Estimating the density of a possibly missing response variable in nonlinear regression

Ursula U. Müller[1]

Department of Statistics
Texas A&M University
College Station, TX 77843-3143
USA

Abstract

This article considers linear and nonlinear regression with a response variable that is allowed to be "missing at random". The only structural assumptions on the distribution of the variables are that the errors have mean zero and are independent of the covariates. The independence assumption is important. It enables us to construct an estimator for the response density that uses all the observed data, in contrast to the usual local smoothing techniques, and which therefore permits a faster rate of convergence. The idea is to write the response density as a convolution integral which can be estimated by an empirical version, with a weighted residual-based kernel estimator plugged in for the error density. For an appropriate class of regression functions, and a suitably chosen bandwidth, this estimator is consistent and converges with the optimal parametric rate $n^{1/2}$. Moreover, the estimator is proved to be efficient (in the sense of Hájek and Le Cam) if an efficient estimator is used for the regression parameter.

Keywords: least dispersed estimator, semiparametric regression, empirical likelihood, influence function, gradient.
2000 MSC: 62G07, 62J02, 62G20

1. Introduction

We study regression models of the form $Y = r_\vartheta(X) + \varepsilon$, where r_ϑ is a linear or nonlinear regression function that depends smoothly on a finite-dimensional parameter vector ϑ. We assume that the covariate vector X and the error variable ε are independent, and that the errors have mean zero and finite variance. We will not make any further model assumptions on the distributions of X and ε, in other words our model is a *semiparametric* regression model. Note that this model with an *unknown* error distribution is particularly relevant in situations

Email address: uschi@stat.tamu.edu (Ursula U. Müller)
URL: http://www.stat.tamu.edu/∼uschi/ (Ursula U. Müller)
[1]Ursula U. Müller was supported by NSF Grant DMS-0907014.

where it is not appropriate to assume that the errors are from a normal distribution, or from some other specific distribution, which would allow estimation of the regression function using likelihood techniques.

We are interested in situations where some of the responses Y are missing. More precisely, we will assume that Y is *missing at random* (MAR). This means that we always observe X, but only observe Y in those cases where some indicator δ equals one, and the indicator δ is conditionally independent of Y given X, i.e. $P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X)$. MAR is a common assumption and is reasonable in many situations (see Little & Rubin, 2002, Chapter 1). One example would be the problem of non-responses in survey questions: assume, for example, that additional data about socioeconomic status are available. It is possible that the response probabilities are different for subjects with different socioeconomic backgrounds. It is also possible that subjects from the same status group are equally likely to respond, regardless what the response would be.

Note that the more intuitive notion of randomness for the missing value mechanism is called *missing completely at random* (MCAR). Here the missing value mechanism does not depend on observed or unobserved measurements, i.e. $P(\delta = 1|X, Y)$ is a constant, $P(\delta = 1|X, Y) = \pi(X) = \pi$. (By assuming MAR responses we will also cover the MCAR situation, since $\pi(X) = \pi$ is simply a special case.) The situation when data are *not missing at random* (NMAR) will not be studied here: in this case $P(\delta = 1|X, Y) = \pi(X, Y)$ is a function of X and Y. It therefore depends on data that are missing, which means that inference requires auxiliary information. With the MAR assumption, $P(\delta = 1|X, Y) = \pi(X)$ depends only on observable data, i.e. the mechanism $\pi(X)$ is "ignorable": it need not be modeled since it can be estimated from the data.

In this article we study density estimators, specifically consistent estimators of the density of the response variable Y which converge with the unusual (fast) rate $n^{1/2}$. The simplest (slowly converging) estimator of the density of a variable Y at some point y is a kernel estimator based on observed responses,

$$\frac{1}{N}\sum_{i=1}^{n}\delta_i k_b(y - Y_i), \tag{1.1}$$

where N is the number of completely observed data points, $N = \sum_{j=1}^{n}\delta_j$, and where $k_b(y) = k(y/b)/b$ with kernel function k and bandwidth $b > 0$. If there is additional information available in the form of a single covariate or a covariate vector X, then it is intuitively clear that an estimator which uses the additional information should be better than the kernel estimator above. This idea is, for example, used by Wang (2008) for a related regression model with MAR responses, but without assuming independence of covariates and errors. He introduces a probability weighted estimator and an imputed estimator, and proves local asymptotic normality – but with rates slower than $n^{1/2}$, which is typical for kernel estimators. Also related is Mojirsheibani (2007), who studies partial imputation for response density estimators in a nonparametric regression setting with MAR responses. He also obtains convergence rates that are slower than $n^{1/2}$.

Here we construct an estimator for the response density from a sample $(X_i, \delta_i Y_i, \delta_i)$. Under appropriate conditions on the regression function and the distribution, our estimator will converge with the desired rate $n^{1/2}$, and, beyond that, will be efficient. The case with missing responses is an important generalization of the case with fully observed data, which is covered as a special case with all indicators $\delta_i = 1$. This is a research area where little work has been done, even if we include cases where all data are observed.

In order to introduce the estimator we write M for the covariate distribution and f for the error density. We also suppose that $r_\vartheta(X)$ has a density g. Then the response density, say $q(y)$, can be written as a convolution integral,

$$\begin{aligned} q(y) &= \int f\{y - r_\vartheta(x)\}\, M(dx) = \int f(y-u) g(u)\, du \\ &= \int f(u) g(y-u)\, du = E\{g(y-\varepsilon)\} = f * g(y). \end{aligned}$$

This representation suggests two plug-in estimators of the integral: firstly, a convolution of kernel density estimators,

$$\begin{aligned} \hat{q}(y) = \hat{f} * \hat{g}(y) \quad \text{with } \hat{f}(z) &= \frac{1}{N} \sum_{j=1}^{n} \delta_j k_b(z - \hat{\varepsilon}_j), \\ \hat{g}(z) &= \frac{1}{n} \sum_{j=1}^{n} k_b\{z - r_{\hat{\vartheta}}(X_j)\}, \end{aligned} \tag{1.2}$$

where $\hat{\varepsilon}_i = Y_i - r_{\hat{\vartheta}}(X_i)$ are the residuals based on a $n^{1/2}$-consistent estimator $\hat{\vartheta}$ of ϑ. Another obvious and perhaps even simpler estimator is

$$\int \hat{f}\{y - r_{\hat{\vartheta}}(x)\}\, d\hat{M}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\{y - r_{\hat{\vartheta}}(X_i)\} \tag{1.3}$$

with $\hat{f}$ from above. Here $\hat{M}$ is just the empirical covariate distribution function. For technical reasons we will work with estimator (1.2) in this article. However, it is easy to see that this estimator can always be written in the form (1.3). The reverse does not hold in general. The two estimators are the same if, for example, the kernel k in $\hat{f}$ is the standard normal density. (See Section 5 for more details.)

Note that the two estimators (1.2) and (1.3) use all observations, whereas the usual kernel estimator (1.1) only uses a fraction of the data, namely the responses Y_i in a neighborhood of y, which explains the faster convergence rate. The convolution approach is therefore, in general, better than the usual approach (1.1), and even better than the usual estimator based on complete data pairs. A degenerate case is given if the regression function is a constant, $r_\vartheta(x) = \vartheta$. In this case we do not estimate an integral: $q(y)$ is just a shift of the error density, $q(y) = \int f(y-\vartheta) M(dx) = f(y-\vartheta)$ – which is estimated with the usual slow rates.

Now suppose that the response density can be written as a non-degenerate convolution integral. The estimator (1.2) will, in general, be $n^{1/2}$-consistent but not efficient. In order to make it efficient we have to use an efficient estimator for $\hat{\vartheta}$. We will also have to incorporate the mean zero constraint on the error distribution, which we achieve by adding Owen's empirical likelihood weights. The weighted estimator of the error density, $\hat{f}_w$, is then a weighted version of $\hat{f}$, with weights $\hat{w}_j$ based on residuals $\hat{\varepsilon}_j = Y_j - r_{\hat{\vartheta}}(X_j)$ such that the weighted residuals are centered around zero, $\sum_{j=1}^n \hat{w}_j \delta_j \hat{\varepsilon}_j = 0$. Our final estimator is the weighted version $\hat{q}_w(y)$ of (1.2), namely

$$\hat{q}_w(y) = \hat{f}_w * \hat{g}(y) \quad \text{with} \quad \hat{f}_w(z) = \frac{1}{N} \sum_{j=1}^n \hat{w}_j \delta_j k_b(z - \hat{\varepsilon}_j), \tag{1.4}$$

and with $\hat{g}$ as before. For some auxiliary results we will refer to Müller (2009), where, in the same nonlinear regression setting, we derived efficient estimators for expectations $Eh(X, Y)$. The key idea in that article is to exploit the independence of covariates and errors by writing the expectation as an iterated expectation which can be estimated by empirical plug-in estimators. The construction of an efficient estimator for $\hat{\vartheta}$ will not be discussed here. We refer to Müller (2009) Section 5. For further background on efficient estimation of expectations in various regression settings with MAR responses see also Müller, Schick and Wefelmeyer (2006). Weighted residual-based density estimators for *autoregressive* models are studied in Müller, Schick and Wefelmeyer (2005).

To our knowledge, our proposed response density estimator is the first efficient estimator for nonlinear and even linear regression. The result is new even if we only consider the special case with completely observed data. By allowing for responses *missing at random* we cover a common missing data situation. Our efficiency results are based on the Hájek–Le Cam theory for locally asymptotically normal families. As a consequence, the proposed estimator has a limiting normal distribution with the asymptotic variance determined by the influence function, which provides a basis for further inference.

Our estimator is motivated by Saavedra and Cao (1999), and Schick and Wefelmeyer (2004a, 2004b), who propose $n^{1/2}$-consistent estimators for marginal densities of first order moving average processes using plug-in methods. An early paper introducing a U-statistic estimator that converges faster than the usual kernel estimator is Frees (1994). Other articles have considered $n^{1/2}$-consistent density estimation in various (complete data) settings, for example Schick and Wefelmeyer (2004c, 2007a), who estimate densities of sums of independent random variables, Du and Schick (2007) who derive estimators for derivatives, and Schick and Wefelmeyer (2007b), who study linear processes. Giné and Mason (2007) derive central limit theorems (in various norms) for density estimators of functions of several variables in a general framework. Two recent papers in a nonparametric regression setting (with an *unspecified* regression function r) are Escanciano and Jacho-Chávez (2010) and Støve and Tjøstheim (2011). Escanciano and Jacho-Chávez describe properties of their estimator $\hat{r}(X)$ which entail asymptotic normality of the (unweighted) response density estimator (1.3) (with $\hat{r}$ in place of $r_{\hat{\vartheta}}$). Støve and Tjøstheim study the mean squared error of

the same estimator. Both articles propose the Nadaraya-Watson estimator for estimating the regression function.

As in Giné and Mason (2007), who prove limiting normality of their estimators, we will have to require square-integrability. In the context of regression this is often violated. Consider again the response density $q(y) = f * g(y) = E\{g(y-\varepsilon)\}$ where g is the density of $r_\vartheta(X)$. In order to guarantee square-integrability we need r_ϑ such that $g(y-\varepsilon)$ has a finite second moment. In Section 5 of this article we discuss this assumption and present some simulations. We study both the situation when the regression function is correctly specified and when it is misspecified – with convincing results. We also identify and discuss irregular cases where the $n^{1/2}$ rate (and efficiency) cannot be obtained, for example when r_ϑ is constant or has parts that are constant or close to constant. If X is two-dimensional this means that there are saddle points or local extrema where the curve is close to flat. The assumptions are typically satisfied if the regression function is linear or partly linear. Note that the result by Escanciano and Jacho-Chávez (2010), for the same reasons, only holds for a restricted class of regression functions – which they implicitly assume by requiring square-integrability. Støve and Tjøstheim (2011) explicitly assume a (univariate) regression function that is strictly monotonic.

