitely break any dependence structure already existing within the given data sample.

In order to preserve the same dependence structure within bootstrapped re-samples we will use the blocks bootstrap approach introduced by Politis and Romano (1992). The block-bootstrap can be thought of as a generalization of the nonparametric bootstrap (case sampling with replacement) where unlike the sampling individual pairs the whole blocks of consecutive pairs are re-sampled instead. It is evident, that within all re-sampled blocks the same dependence structure is preserved while between the blocks there is no further dependence assumed as the re-sampled blocks are drown independently. As far as the weak dependence property is defined as an asymptotic independence this plays no important role here however, the length of blocks is crucial similarly, as the choice of the smoothing parameter in the nonparametric regression estimation.

There were many different adaptations of the block-bootstrap proposed (see a nice summary book of Lahiri (2003) or Lahiri (1992)) however, we will only discuss one specific generalization – the Moving block bootstrap approach.

4.3.1 Moving blocks bootstrap

The Moving Blocks Bootstrap (MBB) was independently introduced by Künsch (1989) and Liu and Singh (1992) for the sample mean. Later on, the idea of MBB was extended by Lahiri (1992) and Politis and Romano (1992) to cover strictly stationary processes as well. We will consider a generalization proposed by Fitzenberger (1997) where he fitted MBB on non-stationary processes too.

The main difference between the classical block-bootstrap and the moving blocks bootstrap is engaged within the composition of blocks. For the traditional blocks bootstrap one assumes a set of non-overlapping blocks[31] (see Carlstein (1986)), which are used for subsampling. In the case of the MBB approach once constructs blocks of a fixed length where any observation (but some last ones) can be used as a starting point for a new block. Such blocks are used for bootstrap resampling after that. The set of re-sampled blocks is joint together to form one data sequence again, which is consequently used for the estimation process and obtaining the bootstrap quantity of interest.

The idea behind the bootstrapping under the α-mixing dependence is the same as before: one tries to use the bootstrap distribution instead of the asymptotic distribution, which is computationally quite intensive to obtain. Using the bootstrap approach under the right regularity assumptions one effectively avoids the plug-in techniques while the bootstrap distribution is consistent with the one, which is of interest.

We will again firstly discuss the algorithm, which will be used to simulate the bootstrap distribution and after that we will also provide a theoretical justification of the algorithm and we will prove that the bootstrap distribution can be indeed used as an asymptotic replacement for the true distribution, which is however unknown.

[31] The non-overlapping blocks bootstrap is slightly less effective than MBB as the blocks are more strictly defined and observations within each block are fixed with no chance to occur in some other block.

Algorithm: MOVING BLOCK BOOTSTRAP (α-mixing dependence)

start

B1$_b$ Given the data sequence $(\mathcal{X},\mathcal{Y})$ we define blocks B_k, for $k = 1,\ldots,(N-l+1)$ of l consecutive pairs of observations, such that $B_k = \{(X_k,Y_k),\ldots,(X_{k+l-1},Y_{k+l-1})\}$;

B2$_b$ Resample with replacement from the set of blocks $\{B_k;\ k = 1,\ldots,(N-l+1)\}$ with equal probabilities to obtain bootstrap data blocks $B_1^\star,\ldots,B_m^\star$, where $ml \asymp N$;

B3$_b$ Form a new data sequence by joining new blocks $B_1^\star,\ldots,B_m^\star$ into one block only $\Rightarrow$ obtain the bootstrapped data sequence $(\mathcal{X}^\star,\mathcal{Y}^\star) = \{(X_i^\star,Y_i^\star);\ i = 1,\ldots,ml\}$;

B4$_b$ Use data $(\mathcal{X}^\star,\mathcal{Y}^\star)$ to re-estimate $m(\cdot)$ (or $m_+(\cdot)$ and $m_-(\cdot)$ respectively) at some given point $x \in (0,1) \Rightarrow$ obtain the bootstrap estimate $\widehat{m}^\star(x)$ (or $\widehat{m}_+^\star(x)$ and $\widehat{m}_-^\star(x)$ respectively);

B5$_b$ Repeat steps B2$_b$ $\rightarrow \cdots \rightarrow$ B4$_b$ to obtain new bootstrap estimates $\widehat{m}_b^\star(x)$, $\widehat{m}_{b+}^\star(x)$ and $\widehat{m}_{b-}^\star(x)$ for $b = 1,\ldots,B$ for $B \in \mathbb{N}$ to be sufficiently large;

B6$_b$ Use the bootstrapped quantities $\widehat{m}_b^\star(x)$, $\widehat{m}_{b+}^\star(x)$ and $\widehat{m}_{b-}^\star(x)$ for $b = 1,\ldots,B$ to mimic the unknown distribution of interest;

end of MBB

Comparing now the moving blocks bootstrap algorithm and the smooth residual bootstrap it seems that MBB is somehow simpler and more data-driven. There is no estimation of the residuals involved as well as any estimation of the unknown scale function $\sigma(\cdot)$ neither. We have already mentioned that the block bootstrap procedure can be seen as a more complex generalization of the model based bootstrap where one just resamples data pairs, which is more straightforward approach to use indeed.

THEOREM 4.3 (Moving blocks bootstrap consistency for α-mixing)
Let us assume model $\boxed{2.47}$ *and the same assumptions as in Theorem 2.11. Moreover, let $ml \asymp N$. Then the following convergence in probability is achieved*

$$\sup_{z\in\mathbb{R}}\left\{\mathbf{P}^\star\left[\sqrt{Nh_N}\,(\widehat{m}^\star(x)-\widehat{m}(x)) \le z\right] - \mathbf{P}\left[\sqrt{Nh_N}\,(\widehat{m}(x)-m(x)) \le z\right]\right\} \xrightarrow[N\to\infty]{\mathbf{P}} 0,$$

where $\mathbf{P}^\star[\ \cdot\ ]$ stands for a conditional probability conditioned on the given finite sample data sequence $(\mathcal{X},\mathcal{Y}) = \{(X_i,Y_i);\ i = 1,\ldots,N\}$.

Proof. The proof of this theorem is given in Section 4.3.2 below. ■

Given Theorem 4.3 we again have a convenient property of MBB, which is that the moving blocks bootstrap provides a consistent estimate of the unknown limit distribution function of interest at the given point $x \in (0,1)$ and therefore, using the moving blocks bootstrap algorithm as defined in steps B1$_b$ – B6$_b$ one can effectively approximate the bootstrapped distribution rather than using some plug-in techniques and trying to express the distribution of interest directly.

Corollary 6
Similar theorems can be also formulated to show the bootstrap consistency with respect to the asymptotic distributions stated in Theorems 3.2 and 3.3 however, under the α-mixing dependence assumption. Additionally, the assertion in Theorem 4.3 can be easily extended to cover the asymptotic distributions of the M-smoothers estimates of the corresponding derivatives up to the order $p \in \mathbb{N}$.

Assuming now the moving blocks bootstrap approach, there is one important issue, which should be definitely addressed here. It relates to an appropriate choice of the length $l \in \mathbb{N}$. Given the finite data sample the length of blocks is fixed but once the sample size increases ($N \to \infty$) we also need to assure that $l \to \infty$ however, sufficiently slow. This is due to the form of dependence within the bootstrapped blocks where the dependence structure is fully preserved in each block but between blocks the dependence is cancelled out by an independent re-sampling. This would also mean that under the fixed length even for $N \to \infty$ the variance-covariance structure would be consistently estimated up to the $(l-1)$-degree covariances $\mathbb{C}\mathrm{ov}(X_i, X_{i+l-1})$ at most. Therefore, in order to estimate the true variance-covariance structure one needs to assume that that $l \to \infty$ as $N \to \infty$.

The role of this length parameter is crucial as it also directly affect the bootstrap estimate and its empirical inference too. Given this fact we could refer to this parameter as to some tuning parameter instead. The optimal value of this parameter needs to be however, still defined.

One way proposed to chose the optimal length of blocks is to minimize the Asymptotic Mean Square Error (AMSE) term of the bootstrap estimate with respect to the length of blocks. It was shown in Fitzenberger (1997) that the MBB approach for the sample mean yields an asymptotic optimal length of blocks expressed as $l \asymp N^{1/3}$ (see Theorem 3.4 in Fitzenberger (1997)). Given the fact that the M-smoothers estimates can be also thought of in sense of some form-specific means, we could assume that a similar condition will hold for MBB under the M-smoothers approach as well.

The question on the optimal length of blocks for a finite sample data is however, still left open. There is a nice discussion on this matter presented in Fitzenberger (1997) but we will not discuss this topic in further details. Let us focus our attention on justification of the MBB approach instead.

4.3.2 Justification of the MBB algorithm

Firstly, one needs to realize the main difference between the residual based bootstrap discussed in Sections 4.1 and 4.2 and the moving blocks bootstrap discussed now. Indeed, the MBB algorithm is based on the original data sequence $(\mathcal{X}, \mathcal{Y})$ rather than the set of residuals and it does not requires any estimation of the regression function $m(\cdot)$, the scale function $\sigma(\cdot)$ or the set of residuals either.

Unlike the minimization problem (4.1) we have for MBB the minimization defined by

$$\widehat{\boldsymbol{\beta}}_x^{\star\star} = \underset{(b_0,\ldots,b_p)^\top \in \mathbb{R}^{p+1}}{Argmin} \sum_{i=1}^{\widetilde{N}} \rho \left(Y_i^\star - \sum_{j=0}^{p} b_j (X_i^\star - x)^j \right) \cdot K\left(\frac{X_i^\star - x}{h_N} \right), \tag{4.9}$$

where $(X_i^\star, Y_i^\star)$, for $i = 1, \ldots, \widetilde{N} = lm$ are just bootstrap data pairs given by joining m bootstrapped blocks of length l into one big block only. This can be analogously expressed as the set of equations

$$\frac{1}{\sqrt{\widetilde{N} h_N}} \sum_{i=1}^{\widetilde{N}} \psi \left(\sigma\left(X_i^\star\right) \varepsilon_i^\star - \sum_{j=0}^{p} b_j \left(\frac{X_i^\star - x}{h_N} \right)^j \right) \cdot \left(\frac{X_i^\star - x}{h_N} \right)^l K\left(\frac{X_i^\star - x}{h_N} \right) = 0, \tag{4.10}$$

for $\quad l = 0, \ldots, p,$

which is solved for the vector of parameter estimates $\widehat{\boldsymbol{\beta}}_x^{\bullet\bullet} \in \mathbb{R}^{p+1}$ (the idea of the notation remains the same as in the previous chapters). The random error quantities $\{\varepsilon_i^\star\}_{i=1}^{\widetilde{N}}$ are defined by an equivalent model expression $\varepsilon_i^\star = \left(Y_i^\star - m(X_i^\star)\right) / \sigma(X_i^\star)$, for $i = 1, \ldots, \widetilde{N} = lm$.

The idea of the proof is now very similar to the proof of Theorem 4.2 we just need to assume some additional statistical machinery to correctly deal with the α-mixing dependence structure.

Let us firstly deal with the quantity $\mathbb{E}^\star \left[\psi\left(\sigma(X_i^\star)\varepsilon_i^\star - \sum_{j=0}^p b_j \left(\frac{X_i^\star - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i^\star - x}{h_N}\right)^l K\left(\frac{X_i^\star - x}{h_N}\right)\right]$, for some $i \in 1, \dots, \widetilde{N}$ where $\mathbb{E}^\star[\cdot]$ stands for a conditional expectation conditioned on $(\mathcal{X}, \mathcal{Y})$. It holds that

$$\begin{aligned}
&\mathbb{E}^\star \left[\psi\left(\sigma(X_i^\star)\varepsilon_i^\star - \sum_{j=0}^p b_j \left(\frac{X_i^\star - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i^\star - x}{h_N}\right)^l K\left(\frac{X_i^\star - x}{h_N}\right)\right] = \\
&= \frac{1}{N-l+1} \sum_{i=1}^{N-l+1} \frac{1}{l} \sum_{t=i}^{i+l-1} \psi\left(\sigma(X_t)\varepsilon_t - \sum_{j=0}^p b_j \left(\frac{X_t - x}{h_N}\right)^j\right) \cdot \left(\frac{X_t - x}{h_N}\right)^l K\left(\frac{X_t - x}{h_N}\right) = \\
&= \frac{1}{l(N-l+1)} \left\{ l \sum_{i=1}^{N} \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^p b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) + \right. \\
&\qquad \left. + \left[(l-1)\widetilde{\psi}_1 + (l-2)\widetilde{\psi}_2 + \dots + \widetilde{\psi}_{b-1} + (l-1)\widetilde{\psi}_N + (l-2)\widetilde{\psi}_{N-1} + \dots + \widetilde{\psi}_{N-l+2}\right] \right\} = \\
&= \frac{1}{N} \sum_{i=1}^{N} \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^p b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right) + O_{\mathbf{P}}\left(\frac{l}{N}\right), \qquad \boxed{4.11}
\end{aligned}$$

where for short we have adopted a notation $\widetilde{\psi}_i = \psi\left(\sigma(X_i)\varepsilon_i - \sum_{j=0}^p b_j \left(\frac{X_i - x}{h_N}\right)^j\right) \cdot \left(\frac{X_i - x}{h_N}\right)^l K\left(\frac{X_i - x}{h_N}\right)$, for $i = 1, \dots, N$. Moreover, given the assumption for $l \in \mathbb{N}$ we have that $l/N \to 0$ as $N \to \infty$. Using now the strong law of large numbers for the α-mixing dependence we have that the sum in $\boxed{4.11}$ converges $[\mathbf{P}]$-almost surely to its mean expectation for $N \to \infty$.

An analogous expression can be also derived for the second moment $\mathbb{E}^\star[\cdot]^2$ therefore, using now both results one can analogously proceed along the lines of the proof of Theorem 4.2. To complete the proof we will need one more theorem, which will be applied to show that

$$\mathsf{X}_{\widetilde{N}}^{\star\top} \mathsf{W}_{\widetilde{N}}^\star \boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}^\star) \Big| (\mathcal{X}, \mathcal{Y}) \xleftrightarrow[N \to \infty]{\mathscr{D}(\mathbf{P})} \mathsf{X}_N^\top \mathsf{W}_N \boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}), \qquad \boxed{4.12}$$

where $\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}^\star) = (\psi(\sigma(x)\varepsilon_1^\star), \dots, \psi(\sigma(x)\varepsilon_{\widetilde{N}}^\star))^\top$ and analogously also for $\boldsymbol{\psi}(\sigma(x)\boldsymbol{\varepsilon}) \in \mathbb{R}^N$.

THEOREM 4.4 (Bootstrap central limit theorem for α-mixing)

Let $\{\xi_n\}_{n=1}^\infty$ be a sequence of random variables with a zero mean parameter and the α-mixing dependence coefficients $\alpha(n)$. Let the following conditions are satisfied:

(i) $\sup_{n \in \mathbb{N}} \mathbb{E}|\xi_n|^{4+\omega} < \infty$,

(ii) $\alpha(n) = O(1/n^{1+\delta})$, *as* $n \to \infty$,

both for some $\omega > 0$ and $\delta > 0$, such that $4/\omega < \delta$. Let moreover

$$\widehat{\xi}_n = \frac{1}{n}\sum_{i=1}^n \xi_i, \quad \widehat{\xi}_n^\star = \frac{1}{n}\sum_{i=1}^n \xi_i^\star \quad \text{and} \quad \sigma_n^2 = \frac{1}{n}\sum_{i,j=1}^n \mathsf{Cov}(\xi_i, \xi_j).$$

Then for $l \to \infty$ and $l = o(n^{1/2})$ and under the MBB algorithm for the mean it holds that

$$\sup_{z \in \mathbb{R}} \left| \mathbf{P}^\star\left(\frac{n}{\sigma_n}(\widehat{\xi}_n^\star - \widehat{\xi}_n) \le z\right) - \mathbf{P}\left(\frac{n}{\sigma_n}\widehat{\xi}_n \le z\right) \right| \xrightarrow[N \to \infty]{\mathbf{P}} 0.$$

Proof. The proof of Theorem 4.4 can be found in Fitzenberger (1997, Theorem 3.1). ■

Given the assertion of Theorem 4.4 we easily obtain that (4.12) holds for $N \to \infty$ however, we have to additionally assume that $\mathbb{E}|\psi(\varepsilon_1)|^{4+\omega} < \infty$ for some $\omega > 0$ in order to satisfy *(i)* in Theorem 4.4. The rest of the proof follows the same idea as the proof of Theorem 4.2 and it can be easily completed using the Mallow metric machinery all over again. ■

Having now the justification of the MBB algorithm completed we can conclude that the bootstrap distribution can be again consistently used as an efficient replacement for the unknown limit distribution, which is mostly of the main interest (making test decisions or constructing confidence intervals). In the proposed moving blocks bootstrap algorithm we have a straightforward and quite simple tool in hand how to mimic some unknown distribution of interest while preserving the consistency property required for the whole approach to work.

4.4 Discussion on bootstrap methods

The bootstrap simulation techniques are well known in statistics for many years now however, they were never so intensively used before effective computers and simulation algorithms were available. Taking into account the modern computer age we live in, the bootstrap methods become a very popular choice in statistic (theoretically and practically as well) as they requires much less technical effort than the classical plug-in techniques used instead.

We have discussed in detail mainly two different bootstrap algorithms, which are quite suitable for the model scenarios we were dealing with in this thesis. One can however, propose and investigate many other bootstrap modifications and adaptations as well. There is certainly a way to propose a model based bootstrap for independent random errors and some blocks bootstrap modifications for the weak dependence concept (e.g. a blocks of blocks bootstrap or a circular block bootstrap).

We have proved that under the right regularity conditions the bootstrap idea in the M-smoothers estimation really works same for the model scenarios with independent random error terms (under the both, homoscedastic and heteroscedastic assumption) as well as the model scenarios with the α-mixing dependence concept. We have stated two important theorems, which clearly states that the bootstrap distributions can be indeed used as suitable asymptotic replacements for the true but unknown limit distributions, which are mostly of interest. Moreover, using the proposed bootstrap algorithms we can effectively simulate from such bootstrap distributions and the results of such simulations can be consistently used to mimic the distribution of interest.

Summarizing now all attributes of the proposed bootstrapping idea in the M-smoothers regression approach we have to acknowledge their nice properties and many advantages therefore, we are convinced that the bootstrap methods dominate the plug-in techniques, which may be used alternatively but technically they are quite extensive with highly sophisticated proofs as well. On a final note, we want to refer to Chapter 6 where we have studied the finite sample performance of the proposed bootstrap methods through out an extensive simulation study and a real data example as well.

"Let no man who is not a Mathematician read the elements of my work."

Leonardo da Vinci
(1452 – 1519)

5 Additions and miscellaneous

We have discussed the M-smoothers regression methods under the variety of different situations and we have also derived and proved all their important statistical properties. We have also mentioned the M-smoothers approach under some discontinuity assumptions and the proper statistical background was derived. Similarly, the appropriate bootstrap algorithms were proposed in order to deal with situations when the target distribution of interest is unknown but crucial for some further processing or statistical decisions making. This can occur if one wants to construct confidence intervals or to come up with a decision about some hypothesis testing problem regarding some change-point occurrences.

However, beside the theoretical results, which were already derived and proved one also needs to work with some additional rather technical issues to be able to make practical and liable decisions and conclusions. Using the M-smoothers approach one mostly meets a requirement to have a sufficient knowledge on some additional quantities, which are mostly not directly specified by the M-smoothers estimation approach itself. Indeed, the knowledge of the right value of the smoothing parameter h_N or the corresponding estimate of the scale function $\sigma(\cdot)$ in heteroscedatic scenarios is crucial and one certainly needs to have good estimates for these quantities no matter what approach (plug-in techniques or bootstrap simulations) he uses to successfully complete the whole task.

In this chapter we will focus on some additional issues related to nonparametric M-smoothers, which are rather not crucial for the estimation method itself but they are also important in order to be able to fully exploit all advantages of the proposed M-smoothers methods. We will discuss some robust approaches to the optimal bandwidth parameter selection and some robust methods for the scale function estimation as well. Finally, we will shortly point our attention on some different choices of the loss function $\rho(\cdot)$, which can be effectively used together with the minimization problem (2.1).

5.1 Robust bandwidth selection

It is a well known fact that once we use nonparametric regression approaches it is crucial for the final performance of the proposed estimator to use a good choice for the smoothing parameter,[32] which controls the amount of smoothness in the final fit. However, a problem on a good choice of the bandwidth parameter arises in statistical modelling since Nadaraya (1964) and Watson (1964) when they firstly proposed to use the kernel methods for a curve estimation. The right choice of the bandwidth parameter is essential especially with respect to the asymptotic properties of the estimator as it assures an optimal convergence and all additional statistical properties nevertheless, it also plays a crucial role

[32] We will equivalently refer to the bandwidth parameter as to the smoothing parameter as well as it really takes a role of controlling the amount of smoothness.

in case of finite data samples where it effectively manages the bias and variance balance (bias-variance trade-off) in the produced estimate.

