Local Partitioned Regression

Norbert Christopeit* and Stefan G.N. Hoderlein†

Juni 2004

Abstract

In this paper, we introduce a Kernel based estimation principle for nonparametric models named local partitioned regression. This principle is a nonparametric generalization of the familiar partition regression in linear models. It has several key advantages:

First, it generates estimators for a very large class of semi- and nonparametric models. A number of examples which are particularly relevant for economic applications will be discussed in this paper. This class contains the additive, partially linear and varying coefficient models as well as several other models that have not been discussed in the literature.

Second, LPR based estimators generally achieve optimality criteria: They have optimal speed of convergence and are oracle-efficient. Moreover, they are simple in structure, widely applicable and computationally inexpensive.

The LPR estimation principle involves preestimation of conditional expectations and derivatives of densities. We establish that the asymptotic distribution of the estimator remains unaffected by preestimation if the total number of regressors is smaller than ten, in the sense that we do not require additional smoothness assumptions in preestimation. Finally, a Monte-Carlo simulation underscores these advantages.

Keywords: Nonparametric, Additive Model, Interaction Terms, Varying Coefficient Models, Oracle Efficiency.

1 Introduction

Since economic theory rarely prescribes a linear model or a specific functional form, semi- and nonparametric methods seem to be ideal tools for econometrics and applied economics. The most widely known of these tools is of course the nonparametric regression model,

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1, 2, \ldots, \tag{1.1}$$

*Bonn University, Department of Economics, Institute of Econometrics, Konrad-Adenauer-Allee 24-42, 53113 Bonn, Germany, email: christopeit@wiwi.uni-bonn.de.

†Mannheim University, Department of Economics, Institute for Statistics, L7, 3-5, 68131 Mannheim, Germany, email: stefan_hoderlein@yahoo.com. We have received helpful comments by seminar participants in Berkeley, Berlin, Boston, ESEM, Madrid, Mannheim, Heidelberg, LSE, Stanford, UCL. Financial support by SFB 504 is gratefully acknowledged.

which models the dependence of a random scalar Y_i on a $d+1$dimensional random vector X_i. The noise term ε_i is assumed to be mean independent of X_i, with $\mathbb{E}[\varepsilon_i|X_i] = 0$ and $\mathbb{E}\left[\varepsilon_i^2|X_i\right] = \sigma^2$, and m is the mean regression function, usually assumed to be smooth. Although this model has been used in applied work, it's usage is severely restricted by the curse of dimensionality, i.e. the fact that the precision of any estimator decreases exponentially with the dimensionality of the regressors. Hence, it is imperative that some structure be placed on the model (1.1) to make it useful for most econometric applications. However, these structural assumptions have to be "mild" in the sense that they should not exclude too many economically interesting models, or be even in conflict with economic theory. Perhaps the most popular class of models that place some structure on the mean regression is the class of additive and partially linear models. In the most basic specification, the mean regression takes the form

$$m(X_i) = c + \sum_{j=1}^{d+1} m_j(X_{ij}),\ i = 1, 2, \ldots, \tag{1.2}$$

where X_{ij} denotes the j-th component of X_i, and the m_j are smooth functions. The partially linear model is nested in (1.2), with $\sum_{j=2}^{d+1} m_j(X_{ij}) = \sum_{j=2}^{d+1} \gamma_j X_{ij}$, where $\gamma_j, j = 2, .., d+1$ are fixed parameters. In principle, estimators of m_j may achieve the same speed of convergence as one dimensional nonparametric regression estimators (Stone (1985)), or even root n if $m_j(x)$ is specified parametrically, e.g. as $\gamma_j x$. Moreover, the m_j are easy to visualize and straightforward to interpret. In this model, only the derivatives are identified. Since marginal effects are of paramount importance throughout economics and econometrics, in this paper we will exclusively be concerned with the estimation of derivatives.

Despite being very appealing in principle, the basic additive structure (1.2) is an example of a model that is too restrictive in a number of economic applications. The main reason is that this structure does not allow for interaction terms, i.e. marginal effects of X_{i1} that vary over the X_{ij}, $j = 2, ..$, are being ruled out. One implication of this limitation is that it is generally at odds with consumer theory: It is well-known that demand systems having log income and household observables entering additively separable must be linear in log income (Blundell, Browning and Crawford (2003)). We emphasize this point as it illustrates the need for a large and flexible class of models, where an alternative specification can be chosen by the researcher if the initial specification is found to be at odds with economic theory.

Our main contribution in this paper is a new Kernel based estimation principle that allows constructing rather simple estimators for a great variety of models, taking model (1.2) as one building block. In particular, several types of interaction structures can be considered, bridging the gap between additive and partially linear models on one side and the unrestricted nonparametric (1.1) model on the other. In addition, we will establish that all estimators achieve certain optimality conditions, and are easy to implement.

This estimation principle is called Local Partitioned Regression, and for brevity will be denoted as LPR. It can be seen as a generalization of both the Frisch-Waugh partitioned regression theorem

and Robinson's (1988) estimator. As is well known, taking linear projectors or simple conditional expectations will not be sufficient for the estimation of model (1.2), due to the nonlinearity of all functions involved. Instead, we will establish that conditioning on Kernel multiplied random variables will yield a tool that works in these nonlinear models. Formally, let $W_i = K((X_i - x_0)/h)/h$ denote a specific Kernel weight to be defined below, and consider the simplest model, $Y_i = k(X_i) + l(Z_i) + \varepsilon_i,\ i = 1, 2, \ldots$ Then, our proposed estimator for k' is given by regressing the residuals $W_i Y_i - \mathbb{E}[W_i Y_i | W_i Z_i]$ on the residuals $W_i X_i - \mathbb{E}[W_i X_i | W_i Z_i]$. This basic principle will be retained throughout the paper, but for more involved models more elaborate procedures have to be devised.

Since we are examining a whole class of models, there is no directly related literature. However, certain models within this class have been carefully examined, most prominently the additive model (1.2). Key contributions in the literature on this model are Tjostheim and Auestad (1994), Newey (1994a) and Linton and Nielsen (1995), for the marginal integration estimators and Opsomer and Ruppert (1997) as well as Mammen, Linton and Nielsen (1999) for backfitting. Both are Kernel based estimators and will be discussed in more detail below. Theoretical results for series based estimators that apply to some models within the class of models we consider are given in Andrews and Whang (1990), Wahba (1992) and Newey (1995). Another model that is contained in the class of models we consider is the varying coefficient model. Fan and Zhang (1999), and Chiang, Rice and Wu (2001) consider Kernel, respectively spline, based estimation of this model. A hybrid model between partially linear and additive model which is contained in our class is considered in Heckman, Ichimura, Smith and Todd (1998). Finally, generalized models for a class of additive type models have been considered by Horowitz (2001) and Mammen and Nielsen (2003), but they do not contain most models considered in this paper. We give further references, when discussing the models in detail.

In general, LPR based estimators achieve optimality properties that are, for the additive model, shared by backfitting and series based estimators, but not by standard marginal integration based estimators. In fact, the latter method can be quite inefficient if the explanatory variables are correlated, arguably rather the rule than the exception in economics. An approach that improves upon marginal integration in that respect combines this method with one backfit, see Linton (1997) for a lucid discussion. Horowitz, Klemelä and Mammen (2004) show more generally that two-step procedures help achieve oracle efficiency. However, it is - as of yet - unknown whether backfitting and two-step procedures apply to any other models beyond (1.2), while LPR extends naturally to a large class, as is established in this paper. Moreover, backfitting, as an iterative procedure, and marginal integration are computationally very expensive. This has particular consequences for the implementation of computer intensive methods such as selecting the optimal bandwidth via cross-validation and bootstrapping confidence intervals, see Kim, Linton and Hengartner (1999) on this topic. LPR in contrast is very simple, computationally less expensive and easy to implement. Another advantage of LPR is that through the local polynomial structure the determination of model complexity, i.e. bandwidth choice and degree of local polynomial, may be performed as in the scalar local polynomial literature, and is well understood. The same holds true for imposing economic restrictions. Finally, in contrast to iterative methods, LPR is robust against misspecifications of parts of the model.

Series based estimation methods share oracle efficiency and wide applicability with LPR. However, series estimators invoke strong support conditions, and the determination of model complexity is not fully explored. For instance, adding one additional term to a series can radically change the fitted value at any point. Also, imposing restrictions is not as straightforward.

The structure of this paper will be as follows: In the second section we will present various models of economic relevance that may be considered by this estimation principle. How the LPR estimation principle may guide in their estimation will be discussed in the third section. As in the partially linear model, certain quantities like conditional expectations have to be pre-estimated, and this step may impact the asymptotic behavior of the estimators. This issue is studied in the fourth section for the case of the basic additive model (1.2). To investigate the small sample performance, we include a Monte Carlo study which analyzes the performance of a LPR based estimator for model (1.2), and compares it with other estimators that have been proposed in the literature. Finally, an outlook concludes the paper.

2 Models and Applications

In this section we will give an overview of economically relevant models that are estimable by LPR based estimators. We will always display the model, and give examples of potential economic applications. Of course, this list of examples is subjective and by no means exhaustive.

2.1 The Basic Additive Model

This is the only model that has received a thorough and in depth investigation. Since we concentrate on the estimation of the derivative of a single component at a fixed position, we rewrite the model given by (1.1) and (1.2) as

$$Y_i = k(X_i) + l(Z_i) + \varepsilon_i,\ i = 1, 2, \ldots, \tag{2.1}$$

where X_i now stands for the first component and $Z_i = (X_{2i}, .., X_{d+1,i})$, so that $k = m_1$ and $l = \sum_{j=1}^{d+1} m_j$. The estimation of the derivative $k'(x_0)$ in the third section is investigated for completely general l so that we already allow at this point for a lot of additional generality. The model given by (2.1) will be called *Model I*. It is of considerable economic and econometric importance for the following reasons:

1. It's main economic justification comes from separability assumptions on the utility or the production function. These assumptions are often invoked to keep a theoretical model tractable, and to focus on the effect of some variables in isolation. Examples are ubiquous across economics. E.g. in production they include the workhorses in this literature (Cobb-Douglas, Leontief).

2. Another justification comes from the control function approach, (Heckman and Robb (1986)). The baseline model is as in (1.1), with the exception that $\mathbb{E}[\varepsilon_i|X_i] \neq 0$. However, there exist instruments Z_i which define U_i as $U_i = X_i - \mathbb{E}[X_i|Z_i]$. The core assumption is then: $\mathbb{E}[\varepsilon_i|X_i, U_i] = l(U_i)$ which yields $\mathbb{E}[Y_i|X_i, U_i] = m(X_i) + l(U_i)$. This model has been considered in detail by Newey, Powell and

Vella (1999). A similar approach can be chosen for selection models, see Das, Newey and Vella (1999).
3. It includes nonparametric panel data models. Take model (1.1), but now indexed with t and i, and add an additive individual specific time invariant random variable, a "fixed effect". Then, time differencing yields an additive stricture, with $\Delta Y_{i,t} = k(X_{i,t}) + l(X_{i,t-1}) + \eta_{it}$, where $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$, $l = -k$ and $\eta_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$.

2.2 The Varying Coefficient Model

Another important generalization of the standard linear regression model, is the following

$$Y_i = \alpha(X_i) + \beta(X_i)'Z_i + \varepsilon_i, \quad i = 1, 2, \ldots, \tag{2.2}$$

where α and β are smooth but unrestricted functions of X_i, a s-dimensional random vector, and Z_i is a k-dimensional random vector with $k + s = d + 1$. This model has several economic justifications:
1. It can be seen as generalization of the partially linear model, which is arguably the most popular semiparametric model in econometrics. In contrast to the partially linear model, it allows for marginal effects that vary across covariates. Since it nests both partially linear and linear models directly, it may well be used in a specification search. Equally well it can be seen as a first order approximation to (1.1) in Z_i. Hence it may be suitable in situations where linearization in some variables is acceptable on theoretical grounds.
2. It can be used to generalize standard linear econometric models. For instance, Chen and Tsay (1993) consider functional coefficient autoregressive models which are exactly of this type. Moreover, it is useful for longitudinal analysis with time varying coefficients, i.e. X_i denotes time, see references in Fan and Zhang (1999).
3. It also allows generalizing key models of applied economics. An example is the Almost Ideal Demand System (Deaton and Muellbauer (1980)). Assume that the log cost function is linear in utility, i.e. $\log c = a(p) + b(p)u$, where p are log prices and u is utility. Then, using standard arguments, the vector of budget shares would be given as $w = \alpha(p) + \beta(p)x$, where $\alpha = \nabla_p a + (a/b)\nabla_p b$ and $\beta = \nabla_p b/b$ and x denotes log nominal total expenditure.
4. In the class of HARA preferences widely used in portfolio choice and consumption, portfolio shares are linear in wealth, with coefficients that vary with age (Merton (1971)).
5. Other applications are given by the numerous cases were a known functional form has coefficients varying systematically with covariates. As will become obvious from the discussion below, we may as well allow for Z_i to enter in a known nonlinear fashion.

2.3 Semilinear Interaction

The following three subsections are devoted to models which combine features of the additive and the varying coefficient models. This is done by augmenting the basic additive model with interaction structures. The first model is

$$Y_i = k(X_i) + l(Z_i) + g(X_i)'\lambda(Z_i) + \varepsilon_i, \ i = 1, 2, \ldots, \tag{2.3}$$

where the k and l functions and the regressors are as in the basic additive model. Compared to the additive model (2.1), the novelty is the interaction term $g'\lambda$, where g is assumed to be smooth unrestricted and unknown, but λ is assumed to be a known, vector valued function. Compared to the varying coefficient model (2.2), we allow for an additional unknown function l. This model will be called *Model II*.

Economic examples of this model include the case where $\lambda(Z_i)$ is a pre-estimable quantity.

1. For instance, $\lambda(Z_i)$ may be a Mill's ratio in a nonparametric selection model with normal errors.
2. Another example is the nonparametric switching regression/treatment model defined as

$$\begin{aligned} Y_{0i} &= k_0(X_i) + \varepsilon_{0i},\ i = 1, 2, \ldots, \\ Y_{1i} &= k_1(X_i) + \varepsilon_{1i},\ i = 1, 2, \ldots \end{aligned}$$

Let $Y_i = \mathbf{1}\{D_i = 0\} Y_{0i} + (1 - \mathbf{1}\{D_i = 0\}) Y_{1i}$. In addition assume that there exist Z_i with the following properties: Let X_i be a true subset of Z_i,
$\mathbb{P}\{D_i = 0|Z_i\} = \mathbb{P}_i$, and let $\mathbb{E}\{\varepsilon_{ji}|X_i, Z_i, D_i = j\} = h_j(\mathbb{P}_i)$, $j = 0, 1$. Since $\mathbb{P}_i$ can be estimated separately, it can be treated as known. Then follows

$$\begin{aligned} \mathbb{E}\{Y_i|X_i, Z_i\} &= \mathbb{P}_i k_0(X_i) + (1 - \mathbb{P}_i) k_1(X_i) + \mathbb{P}_i \mathbb{E}\{\varepsilon_{0i}|Z_i, D_i = 0\} + (1 - \mathbb{P}_i) \mathbb{E}\{\varepsilon_{1i}|Z_i, D_i = 1\} \\ &= k_1(X_i) + g(X_i)\mathbb{P}_i + l(\mathbb{P}_i), \end{aligned}$$

where $g(X_i) = k_0(X_i) - k_1(X_i)$ and $l(\mathbb{P}_i)$ in an obvious fashion. This model nests very diverse models such as, inter alia, Ahn and Powell (1993) as well as Heckman and Vytlacil (2003).

3. As generalization of *Model I*, it can also be used in the control function IV approach. It allows to relax $\mathbb{E}[\varepsilon_i|X_i, U_i] = l(U_i)$ to $\mathbb{E}[\varepsilon_i|X_i, U_i] = l(U_i) + g(X_i)'\lambda(U_i)$. If λ is a higher order polynomial, we may arrive at something "close" to a general solution to the hardly tractable nonparametric IV problem.

Model II is further generalizable, if instead of $g(X_i)'\lambda(Z_i)$ we consider $g(X_i)'P_i$, where P_i may be not Z_i measurable. For instance, P_i may be a set of additional regressors, or a known or preestimated function of X_i and Z_i. The model

$$Y_i = k(X_i) + l(Z_i) + g(X_i)'P_i + \varepsilon_i,\ i = 1, 2, \ldots, \tag{2.4}$$

called *Model III*, has similar types of applications, but differs a bit in identification and requires a different estimator, see section 3 below. One application would be if, in the generalized Almost Ideal, one would use log real total expenditure, i.e. divide x, nominal income, by a price index. Other applications include the case where P_i is a known nonlinear function of X_i and Z_i.

2.4 Unrestricted Interaction

The next type of model allows for completely unrestricted interaction terms. For simplicity, we concentrate in the following discussion on pairwise interaction terms. Then, let the model be defined

as

$$Y_i = k(X_i) + \tilde{l}(Z_i) + \sum_{j=1,..,d} g_j(X_i, Z_{ji}) + \varepsilon_i,\ i = 1, 2, \tag{2.5}$$

where all functions are assumed to be smooth and unknown. In this *Model IV*, the only restriction compared to model (1.1) is the absence of higher order interaction terms. A special case of this model is the following additive model with pairwise interaction terms, which has been considered by Sperlich, Tjostheim and Yang (2002),

$$Y_i = k(X_i) + \sum_{j=1,..,d} l_j(Z_{ji}) + \sum_{j=1,..,d} g_j(X_i, Z_{ji}) + 2 \sum_{j=1,..,d} \sum_{l>j} h_{lj}(Z_{li}, Z_{ji}) + \varepsilon_i,\ i = 1, 2, \ldots,$$

where $\tilde{l}(Z_i) = \sum_{j=1,..,d} l_j(Z_{ji}) + 2 \sum_{j=1,..,d} \sum_{l>j} h_{lj}(Z_{li}, Z_{ji})$.
One economic motivation comes from a relaxed separability assumption:
1. In production function estimation, the following functional forms can be nested in (2.5):
generalized Cobb Douglas $\ln y = c + \sum_{j=1,..,d} \sum_{l=1,..,d} c_{jl} \ln((x_j + x_l)/2)$,
Translog $\ln y = c + \sum_{j=1,..,d} c_j \ln(x_j) + \sum_{j=1,..,d} \sum_{l=1,..,d} c_{jl} \ln(x_j) \ln(x_l)$,
The same holds true for the generalized Leontief, the Quadratic and the generalized Concave.
2. Another example comes from economics of household and family. For instance, if a mothers utility depends on the nutrition of her child, then we may expect nutrition demand to exhibit such a pairwise structure, see Chesher (1997).

2.5 Product Interaction

This type of interaction is closely related to the previous. In particular, consider the modification of *Model IV*, called *Model V*, where the pairwise interaction term is multiplicative

$$Y_i = k(X_i) + \tilde{l}(Z_i) + \sum_{j=1,..,d} g_j(X_i) q_j(Z_{ji}) + \varepsilon_i,\ i = 1, 2, \ldots, \tag{2.6}$$

where all functions are assumed to be smooth and unknown. Examples for econometric applications include
1. The nonparametric random coefficient model, which may be defined as

$$\begin{aligned} Y_i &= \xi_i k(X_i) + \varepsilon_i,\ i = 1, 2, \ldots, \\ \xi_i &= \bar{\xi} + V_i = 1 + V_i, \end{aligned}$$

with endogenous regressors, i.e. $\mathbb{E}[\varepsilon_i | X_i] \neq 0$ as well as $\mathbb{E}[V_i | X_i] \neq 0$ and $\bar{\xi} = 1$ as normalization. In the control function framework, we have exactly

$$\mathbb{E}[Y_i | X_i, U_i] = k(X_i) + k(X_i) l(U_i) + h(U_i),$$

where U_i is again defined as above, and it is assumed that $\mathbb{E}[V_i | X_i, U_i] = l(U_i)$ as well as $\mathbb{E}[\varepsilon_i | X_i, U_i] = h(U_i)$.
2. The control function approach with heteroscedasticity, i.e. the model is $Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$, with

$\mathbb{E}[\varepsilon_i|X_i] \neq 0$. As above, assume that there exist instruments Z_i which define U_i as $U_i = X_i - \mathbb{E}[X_i|Z_i]$ such that $\mathbb{E}[\varepsilon_i|X_i, U_i] = l(U_i)$. This yields $\mathbb{E}[Y_i|X_i, U_i] = m(X_i) + \sigma(X_i)l(U_i)$.
3. The model of Florens, Heckman, Meghir and Vytlacil (2004) is similar to the nonparametric random coefficient model, and yields also a similar structure.

2.6 Hybrid Models

Finally, several of these features may be combined. For instance, a model that approaches the unrestricted nonparametric model (1.1), is when interaction terms of lower order are nonparametrically, but higher order interaction is done parametrically.

3 Models - LPR Estimation as a Theoretical Principle

In this section we discuss estimation and identification of the models introduced. The unifying theme will be the LPR principle, which makes use of the basic additive structure of the conditional expectation. The main theoretical results are given in this section, however, the proofs may be found in the appendix.

The theoretical LPR estimators contain conditional expectations which will generally not be known given the data. At the second stage considered in section 4, we will show how these conditional expectations may be replaced by consistent estimators in such a way that the asymptotic behavior of the least squares estimator is not affected. Since this can be done in a variety of ways - of which we single out one specific - this procedure resembles the familiar passage from Aitken/GLS estimators to *feasible* Aitken/FGLS estimators.

3.1 *Model I*

In this subsection we consider model (1.2) in the scenario of iid explanatory variables and with the variable of main interest (namely X) being one dimensional[1]. This model is a building block of all subsequent models, and the basic principle of LPR can best be illustrated here. Hence, we will be more explicit in this section, and will be brief in others, were we focus on the novel elements in each model. Another reason for considering this model in greater detail is that it is the model that has been extensively considered, and where well established estimation methods already exist.

