Estimation of a distribution from data with small measurement errors *

Ann-Kathrin Bott[1,†], Luc Devroye[2] and Michael Kohler[1]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: abott@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *School of Computer Science, McGill University, 3450 University Street, Montreal H3A 2K6, Canada, email: lucdevroye@gmail.com*

March 19, 2013

Abstract

In this paper we study the problem of estimation of a distribution from data that contain small measurement errors. The only assumption on these errors is that the average absolute measurement error converges to zero for sample size tending to infinity with probability one. In particular we do not assume that the measurement errors are independent with expectation zero. Throughout the paper we assume that the distribution, which has to be estimated, has a density with respect to the Lebesgue-Borel measure.

We show that the empirical measure based on the data with measurement error leads to an uniform consistent estimate of the distribution function. Furthermore, we show that in general no estimate is consistent in the total variation sense for all distributions under the above assumptions. However, in case that the average measurement error converges to zero faster than a properly chosen sequence of bandwidths,the total variation error of the distribution estimate corresponding to a kernel density estimate converges to zero for all distributions. In case of a general additive error model we show that this result even holds if only the average measurement error converges to zero. The results are applied in the context of estimation of the density of residuals in a random design regression model where the residual error is not independent from the predictor.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: Density estimation, distribution estimation, total variation error, L_1 error, measurement errors, nonparametric regression, residuals, universal consistency.

1 Introduction

Let X be a real-valued random variable with distribution μ. One of the main problems in statistics is to estimate μ from a sample $X_1, \ldots, X_n$ of X. The well-known theorem

*Running title: *Distribution estimation with small measurement errors*

†Corresponding author. Tel: +49-6151-16-75878

of Glivenko-Cantelli implies that in case X, X_1, X_2, ... are independent and identically distributed we have

$$\sup_{x\in\mathbb{R}} |\mu_n((-\infty,x]) - \mu((-\infty,x])| \to 0 \quad a.s., \tag{1}$$

where

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^n 1_A(X_i)$$

denotes the empirical distribution of X_1, ..., X_n (cf., e.g., Theorem 12.4 in Devroye, Györfi and Lugosi (1996)). So with this estimate we get consistent estimates of the probabilities of all intervals. However, if we are interested in estimation of general sets, we can consider the total variation error

$$\sup_{B\in\mathcal{B}} |\hat{\mu}_n(B) - \mu(B)| \tag{2}$$

and try to construct estimates $\hat{\mu}_n$ such that this total variation error converges to zero almost surely. Unfortunately, as was shown in Devroye and Györfi (1990), no estimate exists with the property

$$\sup_{B\in\mathcal{B}} |\hat{\mu}_n(B) - \mu(B)| \to 0 \quad a.s. \tag{3}$$

for all distributions. But if we assume that a density f of X exists, i.e., if μ is given

$$\mu(B) = \int_B f(x)\,dx \quad (B \in \mathcal{B}),$$

then we can construct estimates which satisfy (3) for all distributions via properly defined density estimates. More precisely, let $f_n(\cdot) = f_n(\cdot, X_1, \dots, X_n)$ be an estimate of f by a density f_n satisfying

$$\int |f_n(x) - f(x)|\,dx \to 0 \quad a.s. \tag{4}$$

for all densities f. E.g., the kernel density estimate (cf., e.g., Rosenblatt (1956), Parzen (1962))

$$f_n(x) = \frac{1}{n \cdot h_n}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

which depends on a density $K : \mathbb{R} \to \mathbb{R}$ (so-called kernel) and a bandwidth $h_n > 0$, has this property if h_n satisfies

$$h_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n \cdot h_n \to \infty \quad (n \to \infty) \tag{5}$$

(cf., e.g., Mnatsakanov and Khmaladze (1981) and Devroye (1983); general results in density estimation can be also found in the books of Devroye and Györfi (1985), Devroye (1987) and Devroye and Lugosi (2000)). In this case Scheffé's Lemma (cf., e.g., Devroye and Györfi (1985)) implies that the estimate

$$\hat{\mu}_n(B) = \int_B f_n(x)\,dx \quad (B \in \mathcal{B})$$

satisfies (3) for all distributions μ which have a density.

In this paper we assume that instead of the sample $X_1, \ldots, X_n$ of X we have available only data $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ such that the average absolute error between X_i and $\bar{X}_{i,n}$ converges to zero almost surely, i.e., we assume that

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \tag{6}$$

Here we do not assume anything on the measurement errors $\bar{X}_{i,n} - X_i$ $(i = 1, \ldots, n)$. In general, those errors do not need to be random and in case that they are random they do not need to be independent or identically distributed and they do not need to have expectation zero, so estimates for convolution problems (see, e.g., Meister (2009) and the literature cited therein) are not applicable in the context of this paper. Note also that our set-up is triangular.

Since we do not assume anything on the nature of the measurement errors besides that they are asymptotically negligible in the sense that (6) holds, it seems to be a natural idea to ignore them completely and to try to use the same estimates as in the case that an independent and identically distributed sample is given. In this paper we investigate whether the above mentioned distribution estimates are in this situation still consistent. As main results we show first that the corresponding empirical distribution satsifies (1) for all distributions μ which have a density with respect to the Lebesgue-Borel measure. Secondly, we show that the kernel density estimate

$$f_n(x) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{x - \bar{X}_{i,n}}{h_n}\right)$$

satisfies (4) whenever (5) and

$$\frac{1}{n \cdot h_n} \sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s.$$

hold. But, if we just assume (6) then our third result implies that there does not exist any estimate satisfying (4) for all distributions and all data with measurement errors satisfying (6). Thus, (6) is in general not strong enough a condition to guarantee total variation convergence. There is a large literature on the recovery of densities from noisy data if the noise is fixed. If the noise distribution is fixed and known, and if the noise is independent, then by deconvolution, it is possible to consistently estimate the density. However, if the noise distribution is fixed and unknown, and if the noise is independent, then it is clearly impossible to recover the density. The situation for independent but variable unknown noise is a bit better. Our fourth result shows that (6) is all that is needed for the above kernel density estimate to satisfy (4).