The earlier sections of this paper are organized as follows. In Section 2 we derive an expansion of the unweighted version of the response density estimator. This result is used in Section 3, where we provide an approximation of the final weighted estimator. Both expansions still involve the difference $\hat{\vartheta} - \vartheta$. Section 4 is on efficiency: we characterize the influence function of an *efficient* estimator of the response density. Comparing the efficient influence function and the approximation in Section 3 yields that the weighted estimator is efficient if $\hat{\vartheta}$ is efficient. The main result is given at the end of Section 4 in Theorem 3.

2. The unweighted estimator

We commence by deriving a first asymptotic expansion for the unweighted estimator (1.2). We begin with some notation. Recall that we are considering the nonlinear regression model $Y = r_\vartheta(X) + \varepsilon$, with covariate vector X independent of the error variable ε. We write M for the distribution function of X, and F and f for the distribution function and the density function of ε. The error ε has mean zero and finite variance. Write $\hat{\varepsilon}_i = Y_i - r_{\hat{\vartheta}}(X_i)$ and $k_b(y) = k(y/b)/b$, where k is a kernel and b a bandwidth. The unweighted estimator (1.2) has the form $\hat{q} = \hat{f} * \hat{g}$ with $\hat{f}(z) = N^{-1} \sum_{j=1}^n \delta_j \, k_b(z - \hat{\varepsilon}_j)$ and $\hat{g}(z) = n^{-1} \sum_{j=1}^n k_b\{z - r_{\hat{\vartheta}}(X_j)\}$, where $\hat{\vartheta}$ is a $n^{1/2}$-consistent estimator. We work with a second order kernel. This means $\int k(u)\,du = 1$, $\int uk(u)\,du = 0$ and $0 < \int u^2 k(u)\,du < \infty$.

We now state the main conditions under which the asymptotic expansion of the response density estimator is proved. We will assume throughout this article that $P(\delta = 1) = E\delta$ is positive to exclude the degenerate case that a response is (almost surely) never observed.

Condition K. The kernel k is of order two and is three times continuously

differentiable with $\|k\|_2$, $\|k'\|_2$, $\|k''\|_2$ and $\|k'''\|_2$ finite.

Condition F. The error density f has an integrable derivative f' that is Lipschitz, $|f'(x) - f'(y)| \le |x - y|L$ for some positive finite constant L.

Condition G. The density g of $r_\vartheta(X)$ is of bounded variation.

Condition R. The regression function $\tau \mapsto r_\tau(x)$ is differentiable with a p-dimensional gradient $\dot{r}_\vartheta(x)$ which satisfies $E|\dot{r}_\vartheta(X)|^4 < \infty$ and the Lipschitz condition

$$|\dot{r}_\tau(x) - \dot{r}_\vartheta(x)| \le |\tau - \vartheta| a(x)$$

with $a \in L_2(M)$.

Condition T. The estimator $\hat{\vartheta}$ is $n^{1/2}$-consistent.

As a consequence of conditions F and G, the densities f and g are bounded and therefore square-integrable. The response density $q = f * g$ thus belongs to $C_0(\mathbb{R})$, where $C_0(\mathbb{R})$ is the set of all continuous functions h from $\mathbb{R} \to \mathbb{R}$ that vanish at infinity, $\sup_{|y|>M} |h(y)| \to 0$ as $M \to \infty$. Note that $C_0(\mathbb{R})$ equipped with the sup-norm $\|\cdot\|_\infty$ is a separable Banach space. We will consider a central limit theorem for our response density estimator in this space. It also follows from Condition F that $q = f * g$ has a Lipschitz continuous derivative $q' = f' * g$.

Condition R implies that $|r_{\vartheta+\tau} - r_\vartheta - \dot{r}_\vartheta^\top \tau| \le |\tau^2| a$ for τ close to ϑ. From this we immediately obtain for a $n^{1/2}$-consistent estimator $\hat{\vartheta}$

$$\frac{1}{n}\sum_{i=1}^n \left| r_{\hat{\vartheta}}(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^\top(\hat{\vartheta} - \vartheta) \right| = O_p(n^{-1}), \tag{2.1}$$

$$\frac{1}{n}\sum_{i=1}^n \left\{ r_{\hat{\vartheta}}(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^\top(\hat{\vartheta} - \vartheta) \right\}^2 = O_p(n^{-2}).$$

Also, since $|\dot{r}_\vartheta(X)|$ has a finite fourth moment, we have

$$\max_{1\le j\le n} |\dot{r}_\vartheta(X_j)| = o_p(n^{1/4}).$$

A further consequence of Condition R is

$$\max_{1\le j\le n} \delta_j |\hat{\varepsilon}_j - \varepsilon_j| = o_p(1),$$

which was proved in Lemma 7.1 in Müller (2009).

In Theorem 1 we will derive an approximation for the difference between the unweighted estimator $\hat{q}$ and the density q, which will involve the difference $\hat{\vartheta} - \vartheta$ where, for now, $\hat{\vartheta}$ is some $n^{1/2}$-consistent estimator of ϑ. An expansion

for $\hat{\vartheta} - \vartheta$ is provided in Section 4. We introduce some notation and write

$$\begin{aligned} \varrho &= E\{\dot{r}_\vartheta(X)|\delta = 1\} \\ A_1(y) &= \frac{1}{N}\sum_{i=1}^{n} \delta_i\{g(y-\varepsilon_i) - q(y)\}, \\ A_2(y) &= \frac{1}{n}\sum_{i=1}^{n}[f\{y - r_\vartheta(X_i)\} - q(y)], \\ \dot{q}(y) &= -\int f'\{y - r_\vartheta(x)\}\dot{r}_\vartheta(x)\, M(dx). \end{aligned}$$

The proof of Theorem 1 rests on the following lemmas, which are proved in Section 6 at the end of this article.

Lemma 1. *Suppose that Conditions K, F, R and T are satisfied and set*

$$\tilde{f}(z) = \frac{1}{N}\sum_{j=1}^{n} \delta_j k_b(z - \varepsilon_j).$$

Then, as $nb^\alpha \to \infty$ for some $\alpha \geq 7/2$ and $b \to 0$, the following properties hold:

$$\|\hat{f} - \tilde{f} - f'\varrho^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2}), \tag{2.2}$$

$$\|\tilde{f} - f_b\|_2 = O_p((nb)^{-1/2}), \tag{2.3}$$

$$\|f_b - f\|_2 = o(b^{3/2}), \tag{2.4}$$

with $f_b = f * k_b$ differentiable with derivative $f_b' = f' * k_b$ which is bounded and Lipschitz.

Lemma 2. *Suppose that Conditions K, G, R and T are satisfied and set*

$$\tilde{g}(z) = \frac{1}{n}\sum_{j=1}^{n} k_b\{z - r_\vartheta(X_j)\}, \qquad \bar{\Gamma}_1(z) = E[k_b'\{z - r_\vartheta(X)\}\dot{r}_\vartheta(X)].$$

Let $nb^\alpha \to \infty$ for some $\alpha \geq 7/2$ and $b \to 0$. Then, with $g_b = g * k_b$, the following properties hold:

$$\|\hat{g} - \tilde{g} + \bar{\Gamma}_1^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2}), \tag{2.5}$$

$$\|\bar{\Gamma}_1\|_2 = O(b^{-1}), \tag{2.6}$$

$$\|\tilde{g} - g_b\|_2 = O_p((nb)^{-1/2}), \tag{2.7}$$

$$\|g_b - g\|_2 = O(b^{1/2}), \quad \|g_b - g\|_1 = O(b), \quad \|g_b'\|_1 = O(1). \tag{2.8}$$

Lemma 3. *Assume that Conditions F and G are satisfied and let $n \to \infty$ and $b \to 0$. Then we have, for any kernel k,*

$$\|A_i * k_b - A_i\|_\infty = o_p(n^{-1/2}) \quad \text{for } i = 1, 2,$$

and $A_1 * k_b$ and A_2 belong to $C_0(\mathbb{R})$ and converge in distribution in $C_0(\mathbb{R})$ to a Gaussian process.

Theorem 1. *Assume that Conditions K, F, G, R and T are satisfied. Then, as $nb^\alpha \to \infty$ for some $\alpha \geq 7/2$ and $nb^4 \to 0$, the following expansion holds,*

$$\|\hat{q} - q - A_1 - A_2 - (q'\varrho + \dot{q})^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2}).$$

Proof. In order to prove the statement we write

$$\hat{q} = \hat{f} * \hat{g} = f_b * g_b + (\hat{f} - f_b) * g_b + f_b * (\hat{g} - g_b) + (\hat{f} - f_b) * (\hat{g} - g_b).$$

We begin with the last term on the right-hand side. In order to show that it is asymptotically negligible, we will use the properties given in Lemmas 1 and 2. Lemma 2 yields $\|\hat{g} - g_b\|_2 = O_p((n^{-1/2}b^{-1})$. In view of the inequality $\|(\hat{f} - f_b) * (\hat{g} - g_b)\|_\infty \leq \|\hat{f} - f_b\|_2 \|\hat{g} - g_b\|_2$ and since $nb^3 \to \infty$, we obtain the rate

$$\|(\hat{f}-f_b)*(\hat{g}-g_b)\|_\infty = O_p((nb)^{-1/2})O_p(n^{-1/2}b^{-1}) = O_p(n^{-1}b^{-3/2}) = o_p(n^{-1/2}).$$

Now consider $f_b * g_b = q * k_b * k_b$. The derivative q' is Lipschitz and k is of order two. Hence, by a standard argument, $\|q * k_b - q\|_\infty = O(b^2)$, which yields $\|(q * k_b - q) * k_b\|_\infty \leq \|q * k_b - q\|_\infty \|k_b\|_1 = O(b^2)$. This and $nb^4 \to 0$ give

$$\|f_b * g_b - q\|_\infty = O(b^2) = o_p(n^{-1/2}).$$

We conclude the proof by showing

$$\|(\hat{f} - f_b) * g_b - A_1 - q'\varrho^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2}), \tag{2.9}$$

$$\|f_b * (\hat{g} - g_b) - A_2 - \dot{q}^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2}). \tag{2.10}$$

In order to prove (2.9) we use (2.2) which yields

$$(\hat{f} - f_b) * g_b = (\tilde{f} - f_b) * g_b + f' * g_b\varrho^\top(\hat{\vartheta} - \vartheta) + T_1 * g_b$$

with $\|T_1\|_2 = o_p(n^{-1/2})$. Since g is bounded it has finite L_2 norm and $\|g_b - g\|_2 \to 0$. This gives $\|T_1 * g_b\|_\infty \leq \|T_1\|_2\|g_b\|_2 = o_p(n^{-1/2})$. A standard argument, using the fact that q' is Lipschitz, yields $\|f' * g_b - q'\|_\infty = \|q' * k_b - q'\|_\infty = O(b)$. Hence we have $\|f' * g_b\varrho^\top(\hat{\vartheta} - \vartheta) - q'\varrho^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2})$. It is easy to see that $(\tilde{f} - f_b) * g_b = A_1 * k_b * k_b$. It follows from Lemma 3 that $\|A_1 * k_b - A_1\|_\infty = o_p(n^{-1/2})$ and therefore $\|A_1 * k_b * k_b - A_1\|_\infty = o_p(n^{-1/2})$. This proves (2.9).