Turning to the basic additive model, the only identification restriction is that all component functions are only identified up to a constant, or, put differently, only the marginal effects are identified. Since the model is completely symmetric, we shall concentrate on the estimation of one derivative $k'(x_0)$. Our objective is then to find an estimator for $k'(x_0)$ that is asymptotically normal at rate $\sqrt{nh^3}$, where h is the bandwidth. This yields, under certain smoothness assumptions like (A3), the optimal rate of convergence (see Stone (1985)).

[1]Extensions to the α-mixing case may be performed as in Christopeit and Hoderlein (2002).

General remark: Throughout the paper, we shall use the same symbol f to denote densities, the kind of density being indicated by the arguments. E.g., $f(x,z,p)$ is the joint density of (X,Z,P), $f(x,z)$ the joint density of (X,Z), etc.. Partial derivatives will be denoted by ∂_x, ∂_x^2,... Now, we introduce the LPR estimation principle and establish the asymptotic properties of LPR in the additive model.

Assumptions for *Model I*

(A1-I) The (X_i, Z_i) are iid $\mathbb{R}\times\mathbb{R}^d$-valued random variables with continuous joint density $f(x,z)$. f is twice continuously differentiable with respect to x in a neighborhood of x_0 for all z, and there exists a bounded nonnegative Borel function $\gamma(z)$ with $\int \gamma(z)dz < \infty$ such that

$$\sup_{|x-x_0|\le h/2} \left[|f(x,z)| + |\partial_x f(x,z)| + \left|\partial_x^2 f(x,z)\right|\right] \le \gamma(z)$$

for h sufficiently small. The set $\{z : f(x_0,z) = 0 \text{ and } \partial_x f(x_0,z) \neq 0\}$ has Lebesgue measure zero. Finally, $f(x_0) > 0$.

(A2-I) The ε_i are iid with zero mean and variance σ^2. For every i, ε_i is independent of the σ-algebra $\mathcal{F}_{i-1} = \sigma\{X_1, Z_1, \ldots, X_i, Z_i; \varepsilon_1, \ldots, \varepsilon_{i-1}\}$.

(A3-I) k is three times continuously differentiable in a neighborhood of x_0.

(A4) $\mathbb{P}(Z_i = 0) = 0$

(A5) $K(x) = 1_{[-1/2,1/2]}(x)$.

(A6) $nh_n^3 \to \infty$.

Most of these assumptions are common technicalities. (A1) is a common boundedness assumption. (A2) can be relaxed to allow for heteroscedasticity and serial dependence, the latter being discussed in Christopeit and Hoderlein (2002). (A4) may seem a rather uncommon assumption at first sight. It is, however, unrestrictive since it is automatically fulfilled in the case of continuously distributed Z_i. In the case of a mixed distribution where a positive probability is placed on $Z_i = 0$ we may simply shift the distribution of Z_i to $Z_i + c$, so that $\mathbb{P}(Z_i + c = 0) = 0$, estimate the function and than shift the curves back again. The choice of a uniform kernel in (A5) is made to simplify the proofs substantially. It is discussed in more detail in the proofs below.

Remark 3.1: An inspection of the proofs (actually only the proof of Lemma A1.6 is concerned) shows that (A2) can be weakened to independence up to second order (of ε_i and $\mathcal{F}_{i-1}$) plus conditional homoscedasticity together with a conditional uniform integrability condition.

To begin with, expand (2.1) in the form

$$Y_i = k(x_0) + k'(x_0)(X_i - x_0) + l(Z_i) + r_i + \varepsilon_i, \tag{3.1}$$

with $r_i = \frac{1}{2}k''(x_0)(X_i - x_0)^2 + \frac{1}{3!}k^{(3)}(x_0 + \eta_i(X_i - x_0))(X_i - x_0)^3$, $\eta_i = \eta(X_i) \in (0,1)$. In the semiparametric model $k(x) = x'\beta$, it is sufficient to take the conditional expectation of Y_i with respect to Z_i and subtract it from (3.1) to remove the nonlinear $l(\cdot)$ function. Here, however, things are a bit more involved due to the bias expression r_i. If we were to proceed as in the simple partially linear model, we would end up with a complicated and nonvanishing asymptotic bias. This complication arises from the fact that the conditional expectation of the bias does not vanish. For applications it may be sufficient to take means to make this bias "small" (cf. Hoderlein (2002) for a possible estimation method). The route we are going to pursue here is to premultiply (3.1) and the conditioning variables with a kernel *before* taking conditional expectations and differences. Performing this operation which we call quasidifferencing, (3.1) becomes

$$\begin{aligned} W_iY_i - \mathbb{E}\{W_iY_i|W_iZ_i\} &= W_ik(x_0) - \mathbb{E}\{W_ik(x_0)|W_iZ_i\} \\ &\quad + k'(x_0)\left[W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_iZ_i\}\right] \\ &\quad + W_il(Z_i) - \mathbb{E}\{W_il(Z_i)|W_iZ_i\} \\ &\quad + W_ir_i - \mathbb{E}\{W_ir_i|W_iZ_i\} + W_i\varepsilon_i - \mathbb{E}\{W_i\varepsilon_i|W_iZ_i\}, \end{aligned} \tag{3.2}$$

where $W_i = W_{ni} = h^{-1}K\left((X_i - x_0)/h\right)$, $K(x)$ is the uniform kernel (A6) and $h = h_n$ the bandwidth. The main reason for doing so is that then the constant and the $l(Z_i)$ - terms cancel out, as is shown by the following trivial

Lemma 3.1 *Let ϕ be continuous and ψ measurable. Then, under assumption (A5),*

$$\mathbb{E}\{W_i\phi(X - x_0)\psi(Z_i)|W_iZ_i\} = W_i\psi(Z_i)\mathbb{E}\{\phi(X - x_0)|W_iZ_i\} \quad (a.s.),$$

for functions ϕ and ψ.

Proof. Suppressing the index i, denote $L = WZ$. Obviously, $W\psi(Z) = h^{-1}\psi(hL)1_{\{L\neq 0\}}$ $+W\psi(0)1_{\{Z=0\}}$. Therefore $W\psi(Z)$ coincides a.s. with the L-measurable random variable $h^{-1}\psi(hL)1_{\{L\neq 0\}}$. ■

As a consequence, we may as well write $\mathbb{E}\{W_i\phi(X - x_0)\psi(Z_i)|W_iZ_i\} =$ $\psi(Z_i)\mathbb{E}\{W_i\phi(X - x_0)|W_iZ_i\}$. An exact analogue of Lemma 3.1 can be found for certain more general kernels, but only at the price of imposing rather artificial conditions on the joint distribution of X_i and Z_i. It remains to be determined by future research whether such an equality holds up to orders of h for general kernels.

To simplify the rather tedious (3.2) we introduce the following notation:

$\theta = \theta(x_0) = k'(x_0)$,

$U_i = U_{ni}(x_0) = W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_iZ_i\}$,

$V_i = V_{ni} = W_iY_i - \mathbb{E}\{W_iY_i|W_iZ_i\}$,

$\widetilde{r}_i = \widetilde{r}_{ni} = W_ir_i - \mathbb{E}\{W_ir_i|W_iZ_i\}$,

$\widetilde{\varepsilon}_i = \widetilde{\varepsilon}_{ni} = W_i\varepsilon_i - \mathbb{E}\{W_i\varepsilon_i|W_iZ_i\}$.

Throughout, $\kappa_p = \int s^p K(s)ds$ will denote the p- th moment of the uniform kernel. In view of Lemma 3.1, (3.2) can then be rewritten as

$$V_i = \theta U_i + \widetilde{r}_i + \widetilde{\varepsilon}_i. \tag{3.3}$$

Now the similarities with the Robinson estimator for θ in the partially linear model are obvious. The differences to the Robinson model are that $\theta = k'(x_0)$ may vary with position, and the existence of a bias term $\widetilde{r}_i$. Moreover, the "disturbances" $\widetilde{r}_i + \widetilde{\varepsilon}_i$ are correlated with the regressors.

Furthermore, like with Robinson's estimator, the variables U_i, V_i are not a priori known given the data. Rather, the conditional expectations have to be estimated, so that the actually calculable regression would be that of $\hat{V}_i$ on $\hat{U}_i$, $\hat{V}_i$ and $\hat{U}_i$ being appropriate estimates of V_i and U_i, respectively. As already mentioned, our focus in this section is on the regression (3.3) taking the U_i and V_i *as known*, and we postpone the discussion of preestimation issues until section four.

The OLS estimator for model (3.3) is

$$\tilde{\theta}_n = \left[\sum_{i=1}^{n} U_i^2\right]^{-1} \sum_{i=1}^{n} U_i V_i. \tag{3.4}$$

The following result is shown in Appendix A1.

Theorem 3.1 *In Model I, $\tilde{\theta}_n$ is asymptotically normal at rate $\sqrt{nh^3}$. More precisely,*

$$\sqrt{nh^3}(\widetilde{\theta}_n - \theta - h^2 A^{-1} b^I + o_P(h^2)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 A^{-1}\right)$$

with $A = \kappa_2 f(x_0)$ and

$$b^I = \frac{1}{2}(\kappa_4 - \kappa_2^2) k''(x_0) \partial_x f(x_0) + \frac{1}{6}\kappa_4 k^{(3)}(x_0) f(x_0).$$

Remark 3.2: From this result it is obvious that LPR achieves the optimal speed of convergence $\sqrt{nh^3}$, and is oracle efficient because it achieves exactly the one dimensional variance. The bias, however, is not oracle since it contains the term $\kappa_2^2 k''(x_0)\partial_x f(x_0)$. This means that generally the bias may be larger or smaller than that of an oracle estimator. However, in the case of $k^{(3)}(x_0) = 0$ the bias is actually smaller in absolute values as $0 < (\kappa_4 - \kappa_2^2) < \kappa_4$. Note that if, instead of a locally linear model we consider a local quadratic, the first bias term on the first derivative estimator vanishes, and we are left with oracle bias and variance.

More generally, due to the obvious similarities with local polynomial estimators (LPE), the choice of bandwidth and the degree of local polynomial can be treated exactly as for LPEs. It is also intuitively clear that estimators pertaining to distinct x-values should be asymptotically independent. A proof of this is available from the authors upon request. Summarizing, LPR seems particularly well suited for estimating marginal effects in the additive model.

3.2 *Model II*

Now we consider a class of models which can be equally well characterized as generalizations of varying coefficient model (2.2) and as additive models with general semilinear interaction structure. To make

the discussion of the asymptotic properties of an estimator for this model more transparent, we focus on the case when λ is a scalar valued function. We emphasize that this is done for simplicity, the more general case follows straightforwardly. Then, let the model be defined as

$$Y_i = k(X_i) + l(Z_i) + g(X_i)\lambda(Z_i) + \varepsilon_i,\ i = 1, 2, \ldots, \tag{3.5}$$

To analyze this model, we need the following set of

Assumptions for *Model II*

(A1-IIa) The (X_i, Z_i) are iid $\mathbb{R} \times \mathbb{R}^d$-valued random variables with continuous joint density $f(x, z)$. f is twice continuously differentiable with respect to x in a neighborhood of x_0, and there exists a nonnegative Borel function $\gamma(z)$ such that $\int \gamma(z)dz < \infty$ and $\int |\lambda(Z_i)|^4\gamma(z)dz < \infty$ as well as

$$\sup_{|x-x_0|\leq h/2} \left[|f(x, z)| + |\partial_x f(x, z)| + \left|\partial_x^2 f(x, z)\right|\right] \leq \gamma(z)$$

for h sufficiently small. The set $\{z : f(x_0, z) = 0 \text{ and } \partial_x f(x_0, z) \neq 0\}$ has Lebesgue measure zero. The marginal density $f(x_0) > 0$.

(A1-IIb) Let $\pi_\nu(x)$ be a (continuous) version of the conditional expectation $\mathbb{E}\{\lambda(Z_i)^\nu | X_i = x\}$. Then the *identifiability condition* $\pi_2(x) > \pi_1(x)^2$ holds.

(A2-II) = (A2-I)

(A3-III) k and g are three times continuously differentiable in a neighborhood of x_0.

(A4) - (A6) as in section 3.1.

Note the obvious similarities between these assumptions and those discussed previously. The only material novelty is the identifiability condition $\pi_2(x) > \pi_1(x)^2$. This rules out that $\lambda(Z_i)$ is a direct function of X_i only, in which case the whole model would be ill-defined. To analyze the model, we expand again the right hand side of (3.5) about x_0 :

$$Y_i = k(x_0) + k'(x_0)(X_i - x_0) + l(Z_i) + \left[g(x_0) + g'(x_0)(X_i - x_0)\right]\lambda(Z_i) + r_i + \varepsilon_i,$$

where now

$$\begin{aligned} r_i &= \frac{1}{2}k''(x_0)(X_i - x_0)^2 + \frac{1}{3!}k^{(3)}(x_0 + \eta_i(X_i - x_0))(X_i - x_0)^3 \\ &\quad + \left[\frac{1}{2}g''(x_0)(X_i - x_0)^2 + \frac{1}{3!}g^{(3)}(x_0 + \eta_i(X_i - x_0)x_0)(X_i - x_0)^3\right]\lambda(Z_i). \end{aligned}$$

Quasidifferencing yields

$$\begin{aligned} W_iY_i - \mathbb{E}\{W_iY_i|W_iZ_i\} &= k'(x_0)\left[W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_iZ_i\}\right] \\ &\quad + g'(x_0)\left[W_i(X_i - x_0)\lambda(Z_i) - \mathbb{E}\{W_i(X_i - x_0)\lambda(Z_i)|W_iZ_i\}\right] \\ &\quad + W_ir_i - \mathbb{E}\{W_ir_i|W_iZ_i\} + W_i\varepsilon_i - \mathbb{E}\{W_i\varepsilon_i|W_iZ_i\} \end{aligned}$$

since, by Lemma 3.1, the $k(x_0), l(Z_i)$ and $g(x_0)$ terms cancel out. We introduce the following notation:
$\theta = \theta_n(x_0) = (\theta_1, \theta_2)' = (k'(x_0), g'(x_0))'$,
$U_i = U_{ni} = W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_i Z_i\}$,
$\Phi_i = \Phi_{ni} = (U_i, U_i\lambda(Z_i))'$, and $V_i, \widetilde{r}_i, \widetilde{\varepsilon}_i$ are similar to section 3.1. We may then write (3.5) as linear regression

$$V_i = \theta'\Phi_i + \widetilde{r}_i + \widetilde{\varepsilon}_i$$

and consider the OLS estimator

$$\tilde{\theta}_n = \left[\sum_{i=1}^{n} \Phi_i \Phi_i'\right]^{-1} \sum_{i=1}^{n} \Phi_i V_i.$$

The following result is established in Appendix A1.

Theorem 3.2 *In Model II, $\tilde{\theta}_n$ is asymptotically normal at rate $\sqrt{nh^3}$. More precisely,*

$$\sqrt{nh^3}(\tilde{\theta}_n - \theta - h^2 A^{-1} b^{II} + o_P(h^2)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 A^{-1}\right),$$

with

$$A = \kappa_2 f(x_0) \begin{pmatrix} 1 & \pi_1(x_0) \\ \pi_1(x_0) & \pi_2(x_0) \end{pmatrix}$$

and the bias b^{II} being given by (A1.14-II) in Appendix 1.

Remark 3.3: Again, LPR achieves the optimal speed of convergence $\sqrt{nh^3}$. The oracle property is less clear: Certainly, $k'(x_0)$ and $g'(x_0)$ are not estimable separately in general, the exception being the case when $\mathbb{E}\{\lambda(Z_i)|X_i = x\} = 0$, i.e. $\lambda(Z_i)$ and X_i are independent. To deal with this case, we define the restricted oracle property as follows: Suppose an oracle had given us the behavior of all terms *not involving* x. Then it is obvious that our estimator has this restricted oracle property. Extending this result to the multivariate case, we obtain something which is qualitatively very similar to the result obtained by Fan and Zhang (1999). Similar remarks as above apply to the bias term, also regarding second order expansion. Again bandwidth choice and higher order bias reduction are as in the scalar LPEs.

It should be emphasized that we deal here only with the estimation of the derivatives of the functions. In the varying coefficient model, interest may really be centered at the slope coefficient. Hence, *Model II* is perhaps better suited for the estimation of the derivatives of additive models with this type of interaction, where the marginal effects of the Z are already captured by another term.

3.3 *Model III*

As already mentioned, the LPR framework allows to estimate a large class of models out of which we consider a number of specific submodels. For instance, consider the case where in the varying coefficient model (2.2) we allow for an additional additive component. Then,

$$Y_i = k(X_i) + l(Z_i) + g(X_i)P_i + \varepsilon_i, \ i = 1, 2, \ldots, \tag{3.6}$$

This model is, in our nomenklatura, *Model III*. We will discuss its estimation under the following set of assumptions:

Assumptions for *Model III*

(A1-IIIa) The (X_i, Z_i, P_i) are iid $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$-valued random variables with joint density $f(x, z, p)$. f is twice continuously differentiable with respect to x in a neighborhood of x_0, and there exists a nonnegative Borel function $\int\int \gamma(z,p)dpdz < \infty$ such that

$$\sup_{|x-x_0|\leq h/2} \left[|f(x,z,p)| + |\partial_x f(x,z,p)| + \left|\partial_x^2 f(x,z,p)\right|\right] \leq \gamma(z,p)$$

for h sufficiently small. The set $\{z : f(x_0, z) = 0 \text{ and } \partial_x f(x_0, z) \neq 0\}$ has Lebesgue measure zero. The marginal density $fx_0) > 0$, and the joint density $f(x, z)$ is continuous.

(A1-IIIb) P_i has compact support. Let $\pi_\nu(x)$ and $\pi_\nu(x, z)$ be versions of the conditional expectations $\mathbb{E}\{P_i^\nu | X_i = x\}$ and $\mathbb{E}\{P_i^\nu | X_i = x, Z_i = z\}$, resp., and $\sigma_P^2(x) = \mathbb{E}\{[P_i - \pi_\nu(X_i, Z_i)]^2 | X_i = x\}$ (all versions can be chosen as continuous functions if (A1-IIIa) is satisfied). Then

$$\pi_2(x) > \pi_1(x)^2 \text{ and } \sigma_P^2(x) > 0.$$

(A2-III) The ε_i are iid with zero mean and variance σ^2. For every i, ε_i is independent of the σ-algebra $\mathcal{F}_{i-1} = \sigma\{X_1, Z_1, P_1, \ldots, X_i, Z_i, P_i; \varepsilon_1, \ldots, \varepsilon_{i-1}\}$.

(A3-III) k and g are three times continuously differentiable in a neighborhood of x_0.

(A4) - (A6) as in section 3.1.

These are by and large qualitatively similar assumptions as in the previous models. To treat the new variables P_i we have to invoke additional assumptions. In particular, the identifiability condition have to be extended. Note that $\sigma_P^2(x) > 0$, which rules out that P_i is a direct function of X_i and Z_i.

Taking the now familiar steps, we obtain

$$\begin{aligned} W_iY_i - \mathbb{E}\{W_iY_i|W_iZ_i\} &= k'(x_0)\left[W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_iZ_i\}\right] \\ &\quad + g(x_0)[W_iP_i - \mathbb{E}\{W_iP_i|W_iZ_i\}] \\ &\quad + g'(x_0)\left[W_iP_i(X_i - x_0) - \mathbb{E}\{W_iP_i(X_i - x_0)|W_iZ_i\}\right] \\ &\quad + W_ir_i - \mathbb{E}\{W_ir_i|W_iZ_i\} + W_i\varepsilon_i - \mathbb{E}\{W_i\varepsilon_i|W_iZ_i\}, \end{aligned}$$

where now

$$\begin{aligned} r_i &= \frac{1}{2}k''(x_0)(X_i - x_0)^2 + \frac{1}{3!}k^{(3)}(x_0 + \eta_i(X_i - x_0))(X_i - x_0)^3 \\ &\quad + \left[\frac{1}{2}g''(x_0)(X_i - x_0)^2 + \frac{1}{3!}g^{(3)}(x_0 + \zeta_i(X_i - x_0))(X_i - x_0)^3\right] P_i, \end{aligned}$$

with $\eta_i = \eta(X_i) \in (0,1), \zeta_i = \zeta(X_i) \in (0,1)$. Note in particluar that - in remarkable contrast to before - $g(x_0)$ itself is also identified. However, if we were to attempt to only estimate $k'(x_0)$ and $g(x_0)$, and

treat the third expression as the leading bias term, we would end with a persistently biased estimator for $k'(x_0)$. Hence, all three quantities have to estimated jointly. To this end, introduce the following notation:
$\theta = \theta_n(x_0) = (\theta_1, \theta_2, \theta_3)' = (k'(x_0), h^{-1}g(x_0), g'(x_0))'$
$U_i = U_{ni} = W_i(X_i - x_0) - \mathbb{E}\{W_i(X_i - x_0)|W_iZ_i\}$,
$Q_{0i} = Q_{0ni} = h[W_iP_i - \mathbb{E}\{W_iP_i|W_iZ_i\}]$,
$Q_{1i} = Q_{1ni} = W_iP_i(X_i - x_0) - \mathbb{E}\{W_iP_i(X_i - x_0)|W_iZ_i\}$,
$\Phi_i = \Phi_{ni} = (U_i, Q_{0i}, Q_{1i})'$,
and $V_i, \widetilde{r}_i, \widetilde{\varepsilon}_i$ are as in section 3.1. We may then write (3.6) write as linear regression

$$V_i = \theta'\Phi_i + \widetilde{r}_i + \widetilde{\varepsilon}_i.$$

Again, the OLS estimator

$$\tilde{\theta}_n = \left[\sum_{i=1}^n \Phi_i\Phi_i'\right]^{-1} \sum_{i=1}^n \Phi_i V_i.$$

turns out to be asymptotically normal at rate $\sqrt{nh^3}$.