Finally, we apply our results in the context of estimation of the density of residuals in a random design regression model where we do not assume that the predictor and the residual error are independent.

The outline of the paper is as follows: The main results are formulated in Section 2 and proven in Section 4. In Section 3 we describe the application of our main results to the problem of estimation of the density of residual errors in a regression model.

2 Main results

The empirical distribution function is possibly the simplest way to estimate a distribution function. Even if there is no sample $X_1, \dots, X_n$ of X available, we obtain a Glivenko-Cantelli result with adequate assumptions on the available data $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$ in case that the distribution of X_1 has a density with respect to the Lebesgue-Borel measure.

Theorem 1. *Let $X_1, X_2 \dots$ be independent and identically distributed real valued random variables with density f (with respect to the Lebesgue-Borel-measure) and let $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$ be random variables which satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \quad (n \to \infty). \tag{7}$$

Then the empirical distribution function

$$\hat{\mu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(\bar{X}_{i,n})$$

of $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$ satisfies

$$\sup_{x\in\mathbb{R}} |\hat{\mu}_n((-\infty, x]) - \mu((-\infty, x])| \to 0 \quad a.s. \quad (n \to \infty).$$

The total variation error of the above estimate does not converge to zero, since by definition of $\hat{\mu}_n$ we have $\hat{\mu}_n(\{\bar{X}_{1,n}, \dots, \bar{X}_{n,n}\}) = 1$, and since $\mu(\{\bar{X}_{1,n}, \dots, \bar{X}_{n,n}\}) = 0$ in case that μ has a density with respect to the Lebesgue-Borel measure. However, our next theorem shows that if we choose a proper sequence $(h_n)_n$ of bandwidths satisfying

$$\frac{1}{n \cdot h_n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \quad (n \to \infty),$$

then we can construct a density estimate which is universally consistent in the L_1-sense and hence for which by Scheffé's Lemma the total variation error of the corresponding distribution estimate converges to zero regardless of the density f. To do this, we ignore the measurement errors again completely for estimation and define a standard kernel density estimate applied to the data with measurement errors via

$$f_n(x) = \frac{1}{n \cdot h_n}\sum_{i=1}^{n} K\left(\frac{x - \bar{X}_{i,n}}{h_n}\right).$$

Theorem 2. *Let K be any density on $\mathbb{R}_+$, let $h_n > 0$ and let f_n be defined as above. Assume that*

$$h_n \to 0 \quad \textit{and} \quad n \cdot h_n \to \infty \quad (n \to \infty). \tag{8}$$

Then

$$\frac{1}{n \cdot h_n} \sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad \textit{in } L_1 \textit{ or a.s., resp.} \tag{9}$$

implies

$$\int |f_n(x) - f(x)| \, dx \to 0 \quad \textit{in } L_1 \textit{ or a.s., resp.}$$

As shown in Devroye et al. (2012) (cf., proof of Theorem 2 in Devroye et al. (2012)), Theorem 1 is no longer valid if we replace (9) by

$$\frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \tag{10}$$

But if (10) holds we can always find $h_n = h_n(X_1, \bar{X}_{1,n}, \dots X_n, \bar{X}_{n,n})$ such that (8) and (9) hold, and consequently the resulting estimator f_n is strongly universally L_1-consistent. However, this estimator depends on the non-observable $X_1, \dots, X_n$. Surprisingly, as our next theorem shows, it is in general not possible to construct an estimate which is consistent for all densities and all samples satisfying (10), even if our sample with measurement errors does not change each time completely when the sample size changes, i.e., if we have given data $\bar{X}_1, \dots, \bar{X}_n$ instead of $\bar{X}_{1,1}, \dots, \bar{X}_{n,n}$.

Theorem 3. *There does not exist a sequence $(f_n)_n$ of density estimates satisfying*

$$\int |f_n(x, \bar{X}_1, \dots, \bar{X}_n) - f(x)| \, dx \to^P 0 \quad (n \to \infty)$$

for all densities f and all random variables $\bar{X}_1$, $\bar{X}_2$, ... satisfying

$$\frac{1}{n} \sum_{i=1}^{n} |\bar{X}_i - X_i| \to 0 \quad a.s. \tag{11}$$

for some independent and identically distributed X_1, X_2, ... with density f.

Remark 1. Assume that $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$ changes with every $n \in \mathbb{N}$ such that

$$\max_{i=1,\dots,n} |X_i - \bar{X}_{i,n}| \to 0 \quad \text{a.s.} \tag{12}$$

Then there does not exist a sequence $(f_n)_n$ of density estimates satisfying

$$\int |f_n(x, \bar{X}_{1,n}, \dots, \bar{X}_{n,n}) - f(x)| \, dx \to^P 0 \quad (n \to \infty)$$

for all densities f and all random variables $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$, which satisfy the condition (12). This can be proven as Theorem 3 above, if we set $\bar{X}_{i,n} = X_i^{(n)}$ in the proof of

Theorem 3.