Standard methods for the bandwidth selection are based on evaluating a quality of a statistical estimate in order to determine how close to the true curve, which is of interest, the given estimate really is. Some popular criteria used here are Integrated Squared Error, Mean Integrated Squared Error, Average Squared Error and Mean Average Squared Error, which all heavily depend on the bandwidth parameter itself. The optimal bandwidth parameter is then defined as the one that minimizes some of the measure of error criterion above with respect to all possible bandwidth choices and given the finite data sample under the consideration.

All of the criteria above can be thought to be asymptotically similar (see Härdle (1990)) however, all of them were designed for models with normally distributed random error terms. It was shown in simulations and some theoretical calculations as well that using such criteria with even a small amount of outlying observations one can get an extremely biased bandwidth selection (e.g. see Leung et al. (1993) or Leung (2005) for some brief overviews). Not to mention even some heavy-tailed random error distributions. It is important to notice, that using the given criteria with outlying observations or heavy-tailed distributions of random errors will fail in producing a proper estimate even in the case the curve estimate is defined by a proper robust minimization. This is due to the fact that such criteria do not produce reasonable estimates for random errors anymore therefore, one has to use some appropriate robust equivalents for these criteria as well to get a fully consistent estimation machinery for the M-smoothers estimation method.

We will base our approach on a suggestion proposed by Ronchetti et al. (1997) where the authors used Cross-Validation (CV) method together with nonparametric regression to get an optimal bandwidth parameter. The Cross-Validation quantity was shown to be an efficient estimate for the Asymptotic Mean Squared Error term and the bandwidth parameter, which minimizes the CV function is said to be an optimal choice under the given finite data sample.
Some further generalizations as General Cross-Validation (GCV) were proposed afterwards however, we will rather stay focussed on the Cross-Validation methods only and we will discuss some of their robust generalizations introduced to account for outlying observations and heavy-tailed distributions too.

Boente et al. (1997) introduced another bandwidth selection method for the robust nonparametric regression estimation however, their method is not fully data-driven as a pivotal bandwidth parameter is required to be given in advance. In our bandwidth selection approach we would like to stay fully automatic therefore, we will adopt and we will further discuss the M-Cross-Validation method (Robust Cross Validation (RCV) respectively) proposed by Lee and Cox (2010), which was shown to produce a consistent estimate for the AMSE quantity, which we have already discussed in Chapter 2.

The original leave-one-out Cross-Validation is the most often used for the smoothing parameter selection in statistic in general (for a nice summary book see Simonoff (1996)). It is defined as

$$\mathsf{CV}(h_N) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i - \widehat{m}_{h_N}^{(-i)}(X_i)\Big)^2, \tag{5.1}$$

where $\widehat{m}_{h_N}^{(-i)}(\cdot)$ is a corresponding estimate of the unknown but true regression function $m(\cdot)$ given the bandwidth parameter h_N while leaving the i^{th} observation out of data used for the estimation. Many extensive discussions and examples can be found on this criterion in literature.

To deal with some inconsistency issues involved in the smoothing parameter selection when using classical criteria with outlying observations or heavy-tailed distributions Lee and Cox (2010) suggested to replace the square loss function in (5.1) with a more general loss function $\rho(\cdot)$ – the same one as we have used for the estimation processes in Chapter 2. This gives us a general Cross-Validation expression

$$\mathsf{RCV}(h_N) = \frac{1}{N}\sum_{i=1}^{N} \rho\left(Y_i - \widehat{m}_{h_N}^{(-i)}(X_i)\right), \tag{5.2}$$

where again $\widehat{m}_{h_N}^{(-i)}(\cdot)$ is the corresponding estimate of the unknown regression function $m(\cdot)$ given the specific value of the bandwidth parameter h_N however, calculated with respect to general loss function $\rho(\cdot)$ instead. Function RCV$(\cdot)$ is called the Robust Cross-Validation function or M-Cross-Validation respectively. The optimal value of the bandwidth parameter is of course again defined as the value, which minimizes the RCV(h_N) quantity (5.2).

One special case of the Robust Cross-Validation function we can get for the choice $\rho(\cdot) = |\cdot|$, which gives us so called Absolute Cross-Validation (ACV) proposed by Wang and Scott (1994) and also discussed in Yang (2006). Another special case of the RCV$(\cdot)$ function was studied by Boente et al. (1997) where the authors plugged the Huber function into (5.2) instead.

To conclude, Robust Cross-Validation approach as defined in (5.2) can be effectively used together with the proposed M-smoothers estimation techniques in order to present a complete and statistically consistent nonparametric regression estimation machinery, which is capable of accounting for some presence of outlying observations or heavy-tailed distributions of random errors as well. As far as the robust bandwidth selection techniques have been already extensively discussed in many literature we will not further discuss this topic in this thesis.

5.2 Robust scale estimation

Another important issue, which we also need to think of in case of the M-smoothers approaches (at least those with heteroscedastic variance structure) is the scale function $\sigma(\cdot)$ involved in model (2.39). The scale function $\sigma(\cdot)$ is same as function $m(\cdot)$ unknown and under some circumstances it also needs to be estimated. Especially, once we consider the asymptotic distribution of the M-smoothers estimates under the heteroscedastic model scenario we need to plug-in the estimate of the scale function $\sigma(\cdot)$ into the variance expression in order to be able to construct confidence intervals or to draw any conclusions regarding some hypothesis testing problems.

Of course, one could also avoid plugging-in the scale estimate into the variance expression and to rather turn his attention to bootstrap methods discussed in Chapter 4 as the rate of convergence was shown to be much faster for simulation methods than for plug-in techniques however, the scale function $\sigma(\cdot)$ needs to be estimated in the case of bootstrap approximations anyway therefore, estimation of the scale function within heteroscedastic model structures is an important issue by its own.

There was already one approach mentioned how to introduce a scale function estimate where the estimator came directly from the simultaneous M-smoothers estimation problem where one estimates the regression function and the scale at once. We have however, discussed a two-stage estimation process where the scale function $\sigma(\cdot)$ was to be estimated separately. We will therefore focus on this second stage of estimation – the estimation of the scale function – now.

Let us start with a scale estimation approach, which comes as a generalization of a class of estimators based on differences (see Hall et al. (1990)) defined by the expression

$$\widehat{\sigma}^2 = \frac{1}{N-1} \cdot \sum_{i=k_1+1}^{N-k_2} \left(\sum_{k=-k_1}^{k_2} d_k Y_{i+k} \right)^2 , \tag{5.3}$$

which are commonly used for homoscedastic model scenarios with regressor quantities $X_1, \dots, X_N$ to be fixed constants. Here in (5.3) the sequence $\{d_k\}_{k=-k_1}^{k_2}$ stands for a difference sequence, which satisfies that $\sum_{k=-k_1}^{k_2} d_k = 0$ and $\sum_{k=-k_1}^{k_2} d_k^2 = 1$, where $d_{-k_1} \neq 0 \neq d_{k_2}$ for some k_1 and k_2 to be non-negative integers. Referring now to the quantity $k_1 + k_2$ we have the order of the scale estimator (5.3). The simplest class of this estimators comes for the order $k_1 + k_2 = 1$, which are well known scale estimators proposed by Rice (1984) and they are defined as

$$\widehat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=1}^{N} (Y_{i+1} - Y_i)^2 , \tag{5.4}$$

for quantities $X_1, \dots, X_N$ to be fixed again. Some additional extensions were proposed by Müller and Stadtmüller (1987) and Brown and Levine (2007) in order to fit these estimators to heteroscedastic regression cases as well as random design cases too.

The problem however, which comes with this class of estimators is similar to the problem related to a bandwidth parameter selection in a classical nonparametric least squared regression, which means that even a small presence of outliers can cause failing of these methods. Indeed, estimators based on squared differences in general do lack a property of being robust with respect to outliers.

To avoid such problems one need to think of some alternative approaches or additional generalizations of classical methods, which could be plausible under the model with outlying observations or even heavy tailed distributions of random errors. One possible generalization was proposed by Boente et al. (2010) and it can be considered as a step towards heteroscedastic regression scenarios with outlying observations and heavy-tailed distributions and it is defined as

$$\widehat{\sigma}(x) = \inf \left\{ z > 0; \ \sum_{i=1}^{N-1} w_{Ni}(x) \cdot \chi \left(\frac{Y_{i+1} - Y_i}{\alpha_1 \cdot z} \right) \leq \alpha_2 \right\} , \tag{5.5}$$

where $w_{Ni}(x)$ are weights (e.g. kernel weights) for $i = 1, \dots N-1$ and function $\chi(\cdot)$ is some score function. The constants $\alpha_1 > 0$ and $\alpha_2 \in (0,1)$ are chosen such that they satisfy that $\mathbb{E}\chi(Z_1) = \alpha_1$ and $\mathbb{E}\chi\left(\frac{Z_2 - Z_1}{\alpha_1}\right) = \alpha_2$, where Z_1 and Z_2 are independent random variables with the distribution function, which corresponds with the major distribution function or the random error terms[33]. Moreover, for the score function $\chi(\cdot)$ we assume it is continuous, bounded and also strictly increasing moreover, such that $\chi(0) = 0$ and $0 < \sup_{x \in \mathbb{R}} \chi(x)$.

The estimator (5.5) is called the local M-estimator of the scale function $\sigma(\cdot)$ at the point $x \in (0,1)$ based on successive differences of the response variable Y_i, for $i = 1, \dots, N$. Obviously, for $\chi(x) = x^2$, $\alpha_1 = \sqrt{2}$ and $\alpha_2 = 1$ the Boente estimator reduces to the classical Rice estimator (5.4).

[33] In case of outlying observations we can describe the distribution function of random errors as a kind of mixture of at least two different distributions where the first one (where the majority of random error terms come from) is referred to as a major distribution and all remaining distributions responsible for any outliers, which contribute with much smaller weights are referred to as minor distributions.

Under some regularity conditions (see Boente et al. (2010) for further details and exact proofs) it was derived that the estimate of the scale function $\sigma(\cdot)$ defined by 5.5 for some given point of interest $x \in (0,1)$ yields a strong consistency and asymptotic normality property once the number of the sample size N tends to infinity.

5.3 Robust loss functions

Finally, we will shortly mention some other choices of the loss function $\rho(\cdot)$, which can be possibly used with the minimization problem 2.1. Until now we have mainly discussed three options here: the L_2 norm, which produces an estimate in sense of a conditional expectation, the L_1, norm which leads to an estimate in sense of a conditional median estimation and finally, the Huber function, which can be thought of as a bridge between the L_2 and L_1 norm and it is in general considered to be the most common representative used for the robust estimation procedures.

On the other hand, given the assumption A4 where the loss function $\rho(\cdot)$ is assumed to be symmetric, convex and Lipschitz such that its derivative (or one-sided derivatives at least) exists one can possibly think of many different choices of the loss function $\rho(\cdot)$. One just needs to make sure, that the given choice of the loss function satisfies all necessary assumptions namely, the assumption A4 and together with the distribution function $G(\cdot)$ the assumption A5 as well.

In Table 5.1 we present a small overview of some different loss functions, which are also commonly used for the modern robust estimation procedures. Three of them have been already discussed in detail in the previous chapters but the remaining functions can be also found in literature or statistical software packages.

Finally, let us give a short note on different choices of the loss functions $\rho(\cdot)$. One has to be aware of the fact that a different choice of the loss function $\rho(\cdot)$ brings also in hand a different interpretation for the true model. Indeed, as we have already said, using the most common L_2 norm we obtain an estimate in sense of a conditional expectation, which is $m(x) = \mathbb{E}\left[Y|X=x\right]$. Similarly, for the L_1 norm we have an estimate in sense of a conditional median, which is $m(x) = \mathbb{M}\text{ed}\left[Y|X=x\right]$.

Unlike this two cases the identification of the true but unknown regression function $m(x)$ rapidly changes once we use some other loss function $\rho(\cdot)$ to produce and estimate. Indeed, for a general loss function $\rho(\cdot)$ the conditional mean (or conditional median respectively) interpretation does not hold any more. Instead, we have to identify the unknown regression function $m(\cdot)$ in a more general way as

$$m(x) = Argmin \quad \mathbb{E}\left[\rho\left(Y - \mu(X)\right)|X=x\right], \tag{5.6}$$

where the minimization takes place with respect to all functions $\mu \in \mathscr{L}_{p+1}(0,1)$, where $\mathscr{L}_{p+1}(0,1)$ is a set of $(p+1)$-times Lipschitz function defined over the interval $(0,1)$. Using now the identification 5.6 for some unknown regression function $m(\cdot)$ we can also correctly define the estimate of $m(\cdot)$ in terms of the M-smoothers approach while considering any general loss function, which satisfies all assumptions. The functional identification makes indeed an important aspect of the whole concept of the robust M-smoothers regression approach and one needs to be fully aware of different interpretation options, which originate in different loss functions, which can be used for the estimation process.

Referring in this thesis to the robust estimation approaches it is expected that the proposed methods will be less sensitive with respect to outlying observations and heavy-tailed random error distributions than their classical counterparts mostly based on the L_2 norm. It is worth of mention here that

Loss function type *(common name)*	**Loss function expression** *function* $\rho(\cdot)$	**Loss function derivative** *function* $\psi(\cdot)$	**Domain** *for* $x \in \mathbb{R}$
▪ L_2 norm	$\frac{x^2}{2}$	x	$x \in \mathbb{R}$
▪ L_1 norm	$\|x\|$	$sgn(x)$	$x \in \mathbb{R}$
▪ $[L_1 - L_2]$ norm	$2\left(\sqrt{\frac{1+x^2}{2}} - 1\right)$	$\frac{x}{\sqrt{\frac{1+x^2}{2}}}$	$x \in \mathbb{R}$
▪ L_p norm	$\frac{\|x\|^p}{p}$	$sgn(x)\|x\|^{p-1}$	$x \in \mathbb{R}$
▪ “Fair” function	$c^2\left[\frac{\|x\|}{c} - \log\left(1 + \frac{\|x\|}{c}\right)\right]$	$\frac{x}{1+\|x\|/c}$	$x \in \mathbb{R}$
▪ Huber's function	$\begin{cases} \frac{x^2}{2} \\ k\left(\|x\| - \frac{k}{2}\right) \end{cases}$	$\begin{cases} x \\ k \cdot sgn(x) \end{cases}$	$\begin{cases} \|x\| \leq k \\ \|x\| > k \end{cases}$
▪ Cauchy's function	$\frac{c^2}{2}\log\left(1 + \frac{x^2}{c^2}\right)$	$\frac{x}{1+(x/c)^2}$	$x \in \mathbb{R}$
▪ Geman-McClure	$\frac{x^2/2}{1+x^2}$	$\frac{x}{(1+x^2)^2}$	$x \in \mathbb{R}$
▪ Welsch's function	$\frac{c^2}{2}\left[1 - \exp\left\{-\frac{x^2}{c^2}\right\}\right]$	$x \exp\left\{-(x/c)^2\right\}$	$x \in \mathbb{R}$
▪ Tuckey's function	$\begin{cases} \frac{k^2}{6}\left[1 - \left(1 - (x/k)^2\right)^3\right] \\ (k^2)/6 \end{cases}$	$\begin{cases} x\left[1 - (x/k)^2\right]^2 \\ 0 \end{cases}$	$\begin{cases} \|x\| \leq k \\ \|x\| > k \end{cases}$

Figure 5.1: Some additional options for the loss function $\rho(\cdot)$, which can be also used for the proposed M-smoothers estimation methods. Some loss functions are defined for the appropriate trimming constants $k > 0$ and $c > 0$.

robustness property or sensitivity with respect to outliers can be measured in a quite objective way, using some measures of robustness. A popular measure of robustness at least for the location models is a finite sample break-down point, which is defined as a proportion of outliers within the original finite sample data such that this proportion can already cause a failure of the whole estimator. However, the idea of the break-down points can be also extended to higher order models as well.
It is a well known fact that the finite sample break-down point for the estimators based on a classical mean approach or squared differences is equal to $1/N$, which is actually the smallest one possible therefore, such methods are not consider to be robust[34] at all.

On the other hand, estimators based on L_1 norm – so called median estimators – yield the property of being very robust with respect to outlying observations and even heavy-tailed distributions and the actual finite sample break-down point for the L_1 based methods is equal to $1/2$, which is the maximal possible proportion for all common estimation methods[35]. To investigate the final sample break-down points for different choices of the loss function $\rho(\cdot)$ one needs to incorporate some additional computations. If a reader is interested in more details we refer to Jurečková (2001).

[34] Finite sample break-down point equal to $1/N$ means that just one outlying observation out of N can cause that the whole estimate produced by some non-robust procedure will fail.

[35] In this case the whole half of data points can be produced by some artificial system (e.g. outliers, or data errors) but using the robust approach based on the L_1 norm the final estimate will still have a property of being consistent.

*"God does not care about
our mathematical difficulties.
He integrates empirically."*

Albert Einstein
(1879 – 1955)

6 Computational Aspects of M-smoothers

Until now, we have discussed the M-smoothers regression estimates and the corresponding statistical tests from the theoretical point of view – asymptotically – as we have considered the sample size N to tend to infinity. However, this is never the case in real data problems therefore, it is always important to investigate some finite sample properties for a proposed statistical method as well.

We have derived all important statistical properties and qualities for the proposed M-smoothers approach under the variety of different scenarios. We have proposed the estimation methods and testing procedures as a complex statistical set of tools for dealing with models that will be flexible with respect to the considered shape of the functional dependence, the variance-covariance structure as well as the assumed dependence concept for the random error terms while also allowing for some discontinuity points and structural breaks.

In this chapter we will focus on some computational aspects of the M-smoothers techniques and we will investigate the effect of different model settings to the performance of the proposed estimator. We will firstly discuss some results from an extensive simulation study, carefully designed to study the given estimator and after that, we will also present some results obtained when applying these testing and estimating procedures to a real data problem. All programming work was provided using the statistical software R 2.11.1 and our own procedures.

6.1 Simulation study

For the simulation purposes we have proposed to use the following regression function defined by

$$m(x) = -8x\sin(2\pi x) + 7.2x\sin(2\pi x)\cdot \mathbb{I}_{\{x>0.5\}} + 0.2\cdot \mathbb{I}_{\{x>0.8\}}, \tag{6.1}$$

for $x \in (0,1)$ and $\mathbb{I}_{\{\cdot\}}$ to be an identifier function of the corresponding event of interest.

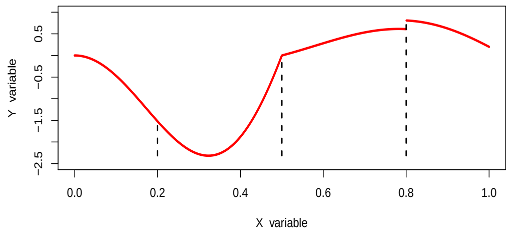


Figure 6.1: Regression function used for simulations.

For function (6.1) one can easily verify, that function $m(\cdot)$ is continuous and even smooth up to the third order at least at the point $x_1 = 0.2$, it has a first order discontinuity at the point $x_2 = 0.5$ (a jump in the first derivative) and finally, it has a zero order jump at the point $x_3 = 0.8$ (a jump in the regression function itself) with one-sided derivatives both to be equal (see Figure 6.1 on the left for a brief overview of the chosen function).

We will carefully focus on the performance of the M-smoothers estimation approach for these three specific points while assuming a variety of different regression options.
Firstly, we will deal with the M-smoothers estimator given for a symmetric kernel function and after that we will also mention one-sided M-smoothers with the corresponding test statistics.
Similarly, we will also discuss the performance of the bootstrap algorithms when considered together with the corresponding test statistics, which is of interest.

6.1.1 Simulation settings

Let us firstly state the list of different options we have proposed to use for the simulation study:

❒ **RANDOM ERROR DISTRIBUTION** (four different choices)

$\mathscr{D}_1 : \mathbb{N}(0,1)$;
$\mathscr{D}_2 : 0.90 \times \mathbb{N}(0,1) + 0.10 \times N(0,\sigma^2 = 25)$;
$\mathscr{D}_3 : 0.95 \times \mathbb{N}(0,1) + 0.05 \times N(0,\sigma^2 = 225)$;
$\mathscr{D}_4 : \mathbb{C}(0,1)$;

❒ **LOSS FUNCTION** (four different choices)

- classical L_2 norm (squared loss function);
- robust L_1 norm (absolute loss function);
- the Huber function;
- the Tuckey function;

❒ **SAMPLE SIZE** (three different choices)

$N = 50$;
$N = 200$;
$N = 1000$;

❒ **APPROXIMATION ORDER** (three different choices)

- local constant approximation $(p = 0)$;
- local linear approximation $(p = 1)$;
- local cubic approximation $(p = 3)$;

Four different choices for the distribution function can be seen as a certain bridge coming from the classical normality assumption required for the L_2-based methods towards to the regression with some robust flavour where we are firstly adding some outliers and finally, we go for a heavy-tailed distribution at the end. The given four choices for the loss function are quite obvious as well at least for the first three functions as all three of them were already discussed in this thesis for several times. Additionally, we will also consider the Tuckey function, which is somehow the most common representative of robust loss functions, which does not assign positive weights along the whole real line. All three functions before does.

For the sample size we will consider $N = 50$, which is a common sample size for classical data problems while $N = 1000$ is in our opinion already a quite sufficient sample size to show an asymptotic behaviour of the proposed estimators while also managing the time costs and computational efficiency. Finally, for the order of the local polynomial approximation, we have chosen the simplest scenario for $p = 0$ however, in this case the estimates tend to be inconsistent in boundary regions. Given the same reason we have not considered any even orders and given some interpretability options (which we have already mentioned before) we did not consider any higher orders of approximation than $p = 3$.