Theorem 3.3 *In Model III, $\tilde{\theta}_n$ is asymptotically normal at rate $\sqrt{nh^3}$. More precisely,*

$$\sqrt{nh^3}(\tilde{\theta}_n - \theta - hA^{-1}b^{III} + o_P(h)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2A^{-1}\right),$$

with

$$A = \kappa_2 f_X(x_0) \begin{pmatrix} 1 & 0 & \pi_1(x_0) \\ 0 & \kappa_2^{-1}\sigma_P^2(x_0) & 0 \\ \pi_1(x_0) & 0 & \pi_2(x_0) \end{pmatrix}$$

and the bias b^{III} being given by (A1.14-III) in Appendix 1.

Remark 3.4: The reader will notice that the expression $\sigma_P^2(x_0)\left[\pi_2(x_0) - \pi_1^2(x_0)\right]$ is (up to a nonzero constant) just the determinant of the matrix A, so that the requirements in assumption (A2-IIIb) constitute the identifiability condition. A consequence is that P cannot be a function of X_i and Z_i only. Hence, in a selection scenario where P is for instance the Mill's ratio, this constitutes an exclusion restriction. As above, LPR achieves the optimal speed of convergence, i.e. $\sqrt{nh^3}$ for the derivatives and $\sqrt{nh}$ for the function. For the derivatives, as in *Model II* the restricted oracle property holds. In contrast, the oracle property for the direct estimator of the function g is not clear: $\sigma^2\left(f_X(x_0)\sigma_P^2(x_0)\right)^{-1}$ is certainly greater than the variance of an one dimensional function. Again, bandwidth choice and higher order bias reduction are as in the scalar LPEs.

For applications, there are two key differences compared to *Model II.* First, we obtain an estimator for the function g itself. Second, we need to condition on Z only. This might be useful in cases where we have a large number of regressors. Since this model offers a way to estimate the function g, it may be directly compared to the result of Fan and Zhang (1999), who consider estimation of the model

without the additive component l. Indeed, these models are better compared using a hybrid model, where we allow not just for $g(X_i)P_i$ as in *Model III* but also for a $d(X_i)Z_i$ interaction term as in *Model II*. Since estimation of this term only affects the derivatives, the asymptotic distribution of the function g is not affected. Then, the variance of the Fan and Zhang estimator depends on the inverse of the second moment matrix of Z_i and P_i, conditional on X_i. In case of scalar random variables, the key difference is between $\left(\sigma_P^2(x_0)\right)^{-1}$ and $\left(\pi_2(x_0) - \mathbb{E}\{Z_iP_i|X_i = x_0\}^2/\mathbb{E}\{Z_i^2|X_i = x_0\}\right)^{-1}$, and there is no clear ranking between the two

3.4 *Model IV*

In this subsection we consider the case of unrestricted pairwise interaction terms, i.e. the model is given by

$$Y_i = k(X_i) + l(Z_i) + g(X_i, Z_{1i}) + \varepsilon_i,\ i = 1, 2, \ldots, \tag{3.7}$$

where all functions are assumed to be smooth and unknown. Again we concentrate on the X_i without loss of generality, since the model is completely symmetric. Moreover, we shall focus for simplicity on a model which has only one of these d interaction terms, say, the first (denoted by $g(X_i, Z_{1i})$). Since only the second cross derivative (denoted by g_{xz}) is identified, we have to expand all functions to higher order. In our discussion we focus on the estimation of this cross derivative.

Remark 3.5: In this paragraph, in order to keep the formulas more transparent, we shall use the notation g_x, g_z, g_{xx}, etc. for the partial derivatives and g_x^0, g_z^0, g_{xx}^0, etc., for the partial derivatives evaluated at (x_0, z_{10}).

Assumptions for *Model IV*:

(A1-IVa) The (X_i, Z_i) are iid $\mathbb{R} \times \mathbb{R}^d$-valued random variables with continuous joint density $f(x, z)$. f is twice continuously differentiable with respect to x in a neighborhood of x_0 (for all z) and there exists a nonnegative Borel function $\gamma(z)$ such that $\int \gamma(z)dz < \infty$ and

$$\sup_{|x-x_0|\leq h/2} \left[|f(x, z)| + |\partial_x f(x, z)| + \left|\partial_x^2 f(x, z)\right|\right] \leq \gamma(z)$$

for h sufficiently small. The set $\{z : f(x_0, z) = 0 \text{ and } \partial_x f(x_0, z) \neq 0\}$ has Lebesgue measure zero. The marginal density $f(x_0, z_{10}) > 0$.

(A1-IVb) $f(x, z_1)$ (the joint density of (X_i, Z_{1i})) is three times continuously differentiable with respect to z_1 in a neighborhood of (x_0, z_{10}).

(A2-IV) = (A2-I)

(A3-IV) k and g are three times continuously differentiable in a neighborhood of x_0 and (x_0, z_{10}), respectively.

(A4) - (A5) as in section 3.1.

(A6-IV) $nh_n^4 \to \infty$.

We start with a third order expansion about (x_0, z_{10}) :

$$\begin{aligned} Y_i &= k(x_0) + k'(x_0)(X_i - x_0) + \frac{1}{2}k''(x_0)(X_i - x_0)^2 + l(Z_i) + g^0 \\ &+ g_x^0(X_i - x_0) + g_z^0(Z_{1i} - z_{10}) + \frac{1}{2}g_{xx}^0(X_i - x_0)^2 \\ &+ g_{xz}^0(X_i - x_0)(Z_{1i} - z_{10}) + \frac{1}{2}g_{zz}^0(Z_{1i} - z_{10})^2 + r_i + \varepsilon_i. \end{aligned} \tag{3.8}$$

r_i contains all third order terms $(X_i - x_0)^\mu (Z_{1i} - z_{10})^\nu$ with $\mu + \nu = 3$. Next, we perform the by now familiar quasidifferencing step. This time, however, we have to multiply (3.8) by a joint kernel in (x, z), i.e. by

$$W_i = \frac{1}{h}K\left(\frac{X_i - x_0}{h}\right)\frac{1}{h}K_1\left(\frac{Z_{1i} - z_{10}}{h}\right) = W_i^x W_i^z$$

(with $K(u)$ again the uniform kernel). $K_1(z)$ may be any symmetric kernel with, for simplicity, compact support. Denote the moments of the squared kernel by κ'_m :

$$\kappa'_m = \int s^m K_1(s)^2 ds.$$

In particular, $\kappa'_m = 0$ for m odd. Conditional expectations continue to be taken w.r. to $W_i Z_i$. As a result of quasidifferencing we obtain

$$V_i = \left[k'(x_0) + g_x^0\right] U_i + \frac{1}{2}\left[k''(x_0) + g_{xx}^0\right] U_{2i} + g_{xz}^0 U_i (Z_{1i} - z_{10}) + \widetilde{r}_i + \widetilde{\varepsilon}_i, \tag{3.9}$$

where U_i, V_i are defined as in the previous models and
$U_{2i} = W_i(X_i - x_0)^2 - \mathbb{E}\{W_i(X_i - x_0)^2 | W_i Z_i\}$. Introducing
$\theta = (\theta_1, h\theta_2, h\theta_3)' = (\ k'(x_0) + g_x^0, \frac{h}{2}\left[k''(x_0) + g_{xx}^0\right], h g_{xz}^0)'$,
$\Phi_i = (U_i, h^{-1}U_{2i}, h^{-1}U_i(Z_{1i} - z_{10}))'$
(note the dependence of θ on n), (3.9) may be written as linear regression model

$$V_i = \theta' \Phi_i + \widetilde{r}_i + \widetilde{\varepsilon}_i. \tag{3.10}$$

Our interest is again in the OLS estimator for (3.10):

$$\tilde{\theta}_n = \left[\sum_{i=1}^n \Phi_i \Phi_i'\right]^{-1} \sum_{i=1}^n \Phi_i V_i. \tag{3.11}$$

The following result is established in Appendix 1.

Theorem 3.4. *In Model IV, the OLS-estimator is asymptotically normal at rate* $\sqrt{nh^4}$*.More precisely,*

$$\sqrt{nh^4}\left(\tilde{\theta}_n - \theta - O_P(h^2)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 A^{-1})$$

with

$$A = \begin{pmatrix} \kappa_2\kappa_0' f(x_0, z_{10}) & 0 & 0 \\ 0 & \left(\kappa_4 - \kappa_2^2\right)\kappa_0' f(x_0, z_{10}) & 0 \\ 0 & 0 & \kappa_2\kappa_2' f(x_0, z_{10}) \end{pmatrix}.$$

Remark 3.6: The features of the result show the familiar pattern: First, the optimal speed of convergence is obtained, i.e. $\sqrt{nh^4}$ for the partial derivative of a two dimensional function and $\sqrt{nh^6}$ for the second derivatives. Moreover, the variances are of optimal order and depend only on the joint distribution of X_i and Z_{1i}. The first and third coefficient hence exhibit oracle efficiency. However, the constant in the second expression is smaller (due to κ_2^2), so that the variance of the second coefficient appears to be inflated. We do not state the bias, as it is a complicated and intransparent expression. However, the key features remain preserved, namely that it has the optimal rate, and with appropriate degree of bias reduction we may actually achieve a bias which is oracle.

3.5 Model V

In this section we give a short sketch on how *Model V* could be estimated. Having obtained an estimate for $g(X_i, Z_{1i})$ in the previous subsection, there is only one question remaining, namely how to obtain estimates for the two multiplicative components. Note that $g'(x_0)$ and $q'(z_{1,0})$ are only identified up to a multiplicative constant. If economic theory does not prescribe otherwise, let $\mathbb{E}\left[g'(X_i)\right] = 1$. In this case, the estimation procedure of the previous section may simply be augmented by another step, which is motivated by forming the sample analogue to
$\mathbb{E}\left[g_{xz_1}(X_i, z_{1,0})\right] = \mathbb{E}\left[g'(X_i)\right] q'(z_{1,0}) = q'(z_{1,0})$,
i.e.

$$\widehat{q'(z_{1,0})} = \frac{1}{n}\sum_i h^{-1}\hat{\theta}_3(X_i, z_{1,0}).$$

Note that this integration step opens the way for a slight inefficiency. In particular, the variance may depend upon the joint distribution of X_i and $Z_{1,i}$. More involved estimation procedures that circumvent this disadvantage are imaginable using LPR, but we do not pursue this issue further.

4 Models - LPR based Feasible Estimation

Throughout this section we focus on preestimation of the conditional expectations in the basic additive model. This is however not restrictive: All other estimators share these key parts of the basic additive model, and hence effects of preestimation on their asymptotic behavior can be analyzed using very similar arguments.

In comparison to section 3, we will invoke the following additional assumptions:

(A7) The regressors have compact support.

(A8)

$$\gamma(x, z) = \frac{\partial_x f_{XZ}(x, z)}{f_{XZ}(x, z)} + \frac{x\partial_x^2 f_{XZ}(x, z)}{2f_{XZ}(x, z)}$$

is bounded on the compact support, i.e. uniformly bounded. Also, the second partial derivative of γ w.r.t. x, and cross derivatives $\partial_x^2\partial_z$ are bounded.

Returning to the basic specification, the resulting *feasible LPR estimator* is given by

$$\hat{\theta}_n = \left[\sum_{i=1}^n \hat{U}_i^2\right]^{-1} \sum_{i=1}^n \hat{U}_i\hat{V}_i,$$

with $\hat{U}_i = W_iX_i - \widehat{\mathbb{E}}[W_iX_i|W_iZ_i]$ and $\hat{V}_i = W_iY_i - \widehat{\mathbb{E}}[W_iY_i|W_iZ_i]$, where the hat denotes now appropriate estimators for the theoretical quantities $\mathbb{E}[W_iX_i|W_iZ_i]$ and $\mathbb{E}[W_iY_i|W_iZ_i]$. The question is now how the properties of an estimator that uses suitable replacements for $\mathbb{E}[W_iX_i|W_iZ_i]$ and $\mathbb{E}[W_iY_i|W_iZ_i]$ compare to those of

$$\tilde{\theta}_n = \left[\sum_{i=1}^n U_i^2\right]^{-1} \sum_{i=1}^n U_iV_i,$$

the theoretical estimator. To start with, rewrite

$$\begin{aligned}\hat{U}_i\hat{V}_i &= \Big((W_iX_i - \mathbb{E}[W_iX_i|W_iZ_i]) + \Big(\mathbb{E}[W_iX_i|W_iZ_i] - \widehat{\mathbb{E}}[W_iX_i|W_iZ_i]\Big)\Big) \times \\ &\quad \Big((W_iY_i - \mathbb{E}[W_iY_i|W_iZ_i]) + \Big(\mathbb{E}[W_iY_i|W_iZ_i] - \widehat{\mathbb{E}}[W_iY_i|W_iZ_i]\Big)\Big).\end{aligned}$$

Then, using Lemmas in the appendix, we specify $\widehat{\mathbb{E}}[W_iY_i|W_iZ_i]$ as $W_i\widehat{m}_{-i}(x_0, Z_i)$, where $\widehat{m}_{-i}(x,z) = \widehat{m}_{-i,n}(x,z)$ is a *leave one out estimator* Nadaraya Watson estimator, i.e. $\widehat{m}_{-i}(x_0, Z_i) = \left(\sum_{j\neq i} K_j\right)^{-1} \sum_{j\neq i} K_jY_j$, where $K_j = K\left((X_j - x_0)/H_0, (Z_j - Z_i)/H_0\right)/H_0^{d+1}$, where H_0 is the bandwidth in this regression, and r_0 denotes the order of the Kernel in this regression. It is straightforward that $\widehat{\mathbb{E}}[W_iY_i|W_iZ_i]$ is a function of all the data save the i-th, and of Z_i (Recall that we are interested in $\widehat{m}_{-i}(x_0, Z_i)$). Hence, by the same arguments as in Lemma 3.1, $S_i = W_i\left(m(x_0, Z_i) - \widehat{m}_{-i}(x_0, Z_i)\right)$ is $\mathcal{F}_n$ measurable, where

$$\mathcal{F}_n = \sigma(W_iX_1, .., W_iX_{i-1}, W_iX_{i+1}, .., W_iX_n, W_iZ_1, .., W_iZ_n, W_iY_1, .., W_iY_{i-1}, W_iY_{i+1}, .., W_iY_n).$$

Moreover, using Lemma A1.1, we obtain $\mathbb{E}[W_iX_i|W_iZ_i] = W_i\left[x_0 + h^2\gamma(x_0, Z_i) + O_p(h^4)\right]$, where γ is defined in A8 above. This suggests using $\widehat{\mathbb{E}}[W_iX_i|W_iZ_i] = W_i\left[x_0 + h^2\widehat{\gamma}(x_0, Z_i)\right]$, where $\widehat{\gamma}(x_0, Z_i)$ is given by

$$\hat{\gamma}(x,z) = \frac{\widehat{\partial_x f_{XZ}}(x,z)}{\widehat{f_{XZ}}(x,z)} + \frac{x\widehat{\partial_x^2 f_{XZ}}(x,z)}{2\widehat{f_{XZ}}(x,z)},$$

and the hats denote standard estimators for a density and their derivative, e.g. $\widehat{\partial_x f_{XZ}}(x,z) = (nH_1)^{-1}\sum_{j\neq i}\partial_x K_j$, where $\partial_x K_j = \partial_x K\left((X_j - x_0)/H_1, (Z_j - Z_i)/H_1\right)/H_1^{d+1}$, H_1 is again the bandwidth, and the order of the Kernel here is denoted as r_1. Again we employ a leave one out estimator, so that

$$\mathbb{E}[W_iX_i|W_iZ_i] - \widehat{\mathbb{E}}[W_iX_i|W_iZ_i] = h^2G_i + O_p(h^4),$$

where $G_i = h^2 W_i \left[\gamma(x_0, Z_i) - \widehat{\gamma}(x_0, Z_i)\right]$. As above, G_i is also $\mathcal{F}_n$ measurable. Then, rewrite

$$\frac{1}{nh}\sum_{i=1}^{n}\left[\hat{U}_i\hat{V}_i - U_iV_i\right] = \frac{1}{nh}\sum_{i=1}^{n}A_{ni} + \frac{h}{n}\sum_{i=1}^{n}G_iS_i + \frac{1}{nh}\sum_{i=1}^{n}\xi_i,$$

where $A_{ni} = \left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i + h^2\left(W_iY_i - \mathbb{E}\left[W_iY_i|W_iZ_i\right]\right)G_i$, and ξ_i contains terms of higher order in h. Similarly,

$$\frac{1}{nh}\sum_{i=1}^{n}\left[\hat{U}_i^2 - U_i^2\right] = \frac{2}{nh}\sum_{i=1}^{n}\left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i + \frac{1}{nh}\sum_{i=1}^{n}S_i^2 + \frac{1}{nh}\sum_{i=1}^{n}\psi_i,$$

where ψ_i contains now terms of higher order in h. Our argumentation will proceed as follows: Start by concentrating on the numerator, more precisely on the question under what conditions $(nh)^{-1}\sum_{i=1}^{n}\left[\hat{U}_i\hat{V}_i - U_iV_i\right] = o_p\left(h^2\right)$. This condition arises, because then the error made in preestimation converges faster than the leading bias term. We first establish the conditions under which $B_n = hn^{-1}\sum_{i=1}^{n}G_iS_i = o_p\left(h^2\right)$, i.e. this term vanishes faster than the leading bias term. This will be done using a von Mises type functional expansion. Then we will apply alternative arguments to $(nh)^{-1}\sum_{i=1}^{n}A_{ni}$ to show that this term is $o_p\left(h^2\right)$ under general conditions. The result is summarized in the following Lemma, whose proof is transferred to the appendix.

Lemma 4.1: *Let the model be as defined above and assumptions A1-A8 hold. Then follows that*

$$\frac{1}{nh}\sum_{i=1}^{n}\left[\hat{U}_i\hat{V}_i - U_iV_i\right] = o_p\left(h^2\right),$$

provided $n^{2/7}H_0^{r_0}H_1^{r_1-2} + n^{-5/7}H_0^{-(d+1)/2}H_1^{-(d+5)/2}\ln(n) = o(1)$.

Remark 4.1: We allow for different degrees of bias reduction r_0 and r_1. The background is that in the case of $\gamma(x_0, Z_i)$, we have to estimate second derivatives and need therefore at least $r_1 = 4$. In contrast, the other error element results from estimating the m function, and hence $r_0 \geq 2$ is possible. Choosing $r_0 = 2$ and $r_1 = 4$, for the above condition to hold Z must be less than 10.

By closer inspection of the denominator, we see that this expression is exactly of the same structure. The first rhs term has already been established in Lemma 4.1. The main difference comes from the second term which contains a squared error S_i. This implies that the term that is slowest to converge, i.e. $(nh)^{-1}\sum_{i=1}^{n}S_i^2$, converges roughly h^{-2} slower than $hn^{-1}\sum_{i=1}^{n}G_iS_i$. However, in the denominator all that we have to require is really consistency, i.e. $(nh)^{-1}\sum_{i=1}^{n}\left[\hat{U}_i^2 - U_i^2\right] = o_p(1)$, so that this effect balances. In fact, the faster uniform speed of convergence of the squares of an estimator for the mean, compared to product of estimators for mean and derivatives in the numerator guarantees that the denominator is less problematic. Collecting the results in the following theorem

Theorem 4.2: *Let the model be as defined above and assumptions A1-A8 hold. Then follows that*

$$\tilde{\theta}_n - \hat{\theta}_n = o_p\left(h^2\right),$$

provided $n^{2/7}H_0^{r_0}H_1^{r_1-2} + n^{2/7}H_0^{2r_0} + n^{-5/7}H_0^{-(d+1)/2}H_1^{-(d+5)/2}\ln(n) + n^{-5/7}H_0^{-(d+1)}\ln(n) = o(1)$.

5 Monte Carlo

The small sample performance of LPR based estimators is best analyzed by a Monte Carlo simulation study. We focus on two things. First, we give a performance comparison between LPR and Marginal Integration. Second, we scrutinize the performance of LPR based estimation methods for varying data sizes. Our focus will be on the additive model. For sake of concreteness, we specify the DGP as follows

$$\begin{aligned} Y_i &= m(X_i, Z) + U_i \\ &= X_i + \frac{1}{3}X_i^3 + 2\ln\left((X_i+1)^2+1\right) - 1,5Z_{1i} + |Z_{2i}| + Z_{3i}Z_{4i} + 1,5Z_{5i} + Z_{6i} + U_i, \end{aligned}$$

and let $\beta(X_i) = 1 + X_i^2 + 4\left(X_i+1\right) / \left((X_i+1)^2+1\right)^{-1}$ denote the partial derivative wrt x. To begin with, the (X_i, Z_i, U_i) are assumed to be iid from $N(0,5)$, and are mutually independent. Note that this scenario is favorable to Marginal Integration estimators, as their performance is known to suffer extremely, if the regressors are dependent. In selecting the models, we chose the following sum of squared distances as criterion: Let k denote the positions at which the estimator was evaluated. Then

$$SSD = \sum_{k=1,..,K} \left(\hat{\beta}(x_k) - \beta(x_k)\right)^2.$$

Since the points were chosen to be equidistant, this corresponds roughly to the ISE for the derivative. K, the total number of points, will be set to 50. As means of displaying the results, we will show graphs of LPR based estimators for the derivative for $n = 500, 2000$ and 5000 observations. and performance comparison with a Marginal Integration estimator for $n = 1500$. The MI estimator is specified as

$$\begin{aligned} \hat{\beta}(x_k, z_k) &= \frac{1}{h}\hat{f}(x_k, z_k)^{-2}\left(\hat{f}(x_k, z_k)\hat{g}'(x_k, z_k) - \hat{f}'(x_k, z_k)\hat{g}(x_k, z_k)\right), \\ \hat{\beta}_{MI}(x_k) &= \frac{1}{n}\sum_{i=1,,.,n} \hat{\beta}(x_k, Z_i) \end{aligned}$$

where $\hat{f}(x_k, z_k) = \sum_i K_i$, $\hat{g}(x_k, z_k) = \sum_i K_i Y_i$, $\hat{f}'(x_k, z_k) = \sum_i \partial_x K_i$ and $\hat{g}(x_k, z_k) = \sum_i \partial_x K_i Y_i$, where K_i is a 7-variate product kernel, and $\partial_x K_i$ is the product of a sixth variate product kernel, and one derivative of a kernel. Moreover, we use leave one out estimators at every point.