In the sequel we show that under a particular noise model where independent noise is added to the true data such that the average noise is small, we can obtain weak consistency of our kernel estimate under an even weaker assumption than (11). More precisely, assume that the given data $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ is of the following form

$$\bar{X}_{i,n} = X_i + Y_{i,n} \quad (i = 1, \ldots, n),$$

where the additive noise $Y_{i,n}$ is independent of $X_1, \ldots, X_n$ and where $(X_i, Y_{i,n})$, $1 \leq i \leq n$ are independent. Additionally, we presume that $Y_{1,n}, \ldots, Y_{n,n}$ have probability measures on the Borel sets of the real line. We don't need to make any structural conditions on these probability measures.
The sequence $(Y_{i,n})_i$ of random variables is called diminishing additive noise when

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{P}_{Y_{i,n}} \to 0 \quad (n \to \infty) \tag{13}$$

weakly. We define the function $K_h(x) := (1/h)K(x/h)$ for $x \in \mathbb{R}$. For the kernel estimate

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}\left(x - \bar{X}_{i,n}\right) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}\left(x - X_i - Y_{i,n}\right)$$

we obtain the following result.

Theorem 4. *Let K be a square integrable function that integrates to one, assume that*

$$h_n \to 0 \quad \text{and} \quad n \cdot h_n \to \infty \quad (n \to \infty)$$

and define f_n as above. If the data satisfies the above diminishing noise condition, then

$$\lim_{n \to \infty} \mathbf{E}\left\{ \int |f_n(x) - f(x)| \, dx \right\} = 0.$$

If we drop the adjective "additive", and assume merely that the pairs $(X_i, Y_{i,n}), n \geq 1$, $i \leq n$ are independent [but $Y_{i,n}$ is not independent of X_i] and that the noise is diminishing, then, as shown previously, the density f cannot be consistently estimated by any estimator. If we keep the additivity but drop the diminishing noise condition then f can also not be estimated, although we will not show that in this paper.

3 Estimation of the density of residuals

Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random vectors such that $\mathbf{E}Y^2 < \infty$. Set $m(x) = \mathbf{E}\{Y|X = x\}$ and assume that a density f of

$$\epsilon = Y - m(X)$$

exists. Here we do not assume that ϵ and m are independent. Given (X_1, Y_1), ..., (X_n, Y_n) we are interested in an estimation of f.

Estimating the density of the error distribution in nonparametric regression models has been dealt with by several researchers. Ahmad (1992) showed that under a Lipschitz-condition of the kernel function, the kernel density estimator converges in probability at every continuity point to the real density of the residuals. In case of a continuous error density, the same estimator is pointwise and uniformly consistent (Cheng (2004)), and, in addition, the histogram error density estimator is uniformly and in L_1 consistent (Cheng (2002)). Efromovich (2005) investigated in a homeoscedastic regression model estimates which are as good as estimates using an oracle that knows the underlying regression errors. In the heteroscedastic nonparametric regression model, where the Y_i's have different variances, Efromovich (2006) generalized his optimal estimation for a twice differentiable error density with finite support. Estimators of the residual distribution function include that of Akritas and Van Keilegom (2001), who extended the results of Durbin (1973) and Loynes (1980) to a weak convergence result for a distribution function estimator in a nonparametric heteroscedastic regression model. The empirical distribution function of residuals was used as an estimator in an heteroscedastic model with multivariate covariates by Neumeyer and Van Keilgom (2010).

The L_1 error of estimates of the density of residual errors was considered in the papers Devroye, Felber and kohler (2013) and Györfi and Walk (2012, 2013). In the first one it is assumed that the residual error is independent of the predictor, while the latter papers make the weaker assumption that a conditional density of Y given $X = x$ exists. In our setting both kind of assumptions are not satisfied.

In the sequel, we estimate f from (X_1, Y_1), ..., (X_n, Y_n) by the following procedure: In a first step we compute a regression estimate

$$m_{\lfloor n/2 \rfloor}(\cdot) = m_{\lfloor n/2 \rfloor}(\cdot, (X_1, Y_1), \dots, (X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor})).$$

using the first half of the data. Then compute

$$\hat{\epsilon}_i = Y_i - m_n(X_i) \quad (i = \lfloor n/2 \rfloor + 1, \dots, n)$$

and estimate f by

$$f_n(x) = \frac{1}{(n - \lfloor n/2 \rfloor) \cdot h_n} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} K\left(\frac{x - \hat{\epsilon}_i}{h_n}\right).$$

From Theorem 2 we can conclude

Corollary 1. *Let K be any density on $\mathbb{R}_+$, let $h_n > 0$ and let f_n be defined as above. Assume that*

$$h_n \to 0 \quad \textit{and} \quad n \cdot h_n \to \infty \quad (n \to \infty) \tag{14}$$

and

$$\frac{1}{h_n} \cdot \mathbf{E} \int |m_{\lfloor n/2 \rfloor}(x) - m(x)| \mathbf{P}_X(dx) \to 0 \quad (n \to \infty) \tag{15}$$

holds. Then

$$\int |f_n(x) - f(x)|\, dx \to 0 \quad a.s.$$

Proof. A standard application of the inequality of McDiarmid (McDiarmid (1989)) yields

$$\int |f_n(x) - f(x)|\, dx - \mathbf{E} \int |f_n(x) - f(x)|\, dx \to 0 \quad a.s.$$

(cf., e.g., proof of Theorem 1 in Devroye, Felber and Kohler (2013)), hence it suffices to show

$$\mathbf{E} \int |f_n(x) - f(x)|\, dx \to 0 \quad (n \to \infty).$$

Set

$$\bar{\epsilon}_i = Y_i - m_{\lfloor n/2 \rfloor}(X_i, (X_{n-\lfloor n/2 \rfloor+1}, Y_{n-\lfloor n/2 \rfloor+1}), \dots, (X_n, Y_n)) \quad (i = 1, \dots, n - \lfloor n/2 \rfloor).$$