For the proof of (2.10) it suffices to study $f * (\hat{g} - g_b)$ instead of $f_b * (\hat{g} - g_b)$ since $\|f_b - f\|_2 = o(b^{3/2})$ by (2.4) and $\|\hat{g} - g_b\|_2 = O_p((n^{-1/2}b^{-1})$ as shown above. From Lemma 2 we have

$$f * (\hat{g} - g_b) = f * (\tilde{g} - g_b) - f * \bar{\Gamma}_1^\top(\hat{\vartheta} - \vartheta) + T_2 * f$$

with $\|T_2\|_2 = o_p(n^{-1/2})$ and thus $\|T_2 * f\|_\infty = o_p(n^{-1/2})$. Now use $f * \bar{\Gamma}_1 = -\dot{q} * k_b$ and $\|\dot{q} - \dot{q} * k_b\|_\infty \to 0$, which holds by the Lipschitz property of f', to obtain $\|f * \bar{\Gamma}_1^\top(\hat{\vartheta} - \vartheta) + \dot{q}^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2})$. Finally the identity $f * (\tilde{g} - g_b) = A_2 * k_b$ and Lemma 3 yield $\|A_2 * k_b - A_2\|_\infty = o_p(n^{-1/2})$. This completes the proof of (2.10). □

3. Expansion of the weighted estimator

We now consider the weighted density estimator which uses residual-based weights $\hat{w}_j$ constructed by adapting the empirical likelihood approach introduced by Owen (1988, 2001). The weights are given by

$$\hat{w}_j = \frac{1}{1 + \hat{\lambda}\delta_j\hat{\varepsilon}_j},$$

where the Lagrange multiplier $\hat{\lambda}$ solves

$$\sum_{j=1}^{n} \frac{\delta_j\hat{\varepsilon}_j}{1 + \hat{\lambda}\delta_j\hat{\varepsilon}_j} = 0$$

and satisfies $1 + \hat{\lambda}\delta_j\hat{\varepsilon}_j > 0$ for $j = 1, \ldots, n$. As shown by Owen (1988, 2001), $\hat{\lambda}$ exists and is unique on the event $\min_{1\le j\le n} \delta_j\hat{\varepsilon}_j < 0 < \max_{1\le j\le n} \delta_j\hat{\varepsilon}_j$. This event has probability tending to one. Off this event we set $\hat{\lambda} = 0$. It was shown in Müller (2009, Lemma 3.1) that $\max_{1\le j\le n} |\hat{w}_j - 1| = o_p(1)$ and that the Lagrange multiplier satisfies the expansion

$$\hat{\lambda} \;=\; \frac{1}{\sigma^2}\frac{1}{N}\sum_{j=1}^{n} \delta_j\varepsilon_j - \frac{1}{\sigma^2}\varrho^\top(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}) \;=\; O_p(n^{-1/2}). \quad (3.1)$$

In Theorem 2 we provide an asymptotic expansion for the weighted density estimator. The proof uses the above expansion and also Lemma 4 below, which is proved in Section 6. This also requires that

$$\int u^2\{|k(u)|^2 + |k'(u)|^2\}\, du < \infty. \quad (3.2)$$

Lemma 4. *Suppose that Conditions K, F, R, T and (3.2) are satisfied and let $\psi(z) = zf(z)$. Then, as $nb^3 \to \infty$ and $b \to 0$,*

$$\|\hat{f}_w - \hat{f} + \hat{\lambda}\psi\|_2 = o_p(n^{-1/2}), \quad (3.3)$$

and

$$\int z^2\{\hat{f}(z) - f(z)\}^2\, dz = o_p(1). \quad (3.4)$$

We now state the expansion for the weighted estimator. In addition to the notation from Lemma 4 and the previous section we introduce

$$\bar{q}(y) = E\{g(y - \varepsilon)\varepsilon\} = \psi * g(y), \qquad A_3(y) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \frac{\varepsilon_j}{\sigma^2}\bar{q}(y).$$

Theorem 2. *Suppose that the conditions from Theorem 1 and (3.2) are satisfied. Then the weighted response density estimator $\hat{q}_w$ has the expansion*

$$\|\hat{q}_w - q - A_1 - A_2 + A_3 - (q'\varrho + \dot{q} + \sigma^{-2}\bar{q}\varrho)^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2}).$$

Proof. In view of Theorem 1 it suffices to verify

$$\|\hat{q}_w - \hat{q} + A_3 - \sigma^{-2}\varrho^\top(\hat{\vartheta} - \vartheta)\bar{q}\|_\infty = o_p(n^{-1/2}).$$

This is implied by

$$\big\|\hat{\lambda}\bar{q} - \frac{1}{\sigma^2}\{\frac{1}{N}\sum_{j=1}^{n}\delta_j\varepsilon_j - \varrho^\top(\hat{\vartheta} - \vartheta)\}\bar{q}\big\|_\infty = o_p(n^{-1/2}), \tag{3.5}$$

and

$$\|\hat{q}_w - \hat{q} + \hat{\lambda}\bar{q}\|_\infty = o_p(n^{-1/2}). \tag{3.6}$$

Note that the function ψ introduced in Lemma 4 is square-integrable since f is bounded and has a second moment. As a convolution of two square-integrable functions, $\bar{q} = \psi * g$ belongs to $C_0(\mathbb{R})$ and is therefore bounded. Thus (3.1) implies (3.5). In view of the identity

$$\hat{q}_w - \hat{q} + \hat{\lambda}\bar{q} = \hat{\lambda}(\bar{q} - \psi * \hat{g}) + (\hat{f}_w - \hat{f} + \hat{\lambda}\psi) * \hat{g}$$

we obtain the bound

$$\|\hat{q}_w - \hat{q} + \hat{\lambda}\bar{q}\|_\infty \le |\hat{\lambda}|\|\psi\|_2\|\hat{g} - g\|_2 + \|\hat{f}_w - \hat{f} + \hat{\lambda}\psi\|_2\|\hat{g}\|_2.$$

Statement (3.6) now follows from equations (3.1) and (3.3) and from the fact that $g \in L_2$, which yields $\|g_b - g\|_2 \to 0$ and thus $\|\hat{g} - g\|_2 = o_p(1)$. □

For the efficiency proof in the next section we will need an expansion for $\hat{q}_w - q$ which is asymptotically equivalent to the one given in Theorem 2. More precisely, we will need the expansion with $N = \sum_{j=1}^{n}\delta_j$ replaced by $nE\delta$ in A_1 and A_3. That the two expansions are indeed asumptotically equivalent follows directly from Slutsky's theorem and the law of large number.

Corollary 1. *Suppose that the conditions from Theorem 2 are satisfied and let*

$$\tilde{A}_1(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{E\delta}\{g(y - \varepsilon_i) - q(y)\},$$
$$\tilde{A}_3(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{E\delta}\varepsilon_j\sigma^{-2}\bar{q}(y),$$

for $y \in \mathbb{R}$. Then $\hat{q}_w$ has the expansion

$$\|\hat{q}_w - q - \tilde{A}_1 - A_2 + \tilde{A}_3 - (q'\varrho + \dot{q} + \sigma^{-2}\bar{q}\varrho)^\top(\hat{\vartheta} - \vartheta)\|_\infty = o_p(n^{-1/2}).$$

4. Efficiency and main result

In order to derive the canonical gradient of $q(y)$, $y \in \mathbb{R}$, which characterizes the efficient influence function, one can build on results from Müller (2009) for

estimating $Eh(X,Y)$. In the following we will summarize the key arguments and refer to that article for details. In particular we will point out the main differences and show that the influence function in Theorem 2 is the efficient one, provided that $\hat{\vartheta}$ is efficient for ϑ. Write $\ell = -f'/f$ for the score function for location. By Müller (2009, Lemma 5.1), the efficient $\hat{\vartheta}$ is characterized by the canonical gradient $E\{(\delta\zeta\zeta^\top)^{-1}\}\delta\zeta$ with $\zeta = \zeta(X,Y)$ where

$$\zeta(x,y) = \{\dot{r}_\vartheta(x) - \varrho\}\ell\{y - r_\vartheta(x)\} + \varrho\frac{y - r_\vartheta(x)}{\sigma^2},$$

i.e. $\zeta = \{\dot{r}_\vartheta(X) - \varrho\}\ell(\varepsilon) + \sigma^{-2}\varrho\varepsilon$ with $\varrho = E\{\dot{r}_\vartheta(X)|\delta = 1\}$ as introduced earlier. In the following we assume that f has finite Fisher information for location, $E\ell^2(\varepsilon) < \infty$.

The joint distribution of (X,Y) is specified by the marginal distribution $M(dx)$ of X and the conditional distribution $Q(x,dy)$ of Y given $X = x$. A gradient γ for an arbitrary differentiable functional κ of the distribution of (X,Y) is characterized by

$$\lim_{n\to\infty} n^{1/2}\{\kappa(M_{nu}, Q_{nv}) - \kappa(M,Q)\} = E[\gamma(X,\delta Y,\delta)\{u(X) + \delta v(X,Y)\}]$$

for all $u \in U$ and $v \in V$, where M_{nu} and Q_{nv} are perturbations of M and Q,

$$\begin{aligned} M_{nu}(dx) &\doteq M(dx)\{1 + n^{-1/2}u(x)\}, \\ Q_{nv}(x,dy) &\doteq Q(x,dy)\{1 + n^{-1/2}v(x,y)\}. \end{aligned}$$

The derivatives u and v belong to the tangent space

$$T = \big\{u(X) : u \in U\big\} \oplus \big\{\delta v(X,Y) : v \in V\big\},$$

with U and V given below. The canonical gradient is a gradient that is an element of the tangent space, i.e. it is of the form

$$\gamma_*(X,\delta Y,\delta) = u_*(X) + \delta v_*(X,Y)$$

with the terms of the sum being projections onto the tangent space. As a gradient of κ, γ_* must satisfy the above characterization,

$$\begin{aligned} &\lim_{n\to\infty} n^{1/2}\{\kappa(M_{nu}, Q_{nv}) - \kappa(M,Q)\} \\ &= E\{u_*(X)u(X)\} + E\{\delta v_*(X,Y)v(X,Y)\}. \end{aligned} \tag{4.1}$$

Note that T is the tangent space that is relevant for estimating functionals of M and Q – such as $Eh(X,Y)$ and $q(y) = Ef\{y - r_\vartheta(X)\}$ – but not for functionals of the full joint distribution that also involve the conditional distribution $\pi(x)$ of the indicator variable δ given x (see Müller, 2009, for the complete specification).