The procedure is now as follows. We draw 200 times from the above model. Then we order the realizations according to their SSD. Finally, we will display graphs that are ordered according to quantiles of their SSD. For all estimators, an optimal bandwidth is chosen beforehand by doing a gridsearch with respect to finding the bandwidth that minimizes the average SSD in 100 replications.

A main issue and criticism of nonparametrics is that it is not applicable in datasets usually encountered in applications. To address this issue we show results of an LPR based estimator for varying data sizes. The results can be found in figure 1, which shows the results for $n = 5000$ and figure 2, which show the results for $n = 500$ and $n = 2000$. The results using 5000 observations clearly indicate that this is sufficient data to estimate the model, see fig.1. In the upper lhs picture we see the 0.1 and 0.9 percentile SSD curves. Both track the DGP quite closely. The same is true for the quartiles given in the upper rhs picture. What is even more surprising is that even the worst realization still

captures the whole structure quite well, and shows only little deviations in the local feature build in around $x = -1$. In every realization, the fit in the slopes is almost perfect.

In contrast with $n = 500$ there are some clearly visible differences, see fig. 2. While even the bad realizations still capture the broad picture, deviations in certain areas are quite significant. Nevertheless, the median and 0.75 as well as 0.9 quantile still come relatively close to the true DGP, given the high number of regressors and the sparsity of the data.

Using $n = 2000$ observations generally yields a satisfactory picture. In applications, this seems to be a reasonable lower bound given the number of regressors. Hence, we will do some performance comparison with MI estimator with n=2000. Fig. 3 shows their behavior. Generally speaking, MI estimators still capture the broader picture, but miss out on the details. In this respect their visual performance is similar to the LPR with $n = 500$. However, by closer inspection we find that it is especially in the tails, where the MI estimators are quite inaccurate, which is perhaps not so prominent when looking at the graphs. Note that the graph looks somewhat rugged. However, if we increase the bandwidth, the estimator misses the local feature at $x = -1$.

This problem in the tails is also reflected in the significantly worse performance. The following table shows the SSD at respective quantiles

$Estimator \backslash Quantile$	0.1	0.25	0.5	0.75	0.9
MI	77,917	55,944	44,864	31,089	21,307
LPR	20,579	17,781	13,651	11,521	7,429

As is obvious, the LPR outperforms the MI estimator by a factor of between 3 and 4. Finally, note that even the worse LPR realizations are outperforming MI based estimators, judged by the SSD.

6 Outlook

This paper is concerned with the introduction of a new estimation principle for semi- and nonparametric regression. It is established that this principle yields in a straightforward fashion optimal estimators for a wide class of models. In particular we would like to emphasize the simplicity of the approach as a result of it's local polynomial structure, which allows us also to adapt familiar solutions for higher order bias reduction and choice of bandwidth.

An interesting question is that of extension to other closely related models. The construction of estimators for generalized models of the same class can be performed by noting the approximate methods of moments structure of the estimator. This may yield estimators with link functions suitable for, e.g., binary choice or duration models. Another interesting direction for future research concerns the construction of estimators for regression quantiles, where the conditional quantile is of the form of the models in our class. Additive models containing indices may be another area of application for this principle. Finally, for some applications like control function models, the issue of how generated or pre-estimated regressors impact the asymptotic distribution of the estimator may also be an interesting question left open for future research.

In the second section we gave an extensive, but necessarily incomplete survey of potential applications. Moreover, in earlier work (Hoderlein (2002)), we applied a preliminary version of a LPR based estimator to demand analysis. In this application, the overall performance of LPR estimators was very satisfactory and revealed interesting features parametric models missed. Finding additional fruitful areas of applications will in the end determine the usefulness of this very general nonparametric estimation principle.

7 Appendix 1

We shall first establish the results in section 3.2, of which the corresponding results for the models in considered in (3.1) and (3.3) are obtained as special cases. In the following Lemmas, we shall suppress the index i, so that $(X, Z, P) \sim (X_i, Z_i, P_i)$ and $W \sim W_i$. Assumptions (A1-III) - (A3-III) together with (A4) - (A7) are assumed to be in effect without further mentioning. First we have to establish a series of auxiliary Lemmata.

7.1 Auxiliary Lemmata

Lemma A1.1. *Let $\varphi(x, z, p)$ be a Borel function s.t. $W\varphi(X, Z, P)$ is integrable. Then a version of the conditional expectation of $W\varphi(X, Z, P)$ given WZ is given by*

$$\mathbb{E}\left\{W\varphi(X, Z, P)|WZ = \zeta\right\} = \frac{1}{h}\overline{\varphi}(h\zeta),$$

where

$$\overline{\varphi}(z) = \frac{p_\varphi(z)}{q(z)} 1_{\{z\neq 0\}} 1_{\{q(z)>0\}}, \tag{A1.1a}$$

$$p_\varphi(z) = \int\int_{x_0-h/2}^{x_0+h/2} \varphi(x, z, p) f(x, z, p) dx dp, \tag{A1.1b}$$

$$q(z) = \int\int_{x_0-h/2}^{x_0+h/2} f(x, z, p) dx dp = \int_{x_0-h/2}^{x_0+h/2} f(x, z) dx. \tag{A1.1c}$$

Proof. Let K stand for $K\left(\frac{X-x_0}{h}\right)$, i.e. $K = hW$. For every Borel set B

$$\begin{aligned}
\int_{KZ\in B} \overline{\varphi}(KZ)d\mathbb{P} &= \int_{KZ\in B} 1_{\{KZ\neq 0\}} \frac{p_\varphi(Z)}{q(Z)} 1_{\{q(Z)>0\}} d\mathbb{P} \\
&= \int_{Z\in B} 1_{\{K=1\}} \frac{p_\varphi(Z)}{q(Z)} 1_{\{q(Z)>0\}} d\mathbb{P} \\
&= \int_{Z\in B, |X-x_0|\leq h/2} \frac{p_\varphi(Z)}{q(Z)} 1_{\{q(Z)>0\}} d\mathbb{P} \\
&= \int_B \frac{p_\varphi(z)}{q(z)} 1_{\{q(z)>0\}} \left[\int_{x_0-h/2}^{x_0+h/2} \int f(x,z,p)dpdx\right] dz \\
&= \int_B p_\varphi(z) 1_{\{q(z)>0\}} dz \\
&= \int_B p_\varphi(z)dz \\
&= \int_B \int \int_{x_0-h/2}^{x_0+h/2} \varphi(x,z,p)f(x,z,p)dxdpdz \\
&= \int_{Z\in B} 1_{\{K=1\}} \varphi(X,Z,P)d\mathbb{P} \\
&= \int_{KZ\in B} K\varphi(X,Z,P)d\mathbb{P}.
\end{aligned}$$

where the second equality comes from (A4) and the sixth equality holds since $p_\varphi(z) = 0$ on $\{q(z) = 0\}$. This shows

$$\mathbb{E}\{K\varphi(X,Z,P)|KZ\} = \overline{\varphi}(KZ),$$

from which the assertion follows. ■

As a consequence,

$$\begin{aligned}
\mathbb{E}\left\{W\varphi(X,Z,P)|WZ\right\} &= \frac{1}{h}\frac{p_\varphi(Z)}{q(Z)} 1_{\{|X-x_0|\leq h/2\}} 1_{\{q(Z)>0\}} \\
&= W\frac{p_\varphi(Z)}{q(Z)} 1_{\{q(Z)>0\}} \\
&= W\overline{\varphi}(Z),
\end{aligned}$$

where again use has been made of (A4). Note that $1_{\{|X-x_0|\leq h/2\}} 1_{\{q(Z)>0\}} = 1_{\{|X-x_0|\leq h/2\}}$ a.s. since $\mathbb{P}(|X - x_0| \leq h/2, q(Z) = 0) = 0$.

Remark A.1: Note also that, for φ independent of p, (A1.1b) reduces to

$$p_\varphi(z) = \int_{x_0-h/2}^{x_0+h/2} \varphi(x,z)f(x,z)dx. \qquad \text{(A1.1b')}$$

Henceforth, let us use the shorthand notation $\overline{\zeta} = \mathbb{E}\left\{\zeta|WZ\right\}$.

Lemma A1.2. *Denote $\xi = W(X - x_0)^\mu \psi(Z,P)$, with an integer $\mu \geq 0$* and a measurable function $\psi(z,p)$ s.t. $\int\int |\psi(z,p)|\gamma(z,p)dzdp < \infty$ *(with γ from assumption (A1-IIIa)). Then*

$$\mathbb{E}\xi = \begin{cases} h^\mu \kappa_\mu \int\int \psi(z,p)f(x_0,z,p)dzdp + O(h^{\mu+2}), & \mu \ \textit{even}, \\ h^{\mu+1}\kappa_{\mu+1} \int\int \psi(z,p)\partial_x f(x_0,z,p)dzdp + O(h^{\mu+2}), & \mu \ \textit{odd}, \end{cases}$$

where $\kappa_\mu = \int_{-1/2}^{1/2} s^\mu ds = \frac{1}{\mu+1}\left(\frac{1}{2}\right)^\mu$ *denotes the* μ *-th noncentral moment of the uniform kernel.*

Proof.

$$\begin{aligned}
\mathbb{E}\xi &= \frac{1}{h}\int\int\int K\left(\frac{x-x_0}{h}\right)(x-x_0)^\mu \psi(z,p) f(x,z,p)dxdzdp \\
&= h^\mu \int\int\int_{-1/2}^{1/2} s^\mu \psi(z,p) f(x_0+hs,z,p)dsdzdp \\
&= h^\mu \int\int\int_{-1/2}^{1/2} s^\mu \psi(z,p)\left[f(x_0,z,p)+\partial_x f(x_0,z,p)hs\right. \\
&\qquad\qquad \left. +\frac{1}{2}\partial_x^2 f(x_0+\eta hs,z,p)h^2 s^2\right]dsdzdp \\
&= \begin{cases} h^\mu \kappa_\mu \int\int \psi(z,p) f(x_0,z,p)dzdp + R_h, & \mu \text{ even}, \\ h^{\mu+1}\kappa_{\mu+1}\int\int \psi(z,p)\partial_x f(x_0,z,p)dzdp + R_h, & \mu \text{ odd}. \end{cases}
\end{aligned}$$

The remainder term R_h may be estimated as follows:

$$\begin{aligned}
|R_h| &= \frac{1}{2}h^{\mu+2}\left|\int\int\int_{-1/2}^{1/2} s^{\mu+2}\psi(z,p)\frac{\partial^2}{\partial^2 x}f(x_0+\eta hs,z,p)dsdzdp\right| \\
&\le \frac{1}{2}h^{\mu+2}\left|\int\int\int_{-1/2}^{1/2} |s|^{\mu+2}|\psi(z,p)| \sup_{|x-x_0|\le h/2}\left|\frac{\partial^2}{\partial x^2}f(x,z,p)\right| dsdzdp\right| \\
&\le \frac{1}{2}h^{\mu+2}\int_{-1/2}^{1/2}|s|^{\mu+2}ds\int\int|\psi(z,p)|\gamma(z,p)dzdp \\
&= O(h^{\mu+2}).
\end{aligned}$$

The $\eta \in (0,1)$ in the Taylor expansion actually depends on (z,p) and can be chosen so as to be measurable. (A similar remark applies to all Taylor expansions below without further mentioning.) This shows (A1.1). ■

Remark A.2: For even μ, we may write

$$\mathbb{E}\xi = h^\mu \kappa_\mu \pi_\psi(x_0) f(x_0) + O(h^{\mu+2}),$$

since $\pi_\psi(x) = \int\int \psi(z,p) f(x,z,p)dzdp/f(x)$ is a continuous version of the conditional expectation $\mathbb{E}\{\psi(Z,P)|X=x\}$. For μ odd,

$$\mathbb{E}\xi = h^{\mu+1}\kappa_{\mu+1}\partial_x\left[\pi_\psi(x)f(x)\right]_{x_0} + O(h^{\mu+2}).$$

Lemma A1.3. *Denote* $\xi = W(X-x_0)^\mu\psi(Z,P), \eta = W(X-x_0)^\nu\chi(Z,P)$. *Assume that* $\int|\chi(z,p)|\gamma(z,p)dp < \infty$ *as well as* $\int\overline{\chi}(z)\int|\psi(z,p)|\,\gamma(z,p)dpdz < \infty$, *where* $\overline{\chi}(z) = \sup_{p\in\mathrm{supp}(P)}|\chi(z,p)|$. *Then*

$$\begin{aligned}
&\mathbb{E}\xi\overline{\eta} = \mathbb{E}\overline{\xi}\overline{\eta} = \mathbb{E}\overline{\xi}\eta \\
&= \begin{cases} h^{\mu+\nu}\kappa_{\mu+1}\kappa_\nu B_{odd} + h^{\mu+\nu}o(1) & \text{for } \mu \text{ odd}, \\ h^{\mu+\nu-1}\kappa_\mu\kappa_\nu B_{even} + h^{\mu+\nu-1}o(1) & \text{for both } \mu \text{ and } \nu \text{ even}. \end{cases}
\end{aligned} \tag{A1.2}$$

The constant B_{odd} is given by

$$B_{odd} = \int \pi_\chi(x_0, z)\partial_x \left[\pi_\psi(x, z)f(x, z)\right]_{x_0} dz,$$

and the constant B_{even} by

$$B_{even} = \int \pi_\chi(x_0, z)\pi_\psi(x_0, z)f(x_0, z)dz.$$

Here $\pi_\psi(x, z)$ denotes a version of the conditional expectation $E\{\psi(Z, P)|X = x_0, Z = z\}$.

Remark A.3: The assumptions and hence the formulation of the assertion are asymmetric in μ and ν. Of course, the same result obtains with the roles of μ and ν (and ψ and χ) interchanged.

Proof. Note first that $\mathbb{E}\xi\overline{\eta} = \mathbb{E}\{\overline{\eta}\mathbb{E}\{\xi|WZ\}\} = \mathbb{E}\overline{\xi}\overline{\eta} = \mathbb{E}\overline{\xi}\eta$. By Lemma A.1, with $\varphi(x, z, p) = (x - x_0)^\mu\psi(z, p)$,

$$\begin{aligned}
\overline{\xi} &= W\overline{\varphi}(Z) = \frac{1}{h}\frac{p_\varphi(Z)}{q(Z)}1_{\{|X-x_0|\le h/2\}}1_{\{q(Z)>0\}} \\
&= \frac{1}{h}q(Z)^{-1}1_{\{q(Z)>0\}}1_{\{|X-x_0|\le h/2\}} \\
&\qquad \int\int_{x_0-h/2}^{x_0+h/2}(x - x_0)^\mu\psi(z, p)f(x, z, p)dxdp
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}\overline{\xi}\eta &= \frac{1}{h^2}\int_{\{q(z)>0\}} q(z)^{-1}\left[\int\int_{x_0-h/2}^{x_0+h/2}(x - x_0)^\mu\psi(z, p)f(x, z, p)dxdp\right] \\
&\qquad \left[\int\int_{x_0-h/2}^{x_0+h/2}(x - x_0)^\nu\chi(z, p)f(x, z, p)dxdp\right]dz \\
&= h^{\mu+\nu}\int_{\{q(z)>0\}} q(z)^{-1}\left[\int\psi(z, p)\int_{-1/2}^{1/2}s^\mu f(x_0 + hs, z, p)dsdp\right] \\
&\qquad \left[\int\int_{-1/2}^{1/2}t^\nu\chi(z, p)f(x_0 + ht, z, p)dtdp\right]dz. \qquad \text{(A1.3)}
\end{aligned}$$

For μ odd, expanding $f(x_0 + hs, z, p)$ up to second order,

$$\begin{aligned}
I_\mu(z, p) &= \int_{-1/2}^{1/2}s^\mu f(x_0 + hs, z, p)ds \\
&= \int_{-1/2}^{1/2}s^\mu\left[f(x_0, z, p) + \partial_x f(x_0, z, p)hs\right. \\
&\qquad \left. + \frac{1}{2}\partial_x^2 f(x_0 + \eta hs, z, p)h^2s^2\right]ds \\
&= h\kappa_{\mu+1}\partial_x f(x_0, z, p) + h^2O(1), \qquad \text{(A1.4)}
\end{aligned}$$

with the $O(1)$ - term bounded by $\gamma(z,p)$. On the other hand, by the integral mean value theorem,

$$\begin{aligned} J_\nu(z) &= \int \int_{-1/2}^{1/2} t^\nu \chi(z,p) f(x_0+ht,z,p)dtdp \\ &= \lambda(z) \int \int_{-1/2}^{1/2} f(x_0+ht,z,p)dtdp \\ &= \lambda(z) h^{-1} q(z) \\ &= 1_{\{q(z)>0\}} J_\nu(z), \end{aligned}$$

with $|\lambda(z)| \leq \overline{\chi}(z)$ (note $\lambda(z) = \lambda_h(z)$). Since $I_\nu(z,p) \to \kappa_\nu f(x_0,z,p)$ boundedly (by $\gamma(z,p)$) as $h \to 0$, $\lim_{h\to 0} J_\nu(z) = \lim_{h\to 0} \int \chi(z,p) I_\nu(z,p)dp = \kappa_\nu \int \chi(z,p) f(x_0,z,p)dp$. On the other hand, $h^{-1}q(z) \to f(x_0,z)$. Therefore, on $\mathcal{Z}_0 = \{z : f(x_0,z) > 0\}$,

$$\lim_{h\to 0} \lambda(z) = \kappa_\nu \frac{\int \chi(z,p) f(x_0,z,p)dp}{f(x_0,z)} =: \kappa_\nu C(z).$$

Defining $C(z) = 0$ on $\mathcal{Z}_0^c$, $C(z) = C(x_0,z)$ is a version of the conditional expectation $\mathbb{E}\{\chi(Z,P)|X = x_0, Z = z\}$, and we may write

$$\begin{aligned} J_\nu(z) &= 1_{\{q(z)>0\}} J_\nu(z) \\ &= [\kappa_\nu C(z) + o(1)]\, h^{-1} q(z) 1_{\mathcal{Z}_0}(z) + \lambda(z) h^{-1} q(z) 1_{\mathcal{Z}\setminus\mathcal{Z}_0}(z), \end{aligned} \tag{A1.5}$$

with the $o(1)$ - term bounded by $const \times \overline{\chi}(z)$. Inserting (A1.4) and (A1.5) into (A1.3), we obtain

$$\begin{aligned} \mathbb{E}\bar{\xi}\eta &= h^{\mu+\nu} \int q(z)^{-1} \left[\int \psi(z,p) I_\mu(z,p)dp \right] 1_{\{q(z)>0\}} J_\nu(z)dz \\ &= h^{\mu+\nu} \int_{\mathcal{Z}_0} q(z)^{-1} \int \psi(z,p) \left[h\kappa_{\mu+1} \partial_x f(x_0,z,p) + h^2 O(1) \right] dp \\ &\qquad\qquad [\kappa_\nu C(z) + o(1)]\, h^{-1} q(z) dz \\ &\quad + h^{\mu+\nu} \int_{\mathcal{Z}\setminus\mathcal{Z}_0} q(z)^{-1} \int \psi(z,p) \left[h\kappa_{\mu+1} \partial_x f(x_0,z,p) + h^2 O(1) \right] dp \\ &\qquad\qquad \lambda(z) h^{-1} q(z) dz \\ &= h^{\mu+\nu} \int_{\mathcal{Z}_0} [\kappa_\nu C(z) + o(1)] \\ &\qquad\qquad \times \int \psi(z,p) \left[\kappa_{\mu+1} \partial_x f(x_0,z,p) + hO(1) \right] dpdz \\ &\quad + h^{\mu+\nu} \int_{\mathcal{Z}\setminus\mathcal{Z}_0} \left[\int \psi(z,p) \left[\kappa_{\mu+1} \partial_x f(x_0,z,p) + hO(1) \right] dp \right] \lambda(z)dz \\ &= h^{\mu+\nu} \kappa_{\mu+1} \kappa_\nu \int C(z) \left[\int \psi(z,p) \partial_x f(x_0,z,p)dp \right] dz + h^{\mu+\nu} o(1) \\ &\quad + h^{\mu+\nu} \kappa_{\mu+1} \int_{\mathcal{Z}\setminus\mathcal{Z}_0} \lambda(z) \left[\int \psi(z,p) \partial_x f(x_0,z,p)dp \right] dz \\ &\quad + h^{\mu+\nu+1} \int_{\mathcal{Z}\setminus\mathcal{Z}_0} \lambda(z) \left[\int \psi(z,p) O(1)dp \right] dz. \end{aligned} \tag{A1.6}$$

Since

$$\begin{aligned}\int \psi(z,p)\partial_x f(x_0,z,p)dp &= \partial_x\left[\int \psi(z,p)f(x,z,p)dp\right]_{x_0}\\ &= \partial_x\left[\pi_\psi(x,z)f(x,z)\right]_{x_0}\\ &= \partial_x\pi_\psi(x_0,z)f(x_0,z)+\pi_\psi(x_0,z)\partial_x f(x_0,z),\end{aligned}$$

the first term on the rhs of the last equality in (A1.6) may be written

$$h^{\mu+\nu}\kappa_{\mu+1}\kappa_\nu\int \pi_\chi(x_0,z)\partial_x\left[\pi_\psi(x,z)f(x,z)\right]_{x_0}dz.$$

The third term becomes

$$h^{\mu+\nu}\kappa_{\mu+1}\int_{\mathcal{Z}\setminus\mathcal{Z}_0}\lambda(z)\partial_x\left[\pi_\psi(x,z)f(x,z)\right]_{x_0}dz=0$$

since (by assumption (A1-III)) $\partial_x\left[\pi_\psi(x,z)f(x,z)\right]_{x_0}=0$ on $\mathcal{Z}_0^c$ except on a set of Lebesgue measure zero. The last term is $O(h^{\mu+\nu+1})$. This shows the first part of (A1.2). For both μ and ν even, instead of (A1.4) we have the expansion

$$I_\mu(z,p)=\kappa_\mu f(x_0,z,p)+h^2O(1).$$

Proceeding as in (A1.6), we find that

$$\begin{aligned}\mathbb{E}\bar{\xi}\eta &= h^{\mu+\nu-1}\int_{\mathcal{Z}_0}\left[\kappa_\nu C(z)+o(1)\right]\\ &\qquad\times\int\psi(z,p)\left[\kappa_\mu f(x_0,z,p)+h^2O(1)\right]dpdz\\ &\quad+h^{\mu+\nu-1}\int_{\mathcal{Z}\setminus\mathcal{Z}_0}\left[\int\psi(z,p)\left[\kappa_\mu f(x_0,z,p)+h^2O(1)\right]dp\right]\lambda(z)dz\\ &= h^{\mu+\nu-1}\kappa_\mu\kappa_\nu\int C(z)\left[\int\psi(z,p)f(x_0,z,p)dp\right]dz+h^{\mu+\nu-1}o(1)\\ &\quad+h^{\mu+\nu-1}\kappa_\mu\int_{\mathcal{Z}\setminus\mathcal{Z}_0}\lambda(z)\left[\int\psi(z,p)f(x_0,z,p)dp\right]dz\\ &\quad+h^{\mu+\nu+1}\int_{\mathcal{Z}\setminus\mathcal{Z}_0}\lambda(z)\left[\int\psi(z,p)O(1)dp\right]dz.\end{aligned}$$

Since

$$\int\psi(z,p)f(x_0,z,p)dp=\pi_\psi(x_0,z)f(x_0,z),$$

the first term on the rhs of the second equality may be written

$$h^{\mu+\nu-1}\kappa_\mu\kappa_\nu\int\pi_\chi(x_0,z)\pi_\psi(x_0,z)f(x_0,z)dz.$$

The third term vanishes since, for $z\in\mathcal{Z}_0^c$, $f(x_0,z,p)=0$ for a.a. p. The last term is $O(h^{\mu+\nu+1})$. ■

Versions of Lemma A1.2 and A1.3 that are adapted to the scenario of *Models I* and *II* (in which there is no P - variable) are obtained by simply suppressing the p-component in all functions.