Since our data is independent and identically distributed we know

$$\mathbf{E} \int |f_n(x) - f(x)|\, dx = \mathbf{E} \int \left| \frac{1}{(n - \lfloor n/2 \rfloor) \cdot h_n} \sum_{i=1}^{n-\lfloor n/2 \rfloor} K\left(\frac{x - \bar{\epsilon}_i}{h_n}\right) - f(x) \right| dx.$$

With (15) and the observation

$$|\bar{\epsilon}_i - \epsilon_i| = \left| m_{\lfloor n/2 \rfloor}(X_i, (X_{n-\lfloor n/2 \rfloor+1}, Y_{n-\lfloor n/2 \rfloor+1}), \dots, (X_n, Y_n)) - m(X_i) \right|$$

we can conclude

$$\begin{aligned}
&\frac{1}{(n - \lfloor n/2 \rfloor) \cdot h_n} \mathbf{E} \left\{ \sum_{i=1}^{n-\lfloor n/2 \rfloor} |\bar{\epsilon}_i - \epsilon_i| \right\} \\
&= \frac{1}{h_n} \frac{1}{n - \lfloor n/2 \rfloor} \sum_{i=1}^{n-\lfloor n/2 \rfloor} \mathbf{E} \left| m_{\lfloor n/2 \rfloor}(X_i, (X_{n-\lfloor n/2 \rfloor+1}, Y_{n-\lfloor n/2 \rfloor+1}), \dots, (X_n, Y_n)) - m(X_i) \right| \\
&= \frac{1}{h_n} \mathbf{E} \int |m_{\lfloor n/2 \rfloor}(x) - m(x)| \mathbf{P}_X(dx) \to 0 \quad (n \to \infty).
\end{aligned}$$

Thus, the assertion follows from Theorem 2. □

It is well-known in the literature, that there exists weakly universally consistent nonparametric regression estimates, i.e., estimates m_n with the property

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad (n \to \infty)$$

for all distributions of (X, Y) satisfying $\mathbf{E}Y^2 < \infty$. This was first shown in Stone (1977) in case of nearest neighbor regression estimates, and later also proven for many other nonparametric regression estimates, cf., e.g., Devroye and Wagner (1980) for corresponding results for kernel estimates, Györfi (1981) for corresponding results for partitioning estimates, Lugosi and Zeger (1995) for corresponding results for least squares estimates, and Kohler and Krzyżak (2001) for corresponding results for penalized squares estimates.

If we use such an estimate, the Cauchy-Schwarz inequality implies that for every distribution of (X,Y) with $\mathbf{E}Y^2 < \infty$ we can find a sequence $(h_n)_n$ of bandwidths satisfying $h_n \to 0$ $(n \to \infty)$ and

$$\frac{\mathbf{E}\int |m_n(x) - m(x)|\mathbf{P}_X(dx)}{h_n} \to 0 \quad (n \to \infty).$$

This together with Corollary 1 implies

Corollary 2. *Let K be any density on $\mathbb{R}_+$, and let f_n be defined as above where m_n is one of the above mentioned weakly universally consistent regression estimates. Then for any distribution of (X,Y) with $\mathbf{E}Y^2 < \infty$ there exists a sequence of bandwidths $(h_n)_n$ such that*

$$h_n \to 0 \quad \textit{and} \quad n \cdot h_n \to \infty \quad (n \to \infty)$$

holds and the estimate f_n corresponding to that sequence of bandwidths satisfies

$$\int |f_n(x) - f(x)|\, dx \to 0 \quad a.s.$$

Remark 3. The above estimate depends on the distribution of (X,Y) and hence is not applicable in practice. It is an open problem, whether there exists a weakly universally consistent regression estimate such that we can construct a data-dependent choice of the bandwidth $h_n = h_n((X_1,Y_1),\dots,(X_n,Y_n))$ satisfying (14) and (15) for all distributions of (X,Y) with $\mathbf{E}\,Y^2 < \infty$.

If we impose regularity conditions on (X,Y), in particular smoothness assumptions on X, we can derive rate of convergence results for the expected L_2 error of the regression estimate and choose a fixed sequence of bandwidths satisfying (14) and (15). In this way we can prove results like

Corollary 3. *Let K be any density on $\mathbb{R}_+$, and let f_n be defined as above where*

$$m_n(x) = \frac{\sum_{i=1}^n 1_{[-1,1]}\left(\frac{x-X_i}{\bar{h}_n}\right) \cdot Y_i}{\sum_{j=1}^n 1_{[-1,1]}\left(\frac{x-X_j}{\bar{h}_n}\right)}$$

and $\bar{h}_n = n^{-1/(2+d)}$. Set $h_n = \ln(n) \cdot n^{-1/(d+2)}$. Then

$$\int |f_n(x) - f(x)|\, dx \to 0 \quad a.s.$$

for all distribution of (X,Y) with the properties that m is Lipschitz continuous, X has compact support $supp(X)$ and $\sup_{x\in supp(X)} \mathbf{E}\{Y^2|X=x\} < \infty$.

