Econometric Theory, **18**, 2002, 886–912. Printed in the United States of America.
DOI: 10.1017.S0266466602184040

ESTIMATION IN AN ADDITIVE MODEL WHEN THE COMPONENTS ARE LINKED PARAMETRICALLY

Raymond J. Carroll
Texas A&M University

Wolfgang Härdle
Humboldt-Universität zu Berlin

Enno Mammen
Ruprecht-Karls-Universität Heidelberg

Motivated by a nonparametric GARCH model we consider nonparametric additive autoregression models in the special case that the additive components are linked parametrically. We show that the parameter can be estimated with parametric rate and give the normal limit. Our procedure is based on two steps. In the first step nonparametric smoothers are used for the estimation of each additive component without taking into account the parametric link of the functions. In a second step the parameter is estimated by using the parametric restriction between the additive components. Interestingly, our method needs no undersmoothing in the first step.

1. INTRODUCTION

1.1. Background

Additive nonparametric regression models have found wide use in statistics (Hastie and Tibshirani, 1990) and remain an area of vigorous research (Opsomer and Ruppert, 1997; Opsomer, 2000; Mammen, Linton, and Nielsen, 1999; Linton, 1997; Fan, Härdle, and Mammen, 1998). This paper explores a variant of the problem in which the components of the additive model are linked parametrically.

The recent development of nonlinear time series analysis is primarily due to the efforts to overcome the limitations of linear models such as the autoregres-

The research in this paper was begun when R. Carroll and E. Mammen were visiting the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Humboldt-Universität zu Berlin. Some basic ideas of the paper were developed in discussions with Dag Tjøstheim. Carroll's research was supported by a grant from the National Cancer Institute (CA–57030) and through the Texas A&M Center for Environmental and Rural Health by a grant from the National Institute of Environmental Health Sciences (P30-ESO9106). The research of Enno Mammen was supported by the Deutsche Forschungsgemeinschaft, projects MA1026/6-1 and MA1026/6-2. Address correspondence to: Raymond J. Carroll, Department of Statistics, Texas A&M University, College Station, TX 77843–3143, USA. Wolfgang Härdle, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, D-10178 Berlin, Germany. Enno Mammen, Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany.

sive moving-average (ARMA) models of Box and Jenkins (1976) in real applications. It has long been recognized that financial time series models that incorporate clusters of volatilities are more appropriate than ARMA specifications. We consider here as a motivating example an application of nonlinear time series analysis to foreign exchange high-frequency data.

For these data the autoregressive heteroskedastic models (ARCH) by Engle (1982) have been studied extensively. An ARCH model for time series $\{Y_t\}$ with ARCH error term of order q is defined through $X_t = \sigma_t \xi_t$, where ξ_t are independent mean zero and variance one random variables and $\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \alpha_2 X_{t-2}^2 + \cdots + \alpha_q X_{t-q}^2$, with $\omega > 0, \alpha_i \geq 0, i = 1, \ldots, q$.

In foreign exchange data it has been found that the order q has to be selected quite high to fit the model well (see Bollerslev, 1986). The reason is volatility clusters; i.e., the conditional variances are highly correlated. An ARMA-like model for the squared observations was therefore proposed for σ_t^2 in Bollerslev (1986):

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2. \tag{1}$$

Models of this type are called GARCH(p,q) models. For a general discussion of GARCH models see also Bollerslev, Engle, and Nelson (1994). Although this model class showed better fitting properties it was soon criticized because the dependence on past observations is treated in a symmetric way: positive and negative shocks of $X_{t-1}, \ldots, X_{t-q}$ have the same influence on the volatility of the current period (see Gouriéroux, 1997). Modifications of the GARCH model that allow asymmetries have been proposed by Nelson (1991) (EGARCH models) and by Rabemananjara and Zakoian (1993) and Zakoian (1994) (TGARCH models). In a TGARCH(1,1) model the time series is generated by the dynamics