Lemma A1.2'. *Denote* $\xi = W(X - x_0)^\mu \psi(Z)$*, with an integer* $\mu \geq 0$ and a measurable function $\psi(z)$ s.t. $\int |\psi(z)|\gamma(z)dz < \infty$ *(with* γ *from assumption* (A1-I)*). Then*

$$\mathbb{E}\xi = \begin{cases} h^\mu \kappa_\mu \int \psi(z) f(x_0, z)dz + O(h^{\mu+2}), & \mu \ \ even, \\ h^{\mu+1}\kappa_{\mu+1} \int \psi(z)\partial_x f(x_0, z)dz + O(h^{\mu+2}), & \mu \ \ odd. \end{cases}$$

Note that the integrals may also be expressed as

$$\begin{aligned} \int \psi(z) f(x_0, z)dz &= \mathbb{E}\{\psi(Z)|X = x_0\} f(x_0), \\ \int \psi(z)\partial_x f(x_0, z)dz &= \partial_x \left[\mathbb{E}\{\psi(Z)|X = x\} f(x)\right]_{x_0}. \end{aligned}$$

Lemma A1.3'. *Denote* $\xi = W(X - x_0)^\mu \psi(Z), \eta = W(X - x_0)^\nu \chi(Z)$ *Assume that* $\int |\chi(z)|\,|\psi(z)|\,\gamma(z)dz < \infty$*. Then*

$$\begin{aligned} &\mathbb{E}\xi\overline{\eta} = \mathbb{E}\overline{\xi}\overline{\eta} = \mathbb{E}\overline{\xi}\eta \\ = &\begin{cases} h^{\mu+\nu}\kappa_{\mu+1}\kappa_\nu B_{odd} + h^{\mu+\nu}o(1) & for \ \mu \ odd, \\ h^{\mu+\nu-1}\kappa_\mu\kappa_\nu B_{even} + h^{\mu+\nu-1}o(1) & for \ both \ \mu \ and \ \nu \ even. \end{cases} \end{aligned}$$

The constant B_{odd} *is given by*

$$\begin{aligned} B_{odd} &= \int \chi(z)\psi(z)\partial_x f(x_0, z)dz \\ &= \partial_x \left[\mathbb{E}\{\chi(Z)\psi(Z)|X = x\} f(x)\right]_{x_0} \end{aligned}$$

and the constant B_{even} *by*

$$\begin{aligned} B_{even} &= \int \chi(z)\psi(z) f(x_0, z)dz \\ &= \mathbb{E}\{\chi(Z)\psi(Z)|X = x_0\} f(x_0). \end{aligned}$$

7.2 Overview of the Proof

The rest of this Appendix is divided into three parts, according to the three terms in the decomposition of the OLS-estimator

$$\tilde{\theta}_n - \theta = A_n^{-1}\left[\frac{1}{nh}\sum_{i=1}^n \Phi_i \widetilde{r}_i + \frac{1}{nh}\sum_{i=1}^n \Phi_i \widetilde{\varepsilon}_i\right], \tag{A1.7}$$

where we have put

$$A_n = \frac{1}{nh}\sum_{i=1}^n \Phi_i \Phi_i'.$$

Finally, we will discuss *Models IV* and *V* separately.

7.3 The Moment Matrix A_n

Models I and II

In *Model I*, $\Phi = U$, in *Model II*, $\Phi = (U, \lambda(Z)U)'$ (we again omit indices i in the intermediate calculations). Putting $\xi = W(X - x_0)$, we may write $U = \xi - \overline{\xi}$ and $U^2\lambda(Z)^m = (\xi^2 - 2\xi\overline{\xi} + \overline{\xi}^2)\lambda(Z)^m$. Using Lemmas A1.2 and A1.3 we find that, for $m = 0, 1, 2$,

$$\begin{aligned} \mathbb{E}U^2\lambda(Z)^m &= h\kappa_2 \int \lambda(z)^m f(x_0, z)dz + O(h^2) \\ &= h\kappa_2 \pi_m(x_0) f(x_0) + O(h^2), \end{aligned}$$

where $\pi_m(x_0)$ is a continuous version of the conditional expectation $\mathbb{E}\{\lambda(Z_i)^m | X_i = x_0\}$. Hence

$$\lim_{n\to\infty} \frac{1}{nh} \sum_{i=1}^{n} \mathbb{E}U_i^2 \lambda(Z_i)^m = \kappa_2 \pi_m(x_0) f(x_0). \tag{A1.8}$$

As to the variances, $\mathbb{E}U^4\lambda(Z)^{2m} \le 8[\mathbb{E}\xi^4\lambda(Z)^{2m} + \mathbb{E}\overline{\xi}^4\lambda(Z)^{2m}] = O(h)$ and therefore $\mathrm{var}\big[U^2\lambda(Z)^m\big] = O(h) - O(h^2) = O(h)$. As a consequence,

$$\mathrm{var}\left[\frac{1}{nh}\sum_{i=1}^{n} U_i^2 \lambda(Z_i)^m\right] = O\left(\frac{1}{nh}\right).$$

This together with (A1.8) shows that

$$\mathrm{plim}_{n\to\infty} A_n = \kappa_2 f(x_0) \tag{A1.9-I}$$

in *Model I* and

$$\mathrm{plim}_{n\to\infty} A_n = A$$

in *Model II*, with

$$A = \kappa_2 f(x_0) \begin{pmatrix} 1 & \pi_1(x_0) \\ \pi_1(x_0) & \pi_2(x_0) \end{pmatrix}. \tag{A1.9-II}$$

Model III

Denote $\xi^{\mu,l} = W(X - x_0)^\mu P^l$. Then
$U = W(X - x_0) - \mathbb{E}\{W(X - x_0)|WZ\} = \xi^{1,0} - \overline{\xi^{1,0}}$,
$Q_0 = h[WP - \mathbb{E}\{WP|WZ\}] = h(\xi^{0,1} - \overline{\xi^{0,1}})$,
$Q_1 = WP(X - x_0) - \mathbb{E}\{WP(X - x_0)|WZ\} = \xi^{1,1} - \overline{\xi^{1,1}}$,
$\Phi = (U, Q_0, Q_1)'$.

Let $\pi_m(x_0)$ and $\pi_m(x_0, z)$ be versions of the conditional expectations $\mathbb{E}\{P^m|X = x_0\}$ and $\mathbb{E}\{P^m|X = x_0, Z = z\}$, respectively.
Then every entry of $\Phi\Phi'$ is of the form $h^{2-\mu-\nu}(\xi^{\mu,l} - \overline{\xi^{\mu,l}})(\xi^{\nu,m} - \overline{\xi^{\nu,m}})$ with $\mu, \nu, l, m \in \{0, 1\}$. By Lemmas A1.2 and A1.3, $\mathbb{E}(\xi^{\mu,l} - \overline{\xi^{\mu,l}})(\xi^{\nu,m} - \overline{\xi^{\nu,m}}) = O(h)$ for $\mu + \nu = 1$, hence the corresponding

entries UQ_0 and Q_0Q_1 have expected value $O(h^2)$. For the other entries (i.e. $(\mu,\nu) \in \{(1,1),(0,0)\}$) we find that

$$\begin{aligned}
\mathbb{E}U^2 &= h\kappa_2 f(x_0) + O(h^2), \\
\mathbb{E}UQ_1 &= \mathbb{E}(\xi^{1,0} - \overline{\xi^{1,0}})(\xi^{1,1} - \overline{\xi^{1,1}}) = h\kappa_2\pi_1(x_0)f(x_0) + O(h^2), \\
\mathbb{E}Q_0^2 &= h^2\mathbb{E}(\xi^{0,1} - \overline{\xi^{0,1}})^2 \\
&= h^2\left[h^{-1}\pi_2(x_0)f(x_0) + O(h)\right] + h(h\mathbb{E}\xi^{0,1}\overline{\xi^{0,1}}) \\
&= h\left\{\pi_2(x_0)f(x_0) - \int \left[\pi_1(x_0,z)\right]^2 f(x_0,z)dz\right\} + o(h) \\
&= h\mathbb{E}\left\{P^2 - \mathbb{E}\{P|X=x_0, Z=z\}^2 | X=x_0\right\} f(x_0) + o(h) \\
&= h\sigma_P^2(x_0) + o(h), \\
\mathbb{E}Q_1^2 &= \mathbb{E}(\xi^{1,1} - \overline{\xi^{1,1}})^2 = h\kappa_2\pi_2(x_0)f(x_0) + O(h^2).
\end{aligned}$$

Hence

$$\lim_{n\to\infty} \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}\Phi_i\Phi_i' = A$$

with

$$A = \kappa_2 f(x_0)\begin{pmatrix} 1 & 0 & \pi_1(x_0) \\ 0 & \kappa_2^{-1}\sigma_P^2(x_0) & 0 \\ \pi_1(x_0) & 0 & \pi_2(x_0) \end{pmatrix}. \tag{A1.9-III}$$

As to the variances: since $\mathbb{E}[(\xi^{\mu,l} - \overline{\xi^{\mu,l}})^2(\xi^{\nu,m} - \overline{\xi^{\nu,m}})^2] \leq [\mathbb{E}(\xi^{\mu,l} - \overline{\xi^{\mu,l}})^4\mathbb{E}(\xi^{\nu,m} - \overline{\xi^{\nu,m}})^4]^{1/2}$ and $\mathbb{E}(\xi^{\mu,l} - \overline{\xi^{\mu,l}})^4 \leq 8\mathbb{E}[(\xi^{\mu,l})^4 + \mathbb{E}(\overline{\xi^{\mu,l}})^4] \leq 16\mathbb{E}(\xi^{\mu,l})^4$, it suffices to consider $\mathbb{E}(\xi^{\mu,l})^4$ and $\mathbb{E}(\xi^{\nu,m})^4$. But by Lemma A1.2

$$\begin{aligned}
\mathbb{E}(\xi^{\mu,l})^4 &= \mathbb{E}W^4(X-x_0)^{4\mu}P^{4l} \\
&= h^{-3}\mathbb{E}W(X-x_0)^{4\mu}P^{4l} \\
&= O(h^{4\mu-3}).
\end{aligned}$$

Hence every entry of the $\Phi\Phi'-$ matrix has variance

$$h^{2(2-\mu-\nu)}\left[O(h^{4\mu-3})O(h^{4\nu-3})\right]^{1/2} = O(h).$$

Since the squares of the expected values behave as $O(h^2)$, it follows that the variance of every entry of the matrix A_n behaves as $O\left(\frac{1}{nh}\right)$. Hence, in *Model III,*

$$\text{plim}_{n\to\infty} A_n = A.$$

We gather the results obtained so far in

Lemma A1.4. *In Models I - III,*

$$\text{plim}_{n\to\infty} A_n = A$$

with A given by (A1.9-I) for Model I, (A1.9-II) for Model II and by (A1.9-III) for Model III.

7.4 The Bias Term $\frac{1}{nh}\sum_{i=1}^n \Phi_i \widetilde{r}_i$

Start by considering *Model III*. The term to be analyzed is $\sum_{i=1}^n \Phi_i \widetilde{r}_i$ with $\Phi = (U, Q_0, Q_1)'$ (omitting indices as above) and $\widetilde{r} = Wr - \overline{Wr}$, where

$$\begin{aligned} r &= \frac{1}{2}k''(x_0)(X-x_0)^2 + \frac{1}{3!}k^{(3)}(x_0)(X-x_0)^3 \\ &\qquad + \frac{1}{3!}\left[k^{(3)}(x_0+\eta(X-x_0)) - k^{(3)}(x_0)\right](X-x_0)^3 \\ &\quad + \left[\frac{1}{2}g''(x_0)(X-x_0)^2 + \frac{1}{3!}g^{(3)}(x_0)(X-x_0)^3 \right. \\ &\qquad \left. + \frac{1}{3!}\left[g^{(3)}(x_0+\eta(X-x_0)) - g^{(3)}(x_0)\right](X-x_0)^3\right] P. \end{aligned}$$

Again denoting $\xi^{\mu,l} = W(X-x_0)^\mu P^l$, we may write

$$\begin{aligned} \widetilde{r} &= \frac{1}{2}k''(x_0)(\xi^{2,0} - \overline{\xi^{2,0}}) + \frac{1}{3!}k^{(3)}(x_0)(\xi^{3,0} - \overline{\xi^{3,0}}) \\ &\quad + \frac{1}{2}g''(x_0)(\xi^{2,1} - \overline{\xi^{2,1}}) + \frac{1}{3!}g^{(3)}(x_0)(\xi^{3,1} - \overline{\xi^{3,1}}) + \Delta, \end{aligned}$$

where Δ collects the difference terms. The elements of $\Phi\widetilde{r}$ are then of the following form. To abbreviate the formulas, we shall use the short-hand notation $k_0'' = k''(x_0)$, etc. in the sequel.

1st Component: Since $U = \xi^{1,0} - \overline{\xi^{1,0}}$,

$$\begin{aligned} U\widetilde{r} &= \frac{1}{2}k_0''(\xi^{2,0} - \overline{\xi^{2,0}})(\xi^{1,0} - \overline{\xi^{1,0}}) + \frac{1}{3!}k_0^{(3)}(\xi^{3,0} - \overline{\xi^{3,0}})(\xi^{1,0} - \overline{\xi^{1,0}}) \\ &\quad + \frac{1}{2}g_0''(\xi^{2,1} - \overline{\xi^{2,1}})(\xi^{1,0} - \overline{\xi^{1,0}}) + \frac{1}{3!}g_0^{(3)}(\xi^{3,1} - \overline{\xi^{3,1}})(\xi^{1,0} - \overline{\xi^{1,0}}) \\ &\quad + \Delta(\xi^{1,0} - \overline{\xi^{1,0}}). \end{aligned}$$

By Lemmas A1.2 and A1.3,

$$\begin{aligned} \mathbb{E}U\widetilde{r} &= h^3\left[\frac{1}{2}\left(\kappa_4 - \kappa_2^2\right)k_0''f'(x_0) + \frac{1}{3!}k_0^{(3)}\kappa_4 f(x_0)\right. \\ &\quad + \frac{1}{2}g_0''\left(\kappa_4\partial_x\left[\pi_1(x)f(x)\right]_{x_0} - \kappa_2^2\int \pi_1(x_0,z)\partial_x f(x_0,z)dz\right) \\ &\quad \left. + \frac{1}{3!}g_0^{(3)}\kappa_4\pi_1(x_0)f(x_0) + o(1)\right] + o(h^3), \end{aligned} \tag{A1.10}$$

where the $o(h^3)$ come from the remainder terms Δ,
since $g^{(3)}(x_0+\eta(X-x_0)) - g^{(3)}(x_0) = o_p(1)$ uniformly.

2nd Component: Since $Q_0 = h(\xi^{0,1} - \overline{\xi^{0,1}})$,

$$\begin{aligned} h^{-1}Q_0\widetilde{r} &= \frac{1}{2}k_0''(\xi^{2,0} - \overline{\xi^{2,0}})(\xi^{0,1} - \overline{\xi^{0,1}}) + \frac{1}{3!}k_0^{(3)}(\xi^{3,0} - \overline{\xi^{3,0}})(\xi^{0,1} - \overline{\xi^{0,1}}) \\ &\quad + \frac{1}{2}g_0''(\xi^{2,1} - \overline{\xi^{2,1}})(\xi^{0,1} - \overline{\xi^{0,1}}) + \frac{1}{3!}g_0^{(3)}(\xi^{3,1} - \overline{\xi^{3,1}})(\xi^{0,1} - \overline{\xi^{0,1}}) \\ &\quad + \Delta(\xi^{0,1} - \overline{\xi^{0,1}}). \end{aligned}$$

Applying Lemmas A1.2 and A1.3 we obtain

$$\begin{aligned} h^{-1}\mathbb{E}Q_0\widetilde{r} &= h\left[\frac{1}{2}k_0''\kappa_2\left(O(h^2)-o(1)\right)+\frac{1}{3!}k_0^{(3)}O(h^2)\right. \\ &\quad \left.+\frac{1}{2}g_0''\left(\kappa_2\sigma_P^2(x_0)f(x_0)+o(1)\right)+\frac{1}{3!}g_0^{(3)}\kappa_4 O(h^2)\right] \\ &\quad +o(h^3), \end{aligned}$$

hence

$$\mathbb{E}Q_0\widetilde{r} = \frac{h^2}{2}g_0''\kappa_2\sigma_P^2(x_0)f(x_0)+h^2o(1). \tag{A1.11}$$

3rd component: Since $Q_1 = \xi^{1,1}-\overline{\xi^{1,1}}$,

$$\begin{aligned} \mathbb{E}Q_1\widetilde{r} &= \frac{1}{2}k_0''(\xi^{2,0}-\overline{\xi^{2,0}})(\xi^{1,1}-\overline{\xi^{1,1}})+\frac{1}{3!}k_0^{(3)}(\xi^{3,0}-\overline{\xi^{3,0}})(\xi^{1,1}-\overline{\xi^{1,1}}) \\ &\quad +\frac{1}{2}g_0''(\xi^{2,1}-\overline{\xi^{2,1}})(\xi^{1,1}-\overline{\xi^{1,1}})+\frac{1}{3!}g_0^{(3)}(\xi^{3,1}-\overline{\xi^{3,1}})(\xi^{1,1}-\overline{\xi^{1,1}}) \\ &\quad +\Delta(\xi^{1,1}-\overline{\xi^{1,1}}). \end{aligned}$$

Again by Lemmas A1.2 and A1.3,

$$\begin{aligned} \mathbb{E}Q_1\widetilde{r} &= h^3\left[\frac{1}{2}k_0''(\kappa_4-\kappa_2^2)\partial_x\left[\pi_1(x)f(x)\right]_{x_0}+\frac{1}{3!}k_0^{(3)}\kappa_4\pi_1(x_0)f(x_0)+o(1)\right. \\ &\quad +\frac{1}{2}g_0''\left(\kappa_4\partial_x\left[\pi_2(x)f(x)\right]_{x_0}-\kappa_2^2\int\pi_1(x_0,z)\partial_x\left[\pi_1(x,z)f(x,z)\right]_{x_0}dz\right) \\ &\quad \left.+\frac{1}{3!}g_0^{(3)}\kappa_4\pi_2(x_0)f(x_0)+o(1)\right]+o(h^3). \end{aligned} \tag{A1.12}$$