Proof. Assume that (X,Y) satisfies the assumption at the end of Corollary 3. By Theorem 5.2 in Györfi et al. (2002) we have

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c \cdot n^{-2/(2+d)}$$

for some constant $c \in \mathbb{R}$. Corollary 1 implies the assertion. $\square$

4 Proofs

4.1 Proof of Theorem 1

Let μ_n be the empirical distribution function of $X_1, \ldots, X_n$, i.e., set

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) \quad (A \in \mathcal{B}).$$

We split the expression

$$\hat{\mu}_n\left((-\infty, x]\right) - \mu\left((-\infty, x]\right)$$

in two different ways for $\epsilon > 0$:

$$\begin{aligned}
\hat{\mu}_n\left((-\infty, x]\right) - \mu\left((-\infty, x]\right) = \;& \hat{\mu}_n((-\infty, x]) - \mu_n((-\infty, x+\epsilon]) \\
&+ \mu_n((-\infty, x+\epsilon]) - \mu((-\infty, x+\epsilon]) \\
&+ \mu((-\infty, x+\epsilon]) - \mu((-\infty, x]) \\
= \;& A_{1,n} + A_{2,n} + A_{3,n}
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mu}_n((-\infty, x]) - \mu((-\infty, x]) = \;& \hat{\mu}_n((-\infty, x]) - \mu_n((-\infty, x-\epsilon]) \\
&+ \mu_n((-\infty, x-\epsilon]) - \mu((-\infty, x-\epsilon]) \\
&+ \mu((-\infty, x-\epsilon]) - \mu((-\infty, x]) \\
= \;& B_{1,n} + B_{2,n} + B_{3,n}.
\end{aligned}$$

First we consider

$$A_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\left(1_{(-\infty,x]}(\bar{X}_{i,n}) - 1_{(-\infty,x+\epsilon]}(X_i)\right).$$

The i-th summand becomes one, if

$$\bar{X}_{i,n} \le x \quad \text{and} \quad X_i > x + \epsilon.$$

In this case we have $|\bar{X}_{i,n} - X_i| > \epsilon$. If the i-th summand is not equal to one, it is less than or equal to zero. Hence

$$\begin{aligned}
A_{1,n} &= \frac{1}{n}\sum_{i=1}^{n}\left(1_{(-\infty,x]}(\bar{X}_{i,n}) - 1_{(-\infty,x+\epsilon]}(X_i)\right) \\
&\le \frac{1}{n}\sum_{i=1}^{n} 1_{\{|X_i - \bar{X}_{i,n}| > \epsilon\}} \le \frac{1}{\epsilon}\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_{i,n}|.
\end{aligned}$$

Analogously, we can conclude

$$B_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\left(1_{(-\infty,x]}(\bar{X}_{i,n}) - 1_{(-\infty,x-\epsilon]}(X_i)\right)$$

$$\geq -\frac{1}{n}\sum_{i=1}^{n} 1_{\{|X_i-\bar{X}_i|>\epsilon\}} \geq -\frac{1}{\epsilon}\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}_{i,n}|.$$

Hence, we get

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left(\hat{\mu}_n\left((-\infty,x]\right)-\mu\left((-\infty,x]\right)\right)=\sup_{x\in\mathbb{R}}\left(A_{1,n}+A_{2,n}+A_{3,n}\right)\\
&\leq \frac{1}{\epsilon}\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}_{i,n}|+\sup_{x\in\mathbb{R}}|\mu_n((-\infty,x+\epsilon])-\mu((-\infty,x+\epsilon])|+\sup_{x\in\mathbb{R}}\mu((x,x+\epsilon]).
\end{aligned}$$

By the Glivenko-Cantelli Lemma and (7) it follows

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}}\left(\hat{\mu}_n\left((-\infty,x]\right)-\mu\left((-\infty,x]\right)\right)\leq\sup_{x\in\mathbb{R}}\mu((x,x+\epsilon]).$$

Similarly, we obtain

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left(\mu\left((-\infty,x]\right)-\hat{\mu}_n\left((-\infty,x]\right)\right)=\sup_{x\in\mathbb{R}}\left(-B_{1,n}-B_{2,n}-B_{3,n}\right)\\
&\leq \frac{1}{\epsilon}\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}_{i,n}|+\sup_{x\in\mathbb{R}}|\mu_n((-\infty,x-\epsilon])-\mu((-\infty,x-\epsilon])|+\sup_{x\in\mathbb{R}}\mu((x-\epsilon,x]),
\end{aligned}$$

from which we conclude

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}}\left(\mu\left((-\infty,x]\right)-\hat{\mu}_n\left((-\infty,x]\right)\right)\leq\sup_{x\in\mathbb{R}}\mu((x,x+\epsilon]).$$

Since μ has a density with respect to the Lebesgue-Borel measure, μ is Lebesgue continuous. For the Lebesgue measure λ we know $\sup_{x\in\mathbb{R}}\lambda((x,x+\epsilon])\leq\epsilon$. By the Lebesgue continuity it follows for $\epsilon\to 0$

$$\sup_{x\in\mathbb{R}}\mu((x,x+\epsilon])\to 0.$$

From the above results we conclude

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mu\left((-\infty,x]\right)-\hat{\mu}_n\left((-\infty,x]\right)\right|\\
&\leq\sup_{x\in\mathbb{R}}\left(\mu\left((-\infty,x]\right)-\hat{\mu}_n\left((-\infty,x]\right)\right)+\sup_{x\in\mathbb{R}}\left(\hat{\mu}_n\left((-\infty,x]\right)-\mu\left((-\infty,x]\right)\right)\to 0\quad\text{a.s.}
\end{aligned}$$

The proof is complete. □

4.2 Proof of Theorem 2

Set

$$f_n^*(x)=\frac{1}{n\cdot h_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right).$$

By Devroye and Györfi (1985) we know

$$\int |f_n^*(x) - f(x)|\, dx \to 0 \quad \text{in } L_1 \text{ and a.s.}$$

Hence it suffices to show

$$\int |f_n(x) - f_n^*(x)|\, dx \to 0$$

in expected value or almost surely, respectively. Now, writing $K_h(x) = (1/h)K(x/h)$ and setting $u = (x - X_i)/h_n$ we get

$$\begin{aligned} \int |f_n(x) - f_n^*(x)|\, dx &\leq \frac{1}{n}\sum_{i=1}^{n} \int |K_{h_n}(x - X_i) - K_{h_n}(x - \bar{X}_{i,n})|\, dx \\ &= \frac{1}{n}\sum_{i=1}^{n} \int \left| K(u) - K\left(u - \frac{\bar{X}_{i,n} - X_i}{h_n}\right)\right| du. \end{aligned}$$