The tables with results are stated below: in Tables 1, 2 and 3 there are the corresponding results given for the point $x_1 = 0.2$ (Table 1 for local constant M-smoothers, Table 2 for local linear M-smoothers and finally, Table 3 for local cubic M-smoothers). Analogously, in Tables 4, 5 and 6 there are again the corresponding results given for $x_2 = 0.5$ and in Tables 7, 8 and 9 there are the results stated for $x_3 = 0.8$ (again for $p = 0, 1$ and 3).

Each of 9 tables consists of subtables *(a)*, *(b)*, *(c)* and *(d)*. The first two always refer to the performance of the M-smoothers estimator based on the symmetric Epanechnikov kernel function while subtables *(i)* state the estimated values and subtables *(ii)* give the 95 % confidence interval coverage based on the corresponding bootstrap algorithm. The whole estimation procedure is always repeated independently for 1000 times for each regression setting in order to obtain mean estimates and their standard errors as well. The remaining two subtables refer to the case where one-sided M-smoothers (based on the Epanechnikov kernel) are used instead: subtables *(c)* give the values for the test statistics $T_N^{(0)}(x)$ and $T_N^{(1)}(x)$ respectively, and subtables *(d)* state the average lengths of 95 % confidence intervals for the true value of the corresponding test statistic. For subtables *(b)*, *(c)* and *(d)* we have considered the bootstrap algorithm with $B = 1000$ bootstrap re-samples where the whole bootstrap simulation was independently repeated for 1000 times to get the interval coverage data and the mean interval lengths as well. The results are briefly discussed in the section below.

6.1.2 Simulation results

There are a few important issues we would like to specifically highlight before going to the simulation results[36] below. Firstly, by $T_N^{(j)}(x_k)$ for $j = 0, 1$ and $k = 1, 2, 3$ we will understand the difference[37] of the corresponding one-sided estimates with no standardization with respect to the sample size as we want to get a better insight on the asymptotic behaviour with respect to the variance as well.

Now, in Tables 1–9 one can nicely see the asymptotic performance of the M-smoothers methods, which is reflected in gradually improving estimates (decreasing bias terms) and increasing precision (lowering variance terms), which holds true for most of the considered scenarios. However, for the Cauchy distribution and the classical L_2-based regression this is not true anymore and moreover, the variance term even increases with no restraints once the sample size $N \in \mathbb{N}$ tends to infinity. This directly also effects the corresponding lengths of confidence intervals, which get wider for every larger $N \in \mathbb{N}$ while the lengths for the remaining scenarios evolve in a quite opposite direction.

Another important issue refers to the rate of convergence of the given estimates. Indeed, one can clearly see an obvious difference in behaviour of the M-smoothers estimates of the unknown regression function itself and its derivatives respectively, where the asymptotic performance of the M-smoothers estimate for the derivative is evidently much slower however, always quite evident (except the Cauchy distribution with the L_2-based approach again).
Finally, we also need to keep in mind the specific shape of the function used for simulations and the position of the points used for the estimation where $x_1 = 0.2$ is located close to the global minimum and therefore, some over-estimation (especially for small sample sizes) could be expected. Similarly, $x_2 = 0.5$ is the point where a quite extensive change in direction takes place and $x_3 = 0.8$ is a jump location by itself therefore, some small inconsistencies especially for small sample sizes are natural.

However, the proposed simulation study shows that the finite sample properties of the M-smoothers estimates corresponds quite well with the asymptotic results derived before.

[36] A more detailed description of the tables stated below can be found in the List of Tables at the end the thesis (p.123).

[37] This difference corresponds with the test statistic defined by (3.4), up to the multiplicative constant $\sqrt{Nh_N}$.

Table 1(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local constant estimates)

simulation results for $x_1 = 0.2$

SIMULATION settings				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_1) = m'_-(x_1) = -10.7150$	−1.441	–	−1.442	–	−1.443	–	−1.444	–
				(0.370)	*(0.000)*	*(0.519)*	*(0.000)*	*(0.426)*	*(0.000)*	*(0.369)*	*(0.000)*
		$\mathscr{D}_2$		−1.451	–	−1.496	–	−1.474	–	−1.490	–
				(0.484)	*(0.000)*	*(0.453)*	*(0.000)*	*(0.377)*	*(0.000)*	*(0.370)*	*(0.000)*
		$\mathscr{D}_3$		−1.402	–	−1.472	–	−1.484	–	−1.480	–
				(1.441)	*(0.000)*	*(0.464)*	*(0.000)*	*(0.427)*	*(0.000)*	*(0.395)*	*(0.000)*
		$\mathscr{D}_4$		−0.616	–	−1.313	–	−1.352	–	−1.232	–
				(4.747)	*(0.000)*	*(0.816)*	*(0.000)*	*(0.870)*	*(0.000)*	*(0.852)*	*(0.000)*
	$N = 200$	$\mathscr{D}_1$		−1.498	–	−1.520	–	−1.509	–	−1.498	–
				(0.186)	*(0.000)*	*(0.264)*	*(0.000)*	*(0.201)*	*(0.000)*	*(0.188)*	*(0.000)*
		$\mathscr{D}_2$		−1.458	–	−1.482	–	−1.471	–	−1.473	–
				(0.321)	*(0.000)*	*(0.285)*	*(0.000)*	*(0.223)*	*(0.000)*	*(0.226)*	*(0.000)*
		$\mathscr{D}_3$		−1.488	–	−1.424	–	−1.446	–	−1.452	–
				(1.050)	*(0.000)*	*(0.277)*	*(0.000)*	*(0.225)*	*(0.000)*	*(0.211)*	*(0.000)*
		$\mathscr{D}_4$	$m_+(x_1) = m_-(x_1) = -1.5217$	−1.084	–	−1.511	–	−1.508	–	−1.513	–
				(2.873)	*(0.000)*	*(0.321)*	*(0.000)*	*(0.311)*	*(0.000)*	*(0.319)*	*(0.000)*
	$N = 1000$	$\mathscr{D}_1$		−1.508	–	−1.514	–	−1.506	–	−1.507	–
				(0.113)	*(0.000)*	*(0.137)*	*(0.000)*	*(0.120)*	*(0.000)*	*(0.114)*	*(0.000)*
		$\mathscr{D}_2$		−1.512	–	−1.537	–	−1.534	–	−1.522	–
				(0.215)	*(0.000)*	*(0.164)*	*(0.000)*	*(0.137)*	*(0.000)*	*(0.147)*	*(0.000)*
		$\mathscr{D}_3$		−1.413	–	−1.498	–	−1.494	–	−1.509	-
				(0.461)	*(0.000)*	*(0.147)*	*(0.000)*	*(0.119)*	*(0.000)*	*(0.109)*	*(0.000)*
		$\mathscr{D}_4$		−1.034	–	−1.510	–	−1.504	–	−1.510	–
				(3.967)	*(0.000)*	*(0.204)*	*(0.000)*	*(0.202)*	*(0.000)*	*(0.219)*	*(0.000)*

Table 1(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_1 = 0.2$

SIMULATION settings				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 0$	$N = 50$	$\mathscr{D}_1$		96.7 %	–	96.0 %	–	96.2 %	–	96.5 %	–
		$\mathscr{D}_2$	Repeat: ×1000	94.9 %	–	96.3 %	–	95.8 %	–	95.3 %	–
		$\mathscr{D}_3$		96.7 %	–	95.8 %	–	97.9 %	–	95.3 %	–
		$\mathscr{D}_4$		97.2 %	–	96.3 %	–	96.5 %	–	96.5 %	–
	$N = 200$	$\mathscr{D}_1$		97.3 %	–	96.3 %	–	97.9 %	–	97.7 %	–
		$\mathscr{D}_2$		98.0 %	–	96.0 %	–	97.7 %	–	98.2 %	–
		$\mathscr{D}_3$		97.4 %	–	97.1 %	–	97.9 %	–	97.7 %	–
		$\mathscr{D}_4$	$B = 1000$	97.1 %	–	95.9 %	–	97.9 %	–	97.6 %	–
	$N = 1000$	$\mathscr{D}_1$		95.3 %	–	95.4 %	–	95.9 %	–	95.7 %	–
		$\mathscr{D}_2$		96.0 %	–	96.1 %	–	95.7 %	–	95.5 %	–
		$\mathscr{D}_3$		95.4 %	–	96.0 %	–	95.9 %	–	95.7 %	-
		$\mathscr{D}_4$		93.1 %	–	95.6 %	–	95.9 %	–	95.6 %	–

Table 1(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$T_N^{(1)}(x_1) = (m'_+(x_1) - m'_-(x_1)) = 0$ $T_N^{(0)}(x_1) = (m_+(x_1) - m_-(x_1)) = 0$	−0.300 *(0.596)*	– *(0.000)*	−0.263 *(0.708)*	– *(0.000)*	−0.280 *(0.610)*	– *(0.000)*	−0.294 *(0.597)*	– *(0.000)*
		$\mathscr{D}_2$		−0.241 *(0.946)*	– *(0.000)*	−0.272 *(0.922)*	– *(0.000)*	−0.279 *(0.823)*	– *(0.000)*	−0.275 *(0.809)*	– *(0.000)*
		$\mathscr{D}_3$		−0.153 *(2.211)*	– *(0.000)*	−0.271 *(1.201)*	– *(0.000)*	−0.270 *(1.156)*	– *(0.000)*	−0.293 *(0.673)*	– *(0.000)*
		$\mathscr{D}_4$		0.026 *(4.099)*	– *(0.000)*	−0.264 *(1.707)*	– *(0.000)*	−0.227 *(1.548)*	– *(0.000)*	−0.199 *(1.169)*	– *(0.000)*
	$N = 200$	$\mathscr{D}_1$		−0.097 *(0.363)*	– *(0.000)*	−0.120 *(0.420)*	– *(0.000)*	−0.110 *(0.373)*	– *(0.000)*	−0.097 *(0.363)*	– *(0.000)*
		$\mathscr{D}_2$		−0.186 *(0.571)*	– *(0.000)*	−0.137 *(0.459)*	– *(0.000)*	−0.147 *(0.414)*	– *(0.000)*	−0.156 *(0.435)*	– *(0.000)*
		$\mathscr{D}_3$		−0.031 *(1.389)*	– *(0.000)*	−0.099 *(0.456)*	– *(0.000)*	−0.106 *(0.408)*	– *(0.000)*	−0.098 *(0.381)*	– *(0.000)*
		$\mathscr{D}_4$		0.068 *(4.331)*	– *(0.000)*	−0.154 *(0.607)*	– *(0.000)*	−0.165 *(0.581)*	– *(0.000)*	−0.156 *(0.605)*	– *(0.000)*
	$N = 1000$	$\mathscr{D}_1$		−0.068 *(0.198)*	– *(0.000)*	−0.080 *(0.236)*	– *(0.000)*	−0.075 *(0.203)*	– *(0.000)*	−0.068 *(0.199)*	– *(0.000)*
		$\mathscr{D}_2$		−0.113 *(0.315)*	– *(0.000)*	−0.088 *(0.307)*	– *(0.000)*	−0.093 *(0.274)*	– *(0.000)*	−0.098 *(0.269)*	– *(0.000)*
		$\mathscr{D}_3$		−0.035 *(0.737)*	– *(0.000)*	−0.069 *(0.400)*	– *(0.000)*	−0.073 *(0.385)*	– *(0.000)*	−0.069 *(0.224)*	- *(0.000)*
		$\mathscr{D}_4$		0.104 *(5.366)*	– *(0.000)*	−0.097 *(0.569)*	– *(0.000)*	−0.102 *(0.516)*	– *(0.000)*	−0.098 *(0.389)*	– *(0.000)*

Table 1(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	Repeat: ×1000 $B = 1000$	2.3397	–	2.7784	–	2.3916	–	2.3439	–
		$\mathscr{D}_2$		3.7093	–	3.6156	–	3.2275	–	3.1735	–
		$\mathscr{D}_3$		8.6680	–	4.7082	–	4.5315	–	2.6394	–
		$\mathscr{D}_4$		16.071	–	6.6918	–	6.0685	–	4.5825	–
	$N = 200$	$\mathscr{D}_1$		1.4235	–	1.6464	–	1.4639	–	1.4240	–
		$\mathscr{D}_2$		2.2403	–	1.8004	–	1.6260	–	1.7087	–
		$\mathscr{D}_3$		5.4464	–	1.7909	–	1.5997	–	1.4938	–
		$\mathscr{D}_4$		16.980	–	2.3808	–	2.2801	–	2.3751	–
	$N = 1000$	$\mathscr{D}_1$		0.7798	–	0.9261	–	0.7972	–	0.7813	–
		$\mathscr{D}_2$		1.2364	–	1.2052	–	1.0758	–	1.0578	–
		$\mathscr{D}_3$		2.8893	–	1.5694	–	1.5105	–	0.8798	–
		$\mathscr{D}_4$		21.034	–	2.2306	–	2.0228	–	1.5275	–

Table 2(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local linear estimates)

simulation results for $x_1 = 0.2$											
SIMULATION				L_2 norm		L_1 norm		Huber’s function		Tuckey’s function	
settings				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_1) = m'_-(x_1) = -10.7150$	−1.457 *(0.344)*	−8.117 *(5.567)*	−1.419 *(0.436)*	−7.385 *(8.422)*	−1.453 *(0.396)*	−7.425 *(5.412)*	−1.457 *(0.344)*	−7.930 *(5.515)*
		$\mathscr{D}_2$		−1.439 *(0.484)*	−6.733 *(11.05)*	−1.484 *(0.423)*	−7.168 *(9.799)*	−1.475 *(0.352)*	−6.725 *(7.615)*	−1.474 *(0.350)*	−7.422 *(9.261)*
		$\mathscr{D}_3$		−1.385 *(1.499)*	−7.627 *(16.21)*	−1.466 *(0.431)*	−5.290 *(6.504)*	−1.480 *(0.387)*	−6.718 *(7.610)*	−1.465 *(0.374)*	−6.977 *(8.777)*
		$\mathscr{D}_4$		−0.582 *(4.882)*	2.959 *(45.84)*	−1.279 *(0.801)*	−6.859 *(9.980)*	−1.294 *(0.914)*	−5.320 *(18.53)*	−1.221 *(0.916)*	−6.080 *(15.15)*
	$N = 200$	$\mathscr{D}_1$		−1.492 *(0.178)*	−8.979 *(5.132)*	−1.509 *(0.238)*	−8.368 *(7.546)*	−1.492 *(0.182)*	−8.624 *(6.107)*	−1.491 *(0.178)*	−8.960 *(5.344)*
		$\mathscr{D}_2$		−1.448 *(0.324)*	−10.51 *(8.398)*	−1.493 *(0.263)*	−9.678 *(6.587)*	−1.469 *(0.220)*	−10.66 *(6.136)*	−1.463 *(0.227)*	−10.44 *(5.916)*
		$\mathscr{D}_3$		−1.468 *(1.057)*	−8.339 *(18.31)*	−1.429 *(0.267)*	−9.218 *(5.920)*	−1.446 *(0.226)*	−10.52 *(8.263)*	−1.456 *(0.215)*	−10.18 *(5.023)*
		$\mathscr{D}_4$	$m_+(x_1) = m_-(x_1) = -1.5217$	−1.073 *(2.885)*	−4.982 *(54.38)*	−1.474 *(0.293)*	−8.826 *(6.404)*	−1.489 *(0.285)*	−9.266 *(6.234)*	−1.501 *(0.315)*	−9.498 *(6.343)*
	$N = 1000$	$\mathscr{D}_1$		−1.507 *(0.112)*	−10.12 *(4.255)*	−1.509 *(0.135)*	−10.04 *(5.035)*	−1.506 *(0.117)*	−10.305 *(4.164)*	−1.507 *(0.112)*	−10.09 *(4.253)*
		$\mathscr{D}_2$		−1.516 *(0.214)*	−11.04 *(7.309)*	−1.541 *(0.155)*	−10.97 *(6.265)*	−1.538 *(0.130)*	−10.98 *(5.898)*	−1.527 *(0.143)*	−10.63 *(6.183)*
		$\mathscr{D}_3$		−1.416 *(0.461)*	−11.26 *(16.43)*	−1.494 *(0.140)*	−10.42 *(6.095)*	−1.495 *(0.115)*	−10.74 *(4.852)*	−1.511 *(0.108)*	−10.83 *(5.790)*
		$\mathscr{D}_4$		−1.030 *(3.956)*	10.51 *(74.92)*	−1.506 *(0.187)*	−9.644 *(6.722)*	−1.503 *(0.193)*	−10.03 *(5.945)*	−1.511 *(0.213)*	−9.665 *(6.460)*

Table 2(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_1 = 0.2$											
SIMULATION				L_2 norm		L_1 norm		Huber’s function		Tuckey’s function	
settings				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	Repeat: ×1000	96.7 %	46.7 %	97.1 %	64.1 %	97.1 %	50.2 %	97.1 %	47.4 %
		$\mathscr{D}_2$		98.8 %	52.4 %	97.4 %	62.2 %	98.5 %	51.4 %	98.7 %	48.3 %
		$\mathscr{D}_3$		98.3 %	77.6 %	98.5 %	69.7 %	98.6 %	61.7 %	97.5 %	45.1 %
		$\mathscr{D}_4$		98.3 %	91.9 %	97.7 %	84.1 %	98.2 %	77.5 %	98.8 %	66.7 %
	$N = 200$	$\mathscr{D}_1$		95.0 %	52.9 %	95.1 %	67.5 %	96.4 %	53.0 %	95.1 %	52.9 %
		$\mathscr{D}_2$		96.6 %	61.3 %	95.0 %	67.9 %	95.2 %	54.4 %	95.4 %	54.6 %
		$\mathscr{D}_3$		98.3 %	94.3 %	94.7 %	66.5 %	95.3 %	54.7 %	94.3 %	51.9 %
		$\mathscr{D}_4$	$B = 1000$	97.2 %	98.1 %	96.8 %	76.4 %	97.6 %	67.4 %	98.1 %	66.9 %
	$N = 1000$	$\mathscr{D}_1$		95.3 %	61.0 %	95.8 %	76.8 %	95.3 %	60.2 %	95.3 %	60.0 %
		$\mathscr{D}_2$		95.5 %	69.4 %	96.1 %	75.6 %	95.8 %	61.7 %	95.6 %	62.4 %
		$\mathscr{D}_3$		96.1 %	91.3 %	96.0 %	74.1 %	96.0 %	62.3 %	95.6 %	61.4 %
		$\mathscr{D}_4$		93.2 %	97.5 %	95.9 %	83.7 %	95.9 %	75.3 %	95.8 %	75.4 %

Table 2(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_1) = (m_+(x_1) - m_-(x_1)) = 0$ $T_N^{(1)}(x_1) = (m'_+(x_1) - m'_-(x_1)) = 0$	−0.225 *(0.671)*	0.415 *(5.319)*	−0.228 *(0.829)*	0.524 *(8.755)*	−0.211 *(0.694)*	0.605 *(5.867)*	−0.220 *(0.673)*	0.431 *(5.378)*
		$\mathscr{D}_2$		−0.218 *(1.079)*	0.523 *(6.860)*	−0.242 *(1.070)*	0.585 *(9.810)*	−0.223 *(0.943)*	0.530 *(7.747)*	−0.226 *(0.930)*	0.523 *(6.706)*
		$\mathscr{D}_3$		−0.218 *(2.520)*	0.784 *(13.17)*	−0.232 *(1.357)*	0.396 *(12.47)*	−0.206 *(1.271)*	0.467 *(10.58)*	−0.217 *(0.761)*	0.711 *(5.989)*
		$\mathscr{D}_4$		−0.054 *(4.346)*	−0.575 *(27.02)*	−0.253 *(1.953)*	0.489 *(15.67)*	−0.223 *(1.751)*	0.569 *(13.12)*	−0.197 *(1.349)*	1.001 *(8.722)*
	$N = 200$	$\mathscr{D}_1$		−0.130 *(0.426)*	0.343 *(4.823)*	−0.180 *(0.500)*	0.425 *(6.535)*	−0.139 *(0.439)*	0.407 *(4.908)*	−0.130 *(0.426)*	0.321 *(4.856)*
		$\mathscr{D}_2$		−0.202 *(0.672)*	0.713 *(5.594)*	−0.202 *(0.546)*	0.598 *(6.830)*	−0.172 *(0.489)*	0.433 *(5.160)*	−0.182 *(0.515)*	0.444 *(5.123)*
		$\mathscr{D}_3$		−0.207 *(1.626)*	1.369 *(9.797)*	−0.166 *(0.543)*	0.444 *(7.058)*	−0.138 *(0.478)*	0.552 *(5.228)*	−0.133 *(0.448)*	0.480 *(4.917)*
		$\mathscr{D}_4$		−0.264 *(4.613)*	−1.663 *(32.01)*	−0.191 *(0.717)*	0.748 *(8.273)*	−0.188 *(0.687)*	0.541 *(6.489)*	−0.183 *(0.717)*	0.748 *(6.174)*
	$N = 1000$	$\mathscr{D}_1$		−0.070 *(0.223)*	0.221 *(4.446)*	−0.120 *(0.276)*	0.316 *(5.828)*	−0.079 *(0.231)*	0.295 *(4.495)*	−0.070 *(0.224)*	0.221 *(4.401)*
		$\mathscr{D}_2$		−0.122 *(0.359)*	0.612 *(5.216)*	−0.112 *(0.356)*	0.495 *(6.419)*	−0.102 *(0.314)*	0.316 *(4.813)*	−0.112 *(0.310)*	0.333 *(4.664)*
		$\mathscr{D}_3$		−0.147 *(0.840)*	1.261 *(9.368)*	−0.106 *(0.452)*	0.333 *(6.284)*	−0.078 *(0.423)*	0.443 *(4.794)*	−0.073 *(0.253)*	0.375 *(4.533)*
		$\mathscr{D}_4$		−0.204 *(6.448)*	−1.764 *(81.66)*	−0.130 *(0.651)*	0.631 *(7.895)*	−0.100 *(0.583)*	0.429 *(5.914)*	−0.103 *(0.449)*	0.644 *(5.666)*