Gathering the results in (A1.10) - (A1.12), we obtain

$$\mathbb{E}\Phi\widetilde{r} = h^2b^{III}+h^2o(1) \tag{A1.13}$$

with the components b_i of $b^{III} = (b_1,b_2,b_3)'$ given by

$$\begin{aligned} b_1 &= h\left[\frac{1}{2}\left(\kappa_4-\kappa_2^2\right)k''(x_0)f'(x_0)+\frac{1}{3!}k^{(3)}(x_0)\kappa_4 f(x_0)\right. \\ &\quad +\frac{1}{2}g''(x_0)\left(\kappa_4\partial_x\left[\pi_1(x)f(x)\right]_{x_0}-\kappa_2^2\int\pi_1(x_0,z)\partial_x f(x_0,z)dz\right) \\ &\quad \left.+\frac{1}{3!}g^{(3)}(x_0)\kappa_4\pi_1(x_0)f(x_0)\right], \end{aligned} \tag{A1.14-IIIa}$$

$$b_2 = \frac{1}{2}g''(x_0)\kappa_2\sigma_P^2(x_0)f(x_0), \tag{A1.14-IIIb}$$

$$\begin{aligned} b_3 &= h\left[\frac{1}{2}k''(x_0)(\kappa_4-\kappa_2^2)\partial_x\left[\pi_1(x)f(x)\right]_{x_0}+\frac{1}{3!}k^{(3)}(x_0)\kappa_4\pi_1(x_0)f(x_0)\right. \\ &\quad +\frac{1}{2}g''(x_0)\left(\kappa_4\partial_x\left[\pi_2(x)f(x)\right]_{x_0}-\kappa_2^2\int\pi_1(x_0,z)\partial_x\left[\pi_1(x,z)f(x,z)\right]_{x_0}dz\right) \\ &\quad \left.+\frac{1}{3!}g^{(3)}(x_0)\kappa_4\pi_2(x_0)f(x_0)\right]. \end{aligned} \tag{A1.14-IIIc}$$

As for the variance, $\mathbb{E}U^2\widetilde{r}^2 \sim \mathbb{E}W^4(X-x_0)^6 = O(h^3)$, $\mathbb{E}Q_0^2\widetilde{r}^2 \sim h^2\mathbb{E}W^4(X-x_0)^4 = O(h^3)$, $\mathbb{E}Q_1^2\widetilde{r}^2 \sim \mathbb{E}W^4(X-x_0)^6 = O(h^3)$. Since the expected values are at least $O(h^2)$, this means that all variances are $O(h^3)$. As a consequence, for

$$B = \frac{1}{nh}\sum_{i=1}^{n}\left[\Phi_i\widetilde{r}_i - \mathbb{E}\Phi_i\widetilde{r}_i\right],$$

it holds that

$$\mathbb{E}\left[h^{-2}B\right]^2 = O\left(\frac{1}{nh^3}\right) = o(1), \tag{A1.15}$$

so that $h^{-2}B \xrightarrow{P} 0$. Therefore, finally, by (A1.13) and (A1.15),

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\Phi_i\widetilde{r}_i + B\\ &= hb^{III} + o(h) + h(h^{-1}B)\\ &= hb^{III} + o_P(h).\end{aligned} \tag{A1.16}$$

This settles the bias term for *Model III*. For *Model I*, $\mathbb{E}U\widetilde{r}$ is obtained from (A1.10) by dropping the $g-$ terms, i.e.

$$\mathbb{E}U\widetilde{r} = h^3\left[\frac{1}{2}\left(\kappa_4-\kappa_2^2\right)k''(x_0)f'(x_0) + \frac{1}{3!}k^{(3)}(x_0)\kappa_4 f(x_0) + o(1)\right] + o(h^3).$$

(A1.15) remains valid, so that (A1.16) becomes

$$\frac{1}{nh}\sum_{i=1}^{n}U_i\widetilde{r}_i = h^2b^I + o_P(h^2) \tag{A1.17}$$

with

$$b^I = \frac{1}{2}\left(\kappa_4-\kappa_2^2\right)k''(x_0)f'(x_0) + \frac{1}{3!}k^{(3)}(x_0)\kappa_4 f(x_0). \tag{A1.14-I}$$

For *Model II*, $\mathbb{E}\Phi\widetilde{r}$ is obtained from (A1.10) and (A1.12) by substituting $\lambda(Z)$ for P. Calculation using Lemmas A1.2 and A1.3 gives

$$\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i = h^2b^{II} + o_P(h^2) \tag{A1.18}$$

with the components b_1, b_2 of b^{II} given by

$$\begin{aligned}b_1 &= \frac{1}{2}\left(\kappa_4-\kappa_2^2\right)k''(x_0)f'(x_0) + \frac{1}{3!}k^{(3)}(x_0)\kappa_4 f(x_0)\\ &\quad+\frac{1}{2}g''(x_0)\left(\kappa_4\partial_x\left[\pi_1(x)f(x)\right]_{x_0} - \kappa_2^2\lambda(z)f'(x_0)\right)\\ &\quad+\frac{1}{3!}g^{(3)}(x_0)\kappa_4\pi_1(x_0)f(x_0),\end{aligned} \tag{A1.14-IIa}$$

$$\begin{aligned}b_3 &= \frac{1}{2}k''(x_0)(\kappa_4-\kappa_2^2)\partial_x\left[\pi_1(x)f(x)\right]_{x_0} + \frac{1}{3!}k^{(3)}(x_0)\kappa_4\pi_1(x_0)f(x_0)\\ &\quad+\frac{1}{2}g''(x_0)\left(\kappa_4\partial_x\left[\pi_2(x)f(x)\right]_{x_0} - \kappa_2^2\lambda(z)^2f'(x_0)\right)\\ &\quad+\frac{1}{3!}g^{(3)}(x_0)\kappa_4\pi_2(x_0)f(x_0)\end{aligned} \tag{A1.14-IIc}$$

(note that in *Model II* $\pi_1(x,z) = \mathbb{E}\{\lambda(Z)|X=x, Z=z\} = \lambda(z)$), so that

$$\int \pi_1(x_0,z)\partial_x f(x_0,z)dz = \lambda(z)f'(x_0)$$

and

$$\int \pi_1(x_0,z)\partial_x \left[\pi_1(x,z)f(x,z)\right]_{x_0} dz = \lambda(z)^2 f'(x_0).$$

Gathering the results, we obtain

Lemma A1.5. *In Models I and II,*

$$\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i = h^2 b + o_P(h^2),$$

with $b = b^I$ for Model I given by (A1.14-I) and $b = b^{II} = (b_1, b_2)'$ given by (A1.14-II). For Model III,

$$\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i = hb^{III} + o_P(h),$$

with the components b_i of $b^{III} = (b_1, b_2, b_3)'$ given by (A1.14-III).

7.5 The Error Term $\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{\varepsilon}_i$

Lemma A1.6.

$$\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\Phi_i\widetilde{\varepsilon}_i \xrightarrow{d} \mathcal{N}(0,\sigma^2 A),$$

with A given by (A1.9i) in *Model i, i=I, II ,III.*

Proof. The proof is the same for all models. We show: $(\eta_{ni}, \mathcal{F}_{ni})$, $i = 1,\ldots,n$, $n \geq 1$, with $\eta_{ni} = \sqrt{h/n}\Phi_{ni}\widetilde{\varepsilon}_{ni}, \mathcal{F}_{ni} = \mathcal{F}_i$, is a (vector-valued) martingale difference array such that

$$(i)\quad p\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\left\{\eta_{ni}\eta_{ni}'|\mathcal{F}_{i-1}\right\} = \sigma^2 A,$$

and

$$(ii)\ p\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\left\{\|\eta_{ni}\|^2 1_{\{\|\eta_{ni}\|>\delta\}}|\mathcal{F}_{i-1}\right\} = 0 \text{ for every } \delta > 0.$$

The assertion will then follow from a standard central limit theorem for martingale difference arrays (cf., e.g., Pollard (1984)). Of course, in the present scenario of independent row entries, any other central limit theorem for such arrays will also do. But the above version easily lends itself for extension to certain dependence structures, e.g. mixing processes. Since (by virtue of assumption (A2), $\overline{W_i\varepsilon_i} = \mathbb{E}\{W_i\mathbb{E}\{\varepsilon_i|X_i,Z_i\}|W_iZ_i\} = 0$, $\Phi_i\widetilde{\varepsilon}_i = \Phi_i W_i\varepsilon_i$ and the martingale difference property of (η_{ni}) is plain to see. As to (i), note that $\mathbb{E}\{\Phi_i\Phi_i'\widetilde{\varepsilon}_i^2|\mathcal{F}_{i-1}\} = \sigma^2\Phi_i\Phi_i'W_i^2 = \sigma^2(1/h^2)\Phi_i\Phi_i'$ since $\Phi_i = 0$ on $\{W_i = 0\}$. Therefore, by Lemma A1.4,

$$\sum_{i=1}^{n}\mathbb{E}\left\{\eta_{ni}\eta_{ni}'|\mathcal{F}_{i-1}\right\} = \sigma^2\frac{h}{n}\frac{1}{h^2}\sum_{i=1}^{n}\Phi_i\Phi_i' = \sigma^2\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\Phi_i' \xrightarrow{P} \sigma^2 A.$$

As to (ii), note that

$$\begin{aligned}\{\|\eta_{ni}\| > \delta\} &= \left\{\sqrt{\frac{h}{n}}\,\|\Phi_i W_i\|\,|\varepsilon_i| > \delta\right\} \\ &= \left\{h^2\,\|\Phi_i W_i\| \cdot \frac{|\varepsilon_i|}{\sqrt{nh^3}} > \delta\right\} \\ &\subset \left\{h^4\,\|\Phi_i W_i\|^2 > \delta\right\} \cup \left\{\varepsilon_i^2 > nh^3\delta\right\},\end{aligned}$$

where the last inclusion follows from the simple fact that $|ab| > \delta$ implies that $a^2 > \delta$ or $b^2 > \delta$. Therefore

$$\begin{aligned}&\sum_{i=1}^{n} \mathbb{E}\left\{\|\eta_{ni}\|^2\, 1_{\{\|\eta_{ni}\|>\delta\}} | \mathcal{F}_{i-1}\right\} \\ =\;& \frac{h}{n}\sum_{i=1}^{n} \mathbb{E}\left\{\|\Phi_i W_i\|^2\, \varepsilon_i^2 1_{\{\|\eta_{ni}\|>\delta\}} | \mathcal{F}_{i-1}\right\} \\ \leq\;& \sigma^2 \frac{h}{n}\sum_{i=1}^{n} \|\Phi_i W_i\|^2\, 1_{\left\{h^4\|\Phi_i W_i\|^2>\delta\right\}} + \frac{h}{n}\sum_{i=1}^{n} \|\Phi_i W_i\|^2\, \mathbb{E}\left\{\varepsilon_i^2 1_{\left\{\varepsilon_i^2>nh^3\delta\right\}}\right\} \\ =\;& \sigma^2 \frac{1}{nh}\sum_{i=1}^{n} \|\Phi_i\|^2\, 1_{\left\{h^2\|\Phi_i\|^2>\delta\right\}} + \frac{1}{nh}\sum_{i=1}^{n} \|\Phi_i\|^2\, \mathbb{E}\left\{\varepsilon_i^2 1_{\left\{\varepsilon_i^2>nh^3\delta\right\}}\right\}.\end{aligned}$$

Noting that $\mathbb{E}\left\{\|\Phi_i\|^2\, 1_{\left\{h^2\|\Phi_i\|^2>\delta\right\}}\right\} \leq \left[\mathbb{E}\,\|\Phi_i\|^4\right]^{1/2}\left[\delta^{-2}h^4\mathbb{E}\,\|\Phi_i\|^4\right]^{1/2} = O(h^3)$ (since $\mathbb{E}\,\|\Phi_i\|^4 = O(h)$), the expected value of the first term behaves as $O(h^2)$. Since it is nonnegative, this means that the first term tends to zero in L^1 and hence in probability. The same is true for the second term since $\alpha_n = \mathbb{E}\left\{\varepsilon_i^2 1_{\left\{\varepsilon_i^2>nh^3\delta\right\}}\right\} \to 0$ (independent of i) and $\frac{1}{nh}\sum_{i=1}^{n}\|\Phi_i\|^2$ converges in probability. ■

Remark A.4. The proof of Lemma A1.6 can be extended to the case where the ε_i are conditionally heteroscedastic. Details can be found in Christopeit and Hoderlein (2002).

Theorems 3.1 - 3.3 are now immediate consequences of (A1.7) and Lemmas A1.3 - A1.6:

$$\sqrt{nh^3}\left(\tilde{\theta}_n - \theta - A_n^{-1}\frac{1}{nh}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i\right) = A_n^{-1}\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\Phi_i\widetilde{\varepsilon}_i \xrightarrow{d} \mathcal{N}(0, \sigma^2 A^{-1}),$$

the left hand side being equal to

$$\sqrt{nh^3}\left(\tilde{\theta}_n - \theta - h^2 A^{-1} b + o_P(h^2)\right)$$

in *Models I* and *II* and to

$$\sqrt{nh^3}\left(\tilde{\theta}_n - \theta - hA^{-1} b + o_P(h)\right)$$

in *Model III*.

7.6 *Model IV*

In this subsection we treat exclusively *Model IV*. Making use of the relations

$$\mathbb{E}\{\,(W^z)^2\,(Z_1-z_{10})^m|X=x_0\}=\begin{cases} h^{m-1}\frac{\kappa'_m f(x_0,z_{10})+O(h^2)}{f(x_0)} & \text{for} \quad m \text{ even},\\ h^m\frac{\kappa'_{m+1}f_{z_1}(x_0,z_{10})+O(h^2)}{f(x_0)} & \text{for} \quad m \text{ odd},\end{cases} \tag{A1.19}$$

(which follow from a third order Taylor expansion of $f(x_0,z_1)$ about z_{10}, $m=0,1,2,\ldots,$), we calculate (using Lemmas A1.2' and A1.3')

$$\begin{aligned}
\mathbb{E}U^2 &= \kappa_2\kappa'_0 f(x_0,z_{10})+o(1),\\
\mathbb{E}UU_2 &= h^2\left(\kappa_4-\kappa_2^2\right)\kappa'_0 f_x(x_0,z_{10})+o(h^2),\\
\mathbb{E}U_2^2 &= h^2\left(\kappa_4-\kappa_2^2\right)\kappa'_0 f(x_0,z_{10})+o(h^2),\\
\mathbb{E}U^2(Z_1-z_{10}) &= h^2\kappa_2\kappa'_2 f_{z_1}(x_0,z_{10})+o(h^2),\\
\mathbb{E}UU_2(Z_1-z_{10}) &= h^4\left(\kappa_4-\kappa_2^2\right)\kappa'_2 f_{xz_1}(x_0,z_{10})+o(h^4),\\
\mathbb{E}U^2(Z_1-z_{10})^2 &= h^2\kappa_2\kappa'_2 f(x_0,z_{10})+o(h^2).
\end{aligned}$$

Hence

$$\frac{1}{n}\mathbb{E}\sum_{i=1}^n \Phi_i\Phi_i' = A+o(1)$$

with

$$A=\begin{pmatrix} \kappa_2\kappa'_0 f(x_0,z_{10}) & 0 & 0\\ 0 & \left(\kappa_4-\kappa_2^2\right)\kappa'_0 f(x_0,z_{10}) & 0\\ 0 & 0 & \kappa_2\kappa'_2 f(x_0,z_{10})\end{pmatrix}. \tag{A1.20}$$

As to the variances, note that

$$\begin{aligned}
&\mathbb{E}\left\{(W^x)^4(X-x_0)^{2k}(W^z)^4(Z_1-z_{10})^{2l}\right\}\\
=\;& h^{2k-3}\mathbb{E}\{\,(W^z)^4\,(Z_1-z_{10})^{2l}|X=x_0\}f(x_0)+O(h^{2k-1})\\
=\;& h^{2k-3}h^{2l-3}O(1)+O(h^{2k-1}).
\end{aligned} \tag{A1.21}$$

We have to evaluate (A1.21) for $(k,l)=(2,0),(2,1),(2,2),(3,0),(3,1),(4,0)$ (corresponding to the entries of the $\Phi\Phi'-$ matrix), the combinations and the corresponding values of (A1.21) being distributed as follows:

$$\begin{pmatrix} (2,0) & (3,0) & (2,1)\\ & (4,0) & (3,1)\\ & & (2,2)\end{pmatrix}, \qquad \begin{pmatrix} h^{-2} & 1 & 1\\ & h^2 & h^2\\ & & h^2\end{pmatrix}.$$

Taking account of the scaling performed in the definition of Φ (division by h of the 2nd and 3rd component), the 2nd moments of the $\Phi\Phi'-$ matrix all turn out to behave as $O(h^{-2})$. Since the expected values are at most $O(1)$, this means that the variances of the entries of the $\Phi\Phi'-$ matrix all behave as $O(h^{-2})$ and the entries of

$$\frac{1}{n}\sum_{i=1}^n\left[\Phi_i\Phi_i'-\mathbb{E}\Phi_i\Phi_i'\right]$$

have variance $O\left(\frac{1}{nh^2}\right)$, implying that

$$A_n = \frac{1}{n}\sum_{i=1}^{n} \Phi_i \Phi_i' \xrightarrow{P} A \tag{A1.22}$$

(with A given by (A1.20)). This suggests to write the OLS-estimator (3.11) in the form

$$\sqrt{nh^4}\left(\tilde{\theta}_n - \theta - A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\Phi_i \widetilde{r}_i\right) = A_n^{-1}\sqrt{\frac{h^4}{n}}\sum_{i=1}^{n}\Phi_i\widetilde{\varepsilon}_i \tag{A1.23}$$

(the reason for the appearance of rate $\sqrt{nh^4}$ will become apparent in the discussion of the error term).

As to the bias term, note that $\widetilde{r}$ is composed of terms of the form

$$\varphi(X,Z)W^x(X-x_0)^\mu W^z(Z-z_0)^\nu - \overline{\varphi(X,Z)W^x(X-x_0)^\mu W^z(Z-z_0)^\nu},$$

where $\varphi(x,z)$ is bounded and $\mu+\nu=3$. By Lemma A1.2 and (A1.19),

$$\begin{aligned}
\mathbb{E}\widetilde{r}^2 &\leq const \times \mathbb{E}\left(W^x\right)^2 (X-x_0)^{2\mu}\left(W^z\right)^2 (Z-z_0)^{2\nu} \\
&= h^{2(\mu+\nu-1)}O(1) \\
&= O(h^4).
\end{aligned}$$

Therefore, since $\mathbb{E}\left\|\Phi\right\|^2 = O(1)$, we find that $\mathbb{E}\left\|\Phi\widetilde{r}\right\| \leq \left[\mathbb{E}\left\|\Phi\right\|^2 \mathbb{E}\widetilde{r}^2\right]^{1/2} = O(h^2)$, so that

$$\frac{1}{n}\sum_{i=1}^{n}\Phi_i\widetilde{r}_i = O_P(h^2). \tag{A1.24}$$

We desist from calculating the bias term any further.

Finally, the error term can be handled along the same lines as in the proof of Lemma A1.6, the only difference being the scaling with $\sqrt{h^4/n}$ instead of $\sqrt{h/n}$. In more detail, we now define the martingale difference array $\eta_{ni} = \sqrt{h^4/n}\Phi_{ni}\widetilde{\varepsilon}_{ni}$. Then, since $\overline{W_i\varepsilon_i} = 0$ and hence $\widetilde{\varepsilon}_i = W_i\varepsilon_i$ continues to hold, $\mathbb{E}\{\Phi_i\Phi_i'\widetilde{\varepsilon}_i^2|\mathcal{F}_{i-1}\} = \sigma^2\Phi_i\Phi_i'W_i^2 = \sigma^2(1/h^4)\Phi_i\Phi_i'$ if we assume for simplicity that $K_1(z)$, too, is the uniform kernel. Therefore

$$\sum_{i=1}^{n}\mathbb{E}\{\eta_{ni}\eta_{ni}'|\mathcal{F}_{i-1}\} = \sigma^2\frac{h^4}{n}\frac{1}{h^4}\sum_{i=1}^{n}\Phi_i\Phi_i' = \sigma^2\frac{1}{n}\sum_{i=1}^{n}\Phi_i\Phi_i' \xrightarrow{P} \sigma^2 A$$

by (A1.22). As to the Lindeberg condition (ii), note that

$$\begin{aligned}
\{\left\|\eta_{ni}\right\| > \delta\} &= \left\{\sqrt{\frac{h^4}{n}}\left\|\Phi_i W_i\right\|\left|\varepsilon_i\right| > \delta\right\} \\
&= \left\{h^4\left\|\Phi_i W_i\right\| \cdot \frac{\left|\varepsilon_i\right|}{\sqrt{nh^4}} > \delta\right\} \\
&\subset \left\{h^8\left\|\Phi_i W_i\right\|^2 > \delta\right\} \cup \left\{\varepsilon_i^2 > nh^4\delta\right\},
\end{aligned}$$

where the last inclusion follows from the simple fact that $|ab| > \delta$ implies that $a^2 > \delta$ or $b^2 > \delta$. Therefore

$$
\begin{aligned}
& \sum_{i=1}^{n} \mathbb{E}\left\{ \|\eta_{ni}\|^2 1_{\{\|\eta_{ni}\|>\delta\}} | \mathcal{F}_{i-1} \right\} \\
= & \frac{h^4}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \|\Phi_i W_i\|^2 \varepsilon_i^2 1_{\{\|\eta_{ni}\|>\delta\}} | \mathcal{F}_{i-1} \right\} \\
\leq & \sigma^2 \frac{h^4}{n} \sum_{i=1}^{n} \|\Phi_i W_i\|^2 1_{\left\{h^8 \|\Phi_i W_i\|^2 > \delta\right\}} + \frac{h^4}{n} \sum_{i=1}^{n} \|\Phi_i W_i\|^2 \mathbb{E}\left\{ \varepsilon_i^2 1_{\left\{\varepsilon_i^2 > n h^4 \delta\right\}} \right\} \\
= & \sigma^2 \frac{1}{n} \sum_{i=1}^{n} \|\Phi_i\|^2 1_{\left\{h^4 \|\Phi_i\|^2 > \delta\right\}} + \frac{1}{n} \sum_{i=1}^{n} \|\Phi_i\|^2 \mathbb{E}\left\{ \varepsilon_i^2 1_{\left\{\varepsilon_i^2 > n h^4 \delta\right\}} \right\}.
\end{aligned}
$$

Noting that $\mathbb{E}\left\{ \|\Phi_i\|^2 1_{\left\{h^4 \|\Phi_i\|^2 > \delta\right\}} \right\} \leq \left[\mathbb{E}\|\Phi_i\|^4 \right]^{1/2} \left[\delta^{-2} h^8 \mathbb{E}\|\Phi_i\|^4 \right]^{1/2} = O(h^2)$ (since $\mathbb{E}\|\Phi_i\|^4 = O(h^{-2})$), the expected value of the first term behaves as $O(h^2)$. Since it is nonnegative, this means that the first term tends to zero in L^1 and hence in probability. The same is true for the second term since $\alpha_n = \mathbb{E}\left\{ \varepsilon_i^2 1_{\left\{\varepsilon_i^2 > n h^4 \delta\right\}} \right\} \to 0$ (independent of i) and $\frac{1}{n}\sum_{i=1}^{n} \|\Phi_i\|^2$ converges in probability.