$$X_t = \sigma_t \xi_t; \tag{2}$$

$$\sigma_t^2 = g(X_{t-1}) + \beta \sigma_{t-1}^2, \tag{3}$$

where g is a piecewise linear function with given break points. Clearly, this includes asymmetric choices of g. Furthermore, by increasing the number of breakpoints of g the GARCH(1,1) specification $g(x) = \omega + \alpha_1 x^2$ can be approximated by a TGARCH(1,1) model. In this paper we study a nonparametric generalization of this model. We assume that g is a smooth function, and we do not use any parametric assumptions on g. This generalizes the GARCH(1,1) and the TGARCH(1,1) model. In a slightly modified model we discuss how g can be estimated by smoothing methods. For a motivation of our modified model note that by repeated use of equation (3), we can write $\sigma_t^2 = \sum_{j=1}^{\infty} \beta^{j-1} g(X_{t-j})$. An approximation of this model with a finite number J of lags reads

$$X_t^2 = \sum_{j=1}^{J} \beta^{j-1} g(X_{t-j}) + \epsilon_t, \tag{4}$$

with $\epsilon_t = X_t^2 - \sigma_t^2$. Model (4) was proposed in Hafner (1998). Other non- and semiparametric extensions of ARCH models have been studied by Gouriéroux and Monfort (1992) (QTARCH model) and by Härdle and Tsybakov (1997) and Härdle, Tsybakov, and Yang (1998) (CHARN model). In this paper we will discuss nonparametric estimation of g in model (4). Model (4) is an approximation to model (2) and (3), being implied by it when $J = \infty$. This brings up the obvious question: why not work with (2) and (3) directly? The major reason is that there is no known way to estimate $g(\cdot)$ and β directly in (2) and (3) unless $g(\cdot)$ is low-dimensionally parameterized, as, e.g., in GARCH(1,1) or TGARCH(1,1) models. Furthermore, in general also model (2) and (3) is only an approximation to the stochastic structure of observed data. So a priori it is not clear if (4) or (2) and (3) will provide better approximations in applications. We will study estimates of g and β only under the assumption that model (4) holds. It would be interesting to know how these estimates would behave asymptotically if (2) and (3) hold or more generally another model is correct. However, an asymptotic theory for our estimates in case of model misspecification is out of the scope of this paper. For finite samples, we conjecture that for moderate choices of J our estimates behave similarly in model (4) and model (2) and (3); see also our data example, which follows.

We think that the main advantages of our nonparametric approach are possible applications for model checks and choice. Our estimates of g and β can be compared with alternative parametric proposals. These comparisons give a hint if the shape of g is captured well by a parametric model. We will present a data example where one can see that a classical GARCH model gets the shape of g quite well. Furthermore, it can be checked to see if the parameter β is underestimated in a parametric model. This may happen because an overly restrictive shape is assumed for g in a parametric model. A rigorous statistical analysis of such questions requires development of test procedures. Our theory suggests using test statistics that compare parametric fits of g with our estimate. However theory for such tests will not be developed in this paper. In this paper we will discuss asymptotic theory only for estimates.

We now describe our general methodology. We write $Y_t = X_t^2$ and $\mathbf{X}_t = (X_{t-1}, \ldots, X_{t-J})^T$. Then model (4) can be written as a regression problem with scalar response Y and vector of regressors $\mathbf{X} = (X_1, \ldots, X_J)^T$. Equation (4) is a special case of the ordinary additive model,

$$E(Y|\mathbf{X}) = E(Y) + \sum_{j=1}^{J} m_j(X_j), \tag{5}$$

where for identifiability the component functions satisfy $E\{m_j(X_j)\} = 0$. In our problem the component functions $m_1(\cdot), \ldots, m_J(\cdot)$ are linked by a scalar parameter β_0, and for $j \geq 2$,

$$m_j(x) = \beta_0^{j-1} m_1(x). \tag{6}$$

Our purpose here is to estimate both the parameter β_0 and the base function $m_1(x)$. (Note that $m_1(x) = g(x)$ in the preceding notation.) Among the many possibilities, one stands out as relatively straightforward, namely, to estimate the component functions in the general model (5) and somehow "shrink" them to the model (6). One method we pursue, which is based on considerations from the fields of errors in variables and minimum distance estimation, is computationally straightforward, with the estimate of β_0 having an estimable standard error. In addition, the estimator has the pleasing property that the fit to model (5) can be done in a standard fashion, without the need for any undersmoothing to insure that the estimate of β_0 converges at standard parametric rates. The analysis of this method leads to a second method that is equally simple to compute.