For $\epsilon > 0$, find $\delta > 0$ so small that

$$\sup_{|y|\leq\delta} \int |K(u) - K(u-y)|\, du < \epsilon.$$

Then, by Markov's inequality,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \int \left| K(u) - K\left(u - \frac{\bar{X}_{i,n} - X_i}{h_n}\right)\right| du \\ \leq &\frac{1}{n}\sum_{i=1}^{n} 1_{\left\{\left|(\bar{X}_{i,n}-X_i)/h_n\right|\leq\delta\right\}}\epsilon + \frac{1}{n}\sum_{i=1}^{n} 1_{\left\{\left|(\bar{X}_{i,n}-X_i)/h_n\right|>\delta\right\}} \cdot 2 \\ \leq &\epsilon + \frac{2}{\delta n}\sum_{i=1}^{n}\left|\frac{\bar{X}_{i,n} - X_i}{h_n}\right|, \end{aligned}$$

which is almost surely smaller than 2ϵ for all n large enough by (9) in case that (9) holds almost surely. Otherwise the expectation of the right-hand side above is smaller than 2ϵ for n large enough. This completes the proof. □

4.3 Proof of Theorem 3

Assume to the contrary that there exists a sequence $(f_n)_n$ of density estimates satisfying

$$\int |f_n(x, \bar{X}_1, \ldots, \bar{X}_n) - f(x)|\, dx \to^P 0 \quad (n \to \infty) \tag{16}$$

whenever $\bar{X}_1$, $\bar{X}_2$, ... are such that for some independent and identically distributed X_1, X_2, ... with density f we have

$$\frac{1}{n}\sum_{i=1}^{n} |\bar{X}_i - X_i| \to 0 \quad a.s.$$

Let X_1, X_2, ... be independent and uniformly on $[0,1]$ distributed random variables and let

$$g(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{else,} \end{cases}$$

be the density of X_1. For $k \in \mathbb{N}$, $k \geq 2$ set

$$g_k(x) = \begin{cases} 2 & \text{if } \frac{2\ell}{2k} \leq x < \frac{2\ell+1}{2k} \text{ for some } \ell \in \{0, \dots, k-1\} \\ 0 & \text{else} \end{cases}$$

and

$$X_i^{(k)} = \begin{cases} X_i & \text{if } \frac{2\ell}{2k} \leq X_i < \frac{2\ell+1}{2k} \text{ for some } \ell \in \{0, \dots, k-1\} \\ X_i - \frac{1}{2k} & \text{if } \frac{2\ell+1}{2k} \leq X_i < \frac{2\ell+2}{2k} \text{ for some } \ell \in \{0, \dots, k-1\}. \end{cases}$$

Then $X_1^{(k)}$, $X_2^{(k)}$, ... are independent and identically distributed random variables with density g_k. So if we set $\bar{X}_i = X_i^{(k)}$ for all $i \geq N$ with $N \in \mathbb{N}$ arbitrary we know by (16) that

$$\int |f_n(x, \bar{X}_1, \dots, \bar{X}_n) - g_k(x)| \, dx \to^P 0 \quad (n \to \infty). \tag{17}$$

Next we define for suitable chosen $n_0 := 0 < n_1 < n_2 < \dots$ our data with measurement error by

$$\bar{X}_i = X_i^{(k)} \quad \text{if } n_{k-1} < i \leq n_k \quad (k \in \mathbb{N}).$$

Because of $|X_i^{(k)} - X_i| \leq 1/(2k)$ we have

$$\frac{1}{n} \sum_{i=1}^{n} |\bar{X}_i - X_i| \to 0 \quad a.s.,$$

so our theorem is proven as soon as we can show for some $\epsilon > 0$

$$\limsup_{n \to \infty} \mathbf{P} \left[\int |f_n(x, \bar{X}_1, \dots, \bar{X}_n) - g(x)| \, dx > \epsilon \right] > 0. \tag{18}$$

Next we show that we can choose n_k such that (18) holds. Let $0 < \epsilon < 1$ be fixed and choose n_1 such that

$$\mathbf{P} \left[\int |f_{n_1}(x, X_1^{(1)}, \dots, X_{n_1}^{(1)}) - g_1(x)| \, dx > \epsilon \right] < \frac{1}{2},$$

which is possible because of (16). Given n_1, ..., n_{k-1}, we choose $n_k > n_{k-1}$ such that

$$\mathbf{P} \left[\int |f_{n_k}(x, \bar{X}_1, \dots, \bar{X}_{n_{k-1}}, X_{n_{k-1}+1}^{(k)}, \dots, X_{n_k}^{(k)}) - g_k(x)| \, dx > \epsilon \right] < \frac{1}{2},$$

which is possible because of (17). But if we define n_1, n_2, ... in such a way, we have

$$\mathbf{P} \left[\int |f_{n_k}(x, \bar{X}_1, \dots, \bar{X}_{n_k}) - g_k(x)| \, dx > \epsilon \right] < \frac{1}{2}$$

and accordingly

$$\mathbf{P}\left[\int |f_{n_k}(x, \bar{X}_1, \ldots, \bar{X}_{n_k}) - g_k(x)|\, dx \leq \epsilon\right] \geq \frac{1}{2}$$

for all $k \in \mathbb{N}$. By triangle inequality we know

$$\int |g_k(x) - g(x)|\, dx \leq \int |f_{n_k}(x) - g_k(x)|\, dx + \int |f_{n_k}(x) - g(x)|\, dx.$$