The perturbed distributions $M_{nu}(dx)$ and Q_{nv} must both be probability distributions, i.e. integrate to one. Since we do not assume any model for the distribution of X this yields $U = L_{2,0}(M) = \{u \in L_2(M) : \int u\,dM = 0\}$. The nonlinear regression model constitutes a constraint for the conditional

distribution, $Q(x, dy) = f\{y - r_\vartheta(x)\}dy$. Perturbing f and ϑ gives $Q_{nv}(x, dy) = Q_{nst}(x, dy) = f_{ns}\{y - r_{\vartheta_{nt}}(x)\}\, dy$ with $\vartheta_{nt} = \vartheta + n^{-1/2}t$, $t \in \mathbb{R}^p$, $f_{ns}(y) = f(y)\{1 + n^{-1/2}s(y)\}$ with $s \in S$ where

$$S = \big\{s \in L_2(F) : \int s(y)f(y)\, dy = 0, \int ys(y)f(y)\, dy = 0\big\}.$$

Note that f_{ns} must be a mean zero probability density which explains the two constraints on $s \in S$. As in Müller (2009) we obtain

$$V = \big\{v(x, y) = s\{y - r_\vartheta(x)\} + \ell\{y - r_\vartheta(x)\}\dot{r}_\vartheta(x)^\top t : s \in S, t \in \mathbb{R}^p\big\}.$$

That summarizes the results we shall need from that article. We go on to consider a specific form for $\kappa(M, Q)$ in (4.1), namely $\kappa(M, Q) = q(y) = \int f\{y - r_\vartheta(x)\}M(dx)$. Using again the approximation of $f_{ns}\{y - r_{\vartheta_{nt}}(x)\}$ from Müller (2009) and the formula for the perturbation M_{nu} for M from above, the left-hand side of (4.1) is

$$\begin{aligned}
&\int f_{ns}\{y - r_{\vartheta_{nt}}(x)\}M_{nu}\,(dx) - \int f\{y - r_\vartheta(x)\}\, M(dx)\\
\doteq\ & n^{-1/2}\int f\{y - r_\vartheta(x)\}\{u(x) + v(x, y)\}\, M(dx)\\
=\ & n^{-1/2}\Big(E\big[f\{y - r_\vartheta(X)\}u(X)\big] + E\big[f\{y - r_\vartheta(X)\}v(X, y)\big]\Big)
\end{aligned}$$

with $v \in V$ specified above. Setting $v = 0$ in (4.1) therefore yields

$$E[f\{y - r_\vartheta(X)\}u(X)] = E\{u_*(X)u(X)\}.$$

Since u_* must be centered this is immediately solved by

$$u_*(X) = f\{y - r_\vartheta(X)\} - Ef\{y - r_\vartheta(X)\} = f\{y - r_\vartheta(X)\} - q(y). \tag{4.2}$$

In order to find v_* we set $u = 0$ in (4.1) and obtain

$$E\big[f\{y - r_\vartheta(X)\}v(X, y)\big] \ = \ E\{\delta v_*(X, Y)v(X, Y)\}. \tag{4.3}$$

Analogously to Müller (2009), any element in V can be written $v(x, y) = s\{y - r_\vartheta(x)\} + \zeta(x, y)^\top t$ for some $t \in \mathbb{R}^p$ and $s \in S$, with $\zeta(x, y)$ defined at the beginning of the section. For the canonical gradient we write

$$v_*(X, Y) = s_*(\varepsilon) + \zeta(X, Y)^\top t_*,$$

with $s_* \in S$ and $t_* \in \mathbb{R}^p$ to be determined. With this notation equation (4.3) states

$$\begin{aligned}
&E\big(f\{y - r_\vartheta(X)\}[s\{y - r_\vartheta(X)\} + \zeta(X, y)^\top t]\big)\\
&= E\big[\delta\{s_*(\varepsilon) + \zeta(X, Y)^\top t_*\}\{s(\varepsilon) + \zeta(X, Y)^\top t\}\big].
\end{aligned}$$

In order to determine $s_* \in S$ we set $t = 0$ and use the fact that $\delta\zeta(X,Y)$ is orthogonal to S which yields

$$E\big[f\{y - r_\vartheta(X)\}s\{y - r_\vartheta(X)\}\big] \quad = \quad E\{\delta s_*(\varepsilon)s(\varepsilon)\}. \tag{4.4}$$

Analogously, setting $s = 0$, we obtain a characterization for t_*,

$$E[f\{y - r_\vartheta(X)\}\zeta(X,y)^\top t] \quad = \quad E\{\delta\zeta(X,Y)^\top t_*\zeta(X,Y)^\top t\}. \tag{4.5}$$

Below we will check that $(E\delta)^{-1}g(y-\varepsilon)$ solves (4.4). However, since s_* must be in S, we choose a suitable modification, namely

$$s_*(\varepsilon) = \frac{1}{E\delta}[g(y-\varepsilon) - E\{g(y-\varepsilon)\} - \frac{\varepsilon}{\sigma^2}E\{\varepsilon g(y-\varepsilon)\}]. \tag{4.6}$$

It is easy to see that $s_* \in S$. In particular, s_* solves (4.4):

$$\begin{aligned} E\{\delta s_*(\varepsilon)s(\varepsilon)\} &= E\big(\big[g(y-\varepsilon) - E\{g(y-\varepsilon)\} - \frac{\varepsilon}{\sigma^2}E\{\varepsilon g(y-\varepsilon)\}\big]s(\varepsilon)\big) \\ &= E\{g(y-\varepsilon)s(\varepsilon)\} = \int g(y-u)s(u)f(u)\,du \\ &= E\big[f\{y - r_\vartheta(X)\}s\{y - r_\vartheta(X)\}\big]. \end{aligned}$$

Now consider (4.5) and suppose that $E\{\delta\zeta(X,Y)\zeta(X,Y)^\top\}$ is invertible. It is easy to see that

$$t_*^\top \quad = \quad E[f\{y - r_\vartheta(X)\}\zeta(X,y)^\top]E\{\delta\zeta(X,Y)\zeta(X,Y)^\top\}^{-1} \tag{4.7}$$

solves (4.5). The first expectation equals the factor before $(\hat\vartheta - \vartheta)$ in Theorem 2,

$$E[f\{y - r_\vartheta(X)\}\zeta(X,y)] = \dot q(y) + q'(y)\varrho + \sigma^{-2}\bar q(y)\varrho, \tag{4.8}$$

with $\bar q(y) = E\{g(y-\varepsilon)\varepsilon\}$. To see this consider

$$\begin{aligned} &E[f\{y - r_\vartheta(X)\}\zeta(X,y)] \\ &= E[f\{y - r_\vartheta(X)\}\dot r_\vartheta(X)\ell\{y - r_\vartheta(X)\}] - E[f\{y - r_\vartheta(X)\}\varrho\ell\{y - r_\vartheta(X)\}] \\ &\quad + E[f\{y - r_\vartheta(X)\}\varrho\frac{y - r_\vartheta(X)}{\sigma^2}]. \end{aligned}$$

Since $\ell = -f'/f$ the first term on the right-hand side is

$$-\int f'\{y - r_\vartheta(x)\}\dot r_\vartheta(x)M(dx) = \dot q(y),$$

The same argument and $q'(y) = f' * g = E\{f'\{y - r_\vartheta(X)\}$ shows that the second term equals $q'(y)\varrho$. Apart from the constant vector ϱ/σ^2, the third term is

$$\int f(y-u)(y-u)g(u)\,du = \int f(z)zg(y-z)\,dz = E\{g(y-\varepsilon)\varepsilon\},$$

which is indeed the desired $\bar{q}(y)$ from (4.8). Hence (4.8) holds, and (4.7) can be rewritten as

$$t_*^\top = \{\dot{q}(y) + q'(y)\varrho + \sigma^{-2}\bar{q}(y)\varrho\}^\top E(\delta\zeta\zeta^\top)^{-1} \tag{4.9}$$

We are now in a position to specify the canonical gradient, for which we have derived the form

$$\gamma_*(X, \delta Y, \delta) = u_*(X) + \delta v_*(X, Y) = u_*(X) + \delta\{s_*(\varepsilon) + \zeta(X, Y)^\top t_*\},$$

with $u_*(X)$, $s_*(\varepsilon)$ and t_* given in equations (4.2), (4.6) and (4.9). This result is summarized in the following lemma. Note that we now additionally require that $E(\delta\zeta\zeta^\top)$ is invertible, where $E(\delta\zeta\zeta^\top)$ involves the covariance matrix of $\delta\dot{r}_\vartheta(X)$ and the Fisher information $E\ell^2(\varepsilon)$. We will use that $q = f * g$, i.e. $E\{g(y-\varepsilon)\} = q(y)$ in (4.6).

Lemma 5. *Consider $\zeta = \zeta(X, Y) = \{\dot{r}_\vartheta(X) - \varrho\}\ell(\varepsilon) + \sigma^{-2}\varrho\varepsilon$ with $\ell = -f'/f$. Suppose additionally to the conditions given in Section 2 that $E(\delta\zeta\zeta^\top)$ is invertible and that $E\ell^2(\varepsilon) < \infty$. Then the canonical gradient of the response density $q(y)$, $y \in \mathbb{R}$, is*

$$f\{y - r_\vartheta(X)\} - q(y) + \frac{\delta}{E\delta}\{g(y-\varepsilon) - q(y) - \frac{\varepsilon}{\sigma^2}\bar{q}(y)\} + \{\dot{q}(y) + q'(y)\varrho + \sigma^{-2}\bar{q}(y)\varrho\}^\top E(\delta\zeta\zeta^\top)^{-1}\delta\zeta.$$

Comparing the canonical gradient and the expansion of the weighted estimator in Corollary 1, and taking into account that an efficient estimator $\hat{\vartheta}$ of ϑ has influence function $E(\delta\zeta\zeta^\top)^{-1}\delta\zeta$, we see that our weighted response density estimator is efficient for $q(y)$ if $\hat{\vartheta}$ is efficient. Note that estimators that are efficient in the Hájek–Le Cam sense are asymptotically normal by construction. The lower variance bound is given by the second moment of the canonical gradient, i.e. by $E\gamma_*^2$. We summarize our results in the following main theorem.

Theorem 3. *Assume that Conditions K, F, G, R and (3.2) are satisfied and that the covariance matrices of $\dot{r}_\vartheta(X)$ and of $\delta\dot{r}_\vartheta(X)$ are invertible. Also assume that f has finite Fisher information for location, $E\ell^2(\varepsilon) < \infty$, where $\ell = -f'/f$. Let $\hat{\vartheta}$ be an asymptotically linear estimator of ϑ with influence function $E(\delta\zeta\zeta^\top)^{-1}\delta\zeta$, where $\zeta = \{\dot{r}_\vartheta(X) - \varrho\}\ell(\varepsilon) + \sigma^{-2}\varrho\varepsilon$. Consider the weighted response density estimator (1.4) with bandwidth b satisfying $nb^\alpha \to \infty$ for some $\alpha \geq 7/2$ and $nb^4 \to 0$. Then, uniformly in $y \in \mathbb{R}$, the difference $\hat{q}_w(y) - q(y)$ has the expansion*

$$\frac{1}{n}\sum_{i=1}^n \Big[f\{y - r_\vartheta(X_i)\} - q(y) + \frac{\delta_i}{E\delta}\big\{g(y-\varepsilon_i) - q(y) - \frac{\varepsilon_i}{\sigma^2}\bar{q}(y)\big\} + \{\dot{q}(y) + q'(y)\varrho + \sigma^{-2}\bar{q}(y)\varrho\}^\top E(\delta\zeta\zeta^\top)^{-1}\delta_i\{\dot{r}_\vartheta(X_i) - \varrho\}\,\ell(\varepsilon_i) + \varrho\frac{\varepsilon_i}{\sigma^2}\Big] + o_p(n^{-1/2}).$$

In particular, $\hat{q}_w(y)$ is an efficient estimator of $q(y)$, for every $y \in \mathbb{R}$. Moreover, the process $n^{1/2}(\hat{q}_w - q)$ converges in distribution in the space $C_0(\mathbb{R})$ to a Gaussian process.

The last statement holds in view of Lemma 3 where we show, in particular, that the critical term $A_1 * k_b$ (that involves g, which is only assumed to be of bounded variation) converges in $C_0(\mathbb{R})$.