Table 2(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.6340	20.853	3.2515	34.322	2.7217	23.001	2.6419	21.083
		$\mathscr{D}_2$		4.2319	26.893	4.1978	38.456	3.6966	30.369	3.6455	26.288
		$\mathscr{D}_3$		9.8783	51.639	5.3232	48.903	4.9843	41.492	2.9852	23.478
		$\mathscr{D}_4$		17.039	105.93	7.6586	61.450	6.8669	51.454	5.2899	34.192
	$N = 200$	$\mathscr{D}_1$		1.6704	18.906	1.9600	25.617	1.7243	19.240	1.6736	19.035
		$\mathscr{D}_2$		2.6359	21.931	2.1426	26.776	1.9181	20.229	2.0201	20.085
		$\mathscr{D}_3$		6.3771	38.403	2.1317	27.667	1.8765	20.496	1.7574	19.278
		$\mathscr{D}_4$		18.085	125.50	2.8109	32.433	2.6961	25.439	2.8132	24.201
	$N = 1000$	$\mathscr{D}_1$		0.8780	17.428	1.0838	22.848	0.9072	17.623	0.8806	17.255
		$\mathscr{D}_2$		1.4106	20.447	1.3992	25.164	1.2322	18.870	1.2151	18.286
		$\mathscr{D}_3$		3.2927	36.723	1.7744	24.633	1.6614	18.793	0.9950	17.772
		$\mathscr{D}_4$		25.279	320.13	2.5528	30.950	2.2889	23.184	1.7633	22.212

Table 3(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local cubic estimates)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_1) = m'_-(x_1) = -10.7150$	−1.496 *(0.394)*	−8.325 *(6.313)*	−1.476 *(0.521)*	−7.951 *(9.104)*	−1.478 *(0.446)*	−7.665 *(6.211)*	−1.498 *(0.394)*	−8.147 *(6.354)*
		$\mathscr{D}_2$		−1.524 *(0.537)*	−7.085 *(14.30)*	−1.516 *(0.474)*	−7.621 *(10.03)*	−1.514 *(0.400)*	−7.025 *(10.16)*	−1.523 *(0.410)*	−7.887 *(12.12)*
		$\mathscr{D}_3$		−1.413 *(1.675)*	−8.283 *(18.11)*	−1.520 *(0.464)*	−5.666 *(7.220)*	−1.522 *(0.452)*	−7.321 *(8.619)*	−1.498 *(0.415)*	−7.517 *(9.194)*
		$\mathscr{D}_4$		−0.667 *(4.480)*	2.832 *(48.99)*	−1.297 *(0.907)*	−6.770 *(9.105)*	−1.339 *(0.950)*	−5.754 *(18.98)*	−1.272 *(0.951)*	−5.951 *(21.17)*
	$N = 200$	$\mathscr{D}_1$		−1.499 *(0.205)*	−9.347 *(5.731)*	−1.531 *(0.265)*	−8.852 *(7.958)*	−1.511 *(0.220)*	−8.964 *(6.322)*	−1.501 *(0.206)*	−9.363 *(5.904)*
		$\mathscr{D}_2$		−1.459 *(0.371)*	−10.82 *(9.804)*	−1.500 *(0.284)*	−10.02 *(8.051)*	−1.479 *(0.238)*	−10.75 *(6.720)*	−1.472 *(0.258)*	−10.72 *(7.006)*
		$\mathscr{D}_3$	$m_+(x_1) = m_-(x_1) = -1.5217$	−1.530 *(1.211)*	−8.329 *(23.20)*	−1.441 *(0.274)*	−9.690 *(6.894)*	−1.461 *(0.244)*	−11.23 *(9.868)*	−1.465 *(0.235)*	−10.79 *(6.074)*
		$\mathscr{D}_4$		−1.204 *(3.157)*	−5.848 *(56.55)*	−1.513 *(0.343)*	−9.295 *(7.285)*	−1.532 *(0.340)*	−9.956 *(7.656)*	−1.538 *(0.376)*	−9.986 *(7.715)*
	$N = 1000$	$\mathscr{D}_1$		−1.511 *(0.126)*	−10.09 *(5.013)*	−1.515 *(0.151)*	−10.10 *(5.788)*	−1.510 *(0.131)*	−10.26 *(5.012)*	−1.511 *(0.126)*	−10.08 *(5.011)*
		$\mathscr{D}_2$		−1.521 *(0.244)*	−10.88 *(8.582)*	−1.550 *(0.175)*	−10.81 *(7.453)*	−1.541 *(0.154)*	−11.11 *(6.907)*	−1.528 *(0.170)*	−10.65 *(7.044)*
		$\mathscr{D}_3$		−1.413 *(0.531)*	−11.62 *(19.66)*	−1.499 *(0.153)*	−10.63 *(7.330)*	−1.494 *(0.141)*	−10.82 *(5.691)*	−1.513 *(0.133)*	−10.85 *(6.600)*
		$\mathscr{D}_4$		−1.043 *(4.604)*	11.62 *(76.05)*	−1.513 *(0.209)*	−9.700 *(7.611)*	−1.508 *(0.224)*	−9.911 *(7.236)*	−1.520 *(0.248)*	−9.607 *(8.022)*

Table 3(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$	$\widehat{m}(x_1)$	$\widehat{m}'(x_1)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	Repeat: ×1000	97.5 %	45.9 %	96.5 %	54.4 %	97.3 %	47.8 %	97.4 %	47.6 %
		$\mathscr{D}_2$		97.5 %	50.7 %	96.4 %	51.5 %	97.2 %	46.7 %	97.4 %	47.6 %
		$\mathscr{D}_3$		97.0 %	75.6 %	96.5 %	55.6 %	97.1 %	54.0 %	96.9 %	41.7 %
		$\mathscr{D}_4$		96.8 %	85.2 %	95.9 %	70.6 %	96.5 %	68.5 %	97.1 %	59.2 %
	$N = 200$	$\mathscr{D}_1$		97.1 %	47.3 %	95.6 %	56.4 %	97.0 %	49.8 %	97.3 %	48.4 %
		$\mathscr{D}_2$		97.6 %	57.0 %	96.3 %	56.8 %	96.8 %	48.1 %	96.9 %	51.6 %
		$\mathscr{D}_3$		97.6 %	85.6 %	96.1 %	55.5 %	96.8 %	49.5 %	0.95.7 %	46.9 %
		$\mathscr{D}_4$	$B = 1000$	96.4 %	84.3 %	96.5 %	75.2 %	97.6 %	72.5 %	97.9 %	70.4 %
	$N = 1000$	$\mathscr{D}_1$		95.7 %	63.3 %	95.5 %	76.4 %	95.6 %	69.8 %	95.4 %	66.2 %
		$\mathscr{D}_2$		96.0 %	70.7 %	95.6 %	71.8 %	95.9 %	69.2 %	96.0 %	66.6 %
		$\mathscr{D}_3$		95.9 %	72.6 %	95.7 %	79.0 %	95.5 %	70.5 %	95.6 %	69.3 %
		$\mathscr{D}_4$		93.4 %	85.2 %	96.0 %	78.7 %	95.9 %	72.5 %	95.8 %	70.4 %

Table 3(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_1) = (m_+(x_1) - m_-(x_1)) = 0$ $T_N^{(1)}(x_1) = (m'_+(x_1) - m'_-(x_1)) = 0$	−0.113 *(0.705)*	−0.512 *(5.246)*	−0.146 *(0.845)*	−0.211 *(7.438)*	−0.110 *(0.723)*	−0.814 *(5.430)*	−0.112 *(0.706)*	−0.998 *(5.273)*
		$\mathscr{D}_2$		−0.147 *(1.138)*	−0.463 *(6.777)*	−0.158 *(1.097)*	−0.146 *(7.896)*	−0.133 *(0.992)*	−0.633 *(6.997)*	−0.136 *(0.979)*	−0.609 *(6.540)*
		$\mathscr{D}_3$		−0.161 *(2.671)*	−0.069 *(12.27)*	−0.151 *(1.376)*	−0.008 *(9.782)*	−0.103 *(1.319)*	−0.629 *(8.633)*	−0.118 *(0.800)*	−0.620 *(5.901)*
		$\mathscr{D}_4$		−0.044 *(4.523)*	1.040 *(22.19)*	−0.182 *(2.001)*	−0.249 *(12.67)*	−0.168 *(1.819)*	−0.042 *(10.83)*	−0.156 *(1.423)*	0.244 *(8.372)*
	$N = 200$	$\mathscr{D}_1$		−0.109 *(0.454)*	−0.441 *(5.405)*	−0.107 *(0.513)*	−0.149 *(6.386)*	−0.102 *(0.465)*	−0.400 *(5.379)*	−0.100 *(0.455)*	−0.478 *(5.438)*
		$\mathscr{D}_2$		−0.133 *(0.718)*	−0.106 *(6.273)*	−0.109 *(0.562)*	−0.256 *(6.667)*	−0.106 *(0.518)*	−0.358 *(5.657)*	−0.107 *(0.551)*	−0.352 *(5.782)*
		$\mathscr{D}_3$		−0.156 *(1.741)*	−0.753 *(10.410)*	−0.106 *(0.553)*	−0.262 *(6.726)*	−0.102 *(0.503)*	−0.163 *(5.573)*	−0.101 *(0.477)*	−0.314 *(5.364)*
		$\mathscr{D}_4$		−0.255 *(4.833)*	1.078 *(27.33)*	−0.113 *(0.740)*	−0.291 *(7.972)*	−0.103 *(0.728)*	−0.058 *(6.980)*	−0.114 *(0.767)*	−0.374 *(6.876)*
	$N = 1000$	$\mathscr{D}_1$		−0.089 *(0.235)*	−0.311 *(4.702)*	−0.087 *(0.281)*	−0.220 *(4.371)*	−0.082 *(0.241)*	−0.370 *(4.463)*	−0.080 *(0.235)*	−0.448 *(4.111)*
		$\mathscr{D}_2$		−0.113 *(0.379)*	−0.136 *(5.473)*	−0.089 *(0.365)*	−0.186 *(5.098)*	−0.086 *(0.330)*	−0.228 *(5.348)*	−0.087 *(0.326)*	−0.422 *(4.633)*
		$\mathscr{D}_3$		−0.166 *(0.890)*	−1.683 *(9.621)*	−0.086 *(0.458)*	−0.192 *(5.322)*	−0.082 *(0.439)*	−0.133 *(4.338)*	−0.081 *(0.266)*	−0.384 *(5.064)*
		$\mathscr{D}_4$		−0.346 *(7.507)*	−2.008 *(41.40)*	−0.093 *(0.667)*	−0.121 *(6.912)*	−0.083 *(0.606)*	−0.111 *(5.402)*	−0.094 *(0.474)*	−0.204 *(6.256)*

Table 3(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_1 = 0.2$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$	$T_N^{(0)}(x_1)$	$T_N^{(1)}(x_1)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.7655	20.566	3.3137	29.158	2.8352	21.288	2.7711	20.670
		$\mathscr{D}_2$		4.4646	26.569	4.3018	30.952	3.8887	27.427	3.8406	25.638
		$\mathscr{D}_3$		10.471	48.104	5.3970	38.348	5.1714	33.841	3.1374	23.134
		$\mathscr{D}_4$		17.732	87.004	7.844	49.702	7.1314	42.469	5.5789	32.820
	$N = 200$	$\mathscr{D}_1$		1.7811	21.189	2.0115	25.033	1.8265	21.086	1.7863	21.317
		$\mathscr{D}_2$		2.8173	24.590	2.2057	26.136	2.0324	22.176	2.1606	22.668
		$\mathscr{D}_3$		6.8260	40.808	2.1681	26.365	1.9737	21.849	1.8735	21.027
		$\mathscr{D}_4$		18.947	107.13	2.9019	31.250	2.8566	27.364	3.0078	26.954
	$N = 1000$	$\mathscr{D}_1$		0.9218	18.433	1.1045	17.135	0.9450	17.497	0.9237	16.117
		$\mathscr{D}_2$		1.4882	21.455	1.4339	19.983	1.2962	20.964	1.2802	18.161
		$\mathscr{D}_3$		3.4906	37.717	1.7990	20.862	1.7238	17.006	1.0458	19.850
		$\mathscr{D}_4$		29.430	162.29	2.6147	27.095	2.3771	21.177	1.8596	24.524

Table 4(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local constant estimates)

simulation results for $x_2 = 0.5$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$m_+(x_2) = m_-(x_2) = 0 \quad m'_+(x_2) = 2.5133 \wedge m'_-(x_2) = 25.1327$	−0.195	–	−0.198	–	−0.163	–	−0.193	–
				(0.410)	*(0.000)*	*(0.507)*	*(0.000)*	*(0.430)*	*(0.000)*	*(0.414)*	*(0.000)*
		$\mathscr{D}_2$		−0.291	–	−0.280	–	−0.238	–	−0.315	–
				(0.473)	*(0.000)*	*(0.500)*	*(0.000)*	*(0.460)*	*(0.000)*	*(0.398)*	*(0.000)*
		$\mathscr{D}_3$		−0.109	–	−0.202	–	−0.194	–	−0.230	–
				(1.131)	*(0.000)*	*(0.542)*	*(0.000)*	*(0.482)*	*(0.000)*	*(0.417)*	*(0.000)*
		$\mathscr{D}_4$		0.065	–	−0.228	–	−0.257	–	−0.240	–
				(1.402)	*(0.000)*	*(0.614)*	*(0.000)*	*(0.544)*	*(0.000)*	*(0.576)*	*(0.000)*
	$N = 200$	$\mathscr{D}_1$		−0.146	–	−0.124	–	−0.124	–	−0.140	–
				(0.244)	*(0.000)*	*(0.314)*	*(0.000)*	*(0.267)*	*(0.000)*	*(0.244)*	*(0.000)*
		$\mathscr{D}_2$		−0.149	–	−0.128	–	−0.142	–	−0.156	–
				(0.344)	*(0.000)*	*(0.294)*	*(0.000)*	*(0.272)*	*(0.000)*	*(0.282)*	*(0.000)*
		$\mathscr{D}_3$		−0.266	–	−0.117	–	−0.145	–	−0.182	–
				(0.805)	*(0.000)*	*(0.300)*	*(0.000)*	*(0.265)*	*(0.000)*	*(0.234)*	*(0.000)*
		$\mathscr{D}_4$		−0.253	–	−0.180	–	−0.221	–	−0.203	–
				(2.190)	*(0.000)*	*(0.340)*	*(0.000)*	*(0.324)*	*(0.000)*	*(0.360)*	*(0.000)*
	$N = 1000$	$\mathscr{D}_1$		−0.078	–	−0.091	–	−0.088	–	−0.086	–
				(0.131)	*(0.000)*	*(0.149)*	*(0.000)*	*(0.129)*	*(0.000)*	*(0.129)*	*(0.000)*
		$\mathscr{D}_2$		−0.081	–	−0.098	–	−0.091	–	−0.086	–
				(0.178)	*(0.000)*	*(0.187)*	*(0.000)*	*(0.146)*	*(0.000)*	*(0.150)*	*(0.000)*
		$\mathscr{D}_3$		−0.131	–	−0.102	–	−0.090	–	−0.085	–
				(0.478)	*(0.000)*	*(0.165)*	*(0.000)*	*(0.136)*	*(0.000)*	*(0.130)*	*(0.000)*
		$\mathscr{D}_4$		0.213	–	−0.110	–	−0.087	–	−0.087	–
				(4.205)	*(0.000)*	*(0.179)*	*(0.000)*	*(0.147)*	*(0.000)*	*(0.177)*	*(0.000)*

Table 4(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_2 = 0.5$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	90.8 %	–	86.3 %	–	88.8 %	–	90.3 %	–
		$\mathscr{D}_2$		88.0 %	–	87.9 %	–	88.8 %	–	89.9 %	–
		$\mathscr{D}_3$		78.6 %	–	87.8 %	–	88.7 %	–	89.2 %	–
		$\mathscr{D}_4$		73.6 %	–	87.5 %	–	87.0 %	–	87.0 %	–
	$N = 200$	$\mathscr{D}_1$		97.1 %	–	96.9 %	–	97.7 %	–	96.9 %	–
		$\mathscr{D}_2$		98.1 %	–	97.4 %	–	98.0 %	–	97.9 %	–
		$\mathscr{D}_3$		97.9 %	–	95.5 %	–	96.3 %	–	97.1 %	–
		$\mathscr{D}_4$		98.0 %	–	97.2 %	–	98.6 %	–	98.1 %	–
	$N = 1000$	$\mathscr{D}_1$		95.8 %	–	95.6 %	–	96.4 %	–	95.6 %	–
		$\mathscr{D}_2$		96.8 %	–	96.1 %	–	96.7 %	–	96.6 %	–
		$\mathscr{D}_3$		96.6 %	–	95.4 %	–	95.2 %	–	95.8 %	-
		$\mathscr{D}_4$		94.7%	–	95.9 %	–	96.3 %	–	96.8 %	–

Table 4(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_2) = (m_+(x_2) - m_-(x_2)) = 0 \quad T_N^{(1)}(x_2) = (m'_+(x_2) - m'_-(x_2)) = -22.6194$	0.255 *(0.546)*	– *(0.000)*	0.244 *(0.640)*	– *(0.000)*	0.257 *(0.560)*	– *(0.000)*	0.255 *(0.546)*	– *(0.000)*
		$\mathscr{D}_2$		0.264 *(0.780)*	– *(0.000)*	0.280 *(0.714)*	– *(0.000)*	0.295 *(0.647)*	– *(0.000)*	0.298 *(0.668)*	– *(0.000)*
		$\mathscr{D}_3$		0.156 *(1.231)*	– *(0.000)*	0.259 *(0.707)*	– *(0.000)*	0.264 *(0.644)*	– *(0.000)*	0.251 *(0.593)*	– *(0.000)*
		$\mathscr{D}_4$		−0.117 *(1.554)*	– *(0.000)*	0.238 *(0.987)*	– *(0.000)*	0.277 *(0.918)*	– *(0.000)*	0.250 *(0.898)*	– *(0.000)*
	$N = 200$	$\mathscr{D}_1$		0.094 *(0.360)*	– *(0.000)*	0.114 *(0.410)*	– *(0.000)*	0.088 *(0.368)*	– *(0.000)*	0.095 *(0.360)*	– *(0.000)*
		$\mathscr{D}_2$		0.103 *(0.529)*	– *(0.000)*	0.135 *(0.442)*	– *(0.000)*	0.125 *(0.401)*	– *(0.000)*	0.117 *(0.419)*	– *(0.000)*
		$\mathscr{D}_3$		0.081 *(1.036)*	– *(0.000)*	0.116 *(0.429)*	– *(0.000)*	0.107 *(0.389)*	– *(0.000)*	0.110 *(0.376)*	– *(0.000)*
		$\mathscr{D}_4$		−0.163 *(1.620)*	– *(0.000)*	0.086 *(0.549)*	– *(0.000)*	0.086 *(0.528)*	– *(0.000)*	0.072 *(0.558)*	– *(0.000)*
	$N = 1000$	$\mathscr{D}_1$		0.047 *(0.182)*	– *(0.000)*	0.077 *(0.213)*	– *(0.000)*	0.059 *(0.186)*	– *(0.000)*	0.047 *(0.182)*	– *(0.000)*
		$\mathscr{D}_2$		0.102 *(0.260)*	– *(0.000)*	0.087 *(0.238)*	– *(0.000)*	0.079 *(0.215)*	– *(0.000)*	0.083 *(0.222)*	– *(0.000)*
		$\mathscr{D}_3$		0.120 *(0.410)*	– *(0.000)*	0.083 *(0.235)*	– *(0.000)*	0.073 *(0.214)*	– *(0.000)*	0.065 *(0.197)*	- *(0.000)*
		$\mathscr{D}_4$		−0.181 *(3.518)*	– *(0.000)*	0.093 *(0.329)*	– *(0.000)*	0.77 *(0.306)*	– *(0.000)*	0.076 *(0.299)*	– *(0.000)*