As a consequence, by the central limit theorem cited above (cf. proof of Lemma A1.6),

$$
\sqrt{\frac{h^4}{n}} \sum_{i=1}^{n} \Phi_i \widetilde{\varepsilon}_i \xrightarrow{d} \mathcal{N}(0, \sigma^2 A). \tag{A1.25}
$$

Putting together (A1.22) - (A1.15) we obtain the assertion of Theorem 3.4.

8 Appendix 2

Proof of Lemma 4.1

Step 1: Consider first the behavior of

$$
B_n = \frac{h}{n} \sum_{i=1}^{n} G_i S_i,
$$

and let $\hat{D}_n = \frac{1}{n}\sum_{i=1}^{n} G_i S_i$, meaning that $B_n = h\hat{D}_n$. Rewrite

$$
\hat{D}_n = \int \frac{1}{h^2} K\left(\frac{x - x_0}{h} \right) \left[\widehat{m}(x_0, z) - m(x_0, z) \right] \left[\widehat{\gamma}(x_0, z) - \gamma(x_0, z) \right] d\hat{F}_{XZ},
$$

where $\hat{F}_{XZ}$ is the empirical *c.d.f.* of X_i and Z_i. As mentioned, we use a functional expansion. The strategy will be to establish the behavior of the statistic $\hat{D}_n$ which is a functional $\Gamma\left(\hat{m}, \widehat{\gamma}, \hat{F}\right)$, by studying first the behavior of $D_n = \Gamma\left(\hat{m}, \widehat{\gamma}, F\right)$, and show then that the difference $\hat{D}_n - D_n$ is asymptotically negligible. We analyze $\Gamma\left(\hat{m}, \widehat{\gamma}, F\right)$ using a functional expansion around $\Gamma\left(m, \gamma, F\right)$. Introduce the following notation. Let $f_{YXZ}(y, x, z)$ denote the joint density of (Y_i, X_i, Z_i), and let $f_{XZ}(x, z)$ denote the joint density of (X_i, Z_i). Let

$$
\widehat{f_{YXZ}}(y, x, z) = \frac{1}{n H^{d+2}} \sum_{i=1}^{n} K\left(\frac{Y_i - y}{H}, \frac{X_i - x}{H}, \frac{Z_i - z}{H} \right)
$$

denote a Kernel based estimator, where H is the associated bandwidth. Similarly, let

$$\widehat{f_{XZ}}(x,z) = \frac{1}{nH^{d+1}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{H}, \frac{Z_i - z}{H}\right)$$

and

$$\widehat{m}(x,z) = \frac{\int y\widehat{f_{YXZ}}(y,x,z)dy}{\widehat{f_{XZ}}(x,z)}.$$

Let

$$\begin{aligned}\varphi(t,x,z) &= \frac{\int y f_{YXZ}(y,x,z)dy + t\int y g(y,x,z)dy}{f_{XZ}(x,z)+tk(x,z)} - \frac{\int y f_{YXZ}(y,x,z)dy}{f_{XZ}(x,z)} \\ &= \frac{t\left(\left[\int y g(y,x,z)dy\right] f_{XZ}(x,z) - k(x,z)\int y f_{YXZ}(y,x,z)dy\right)}{f_{XZ}(x,z)\left[f_{XZ}(x,z)+tk(x,z)\right]},\end{aligned}$$

for $t \in [0,1]$ and appropriately defined functions

$$g(x,z) = \widehat{f_{YXZ}}(y,x,z) - f_{YXZ}(y,x,z)$$

and

$$k(x,z) = \widehat{f_{XZ}}(x,z) - f_{XZ}(x,z).$$

Moreover,

$$\begin{aligned}\phi(t,x,z) &= \frac{\partial_x f_{XZ}(x,z) + t\partial_x k(x,z)}{f_{XZ}(x,z)+tk(x,z)} - \frac{\partial_x f_{XZ}(x,z)}{f_{XZ}(x,z)} \\ &\quad + \frac{x\left[\partial_x^2 f_{XZ}(x,z) + t\partial_x^2 k(x,z)\right]}{2\left[f_{XZ}(x,z)+tk(x,z)\right]} - \frac{x\partial_x^2 f_{XZ}(x,z)}{2f_{XZ}(x,z)} \\ &= t\frac{\partial_x k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x f_{XZ}(x,z)}{f_{XZ}(x,z)\left[f_{XZ}(x,z)+tk(x,z)\right]} \\ &\quad + t\frac{x\left(\partial_x^2 k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x^2 f_{XZ}(x,z)\right)}{2f_{XZ}(x,z)\left[f_{XZ}(x,z)+tk(x,z)\right]}\end{aligned}$$

Obviously, $\varphi(0,x,z) = \phi(0,x,z) = 0$. Moreover,

$$\begin{aligned}\frac{\partial\varphi(t,x,z)}{\partial t} &= \frac{\int y g(y,x,z)dy\left(f_{XZ}(x,z)+tk(x,z)\right)}{\left(f_{XZ}(x,z)+tk(x,z)\right)^2} \\ &\quad - \frac{k(x,z)\int y f_{YXZ}(y,x,z)dy + t\int y g(y,x,z)dy}{\left(f_{XZ}(x,z)+tk(x,z)\right)^2} \\ &= \frac{f_{XZ}(x,z)\int y g(y,x,z)dy - k(x,z)\int y f_{YXZ}(y,x,z)dy}{\left(f_{XZ}(x,z)+tk(x,z)\right)^2},\end{aligned}$$

$$\frac{\partial^2\varphi(t,x,z)}{\partial t^2} = -2\frac{\left\{f_{XZ}(x,z)\int y g(y,x,z)dy - k(x,z)\int y f_{YXZ}(y,x,z)dy\right\}k(x,z)}{\left(f_{XZ}(x,z)+tk(x,z)\right)^3}$$

Also

$$\begin{aligned}\frac{\partial\phi(t,x,z)}{\partial t} &= \frac{\partial_x k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x f_{XZ}(x,z)}{\left(f_{XZ}(x,z)+tk(x,z)\right)^2} \\ &\quad + \frac{x\left[\partial_x^2 k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x^2 f_{XZ}(x,z)\right]}{2\left(f_{XZ}(x,z)+tk(x,z)\right)^2},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2\phi(t,x,z)}{\partial t^2} &= -2\frac{\left\{\partial_x k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x f_{XZ}(x,z)\right\} k(x,z)}{\left(f_{XZ}(x,z)+tk(x,z)\right)^3} \\ &\quad -\frac{x\left\{\partial_x^2 k(x,z) f_{XZ}(x,z) - k(x,z)\partial_x^2 f_{XZ}(x,z)\right\} k(x,z)}{\left(f_{XZ}(x,z)+tk(x,z)\right)^3}\end{aligned}$$

Next, define

$$\Psi(t) = \int K\left(\frac{x-x_0}{h}\right)\varphi(t,x_0,z)\phi(t,x_0,z)dF_{XZ},$$

where F_{XZ} is the joint density of X_i and Z_i. This implies that

$$\Psi'(t) = \int K\left(\frac{x-x_0}{h}\right)\left(\frac{\partial\varphi(t,x_0,z)}{\partial t}\phi(t,x_0,z) + \varphi(t,x_0,z)\frac{\partial\phi(t,x_0,z)}{\partial t}\right)dF_{XZ},$$

$$\begin{aligned}\Psi''(t) &= \int K\left(\frac{x-x_0}{h}\right)[\phi(t,x_0,z)\frac{\partial^2\varphi(t,x_0,z)}{\partial t^2} \\ &\quad +2\frac{\partial\phi(t,x_0,z)}{\partial t}\frac{\partial\varphi(t,x_0,z)}{\partial t} + \varphi(t,x_0,z)\frac{\partial^2\phi(t,x_0,z)}{\partial t^2}]dF_{XZ},\end{aligned}$$

Note that $\Psi(0) = \Psi'(0) = 0$ due to $\varphi(0,x,z) = \phi(0,x,z) = 0$. Obviously,

$$D_n = \frac{1}{h^2}\Psi(1)$$

Then, by a Taylor-approximation of Ψ around $t = 0$, we have

$$\Psi(t) = \Psi(0) + \Psi'(0)t + \frac{1}{2}\Psi''(\vartheta(t))t^2,$$

where $0 \le \vartheta(t) \le t$. Hence,

$$\begin{aligned}D_n &= \frac{1}{2h^2}\int K\left(\frac{x-x_0}{h}\right)\phi(\vartheta(t),x_0,z)\frac{\partial^2\varphi(\vartheta(t),x_0,z)}{\partial t^2}dF_{XZ} \\ &\quad +\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial\phi(\vartheta(t),x_0,z)}{\partial t}\frac{\partial\varphi(\vartheta(t),x_0,z)}{\partial t}dF_{XZ} \\ &\quad +\frac{1}{2h^2}\int K\left(\frac{x-x_0}{h}\right)\varphi(\vartheta(t),x_0,z)\frac{\partial^2\phi(\vartheta(t),x_0,z)}{\partial t^2}dF_{XZ}\end{aligned} \tag{A2.1}$$

Consider the behavior of the second term first

$$\begin{aligned}&\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial\phi(\vartheta(t),x_0,z)}{\partial t}\frac{\partial\varphi(\vartheta(t),x_0,z)}{\partial t}dF_{XZ} \\ =\ &\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{f_{XZ}(x_0,z)\int yg(y,x_0,z)dy - k(x_0,z)\int yf_{YXZ}(y,x_0,z)dy}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^2} * \\ &\frac{\left\{\partial_x k(x_0,z) + \frac{x_0}{2}\partial_x^2 k(x_0,z)\right\} f_{XZ}(x_0,z) - k(x_0,z)\left\{\partial_x f_{XZ}(x_0,z) + \frac{x_0}{2}\partial_x^2 f_{XZ}(x_0,z)\right\}}{\left(f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right)^2}dF_{XZ}\end{aligned}$$

$$
\begin{aligned}
= \quad & \frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& +\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\frac{x_0}{2}\partial_x^2 k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& -\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{k(x_0,z)\partial_x k(x_0,z)f_{XZ}(x_0,z)\int yf_{YXZ}(y,x_0,z)dy}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& -\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{k(x_0,z)\frac{x_0}{2}\partial_x^2 k(x_0,z)f_{XZ}(x_0,z)\int yf_{YXZ}(y,x_0,z)dy}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& -\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)\right)\partial_x f_{XZ}(x_0,z)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& -\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)\right)\frac{x_0}{2}\partial_x^2 f_{XZ}(x_0,z)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& +\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\left[k(x_0,z)\right]^2\partial_x f_{XZ}(x_0,z)\int yf_{YXZ}(y,x_0,z)dy}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ} \\
& +\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\left[k(x_0,z)\right]^2\frac{x_0}{2}\partial_x^2 f_{XZ}(x_0,z)\int yf_{YXZ}(y,x_0,z)dy}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ}
\end{aligned}
\tag{A2.2}
$$

All of these eight terms are of the same structure. We take the first one as example, the others follow by similar arguments. Turning to the first term,

$$
\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ},
$$

and bounding the denominator

$$
\frac{1}{\left|f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right|}\leq\frac{1}{\left|f_{XZ}(x_0,z)\right|-\left|k(x_0,z)\right|}\leq\frac{2}{b},
$$

since $\vartheta(t)\in[0,1]$, $|f_{XZ}(x,z)|\geq b$, since we assume continuously distributed RV with compact support. Moreover, $|k(x,z)|\leq b/2$ with probability going to one, if $\hat{f}_{XZ}(x,z)$ consistent. Hence

$$
\begin{aligned}
& \frac{1}{h^2}\int\int K\left(\frac{x-x_0}{h}\right)\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}f_{XZ}(x,z)dxdz \\
\leq \quad & \frac{c_1}{h^2}\int\int K\left(\frac{x-x_0}{h}\right)\left|\partial_x k(x_0,z)\right|\left|\int yg(y,x_0,z)dy\right|\left(f_{XZ}(x_0,z)^2\right)f_{XZ}(x,z)dxdz \\
= \quad & \frac{c_1}{h}\int\int K\left(\psi\right)\left|\partial_x k(x_0,z)\right|\left|\int yg(y,x_0,z)dy\right|\left(f_{XZ}(x_0,z)^2\right)f_{XZ}(\psi h+x_0,z)d\psi dz,
\end{aligned}
$$

where c_j, $j=1,2..$ denote constants, and the last equality is by change of variables. Taking the supremum over f_{XZ}, the this term can be bounded by

$$
\begin{aligned}
& \frac{c_2}{h}\int\left|\partial_x k(x_0,z)\right|\left|\int yg(y,x_0,z)dy\right|\int K\left(\psi\right)f_{XZ}(\psi h+x_0,z)d\psi dz \\
\leq \quad & \frac{c_2}{h}\sup_{x,z}\left|\partial_x k(x,z)\right|\sup_{x,z}\left|\int yg(x,y,z)dy\right|\int\int K\left(\psi\right)f_{XZ}(\psi h+x_0,z)d\psi dz,
\end{aligned}
$$

Since $\int\int K\left(\psi\right)f_{XZ}(\psi h+x_0,z)d\psi dz=f_X(x_0)+O_p\left(h^2\right)$, where the remainder terms are uniformly bounded, it follows for this expression that

$$
\leq\frac{c_3}{h}\sup_{x,z}\left|\partial_x k(x,z)\right|\sup_{x,z}\left|\int yg(x,y,z)dy\right|.
$$

The following standard results, e.g. Haerdle (1990) are useful:

$$\begin{aligned}
\sup_{x,z}|k(x,z)|\,,\sup_{x,z}\left|\int yg(x,y,z)dy\right| &= O_p(H_1^r+n^{-1/2}H_1^{-(d+1)/2}\ln(n)),\\
\sup_{x,z}|\partial_x k(x,z)| &= O_p(H_1^{r-1}+n^{-1/2}H_1^{-(d+3)/2}\ln(n)),\\
\sup_{x,z}\left|\partial_x^2 k(x,z)\right| &= O_p(H_1^{r-2}+n^{-1/2}H_1^{-(d+5)/2}\ln(n)),
\end{aligned}$$

which, for compactness of notation, shall be denoted by $O_p(\|g\|_0), O_p(\|g\|_1)$ and $O_p(\|g\|_2)$. These results are not affected by the fact that we use a leave one out estimator, i.e. they have to be taken over i as well, as can be seen by closer inspection of the proof in Masry (1996). The subscripts in the norms are motivated by the numbers of derivatives. Employing these results, we obtain that

$$\begin{aligned}
&\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}dF_{XZ}\\
=\;& h^{-1}O_p(H_0^r+n^{-1/2}H_0^{-(d+1)/2}\ln(n))O_p(H_1^{r-1}+n^{-1/2}H_1^{-(d+3)/2}\ln(n)),
\end{aligned}$$

where H_0 and H_1 are the associated bandwidths. By closer inspection it becomes obvious that every term in (A2.2) has two preestimation error elements. However, some include second derivatives and hence exhibit a slower speed of convergence. The terms including second derivatives dominate the speed of convergence of the entire expression. Thus

$$\begin{aligned}
&\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial\phi(\vartheta(t),x_0,z)}{\partial t}\frac{\partial\varphi(\vartheta(t),x_0,z)}{\partial t}dF_{XZ}\\
=\;& h^{-1}O_p(\|g\|_0)O_p(\|g\|_2),
\end{aligned}$$

meaning that the behavior of the second element of D_n is clarified. The first and third terms in (A2.1) can be shown by similar arguments to contain three preestimation error elements and converge therefore faster than the second. Hence, $D_n = O_p(h^{-1}\|g\|_0\|g\|_2)$.

To see now that $\hat{D}_n - D_n = o_p\left(h^{-1}\|g\|_0\|g\|_2\right)$, note that since Γ is linear in F,

$$\begin{aligned}
\hat{D}_n - D_n &= \Gamma\left(\hat{m},\widehat{\gamma},\hat{F}\right)-\Gamma\left(\hat{m},\widehat{\gamma},F\right)\\
&= \Gamma\left(\hat{m},\widehat{\gamma},\hat{F}-F\right).
\end{aligned}$$

Therefore, the same expansions as above may be used, with $\hat{F}-F$ in place of F. In particular, $\Psi(0)=\Psi'(0)=0$, and hence in the remainder term we are left with

$$\begin{aligned}
\hat{D}_n - D_n &= \frac{1}{2h^2}\int K\left(\frac{x-x_0}{h}\right)\phi(\vartheta(t),x_0,z)\frac{\partial^2\varphi(\vartheta(t),x_0,z)}{\partial t^2}d\left(\hat{F}_{XZ}-F_{XZ}\right)\\
&+\frac{1}{h^2}\int K\left(\frac{x-x_0}{h}\right)\frac{\partial\phi(\vartheta(t),x_0,z)}{\partial t}\frac{\partial\varphi(\vartheta(t),x_0,z)}{\partial t}d\left(\hat{F}_{XZ}-F_{XZ}\right)\\
&+\frac{1}{2h^2}\int K\left(\frac{x-x_0}{h}\right)\varphi(\vartheta(t),x_0,z)\frac{\partial^2\phi(\vartheta(t),x_0,z)}{\partial t^2}d\left(\hat{F}_{XZ}-F_{XZ}\right)
\end{aligned}$$

As a next step, pick again a typical element out of the analogous eight terms in the second rhs term in (A2.2). Being more explicit about the boundaries,

$$\begin{aligned}
&\frac{1}{h^2}\int_{\underline{z}}^{\overline{z}}\int_{\underline{x}}^{\overline{x}} K\left(\frac{x-x_0}{h}\right) \\
&\times\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4} d\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right) \\
= &\frac{1}{h^2}\int_{\underline{z}}^{\overline{z}}\int_{x_0-h/2}^{x_0+h/2}\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4} d\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right).
\end{aligned}$$

Let

$$b_n(z,x_0,t)=\frac{1}{h^2}\frac{\partial_x k(x_0,z)\int yg(y,x_0,z)dy\left(f_{XZ}(x_0,z)^2\right)}{\left[f_{XZ}(x_0,z)+\vartheta(t)k(x_0,z)\right]^4}.$$

Integration by parts yields for the r.h.s.,

$$\begin{aligned}
&\left[b_n(z,x_0,t)\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right)\right]_{x=x_0-h/2,z=\underline{z}}^{x=x_0+h/2,z=\overline{z}} \\
&-\int_{\underline{z}}^{\overline{z}}\int_{x_0-h/2}^{x_0+h/2}\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right)\partial_x\partial_z b_n(z,x_0,t)dxdz.
\end{aligned}$$

Turning to the first term, this equals

$$\begin{aligned}
&b_n(\overline{z},x_0,t)\left(\hat{F}_{XZ}(x_0+h/2,\overline{z})-F_{XZ}(x_0+h/2,\overline{z})\right) \\
&-b_n(\underline{z},x_0,t)\left(\hat{F}_{XZ}(x_0+h/2,\underline{z})-F_{XZ}(x_0+h/2,\underline{z})\right) \\
&-b_n(\overline{z},x_0,t)\left(\hat{F}_{XZ}(x_0-h/2,\overline{z})-F_{XZ}(x_0-h/2,\overline{z})\right) \\
&+b_n(\underline{z},x_0,t)\left(\hat{F}_{XZ}(x_0-h/2,\underline{z})-F_{XZ}(x_0-h/2,\underline{z})\right).
\end{aligned}$$

Each of these four expressions has the same structure. Since

$$\left(\hat{F}_{XZ}(x_0+h/2,\overline{z})-F_{XZ}(x_0+h/2,\overline{z})\right)=O_p\left(\sqrt{n}\right)$$

by Glivenko-Cantelli,

$$b_n(\overline{z},x_0,t)\left(\hat{F}_{XZ}(x_0+h/2,\overline{z})-F_{XZ}(x_0+h/2,\overline{z})\right)=O_p\left(h^{-2}\left\|g\right\|_0\left\|g\right\|_1\right)O_p\left(\sqrt{n}\right),$$

and the same is true for all other terms. Hence,

$$\left[b_n(z,x_0,t)\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right)\right]_{x=x_0-h/2,z=\underline{z}}^{x=x_0+h/2,z=\overline{z}}=ho_p(\left\|g\right\|_0\left\|g\right\|_1),$$

as $h^{-3}=n^{\frac{3}{7}}<\sqrt{n}$. Now turn to

$$\begin{aligned}
&\int_{\underline{z}}^{\overline{z}}\int_{x_0-h/2}^{x_0+h/2}\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right)\partial_x\partial_z b_n(z,x_0,t)dxdz \\
= &\int_{\underline{z}}^{\overline{z}}\int_{\underline{x}}^{\overline{x}} K\left(\frac{x-x_0}{h}\right)\left(\hat{F}_{XZ}(x,z)-F_{XZ}(x,z)\right)\partial_x\partial_z b_n(z,x_0,t)dxdz
\end{aligned}$$

The rhs is bounded by

$$c_4 \int |\partial_x \partial_z b_n(z, x_0, t)| \left| \hat{F}_{XZ}(x, z) - F_{XZ}(x, z) \right| dxdz.$$

As above, a typical element of $\partial_x \partial_z b_n(z, x_0, t)$ is given by

$$\frac{1}{h^2} \frac{\partial_x^2 \partial_z k(x_0, z) \int y g(y, x_0, z) dy \left(f_{XZ}(x_0, z)^2 \right)}{\left[f_{XZ}(x_0, z) + \vartheta(t) k(x_0, z) \right]^4}$$

The denominator in this terms is again bounded by a constant. Then it is obvious that

$$\begin{aligned} & \frac{1}{h^2} \left| \partial_x^2 \partial_z k(x_0, z) \int y g(y, x_0, z) dy \left(f_{XZ}(x_0, z)^2 \right) \right| \\ \leq & \frac{1}{h^2} \sup_{x,z} \left| \partial_x^2 \partial_z k(x_0, z) \right| \sup_{x,z} \left| \int y g(y, x_0, z) dy \right| \sup_{x,z} f_{XZ}(x_0, z)^2 \\ \leq & \frac{1}{h^2} c_5 \left\| g \right\|_3 \left\| g \right\|_0 \end{aligned}$$

Hence $\sup_{x,z} |\partial_x \partial_z b_n(z, x, t)| = O_p\left(\left\| g \right\|_0 \right) O_p\left(\left\| g \right\|_3 \right)$ and

$$\begin{aligned} & \int_{\underline{z}}^{\overline{z}} \int_{\underline{x}}^{\overline{x}} |\partial_x \partial_z b_n(z, x_0, t)| \left| \hat{F}_{XZ}(x, z) - F_{XZ}(x, z) \right| dxdz \\ \leq & c_6 \sup_{x,z} |\partial_x \partial_z b_n(z, x, t)| \sup_{x,z} \left| \hat{F}_{XZ}(x, z) - F_{XZ}(x, z) \right| \\ = & \frac{c_6}{h^2} O_p\left(\left\| g \right\|_0 \right) O_p\left(\left\| g \right\|_3 \right) O_p\left(n^{-1/2} \right) = o_p\left(h^{-1} \left\| g \right\| \left\| g \right\|_2 \right), \end{aligned}$$

where the last equality follows from

$$O_p\left(\left\| g \right\|_3 \right) O_p\left(h^{-1} n^{-1/2} \right) = o_p\left(\left\| g \right\|_2 \right).$$

Since it was to establish when $(nh)^{-1} h^2 \sum_{i=1}^n G_i S_i = o_p(h^2)$, or

$$\frac{1}{nh} \sum_{i=1}^n G_i S_i = o_p(1),$$

we can give the answer that this the case, if $O_p(n^{2/7} \left\| g \right\|_0 \left\| g \right\|_2) = o_p(1)$. Going into detail this is the case if

$$O_p(n^{2/7} H_0^{r_0} H_1^{r_1 - 2} + n^{-5/7} H_0^{-(d+1)/2} H_1^{-(d+5)/2} \ln(n)) = o_p(1),$$

and the theorem follows.