Our proposed density estimator requires plugging in an efficient estimator $\hat{\vartheta}$ for ϑ in order to be efficient, or at least a $n^{1/2}$-consistent estimator $\hat{\vartheta}$ to achieve the parametric rate of convergence. For the construction of an efficient estimator for ϑ we refer to the discussion in Müller (2009), Section 5, where a "one-step improvement" approach involving a preliminary estimator and an estimator of the influence function is sketched. The idea goes back to Schick (1993) and Forrester et al. (2003) who consider regression models with completely observed data. Since the construction can be quite involved, a simple alternative estimator, with regard to practical applications, would be the ordinary least squares estimator, i.e. the minimizer of the sum of observed squared residuals $\sum_{i=1}^n \delta_i\{Y_i - r_\vartheta(X_i)\}^2$ with respect to ϑ. The estimator solves $\sum_{i=1}^n \delta_i \dot{r}_\vartheta(X_i)\{Y_i - r_\vartheta(X_i)\} = 0$, and therefore has the asymptotic expansion

$$n^{1/2}(\hat{\vartheta} - \vartheta) \;\doteq\; n^{-1/2} \sum_{i=1}^n E\{\delta \dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\}^{-1} \delta_i \dot{r}_\vartheta(X_i)\{Y_i - r_\vartheta(X_i)\}.$$

This i.i.d. representation as a sum of square-integrable variables immediately yields asymptotic normality and $n^{1/2}$-consistency. If the errors happen to be normally distributed, the ordinary least squares estimator can be shown to be efficient: it satisfies the characterization of an efficient estimator $\hat{\vartheta}$ for ϑ from Müller (2009), Lemma 5.1,

$$n^{1/2}(\hat{\vartheta} - \vartheta) \;\doteq\; n^{-1/2} \sum_{i=1}^n E(\delta\zeta\zeta^\top)^{-1} \delta_i \big[\{\dot{r}_\vartheta(X_i) - \varrho\}\, \ell(\varepsilon_i) + \varrho \frac{\varepsilon_i}{\sigma^2}\big].$$

To see this, use the fact that for a normal error density f the score function simplifies to $\ell(\varepsilon) = -f'(\varepsilon)/f(\varepsilon) = \varepsilon/\sigma^2$, and that $\zeta = \{\dot{r}_\vartheta(X) - \varrho\}\ell(\varepsilon) + \sigma^{-2}\varrho\varepsilon = \dot{r}_\vartheta(X)\varepsilon/\sigma^2$ in this case.

5. Simulations and discussion

The proposed response density estimator $\hat{q}_w(y)$ from equation (1.4) is efficient and will therefore, in general, outperform its unweighted version $\hat{q}(y)$ and the simple estimator $N^{-1}\sum_{i=1}^n \delta_i k_b(y - Y_i)$ given in equations (1.2) and (1.1). This is supported by simulations reported in this section (see Tables 1 to 3).

Let us address some computational issues first. Remember that the response density is a convolution of (unknown) densities, $q(y) = f * g(y)$. Our estimator (1.4) uses density estimators for f and g and can, in particular, also be written in the simpler form (1.3),

$$\hat{q}_w(y) = \frac{1}{n}\sum_{i=1}^n \hat{f}_w\{y - r_{\hat{\vartheta}}(X_j)\} = \frac{1}{n}\sum_{i=1}^n \frac{1}{N}\sum_j \hat{w}_j \delta_j\, k_b^*\{y - r_{\hat{\vartheta}}(X_i) - \hat{\varepsilon}_j\},$$

where $k_b^*(y) = k * k(y/b)/b$ is a convolution of two kernels. Since normal distributions are closed under convolutions, i.e. every normal density can be expressed as a convolution of normal densities, one can, for example, take the standard normal density for k^*, which is what we have done for our illustrations.

For the simulations we also chose the error density f to be standard normal. This allows us, by the discussion at the end of the previous section, to use the ordinary least squares estimator as an efficient plug-in estimator for ϑ.

We focus on partially missing responses with $\pi(X) = P(\delta = 1|X) = 1/(1 + e^{-X})$, where X is a one-dimensional covariate which we generated from a uniform distribution with support $(-1, 1)$. This means that $\pi(X)$ takes values between 0.27 and 0.73, and that, on average, half of the responses are missing.

Table 1: Nonlinear regression: simulated mean squared errors

	Estimator	$b = 0.1$	$b = 0.3$	$b = 0.5$	$b = 0.7$	$b = 0.9$	$b = 1.1$
AMSE	U	0.00097	0.00062	**0.00055**	0.00077	0.00133	0.00223
	W	0.00086	0.00052	**0.00046**	0.00070	0.00127	0.00219
	S	0.00849	0.00262	0.00163	**0.00151**	0.00189	0.00287
$y = -1$:	U	0.00072	0.00046	**0.00045**	0.00070	0.00113	0.00164
	W	0.00061	**0.00035**	0.00035	0.00060	0.00106	0.00157
	S	0.00276	0.00091	0.00048	**0.00038**	0.00051	0.00080
$y = 0$:	U	0.00175	0.00099	0.00063	**0.00048**	0.00051	0.00078
	W	0.00146	0.00074	0.00043	**0.00033**	0.00042	0.00071
	S	0.01049	0.00358	0.00234	0.00189	**0.00178**	0.00202
$y = 1$:	U	0.00190	**0.00117**	0.00128	0.00237	0.00450	0.00755
	W	0.00190	**0.00117**	0.00128	0.00237	0.00451	0.00756
	S	0.01617	0.00454	**0.00308**	0.00366	0.00579	0.00866
$y = 2$:	U	0.00100	0.00066	**0.00053**	0.00058	0.00091	0.00159
	W	0.00079	0.00047	**0.00037**	0.00047	0.00084	0.00156
	S	0.01500	0.00397	0.00170	0.00077	**0.00056**	0.00098
$y = 3$:	U	0.00059	0.00047	**0.00043**	0.00047	0.00057	0.00070
	W	0.00046	0.00034	**0.00031**	0.00035	0.00048	0.00062
	S	0.00775	0.00296	0.00218	0.00201	0.00198	**0.00194**
$y = 4$:	U	0.00018	**0.00017**	0.00022	0.00058	0.00063	0.00108
	W	0.00017	**0.00016**	0.00020	0.00047	0.00061	0.00105
	S	0.00194	0.00077	**0.00069**	0.00077	0.00136	0.00201

The figures are the simulated mean squared errors of estimators of $q(y)$ at $y = -1, 0, \ldots, 4$ for different bandwidths b. The top panel shows the simulated *average* mean squared error, which is computed for y-values on a grid with grid size 0.05, ranging from -1.9 to 4.45. The regression function is $r_\vartheta(x) = e^{\vartheta x} = e^x$, i.e. $\vartheta = 1$, the covariates are from the univariate uniform distribution on $(-1, 1)$, and the sample size is $n = 100$. The estimators are the unweighted version "U" of the proposed estimator, the proposed weighted estimator "W", and the simple estimator "S" from equation (1.1). Bold values are the minimum mean squared errors for each estimator.

In Table 1 we consider an exponential regression function and compare the simulated mean squared errors of the three estimators pointwise, and also the respective average mean squared errors computed on a partitioning of the interval $[-1.9, 4.45]$, which is chosen such that $q(y) < 0.01$ for y outside the interval. Note that the average mean squared error is proportional to a Riemann sum, or, more precisely, to an approximation of the mean integrated square error

("MISE") $E[\int\{\hat{q}(y) - q(y)\}^2\,dy]$. Both the weighted and the unweighted versions of our proposed estimator are clearly better than the simple estimator which only uses those data located in a neighborhood of y.

The behavior of the estimator depends strongly on the behavior of the regression function. So far we have only studied a regression function which has a simple form, namely a univariate monotonic regression function. Here the proposed approach works as anticipated. Let us now identify and discuss irregular cases where the $n^{1/2}$ rate (and efficiency) cannot be obtained.

An extreme case is given if the regression function is constant, $r_\vartheta(X) = \vartheta$. In this case the response density is just a shift of the error density, $q(y) = \int f(y-\vartheta)\,M(dx) = f(y-\vartheta)$, and the density g of $r_\vartheta(X) = \vartheta$ is just the point mass in ϑ, $P\{r_\vartheta(X) = \vartheta\} = 1$. The proposed estimator $\hat{q}_w(y)$ with $r_{\hat{\vartheta}}(X) = \hat{\vartheta}$ is therefore just a (weighted) estimator of the shifted error density,

$$\begin{aligned}\hat{q}_w(y) &= \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_j\delta_j\,k_b^*(y-\hat{\vartheta}-\hat{\varepsilon}_j) = \frac{1}{N}\sum_{j=1}^{n}\hat{w}_j\delta_j\,k_b^*(y-\hat{\vartheta}-\hat{\varepsilon}_j)\\ &= \frac{1}{N}\sum_{j=1}^{n}\hat{w}_j\delta_j k_b^*(y-Y_j).\end{aligned}$$

Note that this is a weighted version of the simple kernel density estimator from equation (1.1). In particular, weighting has no effect here, i.e. $\hat{q}_w(y)$ and $\hat{q}(y)$ (and therefore all three estimators) are asymptotically equivalent. This is explained by the different rates of convergence: $\hat{q}_w(y)$ and $\hat{q}(y)$ both converge to $q(y)$ at a rate slower than the parametric rate $n^{1/2}$, whereas the difference between $\hat{q}_w(y)$ and $\hat{q}(y)$ is negligible since it has the order $n^{-1/2}$.

A related case is given when the regression function is a step function, with possible values $\vartheta_1,\ldots,\vartheta_p$. Then the response density is a discrete mixture of shifts of the error density, with weights $P\{r_\vartheta(X) = \vartheta_i\}$. The weights can be estimated empirically, and the response density can be estimated by a mixture of shifts of error density estimators. The rate is similar to that of the usual nonparametric (kernel) estimators for the response density.

We now consider the situation where r_ϑ has no constant parts over intervals, so that the response density is a continuous convolution of the error density. We need square-integrability, i.e., since $q(y) = E\{g(y-\varepsilon)\}$, that $g(y-\varepsilon)$ has a finite second moment, which is often violated.

Suppose first that the covariate is one-dimensional with density m and that the derivative of the regression function is bounded away from zero, $|r'_\vartheta| \geq \eta > 0$. Applying the change of variable theorem to the density g of $r_\vartheta(X)$, the second moment of the random variable $g(y-\varepsilon)$ is

$$\begin{aligned}E\{g^2(y-\varepsilon)\} &= \int \frac{1}{[r'_\vartheta\{r_\vartheta^{-1}(y-z)\}]^2} m^2\{r_\vartheta^{-1}(y-z)\}f(z)\,dz\\ &\leq \frac{1}{\eta^2}Em^2\{r_\vartheta^{-1}(y-\varepsilon)\},\end{aligned}$$

which is finite if $Em^2\{r_\vartheta^{-1}(y-\varepsilon)\} < \infty$. This holds, for example, if the covariate density m is bounded. Then $\hat{q}_w(y)$ has the expansion given in Theorem

2 (for some $n^{1/2}$-consistent $\hat{\vartheta}$), and $\hat{q}_w(y)$ converges at the $n^{1/2}$ rate and is asymptotically normal.