Table 4(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.1409	–	2.5117	–	2.1968	–	2.1420	–
		$\mathscr{D}_2$		3.0595	–	2.8012	–	2.5392	–	2.6211	–
		$\mathscr{D}_3$		4.8287	–	2.7747	–	2.5275	–	2.3256	–
		$\mathscr{D}_4$		6.0952	–	3.8714	–	3.5987	–	3.5211	–
	$N = 200$	$\mathscr{D}_1$		1.4126	–	1.6099	–	1.4459	–	1.4146	–
		$\mathscr{D}_2$		2.0740	–	1.7327	–	1.5754	–	1.6456	–
		$\mathscr{D}_3$		4.0622	–	1.6822	–	1.5277	–	1.4764	–
		$\mathscr{D}_4$		6.3516	–	2.1554	–	2.0724	–	2.1910	–
	$N = 1000$	$\mathscr{D}_1$		0.7136	–	0.8372	–	0.7322	–	0.7140	–
		$\mathscr{D}_2$		1.0198	–	0.9337	–	0.8464	–	0.8737	–
		$\mathscr{D}_3$		1.6095	–	0.9249	–	0.8425	–	0.7752	–
		$\mathscr{D}_4$		13.791	–	1.2904	–	1.1995	–	1.1737	–

Table 5(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local linear estimates)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$m_+(x_2) = m_-(x_2) = 0 \quad m'_+(x_2) = 2.5133 \wedge m'_-(x_2) = 25.1327$	−0.174 *(0.396)*	10.476 *(6.840)*	−0.185 *(0.423)*	8.734 *(6.651)*	−0.169 *(0.384)*	9.635 *(7.252)*	−0.171 *(0.399)*	10.274 *(6.690)*
		$\mathscr{D}_2$		−0.277 *(0.451)*	9.075 *(7.039)*	−0.272 *(0.450)*	9.639 *(8.305)*	−0.246 *(0.434)*	9.748 *(7.365)*	−0.309 *(0.379)*	10.752 *(5.828)*
		$\mathscr{D}_3$		−0.116 *(1.136)*	5.727 *(16.64)*	−0.218 *(0.510)*	10.178 *(8.781)*	−0.227 *(0.443)*	10.638 *(8.148)*	−0.221 *(0.394)*	12.627 *(12.47)*
		$\mathscr{D}_4$		0.087 *(1.390)*	3.165 *(45.82)*	−0.299 *(0.553)*	8.786 *(11.28)*	−0.305 *(0.472)*	10.601 *(16.54)*	−0.237 *(0.539)*	11.196 *(14.54)*
	$N = 200$	$\mathscr{D}_1$		−0.141 *(0.250)*	11.139 *(4.280)*	−0.153 *(0.302)*	10.079 *(5.289)*	−0.125 *(0.262)*	10.572 *(5.425)*	−0.138 *(0.249)*	11.118 *(4.444)*
		$\mathscr{D}_2$		−0.144 *(0.333)*	11.814 *(5.966)*	−0.141 *(0.277)*	12.097 *(7.320)*	−0.155 *(0.250)*	12.337 *(7.966)*	−0.145 *(0.262)*	12.898 *(9.475)*
		$\mathscr{D}_3$		−0.247 *(0.819)*	10.815 *(22.76)*	−0.105 *(0.263)*	10.384 *(8.231)*	−0.152 *(0.262)*	11.013 *(7.801)*	−0.169 *(0.231)*	11.488 *(6.989)*
		$\mathscr{D}_4$		0.169 *(2.176)*	25.444 *(69.24)*	−0.165 *(0.306)*	11.489 *(6.298)*	−0.212 *(0.300)*	11.263 *(6.548)*	−0.180 *(0.352)*	10.958 *(6.075)*
	$N = 1000$	$\mathscr{D}_1$		−0.041 *(0.127)*	13.151 *(3.749)*	−0.037 *(0.137)*	13.281 *(4.541)*	−0.051 *(0.126)*	13.063 *(3.994)*	−0.041 *(0.126)*	13.361 *(4.376)*
		$\mathscr{D}_2$		−0.056 *(0.175)*	13.360 *(10.62)*	−0.003 *(0.169)*	13.030 *(6.387)*	−0.037 *(0.143)*	12.864 *(4.828)*	−0.042 *(0.146)*	12.910 *(4.892)*
		$\mathscr{D}_3$		−0.013 *(0.477)*	11.918 *(18.22)*	−0.019 *(0.165)*	12.565 *(5.210)*	−0.043 *(0.131)*	12.379 *(5.279)*	−0.049 *(0.128)*	12.758 *(4.682)*
		$\mathscr{D}_4$		0.223 *(4.203)*	30.708 *(82.71)*	−0.020 *(0.163)*	11.592 *(9.371)*	−0.018 *(0.143)*	12.353 *(7.576)*	−0.010 *(0.175)*	13.262 *(6.575)*

Table 5(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$B = 1000 \quad$ Repeat: ×1000	85.9 %	41.3 %	89.6 %	40.0 %	87.4 %	47.3 %	87.1 %	41.9 %
		$\mathscr{D}_2$		95.9 %	59.1 %	91.5 %	55.4 %	91.2 %	56.6 %	93.9 %	55.2 %
		$\mathscr{D}_3$		96.7 %	73.5 %	92.2 %	60.0 %	90.8 %	66.0 %	89.4 %	60.3 %
		$\mathscr{D}_4$		98.9 %	91.7 %	97.1 %	72.2 %	97.3 %	64.1 %	97.7 %	54.3 %
	$N = 200$	$\mathscr{D}_1$		97.8 %	49.5 %	96.7 %	48.0 %	98.0 %	56.7 %	97.5 %	50.2 %
		$\mathscr{D}_2$		97.9 %	70.9 %	97.0 %	66.4 %	97.5 %	67.9 %	98.0 %	66.2 %
		$\mathscr{D}_3$		96.8 %	88.2 %	95.8 %	72.0 %	96.8 %	79.2 %	97.1 %	72.3 %
		$\mathscr{D}_4$		93.4 %	85.0 %	96.7 %	86.6 %	96.9 %	76.9 %	976 %	65.1 %
	$N = 1000$	$\mathscr{D}_1$		96.6 %	71.4 %	95.5 %	67.6 %	96.8 %	68.1 %	96.3 %	70.3 %
		$\mathscr{D}_2$		96.7 %	85.1 %	95.8 %	79.7 %	96.3 %	81.5 %	96.9 %	79.4 %
		$\mathscr{D}_3$		94.6 %	90.5 %	95.2 %	86.4 %	95.8 %	65.9 %	95.9 %	86.8 %
		$\mathscr{D}_4$		91.2 %	85.3 %	96.5 %	93.5 %	96.7 %	92.3 %	97.4 %	88.1 %

Table 5(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_2) = (m_+(x_2) - m_-(x_2)) = 0 \quad T_N^{(1)}(x_2) = (m'_+(x_2) - m'_-(x_2)) = -22.6194$	0.108 *(0.617)*	−14.70 *(5.695)*	0.124 *(0.743)*	−14.97 *(9.721)*	0.100 *(0.636)*	−14.80 *(6.507)*	0.105 *(0.618)*	−14.70 *(5.688)*
		$\mathscr{D}_2$		0.086 *(0.867)*	−15.50 *(8.361)*	0.143 *(0.822)*	−15.09 *(11.522)*	0.120 *(0.735)*	−14.95 *(8.937)*	0.122 *(0.761)*	−15.02 *(7.909)*
		$\mathscr{D}_3$		−0.001 *(1.258)*	−18.36 *(18.26)*	0.130 *(0.814)*	−15.47 *(14.07)*	0.098 *(0.730)*	−15.26 *(12.37)*	0.099 *(0.670)*	−15.08 *(6.942)*
		$\mathscr{D}_4$		−0.241 *(1.582)*	−20.49 *(38.83)*	0.114 *(1.101)*	−16.65 *(18.01)*	0.097 *(1.024)*	−16.15 *(14.92)*	0.071 *(1.007)*	−15.83 *(10.97)*
	$N = 200$	$\mathscr{D}_1$		0.095 *(0.424)*	−15.66 *(4.841)*	0.108 *(0.493)*	−16.37 *(6.897)*	0.108 *(0.435)*	−15.08 *(5.054)*	0.095 *(0.425)*	−15.60 *(4.845)*
		$\mathscr{D}_2$		0.113 *(0.621)*	−16.81 *(5.788)*	0.111 *(0.528)*	−15.19 *(7.323)*	0.085 *(0.475)*	−15.92 *(5.398)*	0.097 *(0.499)*	−16.52 *(5.298)*
		$\mathscr{D}_3$		0.096 *(1.144)*	−17.06 *(11.10)*	0.120 *(0.513)*	−16.43 *(7.268)*	0.079 *(0.461)*	−15.96 *(5.289)*	0.070 *(0.444)*	−15.94 *(5.020)*
		$\mathscr{D}_4$		−0.109 *(1.680)*	−10.43 *(43.98)*	0.107 *(0.645)*	−18.52 *(8.632)*	0.115 *(0.623)*	−17.06 *(6.879)*	0.111 *(0.662)*	−18.51 *(6.581)*
	$N = 1000$	$\mathscr{D}_1$		0.065 *(0.205)*	−18.55 *(1.898)*	0.078 *(0.247)*	−18.12 *(3.240)*	0.078 *(0.212)*	−18.36 *(2.169)*	0.065 *(0.206)*	−18.53 *(1.896)*
		$\mathscr{D}_2$		0.067 *(0.289)*	−18.93 *(2.787)*	0.071 *(0.274)*	−18.39 *(3.840)*	0.055 *(0.245)*	−18.34 *(2.979)*	0.067 *(0.253)*	−18.17 *(2.636)*
		$\mathscr{D}_3$		0.066 *(0.419)*	−17.95 *(6.088)*	0.070 *(0.271)*	−18.14 *(4.691)*	0.049 *(0.243)*	−18.65 *(4.125)*	0.040 *(0.223)*	−18.31 *(2.314)*
		$\mathscr{D}_4$		−0.139 *(3.727)*	−8.472 *(92.94)*	0.077 *(0.367)*	−19.50 *(6.004)*	0.045 *(0.341)*	−19.22 *(4.976)*	0.051 *(0.335)*	−19.50 *(3.658)*

Table 5(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.4196	22.324	2.9143	38.107	2.4948	25.510	2.4244	22.297
		$\mathscr{D}_2$		3.4021	32.777	3.2228	45.169	2.8812	35.034	2.9841	31.006
		$\mathscr{D}_3$		4.9344	71.601	3.1928	55.165	2.8621	48.516	2.6270	27.213
		$\mathscr{D}_4$		6.2052	152.21	4.3170	70.612	4.0146	58.519	3.9508	43.017
	$N = 200$	$\mathscr{D}_1$		1.6655	18.978	1.9328	27.039	1.7063	19.814	1.6685	18.994
		$\mathscr{D}_2$		2.4361	22.691	2.0715	28.706	1.8657	21.162	1.9563	20.767
		$\mathscr{D}_3$		4.4878	43.533	2.0115	28.493	1.8090	20.735	1.7405	19.681
		$\mathscr{D}_4$		6.5869	172.40	2.5304	33.839	2.4430	26.968	2.5953	25.799
	$N = 1000$	$\mathscr{D}_1$		0.8065	7.4415	0.9714	12.702	0.8316	8.5033	0.8081	7.4325
		$\mathscr{D}_2$		1.1340	10.925	1.0742	15.056	0.9604	11.678	0.9947	10.335
		$\mathscr{D}_3$		1.6448	23.867	1.0642	18.388	0.9540	16.172	0.8756	9.0711
		$\mathscr{D}_4$		14.612	364.33	1.4390	23.537	1.3382	19.506	1.3169	14.339

Table 6(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local cubic estimates)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_2) = 2.5133 \wedge m'_-(x_2) = 25.1327$	−0.153 *(0.458)*	11.196 *(7.583)*	−0.182 *(0.488)*	9.146 *(7.419)*	−0.140 *(0.459)*	10.081 *(7.988)*	−0.150 *(0.462)*	10.966 *(7.385)*
		$\mathscr{D}_2$		−0.298 *(0.571)*	9.605 *(8.114)*	−0.280 *(0.485)*	10.202 *(9.453)*	−0.233 *(0.486)*	10.492 *(9.940)*	−0.294 *(0.456)*	11.302 *(6.607)*
		$\mathscr{D}_3$		−0.100 *(1.478)*	6.072 *(19.47)*	−0.203 *(0.542)*	11.124 *(12.36)*	−0.193 *(0.475)*	11.929 *(11.64)*	−0.204 *(0.476)*	14.231 *(16.72)*
		$\mathscr{D}_4$		−0.608 *(2.370)*	8.161 *(54.34)*	−0.324 *(0.702)*	8.797 *(10.16)*	−0.263 *(0.607)*	11.889 *(20.38)*	−0.223 *(0.739)*	13.751 *(26.58)*
	$N = 200$	$\mathscr{D}_1$		−0.103 *(0.278)*	11.266 *(5.365)*	−0.131 *(0.343)*	10.526 *(5.946)*	−0.094 *(0.290)*	10.848 *(5.995)*	−0.099 *(0.278)*	11.244 *(5.512)*
		$\mathscr{D}_2$		−0.104 *(0.395)*	12.196 *(7.221)*	−0.101 *(0.311)*	12.714 *(8.153)*	0.102 *(0.288)*	13.154 *(9.467)*	−0.103 *(0.305)*	13.663 *(10.35)*
		$\mathscr{D}_3$		−0.377 *(1.382)*	11.512 *(27.46)*	−0.080 *(0.313)*	10.646 *(9.716)*	−0.103 *(0.288)*	11.600 *(8.860)*	−0.131 *(0.249)*	11.862 *(7.680)*
		$\mathscr{D}_4$	$m_+(x_2) = m_-(x_2) = 0$	−0.658 *(3.956)*	22.852 *(66.37)*	−0.154 *(0.311)*	22.851 *(7.235)*	−0.172 *(0.340)*	11.666 *(7.367)*	−0.131 *(0.413)*	11.120 *(7.885)*
	$N = 1000$	$\mathscr{D}_1$		−0.029 *(0.143)*	13.406 *(4.405)*	−0.032 *(0.153)*	13.864 *(5.102)*	−0.037 *(0.141)*	13.390 *(4.629)*	−0.030 *(0.144)*	13.645 *(5.024)*
		$\mathscr{D}_2$		−0.038 *(0.207)*	13.547 *(11.44)*	−0.009 *(0.198)*	13.158 *(7.208)*	−0.031 *(0.162)*	13.083 *(5.842)*	−0.029 *(0.163)*	13.191 *(5.877)*
		$\mathscr{D}_3$		−0.016 *(0.556)*	10.838 *(21.09)*	−0.021 *(0.185)*	12.753 *(6.282)*	−0.031 *(0.154)*	12.521 *(6.063)*	−0.038 *(0.142)*	12.833 *(5.259)*
		$\mathscr{D}_4$		−0.172 *(5.933)*	34.250 *(86.15)*	−0.006 *(0.192)*	11.875 *(10.04)*	−0.006 *(0.170)*	12.449 *(8.672)*	0.005 *(0.202)*	13.476 *(7.698)*

Table 6(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_2 = 0.5$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$	$\widehat{m}(x_2)$	$\widehat{m}'(x_2)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	90.8 %	65.3 %	90.3 %	68.7 %	90.0 %	65.2 %	90.7 %	65.6 %
		$\mathscr{D}_2$		96.8 %	74.3 %	93.1 %	74.4 %	94.7 %	69.7 %	95.7 %	70.0 %
		$\mathscr{D}_3$		97.9 %	76.3 %	93.2 %	74.8 %	94.2 %	71.4 %	92.9 %	68.1 %
		$\mathscr{D}_4$		99.5 %	77.3 %	96.0 %	79.4 %	95.2 %	78.3 %	95.8 %	77.7 %
	$N = 200$	$\mathscr{D}_1$		96.6 %	74.3 %	96.4 %	71.6 %	96.8 %	73.2 %	96.5 %	74.1 %
		$\mathscr{D}_2$		97.0 %	77.2 %	96.7 %	73.7 %	96.6 %	74.4 %	96.8 %	75.8 %
		$\mathscr{D}_3$		96.1 %	82.6 %	96.2 %	75.2 %	96.3 %	72.3 %	96.3 %	73.0 %
		$\mathscr{D}_4$		94.1 %	82.5 %	96.1 %	77.6 %	77.6 %	97.5 %	78.4 %	97.6 %
	$N = 1000$	$\mathscr{D}_1$		95.6 %	78.0 %	95.6 %	75.8 %	95.8 %	77.7 %	95.5 %	078.8 %
		$\mathscr{D}_2$		95.0 %	82.9 %	95.7 %	78.6 %	95.7 %	78.7 %	95.8 %	80.5 %
		$\mathscr{D}_3$		93.1 %	86.4 %	94.9 %	79.1 %	95.3 %	77.1 %	95.3 %	78.2 %
		$\mathscr{D}_4$		92.1 %	82.9 %	95.7 %	83.9 %	96.5 %	83.2 %	96.6 %	821 %

Table 6(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_2 = 0.5$

SIMULATION settings				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_2) = (m_+(x_2) - m_-(x_2)) = 0 \quad T_N^{(1)}(x_2) = (m'_+(x_2) - m'_-(x_2)) = -22.6194$	0.097 *(0.650)*	−16.48 *(5.128)*	0.144 *(0.758)*	−16.72 *(7.707)*	0.101 *(0.660)*	−16.46 *(5.549)*	0.096 *(0.648)*	−16.46 *(5.138)*
		$\mathscr{D}_2$		0.102 *(0.913)*	−17.14 *(7.306)*	0.170 *(0.836)*	−16.86 *(8.923)*	0.127 *(0.767)*	−16.67 *(7.542)*	0.123 *(0.801)*	−16.76 *(6.935)*
		$\mathscr{D}_3$		0.008 *(1.325)*	−17.89 *(15.03)*	0.139 *(0.833)*	−16.66 *(10.55)*	0.097 *(0.761)*	−16.47 *(9.777)*	0.092 *(0.705)*	−16.67 *(6.349)*
		$\mathscr{D}_4$		−0.216 *(1.648)*	−17.72 *(28.40)*	0.137 *(1.126)*	−17.23 *(13.77)*	0.116 *(1.069)*	−16.95 *(11.81)*	0.090 *(1.055)*	−16.98 *(9.185)*
	$N = 200$	$\mathscr{D}_1$		0.086 *(0.452)*	−19.63 *(5.414)*	0.115 *(0.506)*	−18.53 *(6.416)*	0.100 *(0.460)*	−17.4 *(5.410)*	0.088 *(0.452)*	−19.81 *(5.434)*
		$\mathscr{D}_2$		0.084 *(0.665)*	−19.48 *(6.164)*	0.090 *(0.544)*	−20.05 *(6.850)*	0.114 *(0.503)*	−19.46 *(5.764)*	0.108 *(0.532)*	−19.86 *(5.770)*
		$\mathscr{D}_3$		0.088 *(1.206)*	−19.80 *(9.958)*	0.082 *(0.527)*	−19.27 *(6.727)*	0.083 *(0.488)*	−18.94 *(5.681)*	0.088 *(0.474)*	−18.82 *(5.627)*
		$\mathscr{D}_4$		−0.056 *(2.757)*	0.046 *(34.64)*	0.110 *(0.670)*	−21.01 *(7.978)*	0.099 *(0.663)*	−18.94 *(7.027)*	0.087 *(0.707)*	−18.92 *(6.901)*
	$N = 1000$	$\mathscr{D}_1$		0.056 *(0.216)*	−23.67 *(5.365)*	0.065 *(0.252)*	−20.53 *(6.344)*	0.074 *(0.220)*	−23.44 *(5.386)*	0.058 *(0.216)*	−22.81 *(5.404)*
		$\mathscr{D}_2$		0.062 *(0.304)*	−21.48 *(6.126)*	0.060 *(0.278)*	−22.05 *(6.818)*	0.054 *(0.255)*	−23.46 *(5.724)*	0.058 *(0.267)*	−23.86 *(5.724)*
		$\mathscr{D}_3$		0.068 *(0.441)*	−21.80 *(9.893)*	0.052 *(0.277)*	−21.27 *(6.667)*	0.055 *(0.253)*	−25.04 *(5.636)*	0.058 *(0.235)*	−24.88 *(5.601)*
		$\mathscr{D}_4$		−0.186 *(5.549)*	2.681 *(89.60)*	0.063 *(0.375)*	−24.01 *(7.947)*	0.054 *(0.356)*	−22.94 *(6.989)*	0.049 *(0.351)*	−22.92 *(6.854)*