Step 2: Now we give conditions under which

$$\frac{1}{nh} \sum_{i=1}^n \left(W_i X_i - \mathbb{E}\left[W_i X_i | W_i Z_i \right] \right) S_i + \frac{h}{n} \sum_{i=1}^n \left(W_i Y_i - \mathbb{E}\left[W_i Y_i | W_i Z_i \right] \right) G_i.$$

is $o_p(h^2)$. As mentioned above, S_i is $\mathcal{F}_n$-measurable.

Consider now $\mathbb{E}\left\{ \left(W_i X_i - \mathbb{E}\left[W_i X_i | W_i Z_i \right] \right) S_i \right\}$. First note that this expectation need not exist, as we

are dividing by the sum of the kernels in S_i. without further mentioning, we employ the modified estimator, where the denominator in $\hat{m}_{-i}$ contains $\delta(n)$ which converges to zero fast, e.g., n^{-2}, as in Fan (1992).

To see that $\mathbb{E}\left\{\left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i\right\} = 0$, consider

$$\begin{aligned} & \mathbb{E}\left\{\left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i\right\} \\ = \; & \mathbb{E}\left\{\mathbb{E}\left[\left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i|\mathcal{F}_n\right]\right\} \\ = \; & \mathbb{E}\left\{\left(\mathbb{E}\left[W_iX_i|\mathcal{F}_n\right] - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i\right\}. \end{aligned}$$

But $\mathbb{E}\left[W_iX_i|\mathcal{F}_n\right] = \mathbb{E}\left[W_iX_i|W_iZ_i\right]$, due to the *iid* assumption, and hence

$$\mathbb{E}\left\{\left(W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right)S_i\right\} = 0$$

Next turn to the variance. Let $C_i = W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]$. Then,

$$\begin{aligned} & \mathbb{V}\left\{\frac{1}{nh}\sum_i C_iS_i\right\} \\ = \; & \mathbb{E}\left\{\frac{1}{n^2h^2}\sum_i\sum_j C_iS_iS_jC_j\right\} \\ = \; & \frac{1}{nh^2}\mathbb{E}C_i^2S_i^2 + \mathbb{E}\left\{h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i} C_iS_iS_jC_j\right\}. \end{aligned}$$

The first term poses no problems, as it is smaller than

$$\left(nh^4\right)^{-1}\sup\left|\hat{m}_{-i} - m\right|^2\mathbb{E}C_i^2 = O\left(n^{-1}h^{-3}\left(H_0^{2r_0} + n^{-1}H_0^{-(d+1)}\right)\right),$$

or, using optimal rate of bandwidth $h = n^{-1/7}$, $O\left(n^{-4/7}H_0^{r_0} + n^{-11/7}H_0^{-(d+1)/2}\right)$. Note again that for $\sup\left|\hat{m}_{-i} - m\right|^2$ the standard uniform convergence results for $\sup\left|\hat{m} - m\right|$ extends to the leave one out estimator. The second term in the above expansion is

$$\begin{aligned} & h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i - x_0}{h}\right)K\left(\frac{X_j - x_0}{h}\right)\left[\widehat{m}_{-i}(x_0, Z_i) - m(x_0, Z_i)\right] \\ & \times\left[\widehat{m}_{-j}(x_0, Z_j) - m(x_0, Z_j)\right]\left[W_iX_i - \mathbb{E}\left[W_iX_i|W_iZ_i\right]\right] \\ & \times\left[W_jX_j - \mathbb{E}\left[W_jX_j|W_jZ_j\right]\right]\} \end{aligned} \tag{A2.3}$$

Denote the σ-algebra spanned by

$$W_1X_1, .., W_1X_{i-1}, W_1X_{i+1}, .., W_1X_n, W_1Z_1, .., W_1Z_n, W_1Y_1, .., W_1Y_{i-1}, W_1Y_{i+1}, .., W_1Y_n,$$

$$\vdots$$

$$W_iX_1, .., W_iX_{i-1}, W_iX_{i+1}, .., W_iX_n, W_iZ_1, .., W_iZ_n, W_iY_1, .., W_iY_{i-1}, W_iY_{i+1}, .., W_iY_n,$$

$$\vdots$$

$$W_nX_1, .., W_nX_{i-1}, W_nX_{i+1}, .., W_nX_n, W_nZ_1, .., W_nZ_n, W_nY_1, .., W_nY_{i-1}, W_nY_{i+1}, .., W_nY_n,$$

as $\mathcal{F}_{-i,n}$. Then we can rewrite (A2.3) as

$$\begin{aligned} & h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)[\widehat{m}_{-i}(x_0,Z_i)-m(x_0,Z_i)] \\ & \times\mathbb{E}\left[[\widehat{m}_{-j}(x_0,Z_j)-m(x_0,Z_j)]\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|\mathcal{F}_{-i,n}\right] \\ & \times\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right]\}. \end{aligned} \tag{A2.4}$$

The whole expression would be zero, if $\widehat{m}_{-j}(x_0,Z_j)$ were not a function of X_i and Y_i, since then

$$\mathbb{E}\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]|\mathcal{F}_{-i,n}\right]=0,$$

due to iid. Denote the **leave two out estimator** as $\widehat{m}_{-i,j}(x_0,Z_j)$. Then,

$$\begin{aligned} \widehat{m}_{-j}(x_0,Z_j) &= \widehat{m}_{-i,j}(x_0,Z_j)+\frac{K\left(\frac{X_i-x_0}{h}\right)}{\sum\limits_{s\neq j}K\left(\frac{X_s-x_0}{h}\right)}[Y_i-\widehat{m}_{-i,j}(x_0,Z_i)] \\ &= \widehat{m}_{-i,j}(x_0,Z_j)+H_i\left[m(X_i,Z_i)+\varepsilon_i-\widehat{m}_{-i,j}(x_0,Z_i)\right] \\ &= \widehat{m}_{-i,j}(x_0,Z_j)+H_i\left[m(x_0,Z_i)-\widehat{m}_{-i,j}(x_0,Z_i)\right] \\ &\quad +H_i\partial_x m(x_r,Z_i)(X_i-x_0)+H_i\varepsilon_i, \end{aligned}$$

where $H_i=\frac{K\left(\frac{X_i-x_0}{h}\right)}{\sum\limits_{s\neq j}K\left(\frac{X_s-x_0}{h}\right)}$ and x_r is an intermediate position. Obviously, H_i is $\mathcal{F}_{-i,n}$-measurable. Substituting this expansion into (A2.4) produces

$$\begin{aligned} & h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)[\widehat{m}_{-i}(x_0,Z_i)-m(x_0,Z_i)] \\ & \times H_i\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|\mathcal{F}_{-i,n}\right] \\ & \times\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right] \\ & +\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)[\widehat{m}_{-i}(x_0,Z_i)-m(x_0,Z_i)] \\ & \times H_i\mathbb{E}\left[\varepsilon_i\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|\mathcal{F}_{-i,n}\right] \\ & \times\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right]\}. \end{aligned} \tag{A2.5}$$

The second term in (A2.5) is zero, due to

$$\begin{aligned} & \mathbb{E}\left[\varepsilon_i\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|\mathcal{F}_{-i,n}\right] \\ = \; & \mathbb{E}\left[\varepsilon_i\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right] \\ = \; & \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i|X_i,Z_i\right]\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right]=0. \end{aligned}$$

Hence we are left with the first expression in (A2.5), which in light of *iid* becomes,

$$\begin{aligned} & h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)[\widehat{m}_{-i}(x_0,Z_i)-m(x_0,Z_i)] \\ & \times H_i\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right] \\ & \times\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right] \end{aligned}$$

Taking again conditional expectations, now with respect to $\mathcal{F}_{-j,n}$, produces

$$\begin{aligned}
&h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\\
&\times H_i\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right]\\
&\times H_j\mathbb{E}\left[\partial_x m(x_r,Z_j)(X_j-x_0)\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right]|W_jZ_j\right]\}.
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right]\\
=\;&\mathbb{E}[\partial_x m(x_r,Z_i)(X_i-x_0)\\
&\times\left[K\left(\frac{X_i-x_0}{h}\right)W_i\left(X_i-x_0\right)-\mathbb{E}\left[W_i\left(X_i-x_0\right)|W_iZ_i\right]\right]|W_iZ_i],
\end{aligned}$$

since $W_i=W_iK\left(\frac{X_i-x_0}{h}\right)$. Due to Lemma A.1, the rhs equals now

$$\begin{aligned}
&\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)^2W_i|W_iZ_i\right]\\
&-\mathbb{E}\left[\partial_x m(x_r,Z_i)K\left(\frac{X_i-x_0}{h}\right)(X_i-x_0)|W_iZ_i\right]\mathbb{E}\left[W_i\left(X_i-x_0\right)|W_iZ_i\right]\\
=\;&h\partial_x m(x_0,Z_i)+h^2\zeta_i,
\end{aligned}$$

where

$$\begin{aligned}
\zeta_i\;=\;&\mathbb{E}\left[\partial_x^2 m(x_l,Z_i)(X_i-x_0)^3W_i|W_iZ_i\right]\\
&-\mathbb{E}\left[\partial_x m(x_r,Z_i)K\left(\frac{X_i-x_0}{h}\right)(X_i-x_0)|W_iZ_i\right]\mathbb{E}\left[W_i\left(X_i-x_0\right)|W_iZ_i\right]
\end{aligned}$$

Note that ζ_i contains higher order terms in h and is uniformly bounded by assumption. Then,

$$\begin{aligned}
&h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}\mathbb{E}\{\frac{1}{h^2}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\\
&\times H_i\mathbb{E}\left[\partial_x m(x_r,Z_i)(X_i-x_0)\left[W_iX_i-\mathbb{E}\left[W_iX_i|W_iZ_i\right]\right]|W_iZ_i\right]\\
&\times H_j\mathbb{E}\left[\partial_x m(x_r,Z_j)(X_j-x_0)\left[W_jX_j-\mathbb{E}\left[W_jX_j|W_jZ_j\right]\right]|W_jZ_j\right]\}
\end{aligned}$$

$$
\begin{aligned}
= \quad & h^{-2}\frac{2}{n^2}\mathbb{E}\left\{\frac{\sum_i\sum_{j>i}K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\partial_x m(x_0,Z_i)\partial_x m(x_0,Z_j)}{\left(\sum_{s\neq j}K\left(\frac{X_s-x_0}{h}\right)\right)^2}\right\} \\
& +h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}h\mathbb{E}\left[\frac{K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\partial_x m(x_0,Z_j)\zeta_i}{\left(\sum_s K\left(\frac{X_s-x_0}{h}\right)\right)^2}\right] \\
& +h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}h\mathbb{E}\left[\frac{K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\partial_x m(x_0,Z_i)\zeta_j}{\left(\sum_s K\left(\frac{X_s-x_0}{h}\right)\right)^2}\right] \\
& +h^{-2}\frac{2}{n^2}\sum_i\sum_{j>i}h^2\mathbb{E}\left[\frac{K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)\zeta_i\zeta_j}{\left(\sum_s K\left(\frac{X_s-x_0}{h}\right)\right)^2}\right]
\end{aligned}
$$

The first term is smaller than

$$
h^{-2}\frac{1}{n^2}\sup_{x,z}\left|\partial_x m(x,z)\right|^2\mathbb{E}\left\{\frac{\sum_i\sum_j K\left(\frac{X_i-x_0}{h}\right)K\left(\frac{X_j-x_0}{h}\right)}{\left(\sum_s K\left(\frac{X_s-x_0}{h}\right)\right)^2}\right\}=O\left(\frac{1}{n^2h^2}\right),
$$

while the others converge even faster since they are of higher order in h and ζ_i as well as ζ_j are uniformly bounded. As a consequence, the variance dominates the covariances, and hence

$$
\mathbb{V}\left\{\frac{1}{nh}\sum_i C_iS_i\right\}=O\left(n^{-4/7}H_0^{2r_0}+n^{-11/7}H_0^{-(d+1)}\right),
$$

implying $h^2(nh)^{-1}\sum_i C_iS_i = O_p\left(H_0^{r_0}+n^{-1/2}H_0^{-(d+1)/2}\right)$, i.e. this term converges under any circumstances faster than the fastest bias term. By the same arguments and under weaker conditions, $\frac{1}{nh}h^2\sum_i\left[W_iY_i-\mathbb{E}\left[W_iY_i|W_iZ_i\right]\right]G_i=o_p\left(h^2\right)$, and this concludes Step 2.

Step 3: Finally, the behavior of the higher order terms has to be clarified, i.e. when

$$
\frac{1}{nh}\sum_{i=1}^n\xi_i=o_p\left(h^2\right).
$$

Take a typical element, e.g. $\frac{1}{nh}\sum_{i=1}^n S_i\phi_i$, where ϕ_i is $O_p\left(h^4\right)$ and uniformly bounded. Then, by similar arguments as above

$$
\frac{1}{nh}\sum_{i=1}^n S_i\phi_i=O_p\left(h\left[H_0^{r_0}+n^{-1/2}H_0^{-(d+1)/2}\ln(n)\right]\right),
$$

meaning that this is $o_p\left(h^2\right)$ if $h=o_p\left(H_0^{r_0}+n^{-1/2}H_0^{-(d+1)/2}\ln(n)\right)$. Setting $h=O\left(n^{-1/7}\right)$, we have that this is the case without higher order smoothness assumptions, provided $d+1<10$. This concludes the proof. $\square$

References

[1] AHN, H. AND J. POWELL (1993), "Semiparametric Estimation of Censored Regression Models with a Nonparametric Selection Mechanism", *Journal of Econometrics*, 58, 3-29.

[2] ANDREWS, D.W.K. AND Y.-J. WHANG (1990): "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality," *Econometric Theory*, 6, 466-479

[3] BLUNDELL, R., BROWNING, M., AND I. CRAWFORD (2003): "Nonparametric Engel Curves and Revealed Preference", IFS Working Paper

[4] CHEN, R. AND R. TSAY (1993), "Functional Coefficient Autoregressive Models", *Journal of the American Statistical Association,* 88, 298-308.

[5] CHIANG, C., RICE, J., AND C. WU (2001), "Smoothing Splines Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variable", *Journal of the American Statistical Association,* 96, 605-619.

[6] CHRISTOPEIT, N., AND S.G.N.HODERLEIN (2001): "Further Asymptotic Properties of Local Polynomial Estimators," Working Paper, Bonn University Working Paper.

[7] DAS, M., NEWEY, W.K., AND F.VELLA (1999): "Nonparametric Estimation of Sample Selection Models," Working Paper, Columbia University Working Paper.

[8] DEATON, A., AND J. MUELLBAUER (1980): "An Almost Ideal Demand System," *American Economic Review*, 70, 312-326.

[9] FAN, J. (1992): "Design Adaptive Nonparametric Regression," *Journal of the American Statistical Association,* 87, 998-1004.

[10] FAN, J., AND I. GIJBELS (1996): "Local Polynomial modeling and Its Application", *Monographs on Statistics and Applied Probability*, Chapman and Hall.

[11] FAN, J., AND W. ZHANG (1999), "Statistical Estimation in Varying Coefficient Models", *Annals of Statistics*, 27, 1491-1518.

[12] FLORENS, J.-P., J. HECKMAN, C. MEGHIR AND E. VYTLACIL (2001), "Instrumental Variables, Local Instrumental Variables, and Control Functions", Unpublished Working paper, Chicago.

[13] HECKMAN, J., AND E.VYTALCIL (2004), "Structural Equations, Treatment Effects and Econometric Policy Evaluation", *Econometrica,* forthcoming.

[14] HECKMAN, J., H. ICHIMURA, J. SMITH AND P. TODD (1998), "Charcterizing Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-98.

[15] HECKMAN, J. AND ROBB, (1985): "Alternative Methods for Estimating the Impact of Interventions", *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. New York: Cambridge University Press, 156-245.

[16] HOROWITZ, J. (2001): "Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function," *Econometrica*, 69, 499-514.

[17] HOROWITZ, J., KLEMELÄ, J., AND E. MAMMEN (2004), "Optimal Estimation in Additive Regression Models", Working paper, Mannheim.

[18] HODERLEIN, S., (2002): "Econometric Modelling of Heterogeneous Consumer Behaviour - Theory, Empirical Evidence and Aggregate Implications," *PhD Thesis*, *LSE*.

[19] HODERLEIN, S., (2004): "Nonparametric Demand Systems and a Heterogeneous Population," *Working Paper, Uni Mannheim.*.

[20] KIM, W., O.B. LINTON AND N.W. HENGARTNER (1999): " A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals", *Journal of Computational and Graphical Statistics,* 8, 278-297.

[21] LEWBEL, A., (2001): Demand Systems With and Without Errors, *American Economic Review*, 91, 611-18.

[22] LINTON, O.B., (1997): "Efficient Estimation of Additive Nonparametric Regression Models", *Biometrika,* 84, 468-473.

[23] LINTON, O.B., (2000): "Efficient Estimation of Generalized Additive Nonparametric Regression Models", *Econometric Theory*, 16, 502-523.

[24] LINTON, O.B., AND J.P. NIELSEN (1995): "A Kernel Method of Estimating Structured Non-Parametric Regression based on marginal integration," *Biometrika,* 82, 93-101.

[25] MAMMEN, E., LINTON, O.B., AND J.P. NIELSEN (1999): "The Existence and Asymptotic Properties of a Backfitting Algorithm under Weak Conditions," *Annals of Statistics,* 27, 1443-1490.

[26] MASRY, E. (1996): "Multivariate Local Polynomials for Time Series. Uniform Strong Consistency and Rates" *Journal of Time Series Analysis*, 17, 571-599.

[27] MERTON, R., (1971), "Optimum Consumption and Portfolio Rules in a Continuous Time Model," *Journal of Economic Theory,* 3, 373-413.

[28] NEWEY, W.K., (1994,a): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233-253.

[29] NEWEY, W.K., (1994,b): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica,* 62, 1349-1382.

[30] NEWEY, W.K., (1995): "Convergence Rates and Asymptotic Normality for Series Estimators," MIT, Economics Department Working Paper.

[31] NEWEY, W.K., POWELL, J.L., AND F. VELLA (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models", *Econometrica,* 67, 565-603.

[32] OPSOMER, J.D. AND D. RUPPERT (1997): "Fitting a Bivariate Additive Model by Local Polynomial Regression, " *Annals of Statistics,* 25, 186-211.

[33] POLLARD, D. (1984): *Convergence of Stochastic Processes.* Springer-Verlag, New York.

[34] ROBINSON, P. (1988): "Root- N-Consistent Semiparametric Regression," *Econometrica,* 56, 931-54.

[35] STONE, C.J. (1985): "Additive Regression and Other Nonparametric Models," *Annals of Statistics,* 13, 689-705.

[36] SPERLICH, S. D., TJØSTHEIM AND L. YANG (2002): "Nonparametric Estimation and Testing of Interaction in Additive Models," *Econometric Theory*, 18, 197-251.

[37] TJØSTHEIM, D., AND B. AUESTAD (1994): "Nonlinear Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association*, 89, 1398-1409.

[38] WAHBA, G. (1992): "Spline Methods for Observational Data", Philadelphia: Society for Industrial and Applied Mathematics.