Furthermore we have

$$\int |g_k(x) - g(x)|\, dx = 1.$$

From this we can conclude for any $k \in \mathbb{N}$

$$\begin{aligned}
&\mathbf{P}\left[\int |f_{n_k}(x, \bar{X}_1, \ldots, \bar{X}_{n_k}) - g(x)|\, dx > 1 - \epsilon\right] \\
&\geq \mathbf{P}\left[\int |g_k(x) - g(x)|\, dx - \int |f_{n_k}(x, \bar{X}_1, \ldots, \bar{X}_{n_k}) - g_k(x)|\, dx > 1 - \epsilon\right] \\
&= \mathbf{P}\left[\int |f_{n_k}(x, \bar{X}_1, \ldots, \bar{X}_{n_k}) - g_k(x)|\, dx < \epsilon\right] \\
&\geq \frac{1}{2}.
\end{aligned}$$

The proof is complete. □

4.4 Proof of Theorem 4

We introduce the probability measures ν_n:

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}_{Y_{i,n}}.$$

The diminishing noise condition implies that a random variable Z_n drawn from ν_n tends to 0 in distribution and hence also in probability (cf., e.g., Theorem 18.3 in Jacod and Protter (2000)). We use the notation $*$ for the convolution operation. In general for a function f and a measure μ, we write

$$(f * \mu)(x) = \int f(x - y) d\mu(y).$$

Similarly, for two functions f, g, we have

$$(f * g)(x) = \int f(x - y) g(y)\, dy.$$

The first result we require is the following:

$$\lim_{n\to\infty} \int |f(x) - (f * \nu_n)(x)|\, dx = 0.$$

For an arbitrary $\epsilon > 0$, find a uniformly continuous density g, of compact support, such that

$$\int |f(x) - g(x)|\, dx < \epsilon.$$

Then, omitting (x) and dx in the integrals,

$$\begin{aligned}
\int |f - f * \nu_n| &\leq \int |f - g| + \int |g - g * \nu_n| + \int |(f - g) * \nu_n| \\
&\leq \int |f - g| + \int |g - g * \nu_n| + \int |f - g| * \nu_n \\
&= 2 \int |f - g| + \int |g - g * \nu_n|. \qquad (19)
\end{aligned}$$

Next we consider the second integral on the right hand side of (19): First, since $Z_n \to 0$ in probability, we can find $a_n \downarrow 0$ such that $\mathbf{P}\{|Z_n| \geq a_n\} \to 0$. Thus, if the uniform modulus of continuity of g is ω, and g vanishes off $[-b, b]$,

$$\begin{aligned}
\int |g - g * \nu_n| &= \int |g(x) - \int g(x - z)\, d\nu_n(z)|\, dx \\
&\leq \int \int |g(x) - g(x - z)|\, d\nu_n(z)\, dx \\
&\leq \int_{-b-a_n}^{b+a_n} \omega(a_n)\, dx + \int \int_{|z| \geq a_n} (g(x) + g(x - z))\, d\nu_n(z)\, dx \\
&\leq (2b + 2a_n)\omega(a_n) + 2 \int_{|z| \geq a_n} d\nu_n(z) \\
&= (2b + 2a_n)\omega(a_n) + 2\mathbf{P}\{|Z_n| \geq a_n\} \\
&\to 0 \quad (n \to \infty).
\end{aligned}$$

Hence,

$$\limsup_{n \to \infty} \int |f(x) - (f * \nu_n)(x)|\, dx \leq 2 \int |f(x) - g(x)|\, dx < 2\, \epsilon.$$

Next, we have trivially,

$$\int |(K_{h_n} * f * \nu_n)(x) - (f * \nu_n)(x)|\, dx \leq \int |(K_{h_n} * f)(x) - f(x)|\, dx \to 0 \quad (n \to \infty)$$

when $h_n \to 0$ $(n \to \infty)$. This is a standard result from real analysis [e.g., Theorem 1, Chapter 2 in Devroye and Györfi (1985)]. By the triangle inequality, we thus have

$$\lim_{n \to \infty} \int |(K_{h_n} * f * \nu_n)(x) - f(x)|\, dx = 0.$$

Finally, we are ready for the main argument. Split the L1 error traditionally in bias and variation components:

$$\int |f_n(x) - f(x)|\, dx \le \int |f_n(x) - \mathbf{E} f_n(x)|\, dx + \int |\mathbf{E} f_n(x) - f(x)|\, dx.$$

The last term tends to zero because

$$\mathbf{E}\{K_{h_n}(x - X_i - Y_{i,n})\} = (K_{h_n} * f * \mathbf{P}_{Y_{i,n}})(x),$$

and thus,

$$\mathbf{E}\{f_n(x)\} = \frac{1}{n}\sum_{i=1}^{n}(K_{h_n} * f * \mathbf{P}_{Y_{i,n}})(x) = (K_{h_n} * f * \nu_n)(x).$$

Theorem 4 follows if we can show that

$$\lim_{n\to\infty} \int \mathbf{E}\{|f_n(x) - \mathbf{E} f_n(x)|\}\, dx = 0.$$

For given $\epsilon > 0$, find $a > 0$ such that

$$\int_{|x|\ge a} f(x)\, dx < \epsilon.$$

Note that

$$\begin{aligned}
\int_{|x|\ge a} \mathbf{E} f_n(x)\, dx &= \int_{|x|\ge a} (K_{h_n} * f * \nu_n)(x)\, dx \\
&\le \int_{|x|\ge a} f(x)\, dx + \int |(K_{h_n} * f * \nu_n)(x) - f(x)|\, dx \\
&< 2\epsilon + \int |(K_{h_n} * f * \nu_n)(x) - f(x)|\, dx.
\end{aligned}$$