Table 6(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_2 = 0.5$

SIMULATION settings				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$	$T_N^{(0)}(x_2)$	$T_N^{(1)}(x_2)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.5481	20.104	2.9723	30.212	2.5908	21.751	2.5431	20.142
		$\mathscr{D}_2$		3.5825	28.641	3.2777	34.977	3.0089	29.564	3.1428	27.186
		$\mathscr{D}_3$		5.1945	58.933	3.2667	41.387	2.9864	38.328	2.7638	24.888
		$\mathscr{D}_4$		6.4623	111.33	4.4143	54.002	4.1936	46.323	4.1358	36.005
	$N = 200$	$\mathscr{D}_1$		1.7735	21.224	1.9848	25.154	1.8056	21.206	1.7728	21.303
		$\mathscr{D}_2$		2.6095	24.164	2.1341	26.853	1.9750	22.597	2.0881	22.621
		$\mathscr{D}_3$		4.7302	39.034	2.0673	26.370	1.9152	22.272	1.8586	22.059
		$\mathscr{D}_4$		10.807	135.8	2.6297	31.27	2.6008	27.54	2.773	27.05
	$N = 1000$	$\mathscr{D}_1$		0.8493	21.033	0.9907	24.868	0.8636	21.112	0.8477	21.184
		$\mathscr{D}_2$		1.1941	24.017	1.0925	26.727	1.0029	22.438	1.0476	22.440
		$\mathscr{D}_3$		1.7315	38.780	1.0889	26.136	0.9954	22.096	0.9212	21.957
		$\mathscr{D}_4$		21.753	351.24	1.4714	31.154	1.3978	27.398	1.3786	26.869

Table 7(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local constant estimates)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_3) = m'_-(x_3) = -0.4818$	0.681	–	0.657	–	0.694	–	0.684	–
				(0.380)	*(0.000)*	*(0.398)*	*(0.000)*	*(0.421)*	*(0.000)*	*(0.383)*	*(0.000)*
		$\mathscr{D}_2$		0.618	–	0.616	–	0.616	–	0.605	–
				(0.532)	*(0.000)*	*(0.478)*	*(0.000)*	*(0.419)*	*(0.000)*	*(0.463)*	*(0.000)*
		$\mathscr{D}_3$		0.400	–	0.630	–	0.658	–	0.679	–
				(1.294)	*(0.000)*	*(0.413)*	*(0.000)*	*(0.364)*	*(0.000)*	*(0.349)*	*(0.000)*
		$\mathscr{D}_4$		0.570	–	0.587	–	0.572	–	0.623	–
				(1.461)	*(0.000)*	*(0.660)*	*(0.000)*	*(0.685)*	*(0.000)*	*(0.637)*	*(0.000)*
	$N = 200$	$\mathscr{D}_1$	$m_+(x_3) = 0.8087 \wedge m_-(x_3) = 0.6087$	0.686	–	0.671	–	0.669	–	0.683	–
				(0.175)	*(0.000)*	*(0.249)*	*(0.000)*	*(0.199)*	*(0.000)*	*(0.176)*	*(0.000)*
		$\mathscr{D}_2$		0.688	–	0.699	–	0.684	–	0.692	–
				(0.324)	*(0.000)*	*(0.259)*	*(0.000)*	*(0.216)*	*(0.000)*	*(0.226*	*(0.000)*
		$\mathscr{D}_3$		0.683	–	0.722	–	0.697	–	0.689	–
				(0.901)	*(0.000)*	*(0.267)*	*(0.000)*	*(0.237)*	*(0.000)*	*(0.232)*	*(0.000)*
		$\mathscr{D}_4$		0.777	–	0.687	–	0.685	–	0.691	–
				(2.306)	*(0.000)*	*(0.302)*	*(0.000)*	*(0.287)*	*(0.000)*	*(0.340)*	*(0.000)*
	$N = 1000$	$\mathscr{D}_1$		0.702	–	0.716	–	0.706	–	0.702	–
				(0.101)	*(0.000)*	*(0.133)*	*(0.000)*	*(0.102)*	*(0.000)*	*(0.100)*	*(0.000)*
		$\mathscr{D}_2$		0.686	–	0.693	–	0.687	–	0.681	–
				(0.201)	*(0.000)*	*(0.155)*	*(0.000)*	*(0.127)*	*(0.000)*	*(0.128)*	*(0.000)*
		$\mathscr{D}_3$		0.582	–	0.686	–	0.681	–	0.687	-
				(0.597)	*(0.000)*	*(0.158)*	*(0.000)*	*(0.134)*	*(0.000)*	*(0.115)*	*(0.000)*
		$\mathscr{D}_4$		0.908	–	0.730	–	0.735	–	0.732	–
				(3.996)	*(0.000)*	*(0.185)*	*(0.000)*	*(0.163)*	*(0.000)*	*(0.193)*	*(0.000)*

Table 7(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	Repeat: ×1000	98.5 %	–	96.8 %	–	97.8 %	–	98.4 %	–
		$\mathscr{D}_2$		98.4 %	–	96.6 %	–	98.6 %	–	98.6 %	–
		$\mathscr{D}_3$		96.0 %	–	97.5 %	–	98.2 %	–	97.6 %	–
		$\mathscr{D}_4$		92.3 %	–	96.2 %	–	97.1 %	–	98.6 %	–
	$N = 200$	$\mathscr{D}_1$		97.2 %	–	96.5 %	–	97.3 %	–	96.8 %	–
		$\mathscr{D}_2$		97.9 %	–	96.2 %	–	97.0 %	–	96.9 %	–
		$\mathscr{D}_3$	$B = 1000$	97.4 %	–	96.2 %	–	97.3 %	–	97.0 %	–
		$\mathscr{D}_4$		93.0 %	–	97.0 %	–	97.0 %	–	97.1 %	–
	$N = 1000$	$\mathscr{D}_1$		95.6 %	–	95.5 %	–	96.0 %	–	95.9 %	–
		$\mathscr{D}_2$		96.1 %	–	95.8 %	–	95.9 %	–	95.5 %	–
		$\mathscr{D}_3$		97.0 %	–	96.1 %	–	96.3 %	–	95.0 %	-
		$\mathscr{D}_4$		99.0 %	–	96.3 %	–	95.9 %	–	95.7 %	–

Table 7(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_3 = 0.8$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_3) = (m_+(x_3) - m_-(x_3)) = 0.2$ $T_N^{(1)}(x_3) = (m'_+(x_3) - m'_-(x_3)) = 0$	0.134 *(0.587)*	– *(0.000)*	0.145 *(0.694)*	– *(0.000)*	0.140 *(0.598)*	– *(0.000)*	0.133 *(0.587)*	– *(0.000)*
		$\mathscr{D}_2$		0.170 *(0.929)*	– *(0.000)*	0.139 *(0.865)*	– *(0.000)*	0.141 *(0.780)*	– *(0.000)*	0.148 *(0.784)*	– *(0.000)*
		$\mathscr{D}_3$		0.157 *(1.985)*	– *(0.000)*	0.116 *(0.894)*	– *(0.000)*	0.122 *(0.838)*	– *(0.000)*	0.136 *(0.669)*	– *(0.000)*
		$\mathscr{D}_4$		0.099 *(2.888)*	– *(0.000)*	0.111 *(1.397)*	– *(0.000)*	0.125 *(1.263)*	– *(0.000)*	0.128 *(0.000)*	– *(1.133)*
	$N = 200$	$\mathscr{D}_1$		0.171 *(0.365)*	– *(0.000)*	0.166 *(0.412)*	– *(0.000)*	0.169 *(0.371)*	– *(0.000)*	0.170 *(0.364)*	– *(0.000)*
		$\mathscr{D}_2$		0.162 *(0.567)*	– *(0.000)*	0.159 *(0.442)*	– *(0.000)*	0.159 *(0.408)*	– *(0.000)*	0.161 *(0.433)*	– *(0.000)*
		$\mathscr{D}_3$		0.214 *(1.377)*	– *(0.000)*	0.165 *(0.435)*	– *(0.000)*	0.169 *(0.397)*	– *(0.000)*	0.166 *(0.380)*	– *(0.000)*
		$\mathscr{D}_4$		0.238 *(3.135)*	– *(0.000)*	0.169 *(0.581)*	– *(0.000)*	0.168 *(0.570)*	– *(0.000)*	0.169 *(0.607)*	– *(0.000)*
	$N = 1000$	$\mathscr{D}_1$		0.191 *(0.195)*	– *(0.000)*	0.186 *(0.231)*	– *(0.000)*	0.191 *(0.199)*	– *(0.000)*	0.191 *(0.000)*	– *(0.195)*
		$\mathscr{D}_2$		0.183 *(0.320)*	– *(0.000)*	0.179 *(0.288)*	– *(0.000)*	0.179 *(0.260)*	– *(0.000)*	0.180 *(0.000)*	– *(0.260)*
		$\mathscr{D}_3$		0.234 *(0.863)*	– *(0.000)*	0.185 *(0.298)*	– *(0.000)*	0.189 *(0.279)*	– *(0.000)*	0.186 *(0.223)*	- *(0.000)*
		$\mathscr{D}_4$		0.258 *(4.962)*	– *(0.000)*	0.189 *(0.465)*	– *(0.000)*	0.188 *(0.421)*	– *(0.000)*	0.189 *(0.377)*	– *(0.000)*

Table 7(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_3 = 0.8$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 0$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.3030	–	2.7216	–	2.3476	–	2.3016	–
		$\mathscr{D}_2$		3.6443	–	3.3940	–	3.0597	–	3.0766	–
		$\mathscr{D}_3$		7.7837	–	3.5076	–	3.2856	–	2.6236	–
		$\mathscr{D}_4$		11.322	–	5.4798	–	4.9522	–	4.4422	–
	$N = 200$	$\mathscr{D}_1$		1.4333	–	1.6181	–	1.4576	–	1.4304	–
		$\mathscr{D}_2$		2.2258	–	1.7343	–	1.6013	–	1.6995	–
		$\mathscr{D}_3$		5.3982	–	1.7051	–	1.559	–	1.4913	–
		$\mathscr{D}_4$		12.292	–	2.2781	–	2.2347	–	2.3795	–
	$N = 1000$	$\mathscr{D}_1$		0.7676	–	0.9072	–	0.7825	–	0.7672	–
		$\mathscr{D}_2$		1.2539	–	1.1313	–	1.0199	–	1.0255	–
		$\mathscr{D}_3$		3.3785	–	1.1692	–	1.0952	–	0.8745	–
		$\mathscr{D}_4$		19.454	–	1.8266	–	1.6507	–	1.4807	–

Table 8(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local linear estimates)

simulation results for $x_3 = 0.8$

SIMULATION	settings			L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$m_+(x_3) = 0.8087 \ \wedge \ m_-(x_3) = 0.6087$ $m'_+(x_3) = m'_-(x_3) = -0.4818$	0.679 *(0.377)*	0.807 *(3.241)*	0.648 *(0.401)*	0.435 *(3.042)*	0.689 *(0.412)*	0.683 *(3.364)*	0.680 *(0.379)*	0.797 *(3.235)*
		$\mathscr{D}_2$		0.618 *(0.527)*	1.183 *(3.844)*	0.612 *(0.476)*	0.756 *(3.162)*	0.608 *(0.426)*	1.116 *(3.499)*	0.608 *(0.469)*	1.498 *(3.667)*
		$\mathscr{D}_3$		0.402 *(1.305)*	0.737 *(3.925)*	0.636 *(0.414)*	0.235 *(3.131)*	0.667 *(0.382)*	0.461 *(2.894)*	0.686 *(0.368)*	0.472 *(3.107)*
		$\mathscr{D}_4$		0.577 *(2.448)*	1.111 *(4.346)*	0.560 *(0.660)*	0.170 *(3.336)*	0.574 *(0.661)*	0.462 *(3.394)*	0.624 *(0.630)*	0.506 *(3.619)*
	$N = 200$	$\mathscr{D}_1$		0.685 *(0.176)*	0.869 *(3.139)*	0.671 *(0.241)*	0.886 *(3.171)*	0.669 *(0.196)*	0.968 *(3.284)*	0.682 *(0.177)*	0.967 *(3.162)*
		$\mathscr{D}_2$		0.683 *(0.323)*	1.656 *(3.588)*	0.693 *(0.258)*	1.786 *(3.470)*	0.683 *(0.217)*	1.856 *(3.357)*	0.691 *(0.223)*	1.435 *(3.203)*
		$\mathscr{D}_3$		0.690 *(0.899)*	1.426 *(4.622)*	0.719 *(0.262)*	1.680 *(3.353)*	0.696 *(0.235)*	1.753 *(3.419)*	0.689 *(0.232)*	1.847 *(3.332)*
		$\mathscr{D}_4$		0.784 *(2.306)*	1.992 *(4.687)*	0.695 *(0.306)*	1.437 *(3.435)*	0.685 *(0.294)*	3.639 *(3.684)*	0.692 *(0.345)*	1.334 *(3.898)*
	$N = 1000$	$\mathscr{D}_1$		0.704 *(0.100)*	2.541 *(2.806)*	0.717 *(0.125)*	2.364 *(3.384)*	0.707 *(0.102)*	2.535 *(2.953)*	0.704 *(0.100)*	2.551 *(2.825)*
		$\mathscr{D}_2$		0.686 *(0.201)*	1.424 *(3.704)*	0.695 *(0.151)*	1.764 *(3.548)*	0.688 *(0.125)*	1.979 *(2.736)*	0.681 *(0.128)*	1.965 *(2.865)*
		$\mathscr{D}_3$		0.581 *(0.598)*	2.572 *(4.658)*	0.678 *(0.162)*	2.959 *(3.253)*	0.681 *(0.134)*	2.958 *(2.879)*	0.686 *(0.115)*	2.979 *(2.913)*
		$\mathscr{D}_4$		0.910 *(4.195)*	2.676 *(5.062)*	0.727 *(0.173)*	2.514 *(3.435)*	0.734 *(0.163)*	2.686 *(3.616)*	0.732 *(0.192)*	2.838 *(3.694)*

Table 8(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_3 = 0.8$

SIMULATION	settings			L_2 norm		L_1 norm		Huber's function		Tuckey's function	
				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	98.4 %	51.7 %	96.6 %	46.2 %	97.8 %	52.5 %	98.3 %	51.3 %
		$\mathscr{D}_2$		98.1 %	52.2 %	96.6 %	49.1 %	97.8 %	51.8 %	97.6 %	52.7 %
		$\mathscr{D}_3$		96.5 %	51.3 %	96.0 %	49.7 %	97.7 %	53.4 %	98.3 %	51.8 %
		$\mathscr{D}_4$		92.6 %	46.8 %	95.9 %	47.3 %	96.3 %	55.8 %	97.6 %	50.6 %
	$N = 200$	$\mathscr{D}_1$		97.2 %	54.9 %	96.2 %	52.9 %	96.7 %	56.1 %	99.9 %	55.1 %
		$\mathscr{D}_2$		96.8 %	62.7 %	96.1 %	56.9 %	96.9 %	60.0 %	96.8 %	61.4 %
		$\mathscr{D}_3$		96.1 %	55.4 %	95.8 %	56.3 %	96.6 %	60.2 %	96.9 %	60.0 %
		$\mathscr{D}_4$		90.4 %	46.6 %	96.2 %	59.3 %	96.9 %	60.4 %	96.8 %	61.6 %
	$N = 1000$	$\mathscr{D}_1$		96.0 %	72.0 %	95.6 %	71.8 %	95.7 %	72.1 %	96.0 %	72.1 %
		$\mathscr{D}_2$		95.8 %	67.7 %	95.8 %	71.6 %	95.9 %	75.8 %	95.8 %	76.4 %
		$\mathscr{D}_3$		95.1 %	55.4 %	96.2 %	71.3 %	95.6 %	75.2 %	95.8 %	75.0 %
		$\mathscr{D}_4$		89.4 %	41.6 %	96.2 %	72.3 %	95.4 %	75.4 %	95.8 %	76.6 %

Table 8(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_3) = (m_+(x_3) - m_-(x_3)) = 0.2$ $T_N^{(1)}(x_3) = (m'_+(x_3) - m'_-(x_3)) = 0$	0.140 *(0.643)*	−0.180 *(3.487)*	0.152 *(0.753)*	−0.136 *(3.426)*	0.145 *(0.660)*	−0.184 *(3.493)*	0.139 *(0.645)*	−0.183 *(3.484)*
		$\mathscr{D}_2$		0.147 *(1.006)*	−0.281 *(3.822)*	0.140 *(0.927)*	−0.172 *(3.590)*	0.140 *(0.852)*	−0.250 *(3.683)*	0.142 *(0.860)*	−0.259 *(3.738)*
		$\mathscr{D}_3$		0.089 *(2.047)*	−0.270 *(4.361)*	0.123 *(0.953)*	−0.152 *(3.544)*	0.128 *(0.904)*	−0.233 *(3.658)*	0.143 *(0.728)*	−0.198 *(3.527)*
		$\mathscr{D}_4$		0.103 *(2.924)*	−0.149 *(4.937)*	0.149 *(1.469)*	−0.134 *(4.030)*	0.144 *(1.352)*	−0.214 *(4.195)*	0.164 *(1.233)*	−0.234 *(4.187)*
	$N = 200$	$\mathscr{D}_1$		0.179 *(0.415)*	−0.157 *(3.643)*	0.177 *(0.461)*	−0.137 *(3.562)*	0.178 *(0.423)*	−0.183 *(3.596)*	0.178 *(0.415)*	−0.170 *(3.627)*
		$\mathscr{D}_2$		0.156 *(0.635)*	−0.257 *(3.762)*	0.169 *(0.493)*	−0.155 *(3.583)*	0.165 *(0.465)*	−0.190 *(3.630)*	0.165 *(0.494)*	−0.210 *(3.682)*
		$\mathscr{D}_3$		0.133 *(1.470)*	−0.214 *(4.467)*	0.175 *(0.485)*	−0.115 *(3.573)*	0.177 *(0.452)*	−0.157 *(3.624)*	0.173 *(0.434)*	−0.150 *(3.638)*
		$\mathscr{D}_4$		0.148 *(3.169)*	−0.141 *(5.567)*	0.170 *(0.640)*	−0.126 *(3.770)*	0.164 *(0.639)*	−0.156 *(3.961)*	0.160 *(0.682)*	−0.195 *(4.019)*
	$N = 1000$	$\mathscr{D}_1$		0.194 *(0.214)*	−0.112 *(3.626)*	0.192 *(0.251)*	−0.091 *(3.549)*	0.193 *(0.220)*	−0.122 *(3.582)*	0.193 *(0.215)*	−0.118 *(3.615)*
		$\mathscr{D}_2$		0.161 *(0.355)*	−0.179 *(3.753)*	0.184 *(0.309)*	−0.109 *(3.552)*	0.180 *(0.284)*	−0.147 *(3.604)*	0.180 *(0.286)*	−0.156 *(3.656)*
		$\mathscr{D}_3$		0.138 *(0.782)*	−0.161 *(4.464)*	0.190 *(0.317)*	−0.089 *(3.547)*	0.192 *(0.301)*	−0.130 *(3.611)*	0.188 *(0.242)*	−0.116 *(3.617)*
		$\mathscr{D}_4$		0.106 *(6.9740)*	−0.256 *(7.055)*	0.185 *(0.489)*	−0.086 *(3.745)*	0.179 *(0.450)*	−0.123 *(3.945)*	0.175 *(0.411)*	−0.143 *(4.006)*

Table 8(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 1$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: $\times 1000$	2.5232	13.671	2.9544	13.431	2.5906	13.695	2.5298	13.657
		$\mathscr{D}_2$		3.9452	14.985	3.6343	14.074	14.439	3.3407	14.654	3.3733
		$\mathscr{D}_3$		8.0251	17.097	3.7378	13.894	3.5446	14.339	2.8543	13.826
		$\mathscr{D}_4$		11.462	19.355	5.7620	15.799	5.3000	16.447	4.8353	16.415
	$N = 200$	$\mathscr{D}_1$		1.6289	14.283	1.8090	13.965	1.6602	14.099	1.6285	14.219
		$\mathscr{D}_2$		2.4928	14.749	1.9331	14.045	1.8242	14.231	1.9385	14.434
		$\mathscr{D}_3$		5.7633	17.514	1.9017	14.008	1.7733	14.209	1.7013	14.262
		$\mathscr{D}_4$		12.424	21.824	2.5099	14.781	2.5085	15.530	2.6758	15.754
	$N = 1000$	$\mathscr{D}_1$		0.8410	14.213	0.9848	13.914	0.8635	14.044	0.8432	14.171
		$\mathscr{D}_2$		1.3150	14.711	1.2114	13.925	1.1135	14.130	1.1244	14.332
		$\mathscr{D}_3$		2.6750	17.499	1.2459	13.907	1.1815	14.158	0.9514	14.179
		$\mathscr{D}_4$		27.340	27.656	1.9206	14.681	1.7666	15.464	1.6117	15.706