It is possible that r_ϑ is continuous but only piecewise invertible, with a derivative that vanishes (or does not exist) at certain points. For example, $r_\vartheta(x+t) - r_\vartheta(x)$ may behave approximately like ct^a or $c|t|^a$ for small t and certain values of x. Simple examples are $r_\vartheta(x) = |x|^{1/2}$, $|x|$, x^2 and x^3 at $x = 0$. Again we can apply the change of variable theorem, now to the monotonic parts of r_ϑ, and compute the density g of $r_\vartheta(X)$, which becomes a sum of densities. The behavior of $g(y-\varepsilon)$ depends crucially on the points where r'_ϑ is zero (or does not exist), i.e. on the denominator $[r'_\vartheta\{r_\vartheta^{-1}(y-z)\}]^2$ in the change of variable formula at those points. For example, for a quadratic regression function $r_\vartheta(x) = x^2$ we have $[r'_\vartheta\{r_\vartheta^{-1}(x)\}]^{-2} = 1/(4x)$, i.e. $E\{g^2(y-\varepsilon)\}$ cannot be finite since $\lim_{\epsilon\to 0}\int_\epsilon^1 1/x\,dx$ diverges. Due to the fact that $\lim_{\epsilon\to 0}\int_\epsilon^1 1/x^p\,dx$ converges for $p < 1$ and diverges for $p \geq 1$, we conclude that if $r_\vartheta(x+t) - r_\vartheta(x)$ behaves like ct^a or $c|t|^a$ for one (or more) x then, if $a \geq 2$, the random variable $g(y-\varepsilon)$ does not have a finite second moment. Therefore, $\hat{q}_w(y)$ converges at a slower rate than $n^{1/2}$, depending on a. The slower rate for certain types of regression functions follows as in Schick and Wefelmeyer (2009a), who derive convergence rates for estimators of the density of $|X_1|^a + |X_2|^a$ with i.i.d. observations $X_1, \ldots, X_n$. See also Schick and Wefelmeyer (2009b), who consider density estimators for sums of squared observations. They obtain asymptotic normality (pointwise) with rate $(\log n/n)^{1/2}$, which is slightly slower than $n^{1/2}$.

Table 2: Comparison of regression functions

Estimator		n	$b=0.1$	$b=0.3$	$b=0.5$	$b=0.7$	$b=0.9$	$b=1.1$
(a)	U	50	22.412	14.029	**11.140**	12.269	17.592	26.328
		100	9.730	7.219	**6.748**	9.527	15.695	25.145
		500	1.796	**1.683**	2.894	6.716	13.725	23.443
	W	50	20.875	12.549	**9.880**	11.238	16.792	25.722
		100	8.325	5.916	**5.628**	8.639	14.998	24.629
		500	1.297	**1.222**	2.479	6.398	13.474	23.262
	S	50	163.618	45.057	**23.911**	18.929	21.576	28.921
		100	85.412	25.702	14.969	**14.350**	18.701	27.260
		500	20.281	8.485	**7.041**	9.689	15.892	24.974
(b)	U	50	60.200	28.941	**19.419**	20.329	28.698	42.030
		100	26.937	14.500	**11.470**	15.577	25.575	40.018
		500	4.905	**3.248**	4.953	11.335	22.563	38.168
	W	50	58.773	27.308	**18.032**	19.202	27.815	41.334
		100	25.711	13.321	**10.443**	14.768	24.946	39.544
		500	4.559	**2.874**	4.690	11.084	22.410	38.051
	S	50	175.453	46.209	24.625	**22.319**	29.450	42.446
		100	87.330	23.360	**14.062**	16.554	26.100	40.315
		500	17.314	**4.995**	5.545	11.532	22.717	38.232

The figures are simulated average mean squared errors multiplied by $10,000$, in a set-up similar to Table 1. The regression function is linear in panel (a), $r_\vartheta(X) = \vartheta X$, and the average is computed on the interval $[-3.1, 3.1]$ (outside which $q < 0.01$). In panel (b) the regression function is quadratic, $r_\vartheta(X) = \vartheta X^2$, and the interval on which the average is computed is $[-2.5, 3.2]$. The grid-size is 0.1 for $n = 50$ and $n = 100$, and 0.2 for $n = 500$. In all simulations $\vartheta = 1$.

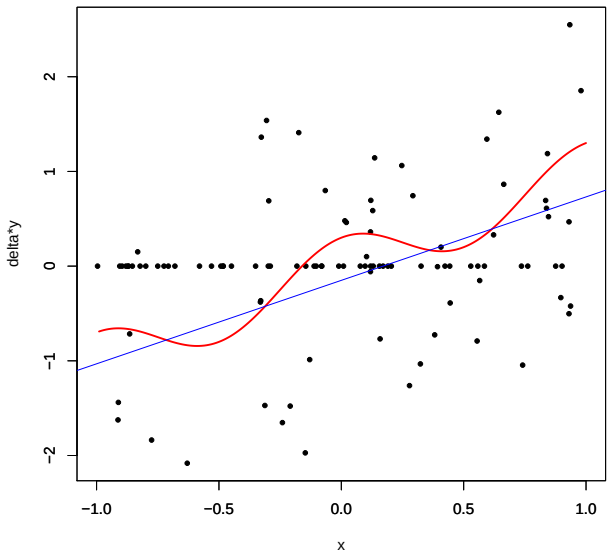

Figure 1: A typical data set with about 50% of the responses missing. The regression function is $X + 0.3\cos(2\pi X)$. The line is the fitted linear regression function $\hat{\vartheta}X$, assuming (erroneously) a linear relationship between Y and X.

To illustrate this, in Table 2 (a) we consider a linear regression function, where the response density can be estimated at the parametric rate, and in panel (b) a quadratic regression function, where we expect a slower rate of convergence. That our estimator in (b) converges at a slower rate than our estimator in (a) is indeed supported by the simulations. Consider, for example, the minimum mean squared errors of our weighted estimator in (a) and (b) when we increase the sample size by the factor 10: for $n = 50$ the minimum average mean squared error in (a) is about 8.1 times as large as for $n = 500$. The improvement is less striking in (b): for $n = 50$ the minimum average mean squared error is about 6.3 times as large as for $n = 500$. Another consequence of the slow rate of convergence of $\hat{q}_w(y)$ and $\hat{q}(y)$ in (b) is that the two estimators are asymptotically equivalent, as explained earlier. The figures for $\hat{q}_w(y)$ and $\hat{q}(y)$ in Table 2 (b) are indeed similar. However, even for $n = 500$ the weighted estimator still performs noticeably better than its unweighted version. It therefore seems reasonable to always work with weights.

In order to study the performance of our estimator when the regression function is not correctly specified, which is typically the case in applications when we assume a *model*, we consider in Table 3 a sinusoidal regression function with a linear trend (see Figure 1), but fit a linear regression line into the data. Perhaps surprisingly, even though our estimator is biased the mean squared errors in Table 3 are very similar to those given in Table 2 (a) where we consider a correctly specified linear regression function. In particular, our estimator again

outperforms the simple estimator, even though that estimator is robust since it uses only the responses Y. This behavior is explained by the fact that in our illustration the mean squared error is dominated by the variance, whereas the bias is very small. Since we are interested in estimating the integral $q(y) = \int f\{y - r_\vartheta(x)\}\, M(dx)$, and not the regression function r_ϑ itself, we expect that our estimator will, in general, work well in the finite sample situation, even if the regression function is not carefully specified – a simple linear regression model will often suffice.

Table 3: Misspecified regression function

Estimator	n	$b = 0.1$	$b = 0.3$	$b = 0.5$	$b = 0.7$	$b = 0.9$	$b = 1.1$
U	50	22.089	14.374	**11.017**	12.195	16.995	25.202
	100	9.473	7.063	**6.834**	9.261	15.306	24.066
W	50	20.562	12.847	**9.782**	11.161	16.207	24.605
	100	8.095	5.775	**5.697**	8.372	14.590	23.524
S	50	165.832	45.018	24.258	**18.939**	21.376	27.903
	100	84.187	25.000	15.407	**13.852**	18.455	26.220

The figures are simulated average mean squared errors multiplied by $10,000$ in a setting related to Table 2(a): again we are fitting a linear regression function $r_\vartheta(X) = \vartheta X$ into the data. However, and in contrast to Table 2(a) where the regression function is correctly specified, the data are generated from a nonlinear regression model, $r_\vartheta(X) = X + 0.3\cos(2\pi X)$, in order to study an example of a misspecified regression function. See Figure 1 for an illustration.

To conclude, let us discuss the impact of the regression function on the behavior of the density estimator in the multivariate covariate case. Here similar but more complex results hold. The density g of the regression function $r_\vartheta(X)$, with covariate vector $X = (X_1, \ldots, X_p)$, is obtained by applying the change of variable formula as follows: first introduce $h(x) = \{r_\vartheta(x), x_2, \ldots, x_p\} = (u_1, u_2, \ldots, u_p)$. If h is one-to-one (and therefore invertible) with non-vanishing Jacobian, i.e. $|\det \partial h(x)/\partial x| \neq 0$, then the density of $h(X)$, $f_{h(X)}(u)$, at $u = (u_1, \ldots, u_p)$ is

$$f_{h(X)}(u) = \Big| \det \frac{\partial h^{-1}(u)}{\partial u} \Big| m\{h^{-1}(u)\}.$$

Due to the simple form of h, invertibility just means that $u_1 = r_\vartheta(x)$ can be solved with respect to x_1. In particular, the Jacobian is simply $\partial r_\vartheta(x)/(\partial x_1)$. The assumptions for change of variable are therefore met if one of the partial derivatives never vanishes, say if $\partial r_\vartheta(x)/(\partial x_1) \neq 0$, and if $u_1 = r_\vartheta(x)$ can be solved with respect to x_1.

The density g of $r_\vartheta(X)$ is the marginal density

$$g(u) = \int \Big| \det \frac{\partial h^{-1}(u)}{\partial u} \Big| m\{h^{-1}(u)\}\, d(u_2, \ldots, u_p).$$

As in the univariate case it is clear that $g(y - \varepsilon)$ has a finite second moment if the marginal density of X_1 is bounded. Now suppose that h is not invertible on the support of X, but there is a partition $\{A_1, \ldots, A_m\}$ of the support where h

is invertible on each of the A_j. Then, if on each of the A_j's the Jacobian does not vanish,

$$f_{h(X)}(u) = \sum_{j=1}^{m} \Big| \det \frac{\partial h_j^{-1}(u)}{\partial u} \Big| m\{h_j^{-1}(u)\}$$

where h_j^{-1} is the inverse of h on A_j. The density of $r_\vartheta(X)$ is computed, analogously to the above, as a marginal density. As in the univariate case, one has to investigate the points where the partial derivatives vanish. Whether $g(y-\varepsilon)$ has a finite second moment or not will again depend on the curvature at those points.

A multivariate case where the assumptions for our main result are in general satisfied is when the regression function has a linear part, $r_\vartheta(X) = \vartheta_1 X_1 + \tilde{r}_{\vartheta_2,\ldots,\vartheta_p}(X_2,\ldots,X_p)$, which includes linear regression as a special case. Let us assume that the linear part of the regression function is not constant, that is, one parameter is non-zero, say $\vartheta_1 \neq 0$. The density of the vector $\{\vartheta_1 X_1 + \tilde{r}_{\vartheta_2,\ldots,\vartheta_p}(X_2,\ldots,X_p), X_2,\ldots,X_p\}$ at $(u_1,u_2,\ldots,u_p)$ is

$$|J| \cdot m\Big\{\frac{u_1 - \tilde{r}_{\vartheta_2,\ldots,\vartheta_p}(u_2,\ldots,u_p)}{\vartheta_1}, u_2,\ldots,u_p\Big\}$$

where $|J| = 1/|\vartheta_1|$ is the Jacobian. Hence the density of $\vartheta^\top X$ is the marginal density,

$$g(u) = \int \frac{1}{|\vartheta_1|} m\Big\{\frac{u_1 - \tilde{r}_{\vartheta_2,\ldots,\vartheta_p}(u_2,\ldots,u_p)}{\vartheta_1}, u_2,\ldots,u_p\Big\}\, d(u_2,\ldots,u_p).$$

The assumptions for our result hold if $g(y-\varepsilon)$ has a finite second moment, which is satisfied if, for example, the density of X_1 is bounded.

This discussion shows that before implementing the density estimator one should check carefully whether $g(y-\varepsilon)$ has a finite second moment. If the regression function does not have a linear (non-constant) part this requires, as a first step, that we check whether the Jacobian vanishes at certain point. If $g(y-\varepsilon)$ does, in fact, have a finite second moment, the weighted density estimator is efficient for $q(y)$. If the assumption is violated, our estimator will, in general, still converge, but at a rate slower than $n^{1/2}$. In that case it is sufficient to work with the unweighted version.