Thus,

$$\limsup_{n\to\infty} \int_{|x|\ge a} \mathbf{E}\{|f_n(x) - \mathbf{E} f_n(x)|\}\, dx \le 2 \limsup_{n\to\infty} \int_{|x|\ge a} \mathbf{E} f_n(x)\, dx < 4\epsilon.$$

We conclude the assertion by showing that

$$\limsup_{n\to\infty} \int_{|x|\le a} \mathbf{E}\{|f_n(x) - \mathbf{E} f_n(x)|\}\, dx = 0.$$

By Jensen's inequality and independence, we have

$$\mathbf{E}\{|f_n(x) - \mathbf{E} f_n(x)|\}^2 \le \mathbf{E}\left\{(f_n(x) - \mathbf{E} f_n(x))^2\right\}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{V}\left\{K_{h_n}(x - X_i - Y_{i,n})\right\} \\
&\le \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}\left\{K_{h_n}(x - X_i - Y_{i,n})^2\right\} \\
&= \frac{1}{n^2 h_n}\sum_{i=1}^{n} (K_{h_n}^2 * f * \mathbf{P}_{Y_{i,n}})(x) \\
&= \frac{1}{nh_n}(K_{h_n}^2 * f * \nu_n)(x).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\int_{|x|\le a} \mathbf{E}\{|f_n(x) - \mathbf{E}f_n(x)|\}\,dx &\le \int_{|x|\le a} \sqrt{\mathbf{E}^2\{|f_n(x) - \mathbf{E}f_n(x)|\}}\,dx \\
&\le \int_{|x|\le a} \sqrt{\frac{1}{nh_n}(K_{h_n}^2 * f * \nu_n)(x)}\,dx \\
&\le \sqrt{\frac{2a}{nh_n}} \times \sqrt{\int_{|x|\le a} (K_{h_n}^2 * f * \nu_n)(x)\,dx} \\
&\le \sqrt{\frac{2a\int K^2}{nh_n}}.
\end{aligned}
$$

This tends to zero if $nh_n \to \infty$. The proof is complete. $\square$

References

[1] Ahmad, I. A. (1992). Residuals density estimation in nonparametric regression. *Statistics and Probability Letters*, **14**, pp. 133–139.

[2] Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric Estimation of the Residual Distribution. *Board of the Foundation of the Scandinavian Journal of Statistics, Blackwell Publishers Ltd*, **28**, pp. 549–567.

[3] Cheng, F. (2002). Consistency of error density and distribution function estimators in nonparametric regression. *Statistics and Probability Letters*, **59**, pp. 257–270.

[4] Cheng, F. (2004). Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. *Journal of Statistical Planning and Inference*, **119**, pp. 95–107.

[5] Devroye, L. (1983). The equivalence in L1 of weak, strong and complete convergence of kernel density estimates. *Annals of Statistics*, **11**, pp. 896–904.

[6] Devroye, L. (1987). A Course in Density Estimation. *Birkhäuser*, Basel.

[7] Devroye, L., Felber, T. and Kohler, M. (2013). Estimation of a density using real and artificial data. *IEEE Transactions on Information Theory* **59**, pp. 1917-1928.

[8] Devroye, L., Felber, T., Kohler, M. and Krzyżak, A. (2012). L_1-consistent estimation of the density of residuals in random design regression models. *Statistics and Probability Letters* **82**, pp. 173-179.

[9] Devroye, L. and Györfi, L. (1985). Nonparametric Density Estimation. The L1 view. *Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. JohnWiley and Sons*, New York.

[10] Devroye, L. and Györfi, L. (1990). No empirical probability measure can converge in the total variation sense for all distributions. *Annals of Statistics*, **18**, pp. 1496-1499.

[11] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[12] Devroye, L. and Lugosi, G. (2000). Combinatorial Methods in Density Estimation. *Springer-Verlag*, New York.

[13] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231–239.

[14] Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics*, **1**, pp. 279–290.

[15] Efromovich, S. (2005). Estimation of the density of regression errors. *Annals of Statistics*, **33**, pp. 2194–2227.

[16] Efromovich, S. (2006). Optimal nonparametric estimation of the density of regression errors with finite support. *AISM*, **59**, pp. 617–654.

[17] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

[18] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. *Springer-Verlag*, New York.

[19] Györfi, L. and Walk, H. (2012). Strongly consistent density estimation of regression residuals. *Statistics and Probability Letters* **82**, pp. 1923-1929.

[20] Györfi, L. and Walk, H. (2013). Rate of convergence of the density estimation of regression residual. *Statistics and Risk Modeling* **30**, pp. 55-73.

[21] Jacod, J. and Protter, P. E. (2000). *Probability essentials.* Universitext - Springer-Verlag, Berlin Heidelberg.

[22] Kohler, M. and Krzyżyk, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054-3058.

[23] Loynes, R. M. (1980). The empirical sample distribution function of residuals from generalized regression. *Annals of Statistics*, **8**, pp. 285–298.

[24] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677–687.

[25] McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics 1989*, vol. 141, pp. 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge.

[26] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics, Vol. 193, Springer.

[27] Mnatsakanov, R. M., and Khmaladze, E. V. (1981). On L_1-convergence of statistical kernel estimators of distribution densities. *Soviet Mathematics Doklady*, **23**, pp. 633–636.

[28] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, **101**, pp. 1067–1078.

[29] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, pp. 1065–1076.

[30] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp. 832–837.

[31] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statististics*, **5**, pp. 595–645.