Table 9(a): FINITE SAMPLE PERFORMANCE OF M-SMOOTHERS (local cubic estimates)

simulation results for $x_3 = 0.8$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$m'_+(x_3) = m'_-(x_3) = -0.4818$	0.691 *(0.416)*	0.858 *(6.577)*	0.653 *(0.421)*	0.541 *(6.035)*	0.702 *(0.429)*	0.613 *(7.071)*	0.692 *(0.416)*	0.854 *(6.606)*
		$\mathscr{D}_2$		0.668 *(0.571)*	2.364 *(9.744)*	0.622 *(0.513)*	0.758 *(7.518)*	0.649 *(0.466)*	1.800 *(7.965)*	0.629 *(0.514)*	2.376 *(8.350)*
		$\mathscr{D}_3$		0.338 *(1.703)*	1.136 *(11.13)*	0.663 *(0.463)*	0.167 *(5.100)*	0.691 *(0.426)*	0.637 *(5.781)*	0.721 *(0.417)*	0.253 *(5.701)*
		$\mathscr{D}_4$		0.311 *(3.069)*	1.345 *(12.99)*	0.595 *(0.746)*	0.112 *(7.714)*	0.593 *(0.726)*	0.620 *(8.191)*	0.668 *(0.705)*	0.761 *(9.062)*
	$N = 200$	$\mathscr{D}_1$		0.709 *(0.217)*	1.363 *(6.387)*	0.699 *(0.279)*	1.231 *(5.821)*	0.694 *(0.240)*	1.150 *(6.120)*	0.707 *(0.218)*	1.514 *(6.364)*
		$\mathscr{D}_2$		0.685 *(0.377)*	2.679 *(8.076)*	0.700 *(0.284)*	2.129 *(6.455)*	0.690 *(0.248)*	2.236 *(5.900)*	0.698 *(0.248)*	1.795 *(5.845)*
		$\mathscr{D}_3$	$m_+(x_3) = 0.8087 \wedge m_-(x_3) = 0.6087$	0.719 *(1.061)*	2.553 *(12.45)*	0.734 *(0.281)*	1.991 *(5.764)*	0.708 *(0.266)*	2.944 *(7.282)*	0.705 *(0.267)*	2.734 *(6.180)*
		$\mathscr{D}_4$		1.226 *(4.183)*	6.099 *(16.32)*	0.693 *(0.357)*	1.288 *(6.254)*	0.701 *(0.340)*	1.281 *(7.020)*	0.711 *(0.406)*	1.014 *(6.833)*
	$N = 1000$	$\mathscr{D}_1$		0.713 *(0.124)*	3.085 *(4.537)*	0.735 *(0.151)*	2.725 *(4.859)*	0.719 *(0.125)*	3.082 *(4.700)*	0.713 *(0.123)*	3.097 *(4.558)*
		$\mathscr{D}_2$		0.694 *(0.228)*	1.191 *(5.971)*	0.700 *(0.171)*	2.111 *(5.397)*	0.698 *(0.143)*	2.106 *(3.828)*	0.698 *(0.150)*	1.892 *(3.861)*
		$\mathscr{D}_3$		0.574 *(0.668)*	5.693 *(12.92)*	0.687 *(0.176)*	4.862 *(7.567)*	0.683 *(0.151)*	3.403 *(4.239)*	0.686 *(0.128)*	3.436 *(4.394)*
		$\mathscr{D}_4$		1.412 *(4.845)*	9.290 *(19.23)*	0.730 *(0.191)*	3.102 *(5.625)*	0.741 *(0.183)*	2.796 *(5.261)*	0.740 *(0.214)*	3.128 *(7.141)*

Table 9(b): SMOOTH RESIDUAL BOOTSTRAP PERFORMANCE (95 % confidence interval coverage)

simulation results for $x_3 = 0.8$

SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	97.9 %	51.7 %	96.5 %	46.2 %	97.5 %	52.5 %	97.9 %	51.3 %
		$\mathscr{D}_2$		97.7 %	51.0 %	96.2 %	51.2 %	97.5 %	51.8 %	97.4 %	52.7 %
		$\mathscr{D}_3$		96.5 %	51.3 %	96.2 %	51.0 %	97.2 %	53.4 %	97.9 %	52.2 %
		$\mathscr{D}_4$		92.8 %	46.8 %	95.6 %	50.9 %	96.3 %	55.6 %	97.3 %	52.6 %
	$N = 200$	$\mathscr{D}_1$		96.7 %	54.9 %	96.1 %	52.9 %	96.5 %	56.1 %	96.6 %	5.51 %
		$\mathscr{D}_2$		96.7 %	53.7 %	96.2 %	56.9 %	96.7 %	60.0 %	96.6 %	61.4 %
		$\mathscr{D}_3$		96.2 %	55.4 %	95.9 %	56.3 %	96.4 %	60.2 %	96.7 %	60.1 %
		$\mathscr{D}_4$		90.6 %	46.6 %	96.0 %	59.3 %	96.6 %	60.4 %	96.5 %	61.6 %
	$N = 1000$	$\mathscr{D}_1$		95.9 %	71.8 %	95.3 %	69.7 %	95.6 %	71.3 %	95.9 %	71.5 %
		$\mathscr{D}_2$		95.7 %	74.2 %	95.3 %	72.4 %	95.8 %	74.9 %	95.7 %	77.4 %
		$\mathscr{D}_3$		92.0 %	66.1 %	95.4 %	72.5 %	95.5 %	76.1 %	95.7 %	75.5 %
		$\mathscr{D}_4$		89.3 %	48.3 %	95.6 %	72.7 %	95.8 %	74.5 %	95.7 %	75.1 %

Table 9(c): FINITE SAMPLE PERFORMANCE OF ONE-SIDED M-SMOOTHERS (test statistic value)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$	$\widehat{m}(x_3)$	$\widehat{m}'(x_3)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$T_N^{(0)}(x_3) = (m_+(x_3) - m_-(x_3)) = 0.2$ $T_N^{(1)}(x_3) = (m'_+(x_3) - m'_-(x_3)) = 0$	0.062 *(0.687)*	−0.123 *(3.005)*	0.059 *(0.786)*	−0.158 *(2.635)*	0.062 *(0.700)*	−0.159 *(2.981)*	0.059 *(0.688)*	−0.117 *(3.010)*
		$\mathscr{D}_2$		0.051 *(1.098)*	−0.124 *(3.322)*	0.054 *(0.971)*	−0.141 *(2.813)*	0.057 *(0.904)*	−0.177 *(3.183)*	0.053 *(0.932)*	−0.141 *(3.251)*
		$\mathscr{D}_3$		−0.034 *(2.225)*	−0.109 *(3.868)*	0.035 *(0.991)*	−0.130 *(2.761)*	0.039 *(0.953)*	−0.162 *(3.168)*	0.054 *(0.781)*	−0.096 *(3.048)*
		$\mathscr{D}_4$		−0.042 *(3.074)*	0.201 *(4.345)*	0.045 *(1.555)*	−0.008 *(3.272)*	0.035 *(1.458)*	−0.058 *(3.707)*	0.067 *1.341)*	−0.090 *(3.685)*
	$N = 200$	$\mathscr{D}_1$		0.060 *(0.449)*	−0.013 *(3.264)*	0.053 *(0.482)*	−0.141 *(2.914)*	0.064 *(0.456)*	−0.055 *(3.219)*	0.061 *(0.448)*	−0.003 *(3.260)*
		$\mathscr{D}_2$		0.035 *(0.698)*	−0.145 *(3.305)*	0.039 *(0.518)*	−0.142 *(2.927)*	0.043 *(0.501)*	−0.121 *(3.237)*	0.045 *(0.538)*	−0.123 *(3.302)*
		$\mathscr{D}_3$		−0.001 *(1.649)*	0.120 *(3.844)*	0.047 *(0.508)*	−0.078 *(2.946)*	0.054 *(0.486)*	−0.017 *(3.267)*	0.054 *(0.470)*	−0.035 *(3.282)*
		$\mathscr{D}_4$		0.119 *(3.347)*	0.378 *(4.816)*	0.044 *(0.680)*	−0.066 *(3.105)*	0.048 *(0.696)*	−0.066 *(3.477)*	0.044 *(0.749)*	−0.034 *(3.517)*
	$N = 1000$	$\mathscr{D}_1$		0.050 *(0.229)*	−0.009 *(2.998)*	0.028 *(0.262)*	−0.064 *(2.608)*	0.054 *(0.233)*	−0.043 *(2.964)*	0.056 *(0.229)*	−0.006 *(3.004)*
		$\mathscr{D}_2$		0.034 *(0.366)*	−0.077 *(3.313)*	0.038 *(0.323)*	−0.082 *(2.810)*	0.031 *(0.301)*	−0.059 *(3.167)*	0.042 *(0.310)*	−0.072 *(3.245)*
		$\mathscr{D}_3$		−0.004 *(0.741)*	0.126 *(3.867)*	0.043 *(0.330)*	−0.068 *(2.760)*	0.040 *(0.317)*	−0.018 *(3.167)*	0.047 *(0.260)*	−0.043 *(3.045)*
		$\mathscr{D}_4$		0.043 *(6.024)*	0.385 *(5.127)*	0.027 *(0.518)*	−0.052 *(3.264)*	0.032 *(0.486)*	−0.013 *(3.701)*	0.028 *(0.447)*	−0.014 *(3.680)*

Table 9(d): BOOTSTRAP LIMIT DISTRIBUTION FOR THE TEST STATISTIC (95 % confidence interval length)

simulation results for $x_3 = 0.8$											
SIMULATION				L_2 norm		L_1 norm		Huber's function		Tuckey's function	
settings				$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$	$T_N^{(0)}(x_3)$	$T_N^{(1)}(x_3)$
$p = 3$	$N = 50$	$\mathscr{D}_1$	$B = 1000$ Repeat: ×1000	2.6956	11.782	3.0814	10.332	2.7451	11.686	2.7001	11.798
		$\mathscr{D}_2$		4.3068	13.022	3.8088	11.028	3.5453	12.480	3.6547	12.743
		$\mathscr{D}_3$		8.7222	15.165	3.8864	10.824	3.7379	12.421	3.0628	11.949
		$\mathscr{D}_4$		12.051	17.033	6.0977	12.828	5.7154	14.534	5.2605	14.448
	$N = 200$	$\mathscr{D}_1$		1.6289	14.283	1.8090	13.965	1.6602	14.099	1.6285	14.219
		$\mathscr{D}_2$		2.4928	14.749	1.9331	14.045	1.8242	14.231	1.9385	14.434
		$\mathscr{D}_3$		5.7633	17.514	1.9017	14.008	1.7733	14.209	1.7013	14.262
		$\mathscr{D}_4$		12.424	21.824	2.5099	14.781	2.5085	15.530	2.6758	15.754
	$N = 1000$	$\mathscr{D}_1$		0.8985	11.754	1.0271	10.224	0.9150	11.621	0.9000	11.775
		$\mathscr{D}_2$		1.4356	12.987	1.2696	11.016	1.1817	12.418	1.2182	12.723
		$\mathscr{D}_3$		2.9074	15.159	1.2954	10.820	1.2459	12.416	1.0209	11.936
		$\mathscr{D}_4$		23.616	20.101	2.0325	12.795	1.9051	14.509	1.7535	14.427

6.2 Real data example

Simulation studies are commonly used to investigate the finite sample properties of some statistical method (M-smoothers estimators in this case) and they are very useful tools as one can adaptively control the initial setting parameters while managing almost all specific issues of whole the method, which is currently under investigation.

In real situations however, one is mostly interested in the finite sample performance of the given statistical method based on some real data sample – optionally a random sample data. We will therefore also discuss a real data problem here but we will firstly give in short a motivation, which was behind the experiment, which we will use to present the performance of M-smoothers with change-point occurrences when applied to a real data exercise.

6.2.1 Motivation on European chub data

The *European chub* (Squallus cephalus; L., 1758), sometimes also called *Round chub* or *Fat chub* is a freshwater fish of the family *Cyprinidae* and it possibly frequents in both, slow and moderate streams mostly in Europe. We have considered an extensive experiment[38] specifically designed to study the behaviour of the European chub individuals within the given group environment.

The European chub fish usually yields a length of 60 – 80 millimeters but the growing process is very slow and it usually takes even more than 10 years to achieve just 50 millimeters in the total length. The main idea within the social behaviour of this fish is that small individuals mostly behave quite different than adult ones while the boundary between immature and adult individuals is assumed to be somewhere close to 50 millimeters with respect to their length (given previous biological researches).

Figure 6.2: European chub (Squallus cephalus; L., 1758)

There are two possible options here: if their habitat features with good ecological quality (clean water, sufficient food sources, safe hideouts) than adult individuals have a tendency to find an ideal location and to settle there, which of course brings their natural activity to a lower level while immature fishes express much more hyperactive behaviour as they are mostly pushed away from ideal locations by adult individuals already settled there.

On the other hand, if there is a lack of ideal conditions in the given environment then adult fishes behave much more active trying to find some spot while immature fishes move as well but their condition factor[39] is much lower therefore, they feature with some less activity in general.

During the experiment we have observed fish specific activities (using especially designed inside body antennas) with a follow up period of a few days while preserving also an information about each individual's length and weight. The idea is to check if the length somewhere close to 50 millimeters can be really assumed to be a kind of boundary, which distinguishes or separates two groups of individuals, immature ones and adult ones or if it does not.

[38] The experiment took place at T.G. Masaryk Water Research Institute in Prague, Czech Republic in 2010.

[39] The condition factor is a numerical quantity computed mainly from the length and the weight of an individual and it expresses the individual's capability to behave hyperactive (something like a BMI for humans).

6.2.2 Real data application

The design of the experiment suggests that there may be some dependence in behaviour as the nature of the European chub fishes is to move within groups – they do preserve so called "social-based behaviour" standards. On the other hand, given the technical limits of the experiment some outlying observations could be recorded as well and moreover, it also seems from Figure 6.3 that some possible heteroscedastic variance structure along the length axe could be present as well.

Given these issues it would be optimal to use the proposed M-smoothers estimation approach while having in mind some possible form of the dependence structure in the observations while also assuming a change in location located somewhere close to the length of 50 millimeters.
We will firstly propose an estimate for the functional relationship and after that we will use one-sided kernel M-smoothers to find a candidate point, which would be treated as a possible jump. Finally, a statistical test is invoked in order to obtain a consistent and statistically relevant decision on the given change-point occurrence.

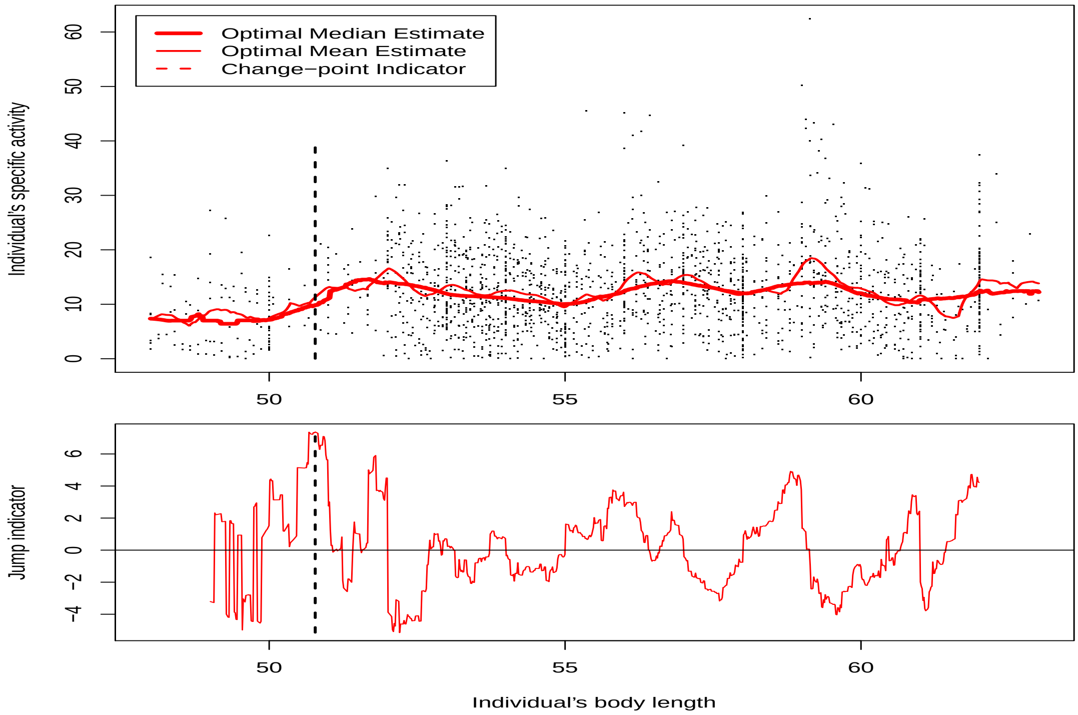
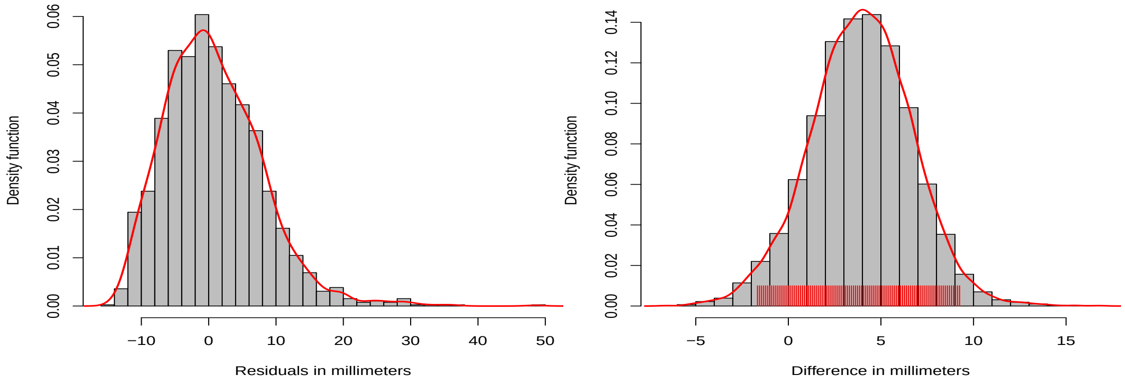


Figure 6.3: The functional dependence of the fish specific behaviour given its length (in millimeters) in the upper plot given for the local linear M-smoothers based on L_2 as well as L_1 norm and the indicator function for a jump occurrence (change in location) in the lower plot given for one-sided local linear M-smoothers based on the L_1 norm.

In Figure 6.3, there is a functional dependence given for the behaviour of a specific individual given its length (in millimeters). Given a presence of outliers (in Figure 6.3 we have used a "censored" axe for the activity measurements as there is also a small amount of activity occurrences close to 100 units) we have used the local linear M-smoothers based on the L_1 norm however, in Figure 6.3 there is also a local linear estimate based on the L_2 norm plotted and the effect of outlying observations is quite obvious as this estimate exhibits much more variability than we would like to. Additionally, we have also provided a plot for an indicator function[40] $T_N(x) = (\widehat{m}_+(x) - \widehat{m}_-(x))$ where the maximum of this function indicates a possible candidate for a jump occurrence. Just for completeness, let us mention that both estimates in Figure 6.3 were computed using the optimal bandwidth parameters bioth defined by the corresponding RCV and CV criterion.

In a correspondence with our expectation there is indeed a candidate for a jump occurrence found by the indicator function $T_N(x)$ at the point $x = 50.778$ (in millimeters). Moreover, in Figure 6.4(a) there is an evident fact that outlying (non-systematic) observations are really an issue in the given data sample therefore, the estimation based on L_1 norm is indeed a good choice to go in this case. We will now use the proposed bootstrap algorithm in order to decide if the jump at this point really occurs or if it does not.



(a) Model residuals plot

(b) Bootstrapped distribution

Figure 6.4: The histogram plots and the corresponding density plots for the vector of model's residuals and the bootstrapped distribution (density function) of the test statistic based on 10 000 bootstrap replicates.

The bootstrapped distribution of the test statistic $T_N(x_0) = (\widehat{m}_+(x_0) - \widehat{m}_-(x_0))$, for $N = 1954$ and $x_0 = 50.778$ is given in Figure 6.4(b) together with a corresponding 95 % confidence interval for the mean parameter of this distribution, which is $\widehat{\mu}_{(B)} = 3.938$ (for $B = 10\ 000$ bootstrap replicates).

The 95 % confidence interval as drawn in Figure 6.4(b) can be equivalently used as a critical region for the statistical test to decide if the jump at the point $x_0 = 50.778$ really occurs or if it does not. Given the fact that the zero value is inside of this 95 % confidence interval (the zero value would be a theoretical quantity for the test statistic under the null hypothesis) we do not reject the null hypothesis given the 95 % confidence level of the test[41].

[40] The indicator function corresponds with the test statistic defined in (3.4) however, up the standardizing term $\sqrt{Nh_N}$, which is constant for the fixed value of $N \in \mathbb{N}$.

[41] The same result would also follow from the test with 90 % confidence level as the corresponding confidence interval in this case would be equal to $(-0.676, 8.333)$

To conclude, there is an obvious evidence for a jump to occur at the point $X_0 = 50.778$ and the size of the jump seems to be somewhere close to 4 units of the corresponding activity measure however, given the result of the statistical test and the fixed sample size $N = 1954$ we can not sufficiently decide if the jump really occur or it does not.

Unlike the theoretical expectation where we have assumed that there would be a jump in activity measurements we did not statistically confirm such hypothesis. On the other hand, such behaviour heavily depends on the natural habitat of the European chubs and we have already mentioned that the experiment the data are from was designed inside the research facility with all simulated conditions however, in a plastic water tank only.

The primary question of interest why the experiment took place was however slightly different and the whole experiment was designed especially to correspond with the primary question of interest. The data we have used to present the actual finite sample performance of M-smoothers on are just the additional outcomes, which came from the experiment and we thought it might be nice to take a look at it and to test the hypothesis, which is commonly referred to as a fact in biological scientific journals and encyclopaedias.

6.3 Discussion on finite sample performance

Using both available tools in hand – the extensive simulation study and the real data example – we have found the performance of the proposed local polynomial M-smoothers approach to be in a quite nice correspondence with our expectations, which comes from the statistical theory derived before.