6. Proofs

Proof of Lemma 1

For the proofs of the lemmas we will repeatedly use the next inequality: if h is absolutely continuous with a square-integrable a.e. derivative h', then

$$\|h(\cdot - t) - h(\cdot - s)\|_2 \leq \|h'\|_2 |t - s|, \quad s,t, \in \mathbb{R}. \tag{6.1}$$

For the proof of (2.2) we introduce the notation

$$\breve{f}(z) \;=\; \frac{1}{N}\sum_{j=1}^{n} \delta_j k_b\{z - \varepsilon_j + \dot{r}_\vartheta(X_j)^\top(\hat{\vartheta} - \vartheta)\}$$

and show

$$\|\hat{f} - \breve{f}\|_2 = o_p(n^{-1/2}), \tag{6.2}$$

$$\|\breve{f} - \tilde{f} - f'\varrho^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2}). \tag{6.3}$$

Using (6.1) with $h = k_b$ we can bound the left-handside of (6.2) by $\|k_b'\|_2 D$, with D the left-hand side of (2.1). Since $\|k_b'\|_2 = O(b^{-3/2})$ and $D = O_p(n^{-1})$, the desired (6.2) follows.

The proof of (6.3) uses a Taylor expansion. For this we introduce the averages

$$B_1(z) = \frac{1}{N}\sum_{j=1}^n \delta_j k_b'(z - \varepsilon_j)\dot{r}_\vartheta(X_j), \quad z \in \mathbb{R},$$

$$B_2(z) = \frac{1}{N}\sum_{j=1}^n \delta_j k_b''(z - \varepsilon_j)\dot{r}_\vartheta(X_j)\dot{r}_\vartheta(X_j)^\top, \quad z \in \mathbb{R},$$

and the conditional expectations $\bar{B}_1(z) = E\{B_1(z)|\delta_1, \ldots, \delta_n\}$ and $\bar{B}_2(z) = E\{B_2(z)|\delta_1, \ldots, \delta_n\}$. The first expectation $\bar{B}_1(z)$ calculates to

$$\frac{1}{N}\sum_{j=1}^n \delta_j \int k_b'(z - y)f(y)\,dy E\{\dot{r}_\vartheta(X_j)|\delta_j = 1\} = 1(N > 0) f * k_b'(z)\varrho.$$

Analogously one obtains $\bar{B}_2 = 1(N > 0) f * k_b''(z) E\{\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top|\delta = 1\}$. The remainder $R = \breve{f} - \tilde{f} - B_1^\top(\hat{\vartheta} - \vartheta) - 1/2\,(\hat{\vartheta} - \vartheta)^\top B_2(\hat{\vartheta} - \vartheta)$ of the expansion is bounded by

$$\|R\|_2 \le \|k_b'''\|_2 \frac{1}{N}\sum_{j=1}^n \delta_j |\dot{r}_\vartheta(X_j)^\top(\hat{\vartheta} - \vartheta)|^3 = O_p(b^{-7/2}n^{-3/2}).$$

Here we used that $\|\dot{r}_\vartheta(X)\|$ has a finite third moment and that $\|k'''\|_2$ is finite. The desired rate $o_p(n^{-1/2})$ follows from $nb^\alpha \to \infty$ for some $\alpha \ge 7/2$. We also have

$$\begin{aligned}
&\int NE(\|B_2(z) - \bar{B}_2(z)\|^2|\delta_1, \ldots, \delta_n)\,dz \\
&\qquad \le \int \frac{1}{N}\sum_{j=1}^n \delta_j E\{|k_b''(z - \varepsilon_j)|^2 \|\dot{r}_\vartheta(X_j)\dot{r}_\vartheta(X_j)^\top\|^2\}\,dz \\
&\qquad \le \|k_b''\|_2^2 \frac{1}{N}\sum_{j=1}^n \delta_j E\{\|\dot{r}_\vartheta(X_j)\dot{r}_\vartheta(X_j)^\top\|^2\} = O_p(b^{-5}).
\end{aligned}$$

This shows that $\|(\hat{\vartheta} - \vartheta)^\top(B_2 - \bar{B}_2)(\hat{\vartheta} - \vartheta)\|_2 = O_p(b^{-5/2}n^{-3/2}) = o_p(n^{-1/2})$ and thus $\|\breve{f} - \tilde{f} - B_1^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2})$. Analogous arguments yield $\|(B_1 - \bar{B}_1)^\top(\hat{\vartheta} - \vartheta)\|_2 = O_p(b^{-3/2}n^{-1}) = o_p(n^{-1/2})$. Now use this, $f * k_b' = f' * k_b$ and $P(N = 0) = (1 - E\delta)^n \to 0$ to obtain $\|B_1^\top(\hat{\vartheta} - \vartheta) - f'\varrho^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2})$. This together with the above rates yields the desired result (6.3).

Statement (2.3) is quickly proved since it is just a consequence of

$$\int NE[\{\tilde{f}(z) - f_b(z)\}^2 | \delta_1, \ldots, \delta_n]\, dz \leq \int k_b^2 * f(z)\, dz = O(b^{-1}).$$

Equation (2.4) finally follows from a standard argument, using the fact that f' is Lipschitz and Condition K (cf. end of proof of Lemma 4 where we treat $f * \tilde{k}_b$). □

Proof of Lemma 2

Equations (2.5), (2.7) and (2.8) can be established analogously to the proofs of equations (2.2)–(2.4) in the previous lemma. For the proof we introduce

$$\breve{g}(z) = \frac{1}{n} \sum_{j=1}^{n} k_b\{z - r_\vartheta(X_j) - \dot{r}_\vartheta(X_j)^\top(\hat{\vartheta} - \vartheta)\},$$

and show

$$\|\hat{g} - \breve{g}\|_2 = o_p(n^{-1/2}),$$
$$\|\breve{g} - \tilde{g} + \bar{\Gamma}_1^\top(\hat{\vartheta} - \vartheta)\|_2 = o_p(n^{-1/2}). \tag{6.4}$$

which immediately yields (2.5). For the proof of (6.4), for example, the roles of B_1 and $\bar{B}_1$ are now played by Γ_1 and $\bar{\Gamma}_1$ where

$$\Gamma_1(z) = \frac{1}{n} \sum_{j=1}^{n} k_b'\{z - r_\vartheta(X_j)\}\dot{r}_\vartheta(X_j).$$

Similarly, Γ_2 is the Hessian matrix and $\bar{\Gamma}_2$ its expectation. Then one verifies $\|\breve{g} - \tilde{g} + \Gamma_1^\top(\hat{\vartheta} - \vartheta) - (\hat{\vartheta} - \vartheta)^\top\Gamma_2(\hat{\vartheta} - \vartheta)\|_2 = O_p(b^{-7/2}n^{-3/2})$ and

$$\int nE(\|\Gamma_2(z) - \bar{\Gamma}_2(z)\|^2)\, dz$$
$$\leq \int E[|k_b''\{z - \dot{r}_\vartheta(X_j)\}|^2 \|\dot{r}_\vartheta(X_j)\dot{r}_\vartheta(X_j)^\top\|^2]\, dz = O_p(b^{-5}),$$

and obtains (6.4) as in the proof of (6.3). Equations (2.7) and (2.8) can be shown analogously to the proofs of the corresponding statements in Lemma 1. In contrast to that proof where we used the Lipschitz assumption on f' we now utilize the fact that g is bounded and L_1-Lipschitz, $\|g(\cdot - t) - g\|_1 = L|t|$ for some L and all $t \in \mathbb{R}$. This holds since g is of bounded variation; see e.g. Schick and Wefelmeyer (2007a). Also use that $g_b' = g * k_b'$ and Condition K.

It remains to verify (2.6). For this we write

$$\bar{\Gamma}_1(z) = \int k_b'(z - y)h(y)g(y)\, dy = b^{-1} \int (hg)(z - bu)k'(u)\, du$$

with $h(y) = E[\dot{r}_\vartheta(X)|r_\vartheta(X) = y)$. Then, with $\bar{\Gamma}_{1i}$ and h_i denoting the i-th component of $\bar{\Gamma}_1$ and h, respectively, we obtain

$$\begin{aligned} b^2 \int |\bar{\Gamma}_{1i}(z)|^2 \, dz &= \int \{ \int h_i g(z - bu) k'(u) \, du \}^2 \, dz \\ &\leq \int \int \{h_i g(z - bu)\}^2 |k'(u)| \, du \int |k'(u)| \, du \, dz \\ &= \|k'\|_1^2 \int \{h_i(z)\}^2 g^2(z) \, dz \\ &\leq \|k'\|_1^2 \|g\|_\infty \int h_i^2(z) g(z) \, dz < \infty. \end{aligned}$$

This shows that $\|\bar{\Gamma}_1\|_2 = O(b^{-1})$, which completes the proof. □

Proof of Lemma 3

By Condition G the densitiy g is of bounded variation. Due to this assumption g can be written as the difference of two bounded monotonic functions. Hence it may be assumed without loss of generality to be of the form

$$g(t) = \int_{(-\infty, t]} \phi \, d\mu, \quad t \in \mathbb{R},$$

for some finite measure μ and some $\phi \in L_1(\mu)$. Hence we can write $q(y) = f * g(y)$ as $F * dg(y) = \int F(y - t)\phi(t) \, d\mu(t)$.

Consider $A_1(z) = N^{-1} \sum_{i=1}^n \delta_i \{g(z - \varepsilon_i) - Eg(z - \varepsilon_j)\}$ and set

$$\begin{aligned} Q(z) &= N^{-1/2} \sum_{i=1}^n \delta_i \{1(\varepsilon_i \leq z) - P(\varepsilon \leq z)\}, \\ w(\delta) &= \sup_{z \in \mathbb{R}} \sup_{|t| \leq \delta} |Q(z + t) - Q(z)|. \end{aligned}$$

Then we can write

$$\begin{aligned} N^{1/2} A_1(z) &= \int Q(z - y) \phi(y) \, d\mu(y), \quad z \in \mathbb{R}, \\ N^{1/2} A_1 * k_b(z) &= \int Q * k_b(z - y) \phi(y) \, d\mu(y), \quad z \in \mathbb{R}. \end{aligned}$$

Thus

$$N^{1/2} \|A_1 * k_b - A_1\|_\infty \leq \|Q * k_b - Q\|_\infty \int |\phi| \, d\mu = o_p(1),$$

which holds since

$$\|Q * k_b - Q\|_\infty \leq w(\sqrt{b}) + 2\|Q\|_\infty \int_{|bx| > \sqrt{b}} |k(x)| \, dx = o_p(1),$$

using known properties of empirical processes and $N/n \to E\delta > 0$. This proves $\|A_1 * k_b - A_1\|_\infty = o_p(n^{-1/2})$.