Especially, in the simulation results we have a nice opportunity to follow the asymptotic performance of the M-smoothers estimators and the effect of an improving precision of the given estimate once the sample size increases. Similarly, we can also see the effect of outlying observations and even more the effect of heavy-tailed random error distributions. Indeed, for the Cauchy distribution and the L_2 norm it does not hold any more that the imprecision improves as the sample size increases as the variability (as well as the length of confidence intervals) will exceed any given boundary. On the other hand, for robust loss functions the imprecision stays within bounds and moreover, it even improves as the sample size increases.

The rate of improvements seems to be much slower for the derivatives estimates however, we have already expected such behaviour given the theoretical results in Chapter 2.

Unlike the designed simulation study, in the case of the European chub experiment we did not have a chance to compare the performance of the M-smoothers method and the bootstrap algorithm with the true outcome however, based on simulations the sample size $N = 1954$ seems to be quite sufficient in order to obtain consistent and reliable results.

Using the M-smoothers approaches (as well as any other modelling techniques in statistics) one has to be always aware of the fact that most of them provide perfect results with respect to the asymptotic assumptions only, which is that the sample size $N \in \mathbb{N}$ tends to infinity. This of course does not mean that such methods could not be used in real data cases or that they would perform badly and their use would be impractical. Just in contrary, one just need to keep in mind the given restrictions of the method and to use an appropriate interpretation at the end to present the final results.

Finite sample data modelling is the only modelling approach a statistician can really meet in real life situations. Modern statistical theories and analysis are widely used in everyday life even without people being aware of it but there should always be a theoretical justification for an any used method and the finite sample properties should be investigated as well.

We have successfully completed our study on the finite sample performance and we have found them to be in a correspondence with the theoretical proves derived before. Given this fact, we think the proposed robust M-smoothers approaches are also well capable of being adopted for real data situations as well. Especially, when there really are many opportunities in practice where to use them.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

Ronald Fisher
(1890 – 1962)

7 Conclusion and final remarks

Robust regression approaches attract more and more interest and popularity in last decades, which could be especially due to the fact that they do have a very nice and convenient property of being robust with respect to outlying observations and even heavy-tailed distributions of random error terms. Given this quality we would consider the M-smoothers regression methods to have in general a huge potential to became popular and important modelling techniques in modern nonparametric regression approaches.

There are always good reasons in hand why to admit outlying observations in the model rather than throwing them away. Indeed, deleting of outliers in data samples was mostly due to some further simplicity rather than some logical or theoretical reasoning behind. There are also many data generating systems, which we could consider where some non-systematic measurements or outlying observations respectively, are implicitly produced by the system and therefore, they should be also taken into account in a statistical analysis. Moreover, in last couple of years heavy-tailed distributions of random errors increase their foundation within statistic and statistical modelling as well.

In order to cover such cases with proper statistical approaches one necessarily needs to consider appropriate regression methods where the presence of outliers will not cause a total failure of the whole estimation process. The M-smoothers approach discussed in this thesis will indeed do the right job here.

From another point of view, one of the most frequent issues related to statistical modelling in general is the *flexibility*. The flexibility can be approached from many different angles however, a light-headed dealing with it may easily cause a biased estimate or even a failure of the estimator. Therefore, one needs to approach the flexibility issues always in a proper statistical way.

In this thesis we have discussed the flexibility options with respect to the unknown functional relationship, which is supposed to be revealed and such flexibility was taken into account by adopting some common local polynomial estimation procedures. Next, we have also treated some flexibility issues with respect to the set of assumptions, which was required to assure that the given results hold.
We have weaken some important assumptions (e.g. distributional assumptions, continuity assumptions, finite moment assumptions) and thereby, we have introduced even more flexibility into the final model. Finally, we have also extended some flexibility with respect to the random sample assumptions as we have allowed for some dependence structures to occur within data while still being able to give a proper M-smoothers estimate and the corresponding statistical inference as well.

Each time we have tried to improve the flexibility of a given model we have done so using a proper statistical argumentation and we have proved that such model generalization really holds.

To conclude, there is actually one more aspect of the "flexibility" approach, which we have also improved considering the M-smoothers regression approaches. This however, does not refer to any statistical assumptions or properties but it rather refers to an "easy-to-use" approach introduced within the proposed simulation algorithms. Indeed, even some quite complicated asymptotic distributional expressions can be easily handled and effectively obtained by adopting a quite simple computer based method and very precise bootstrap simulation algorithms.

Therefore, local polynomial M-smoothers together with the change-point concept can be successfully considered to be an important nonparametric regression modelling technique with a robust flavour and a huge flexibility (in different meanings) in hand.

Additionally, the M-smoothers methods and the change-point idea can both serve as an ideal starting point for some further research and a theoretical development of new statistical approaches and theories in order to introduce even better modelling techniques while always trying to improving their statistical and practical qualities.

7.1 Further research...

At the very end, we would like to propose some additional ideas within the M-smoothers framework with change-points, which could be further extended, developed and properly investigated under some more flexible sets of assumptions.

- one could prove analogous results for some another dependence structures commonly used in statistic specifically, the uniformly strong mixing dependence concept or some others as well;

- one could extend the proposed M-smoothers methods and the change-point theory on multivariate cases as well (with respect to the random variable X as well as the response variable Y);

- one could implement a change-point problem into the scale function $\sigma(\cdot)$ and to develop a set of proper statistical tools for testing and estimating under some common regularity conditions;

- one could also develop a proper statistical theory in order to handle estimation of the unknown regression function and the scale function simultaneously at once;

- one could consider a supremum type test statistics for the hypothesis problems defined in Chapter 3 and to develop a proper extreme value theory to fit the M-smoothers model scenarios as well;

- finally, one could also try to relax the set of assumptions we have proposed for our situations and to give an afford to introduce a minimum set of required assumptions, which will be sufficient for the proofs to hold;

- a.e. – almost everywhere;
- $a.s.$ – almost surely;
- $\alpha(n)$ – α-mixing dependence coefficients for $n \in \mathbb{N}$;
- $\mathcal{A}, \mathcal{B}, \mathcal{F}$ – σ-fields;
- AMSE – Asymptotic Mean Squared Error quantity in general;
- AMSE$_x$ – Asymptotic Mean Squared Error quantity at the point $x \in (0,1)$;
- AR(k) – auto-regressive process of the order $k \in \mathbb{N}$;
- $\mathbb{A}$s.$\mathbb{B}$ias – asymptotic bias term;
- $\mathbb{A}$s.$\mathbb{V}$ar – asymptotic variance term;
- $\boldsymbol{\beta}$, $\mathbf{e}$, $\boldsymbol{\mu}$ – bold symbol for a column vector notation;
- $\mathbb{B}$ias $[X]$ – bias term of a random variable X in sense of a difference;
- $\beth_N, \beth_N^\star$ – /:beth:/ a Hebrew letter for constants which depend on $N \in \mathbb{N}$;
- $\mathbb{C}(0,1)$ – the Cauchy distribution with the location parameter 0 and the scale 1;
- $\mathcal{C}_p(0,1)$ – a set of continuous functions on (0,1) up to the order $p \in \mathbb{N} \cup \{0\}$;
  ($p = 1$ stands a continuity of a function itself and its first derivative as well)
- CLT – Central Limit Theorem;
- $\mathbb{C}$ov – covariance operator;
- CV$(\cdot)$ – Cross-Validation function for bandwidth selection;
- $\mathscr{D}, \mathscr{D}_1, \cdots$ – a notation for some specified distribution functions;
- $d_{m,2}(\cdot,\cdot)$ – the second order Mallow's metric;
- $\Delta \in \mathbb{R}$ – the size of a jump in a change-point model;
- $\Delta(G)$ – dispersion measure of a random variable with G to be a distribution function;
- $\operatorname{diag}\{\cdot\}$ – diagonal matrix consisting of elements in the brackets;
- $\mathbb{E}$ – conditional expectation operator conditioned on X;
- $\mathbb{E}^\star$ – expectation operator with respect to bootstrapped distribution function $G^\star(\cdot)$;
- $G(\cdot)$ – distribution function of random errors;
- $G^{-1}(\cdot)$ – the quantile function which corresponds with distribution $G(\cdot)$;
- $G^\star(\cdot)$ – distribution function of bootstrapped residuals;
- GCV$(\cdot)$ – Generalized Cross-Validation function for bandwidth selection;
- H$_0$, H$_1$ – the null hypothesis and the alternative hypothesis;
- h_N, a_N – nonparametric regression (bootstrap respectively) bandwidth parameter;

Notation

- $\mathbb{I}_{\{\dots\}}$ – identifier function;
- $\boldsymbol{I}_n$ – identity matrix of a type $n \times n$ for some $n \in \mathbb{N}$;
- *iff* – if and only if;
- *i.i.d.* – independent and identically distributed;
- $\mathcal{K}, \mathcal{K}_0, \dots, \mathcal{K}^*$ – general real constants bounded away from zero;
- $K(\cdot)$ – standardized kernel function defined over the interval $(-1, 1)$;
- $\mathcal{L}_p(\mathscr{A})$ – a collection of p^{th}-order, $\mathscr{A}$-measurable random variables, for $p \in \mathbb{N} \cup \{0\}$;
- $\mathscr{L}_p(0,1)$ – a set of smooth (Lispchitz) functions of the order $p \in \mathbb{N} \cup \{0\}$ on $(0,1)$;
  ($p = 1$ stands a Lipschitz property of a function itself and its first derivative)
- $m(\cdot)$ – the unknown regression function;
- $\mathbb{M}\mathrm{ed}(X)$ – median value of a random variable X;
- $N \in \mathbb{N}$ – the sample size (the total number of observations);
- $\mathbb{N}(\mu, \sigma^2)$ – normal distribution with the mean μ and variance σ^2;
- $\mathbb{R}, (\mathbb{N})$ – a field of real (natural) numbers;
- $\rho(\cdot)$ – a general loss function used for minimization (estimation);
- $\varphi(n)$ – φ-mixing dependence coefficients for $n \in \mathbb{N}$;
- $o(\cdot), o_{\mathbf{P}}(\cdot)$ – the Landau symbol - asymptotic negligibility (in probability);
- $O(\cdot), O_{\mathbf{P}}(\cdot)$ – the Landau symbol - asymptotic equivalency (in probability);
- $\mathbf{P}$ – probability measure;
- $\mathbf{P}^{\star}$ – conditional probability conditioned on the original random sample;
- $\Phi(\cdot), \phi(\cdot)$ – distribution function and the corresponding density of $\mathbb{N}(0,1)$ distribution;
- $\mathscr{R}^p$ – a set of basis functions – polynomial functions up to the order $p \in \mathbb{N} \cup \{0\}$;
- RCV$(\cdot)$ – Robust Cross-Validation function for bandwidth selection;
- $\sigma(\cdot)$ – the unknown scale function;
- $\mathbb{V}$ar – variance operator;
- $\mathcal{W}$(t) – standard Wiener process for $t \in (0,1)$;
- WIP – weak invariance principle;
- X$_N$, W$_N$ – the design matrix and the matrix of weights for $N \in \mathbb{N}$;
- $(\mathcal{X}, \mathcal{Y})$ – the random sample $\{(X_i, Y_i);\ i = 1, \dots, N\}$;
- $(\Omega, \mathcal{F}, \mathbf{P})$ – probability space;
- $\|\cdot\|_\infty$ – supremum norm;
- $\wedge, \vee$ – minimum and maximum operators;

- $\xrightarrow{a.s.}$ – convergence almost surely, for $N \to \infty$;
- $\xrightarrow{\mathscr{D}}$ – convergence in distribution, for $N \to \infty$;
- $\xleftrightarrow{\mathscr{D}(\mathbf{P})}$ – approaching each other in limit in distribution in probability, for $N \to \infty$;
- $\xrightarrow{\mathbf{P}}$ – convergence in probability, for $N \to \infty$;
- $\overset{as.}{\approx}, \asymp$ – asymptotic equivalency;
- $\overset{as.}{=}$ – asymptotic equality;
- $\overset{def.}{=}$ – given by definition;
- $\square$ – end of proof of lemma;
- $\blacksquare$ – end of proof of theorem;

List of Tables

Bibliography

Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis* (1st ed.). New York: John Wiley & Sons.

Antoch, J., G. Gregoire, and M. Hušková (2006). Test for continuity of regression function. *Journal for Statistical Planning and Inference 137*, 753–777.

Antoch, J. and P. Janssen (1989). Nonparametric regression m-quantiles. *Statistics & Probability Letters 8*, 355–362.

Baek, J. and T. Wehrly (1993). Kernel estimation for additive models with dependent observations. *Stochastic Process Appl. 47*, 95–112.

Bahadur, R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics 37*, 577–580.

Belyaev, Y. (1995). *Bootstrap, Resampling and Mallows metric*. Lecture Notes 1, Institute of Mathematical Statistics, Umeå University.

Belyaev, Y. and S. Sjöstedt-de Luna (2000). Weakly approaching sequences of random distributions. *Journal of Applied Probability, 37(3)*, 807–822.

Bickel, P. and D. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics 9(6)*, 1196–1217.

Bickel, P. and E. Lehmann (1976a). Descriptive statistics for nonparametric models iii. (dispersion). *The Annals of Statistics 4*, 1139–1158.

Bickel, P. and E. Lehmann (1976b). Descriptive statistics for nonparametric models iv. (spread). *Annals of Statistics 4*, 1159–1178.

Boente, G. and R. Fraiman (1995). Asymptotic distribution of smoothers based on local means and local medians under dependence. *Journal of Multivariate Analysis 54*, 77–90.

Boente, G., R. Fraiman, and J. Meloche (1997). Robust plug-in bandwidth estimators in nonparametric regression. *Journal of Statistical Planning and Inference 57*, 109–142.

Boente, G., M. Ruiz, and R. Zamar (2010). On a robust local estimator for the scale function in heteroscedastic nonparametric regression. *Statistics and probability letters 80*, 1185–1195.

Bose, A. (1998). Bahadur representation of M_m estimates. *The Annals of Statistics 26(2)*, 771–777.

Brown, L. and M. Levine (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics 35*, 2219–2232.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics 14*, 1171–1179.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and their Application* (1st ed.). Cambridge: Cambridge University Press.

Davydov, Y. (1970). The invariance principle for stationary processes. *Theory of Probability and Its Applications 14*, 487–498.

Dobrushin, R. (1970). Prescribing a system of random variables by conditional distribution. *Theory of Probability and its Applications 15*, 458–486.

Dunn, O. (1961). Multiple comparisons among means. *Hournal of the American Statistical Association 56*, 52–64.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL,.

Fan, J. (1992). Design adaptive nonparametric regression. *Journal of American Statistical Association 19*, 1273–1294.

Fan, J. and I. Gijbels (1995). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Comp. Graph. Statistic 4*, 213–227.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications* (1st ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Fitzenberger, B. (1997). The moving blocks bootstrap and robust inference for linear least squares and quantile regression. *Journal of Econometrics 82*, 235–287.

Gao, J., I. Gijbels, and S. Van Bellegem (2008). Nonparametric simultaneous testing for structural breaks. *Journal of Econometrics 143*, 123–142.

Gasser, T. and H. Müller (1979). Kernel estimation of regression functions. in smoothing technique for curve estimation. *Lecture Notes in Mathematics 759*, 23–68.

Hall, P., J. Kay, and D. Titterington (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regreesion. *Biometrica 77*, 521–528.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of American Statistical Association 69*, 383–397.

Hampel, F. (1986). *Contributions to the Theory of Robust Estimation*. Ph.D. Thesis, University of California.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Boston.

He, X. and Q. Shao (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Annals of Statistics 24(6)*, 2608–2630.

H.Rue, C. Chu, F. Godtliebsen, and J. Marron (1998). M-smoother with local linear fit. *Journal of Nonparametric Statistics 14*, 155–168.

Huber, P. (1964). Robust estimation of a location paramerer. *Annals of Mathematical Statistics 35*, 73–101.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Hušková, M. and M. Marušiaková (2009). M-procedures for detection of changes for dependent observations. *Proceedings of the Sixth Workshop on Simulation*, 685–690.

Ibragimov, I. (1959). Some limit theorems for stochastic processes stationary in the strict sense. *Doklady Akademii Nauk SSSR (in Russian) 125*, 711–714.

Ibragimov, I. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications 7*, 349–382.

Ibragimov, I. and Y. Linik (1971). *Independent and Stationary Sequences of Random Variables*. The Netherlands: Wolters-Noordhoff.

J.Fan, I.Gijbels, T. H. and L. Huang (1993). An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation. *Statistica Sinica 6*, 1–19.

Jurečková, J. (2001). *Robust Statistical Methods (in Czech)*. Charles University, Prague: Karolinum Press.

Jurečková, J. and P. Sen (1982). M-estimators and l-estimators of location: uniform integrability and asymptotic risk-efficient sequential versions. *Comm. Statist. Sequential Anal. 1(1)*, 27–56.

Künsch, H. (1989). The jacknife and the bootstrap for general stationary observations. *Annals of Statistics 17*, 1217–1241.

Lahiri, S. (1992). *Second Order Optimality of Stationary Bootstrap*. In Lapage, R. and Billard, L., editors, *Exploring the limits of bootstrap (pages 183 – 214)*, Wiley, New York.

Lahiri, S. (2003). *Resampling Methods for Dependent Data*. Springer-Verlag, New York.

Lee, J. and D. Cox (2010). Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy. *Computational Statistics and Data Analysis 54*, 3131–3143.

Leung, D. (2005). Cross-validation in nonparametric regression with outliers. *The Annals of Statistics 33*, 2291–2310.

Leung, D., F. Marriott, and E. Wu (1993). Bandwidth selection in robust smoothing. *Journal of Nonparametric Statistics 2*, 333–339.

Lin, Z. and C. Lu (1997). *Limit Theory for Mixing Dependent Random Variables*. Springer-Verlag, New York.

Liu, R. and K. Singh (1992). *Moving Blocks Jackknife and Bootstrap Capture Weak Dependence*. In Lapage, R. and Billard, L., editors, *Exploring the limits of bootstrap (pages 225 – 248)*, Wiley, New York.

Loader, C. (1996). Change point estimation using nonparametric regression. *Annals of Statistics 24*, 1667–1678.

Maciak, M. (2007). M-smoothers in testing and estimating. In Šafránková, J. and Pavlů, J., editors. *WDS'07 Proceedings of Contributed Papers [Part I], 73*, 169–174.

Maciak, M. (2008). Spline models with change-points. In Antoch, J. and Dohnal, G., editors. *Robust 2006 Conference Proceedings*, 239–246.

Maciak, M. (2010). Bootstrapping of M-smoothers. *Acta Universitatis Carolinaes, Mathematica et Physica [Suppl.2], AUC 51*, To appear.

Mack, Y. and R. Silverman (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie view. Gebiete 61*, 405–514.

McKean, J. (2004). Robust analysis of linear models. *Statistical Science 19*, 562–570.

Müller, H. (1992). Change points in nonparametric regression analysis. *Annals of Statistics 20*, 737–761.

Müller, H. and U. Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics 15*, 610–625.

Nadaraya, E. (1964). On estimating regression. *Theory Probability Applications 9*, 141–142.

Neumeyer, N. (2006). *Bootstrap Procedures for Empirical Processes of Nonparametric Residuals*. Habilitationsschrift, Ruhr-Universitt Bochum, Germany.

Page, E. (1954). Continuous inspection scheme. *Biometrika 41(1/2)*, 100–115.

Peligrad, M. (1996). On the asymptotic normality of sequences of weak dependent random variables. *Journal of Theoretical Probability 9(3)*, 703–715.

Politis, D. and J. Romano (1992). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *Annals of Statistics 20(4)*, 1985–2007.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics 12*, 1215–1230.

Ronchetti, E., C. Field, and W. Blanchard (1997). Robust linear model selection by cross-validation. *Journal of American Statistical Association 92*, 1017–1023.

Rosenblatt, M. (1956). Central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences 42(1)*, 43–47.

Rosenblatt, M. (1969). Conditional probability density and regression estimates. *Multivariate analysis II ed.Krishnaiah*, 25–31.

Rosenblatt, M. (1971). *Markov Processes: Structure and Asymptotic Behavior*. Berlin: Springer-Verlag.

Ruppert, D. and M. Wand (1994). Multivariate weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Stone, C. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*, 595–620.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press, New York.

Wang, F. and D. Scott (1994). The L_1 method for robust nonparametric regression. *Journal of American Statistical Association 89*, 65–76.

Watson, G. (1964). Smooths regression analysis. *Sankhia Ser. A 26*, 359–372.

Yang, Y. (2006). M-Cross-Validation in Local Median Estimation. *Acta Mathematica Sinica 22 (6)*, 1565–1582.

Yohai, V. and R. Maronna (1979). Asymptotic behavior of M-esimators for the linesr model. *Annals of Statistics 7*, 258–268.

Yokoyama, R. (1980). Moment bounds for stationary mixing sequences. *Z. Wahrsch. verw. Geb. 52*, 45–57.

Zelinka, J. and I. Horová (2001). Kernel estimates of a derivative of a regression function. *In Robust 2000 Conference Proceedings Prague, JČMF (In Czech)*, 382–391.