As a uniformly continuous density, k belongs to $C_0(\mathbb{R})$, and so does $A_1 * k_b$. We shall now establish tightness of $N^{1/2}A_1 * k_b$. In view of the characterization of compact subsets in $C_0(\mathbb{R})$ given in Schick and Wefelmeyer (2004c), we need to show stochastic equicontinuity and stochastic uniformly small tails. We have

$$\sup_{z\in\mathbb{R}} \sup_{|t|\le\delta} N^{1/2}|A_1 * k_b(z+t) - A_1 * k_b(z)| \le w(\delta) \int |\phi|\, d\mu,$$

and $\sup_{|z|>3M} N^{1/2}|A_1 * k_b(z)|$ is bounded by

$$\begin{aligned}
&\sup_{|z|>3M} \int\int |Q(z-y-bu)||k(u)||\phi(y)|\, d\mu(y)\, du \\
&\le \sup_{|z|>3M} \int_{|bu|\le M} \int_{|y|\le M} |Q(z-y-bu)||\phi(y)|\, d\mu(y)|k(u)|\, du \\
&\quad + \|Q\|_\infty \Big\{ \int_{|y|>M} |\phi(y)|\, d\mu(y) \int |k(u)|\, du + \int |\phi|\, d\mu \int_{|bu|>M} |k(u)|\, du \Big\} \\
&\le \sup_{|z|>M} |Q(z)| \int |\phi|\, d\mu \\
&\quad + \|Q\|_\infty \Big\{ \int_{|y|>M} |\phi|\, d\mu + \int |\phi|\, d\mu \int_{|bu|>M} |k(u)|\, du \Big\}.
\end{aligned}$$

By known properties of the empirical distribution function and by our assumptions on g and k, the last term can be chosen arbitrarily small if M is suffiently large.

The statements corresponding to A_2 can be proved similarly. Note that f satisfies stronger assumptions than g, in particular it is uniformly continuous with integrable derivative f'. We refer to Schick and Wefelmeyer (2005, Section 2) who prove for such functions f that $\|A_2 * k_b - A_2\|_\infty = o_p(n^{-1/2})$ and that A_2 converges in distribution in $C_0(\mathbb{R})$ to a Gaussian process, for any kernel k and any bandwidth $b \to 0$. □

Proof of Lemma 4

We verify equation (3.4) first. We do this in two steps by proving $\int z^2\{\hat f(z) - \tilde f(z)\}^2\, dz = o_p(1)$ and $\int z^2\{\tilde f(z) - f(z)\}^2\, dz = o_p(1)$. The second statement follows immediately from

$$\int z^2 N E[\{\tilde f(z) - f_b(z)\}^2 | \delta_1, \dots, \delta_n]\, dz \le \int z^2 k_b^2 * f(z)\, dz = O(b^{-1}),$$

our assumptions on the bandwidth, and

$$\int z^2\{f_b(z) - f(z)\}^2\, dz \to 0.$$

For the proof of the first statement we use the inequalities

$$\{k_b(z-t) - k_b(z-s)\}^2 \le (t-s)^2 \int_0^1 \{k_b'(z-s-\lambda(t-s)\}^2\, d\lambda$$

and $|u+v| \le (1+|u|)(1+|v|)$ which imply that

$$\int z^2\{k_b(z-t)-k_b(z-s)\}^2\,dz \le (t-s)^2(1+|s|)^2(1+|t-s|)^2\int(1+|u|)^2\{k_b'(u)\}^2\,du.$$

The last integral is bounded by $b^{-3}(1+b)^2(\int(1+|u|)^2\{k'(u)\}^2(u)\,du$ and has the order $O(b^{-3})$ since we assume that $\int u^2\{k'(u)\}^2\,du$ is finite. Now consider $[\int z^2\{\hat f(z)-\tilde f(z)\}^2\,dz]^{1/2}$. Using the above arguments we obtain the bound

$$\frac{1}{N}\sum_{j=1}^n \delta_j\Big[\int z^2\{k_b(z-\hat\varepsilon_j)-k_b(z-\varepsilon_j)\}^2\,dz\Big]^{1/2}$$

$$\le O(b^{-3/2})\frac{1}{N}\sum_{j=1}^n \delta_j|\hat\varepsilon_j-\varepsilon_j|(1+|\varepsilon_j|)(1+|\hat\varepsilon_j-\varepsilon_j|) \;=\; o_p(b^{-3/2}n^{-1/2}).$$

This yields the desired order $o_p(1)$ since we require $nb^3\to\infty$. We also used that $\max_{1\le j\le n}\delta_j|\hat\varepsilon_j-\varepsilon_j| = o_p(1)$. This completes the proof of (3.4).

We now prove (3.3). The weights satisfy by construction $\hat w_j(1+\hat\lambda\delta_j\hat\varepsilon_j)=1$. Multiplying both sides with $1-\hat\lambda\delta_j\hat\varepsilon_j$ gives the identity

$$\hat w_j = 1-\hat\lambda\delta_j\hat\varepsilon_j+(\hat\lambda\delta_j\hat\varepsilon_j)^2\hat w_j.$$

Thus we have

$$\hat f_w(z)-\hat f(z)+\hat\lambda\psi(z) = \hat\lambda\{\psi(z)-z\hat f(z)\}+\hat\lambda V_1(z)+\hat\lambda^2 V_2(z) \tag{6.5}$$

with

$$V_1(z)=\frac{1}{N}\sum_{j=1}^n\delta_j(z-\hat\varepsilon_j)k_b(z-\hat\varepsilon_j),\qquad V_2(z)=\frac{1}{N}\sum_{j=1}^n\delta_j\hat\varepsilon_j^2\hat w_j k_b(z-\hat\varepsilon_j).$$

The first term on the right-hand side of (6.5) has the desired order $o_p(n^{-1/2})$ by equation (3.4) proved above and since $\hat\lambda = O_p(n^{-1/2})$ by (3.1). For the third term we have

$$\|V_2\|_2 \le \frac{1}{N}\sum_{j=1}^n\delta_j\hat\varepsilon_j^2\hat w_j\|k_b\|_2 = O_p(b^{-1}),$$

which holds in view of $\|k_b\|_2 = O(b^{-1})$, $\max_{1\le j\le n}\delta_j|\hat\varepsilon_j-\varepsilon_j| = o_p(1)$ and $\max_{1\le j\le n}|\hat w_j-1| = o_p(1)$. This rate suffices since $\hat\lambda^2 = O_p(n^{-1})$.

Now consider the second term on the right-hand side of (6.5) involving V_1. We can rewrite V_1 so that it contains a factor b,

$$V_1(z) = b\frac{1}{N}\sum_{j=1}^n\delta_j\tilde k_b(z-\hat\varepsilon_j)$$

where $\tilde k(x)=xk(x)$. Using inequality (6.1) yields

$$\begin{aligned}\|\frac{1}{N}\sum_{j=1}^n\delta_j(\tilde k_b(z-\hat\varepsilon_j)-\tilde k_b(\cdot-\varepsilon_j)\|_2 &= O(b^{-3/2})\frac{1}{N}\sum_{j=1}^n\delta_j|r_{\hat\vartheta}(X_j)-r_\vartheta(X_j)|\\ &= O_p(n^{-1/2}b^{-3/2}).\end{aligned}$$

This holds since $\int(\tilde{k}'(x))^2\,dx$ can be bounded by $2\int[k^2(x)+x^2\{k'(x)\}^2]\,dx$ which is assumed to be finite. As in the proof of Lemma 1 we obtain

$$\int NE\big[\big\{\frac{1}{N}\sum_{j=1}^{n}\delta_j\tilde{k}_b(z-\varepsilon_j)-f*\tilde{k}_b(z)\big\}^2\big]\,dz\le\int\tilde{k}_b^2*f(z)\,dz=O(b^{-1}).$$

This requires $\int x^2k^2(x)\,dx$ to be finite, which is satisfied since we use a second order kernel. The above and

$$\begin{aligned}f*\tilde{k}_b(z)&=\int\{f(z-bu)-f(z)-f'(z)bu\}uk(u)\,du\\&=b\int_0^1\int\{f'(z-sbu)-f'(z)\}\,ds\,u^2k(u)\,du\end{aligned}$$

gives

$$\begin{aligned}\|f*\tilde{k}_b\|_2^2&\le b^2\int u^2k(u)\,du\int_0^1\iint\{f'(z-sbu)-f'(z)\}^2\,dz\,u^2k(u)\,du\,ds\\&\le Lb^3\int u^2k(u)\,du\int_0^1\iint|f'(z-sbu)-f'(z)|\,dz\,u^2k(u)\,du\,ds\end{aligned}$$

which is of order $o(b^3)$. Now note that V_1 contains an additional factor b which we also have to take into account. This shows that $\|V_1\|_2=O_p(n^{-1/2}b^{-1/2})+o(b^{5/2})$ which is $o_p(1)$ as desired. □

Acknowledgements

The author thanks the referees for constructive comments and extremely helpful suggestions.

References

[1] Du, J. and Schick, A., 2007. Root-n consistency and functional central limit theorems for estimators of derivatives of convolutions of densities. *Int. J. Stat. Manag. Syst.*, 2, 67-87.

[2] Escanciano, J.C. and Jacho-Chávez, D.T., 2010. Root-n uniformly consistent density estimation in nonparametric regression models. Preprint.

[3] Forrester, J., Hooper, W., Peng H. and Schick, A., 2003. On the construction of efficient estimators in semiparametric models. *Statist. Decisions*, 21, 109-138.

[4] Frees, E.W., 1994. Estimating densities of functions of observations. *J. Amer. Statist. Assoc.*, 89, 517-525.

[5] Giné, E. and Mason D.M., 2007. On local U-statistic processes and the estimation of densities of functions of several sample variables. *Ann. Statist.*, 35, 1105-1145.

[6] Little, R.J.A. and Rubin, D.B., 2002. *Statistical analysis with missing data.* Second edition. Wiley-Interscience.

[7] Mojirsheibani, M., 2007. Nonparametric curve estimation with missing covariates: A general empirical process approach. *J. Statist. Plann. Inference*, 137, 2733-2758.

[8] Müller, U.U., 2009. Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, 37, 2245-2277.

[9] Müller, U.U., Schick, A. and Wefelmeyer, W., 2005. Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica*, 15, 177-195.

[10] Müller, U.U., Schick, A. and Wefelmeyer, W., 2006. Imputing responses that are not missing. *Probability, Statistics and Modelling in Public Health* (M. Nikulin, D. Commenges and C. Huber, eds.), 350-363, Springer.

[11] Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

[12] Owen, A.B., 2001. *Empirical Likelihood*. Monographs on Statistics and Applied Probability, 92, Chapman & Hall.

[13] Saavedra, A. and Cao, R., 1999. Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.* 80, 129-155.

[14] Schick, A., 1993. On efficient estimation in regression models. *Ann. Statist.*, 21, 1486-521. Correction and addendum: 23 (1995) 1862-1863.

[15] Schick, A. and Wefelmeyer, W., 2004a. Root n-consistent and optimal density estimators for moving average processes. *Scand. J. Statist.*, 31, 63-78.

[16] Schick, A. and Wefelmeyer, W., 2004b. Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli*, 10, 889-917.

[17] Schick, A. and Wefelmeyer, W., 2004c. Root n-consistent density estimators for sums of independent random variables. *J. Nonparametr. Statist.*, 16, 925-935.

[18] Schick, A. and Wefelmeyer, W., 2007a. Root-n consistent density estimators of convolutions in weighted L1-norms. *J. Statist. Plann. Inference*, 137, 1765-1774.

[19] Schick, A. and Wefelmeyer, W., 2007b. Uniformly root-n consistent density estimators for weakly dependent invertible linear processes. *Ann. Statist.*, 35, 815-843.

[20] Schick, A. and Wefelmeyer, W., 2009a. Convergence rates of density estimators for sums of powers of observations. *Metrika*, 69, 249-264.

[21] Schick, A. and Wefelmeyer, W., 2009b. Non-standard behavior of density estimators for sums of squared observations. *Statist. Decisions*, 27, 55-73.

[22] Støve, B. and Tjøstheim, D., 2011. A Convolution Estimator for the Density of Nonlinear Regression Observations. Preprint.

[23] Wang, Q., 2008. Probability density estimation with data missing at random when covariables are present. *J. Statist. Plann. Inference*, 138, 568-587.