We now illustrate the application of model (4) to foreign exchange (FX) rates. The behavior of FX rates has been the subject of many recent investigations. A correct understanding of the FX rate dynamics has important implications for international asset pricing theories, the pricing of contingent claims, and policy-oriented questions. The FX market is an electronic market, active 24 hours a day. The data set HFDF93 on which the following analysis is based was acquired from Olsen and Associates, Zürich. It contains bid and ask quotes for the rates Deutsche mark against U.S. dollar (DEM/USD), during the time period from October 1, 1992, through February 16, 1993. The quotes are collected from the Reuters FXFX page, which is considered a broad but not "complete" data supply. For more information about this data set, see Dacorogna, Müller, Nagler, Olsen, and Pictet (1993) and, more generally for information about FX rate data suppliers and intradaily FX data, Goodhart and Figliuoli (1991). For October 1, 1992, through February 16, 1993, Figure 1 shows a plot of the DEM/USD log–returns (namely, the logarithm of the ratio of the current to the previous price). Our data set contains 10,000 data values. A kernel density estimate of the (log-) returns is shown in Figure 2. For another nonparametric analysis of DEM/USD exchange rates see Bossaerts, Härdle and Hafner (1996). For an overview of applications of ARCH models in finance see Gouriéroux (1997).

The paper is organized as follows. In Section 2, we define the methods used. Section 3 provides details of the motivating example. Section 5 describes a small simulation. Section 4 discusses our general theory. All proofs are in the Appendix.

2. THE METHODS

The data are $(Y_{J+1}, \mathbf{X}_{J+1}), \dots, (Y_n, \mathbf{X}_n)$, where $Y_i = X_i^2$ and $\mathbf{X}_i = (X_{i-1}, \dots, X_{i-J})^T$. We describe here the two methods used in this paper. Our methods rely on estimates $\{\hat{m}_1(\cdot), \dots, \hat{m}_J(\cdot)\}$ from the model (5), specific examples of which are discussed later in this article.

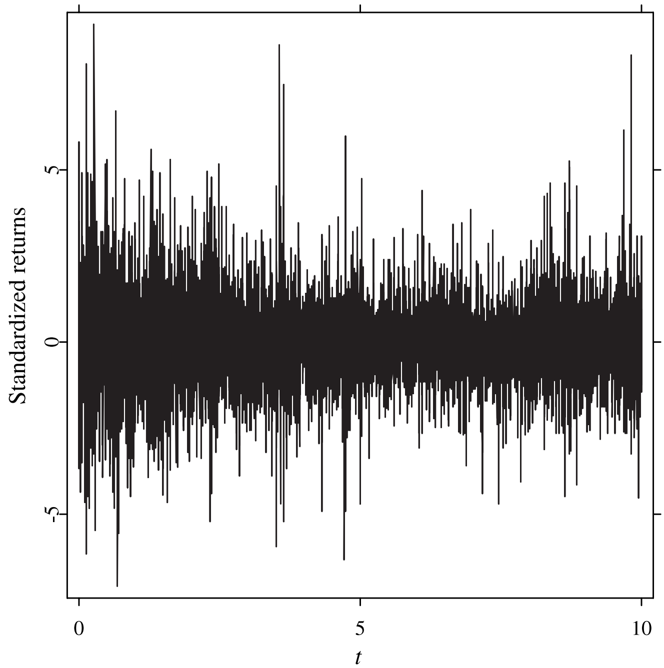

FIGURE 1. Log returns for the rates Deutsche mark against U.S. dollar (DEM/USD), during the time October 1, 1992 through February 16, 1993 (scaled by their empirical standard deviation).

2.1. A Method from Errors in Variables/Minimum Distance

The first method has a natural interpretation as a functional errors-in-variables method (Fuller, 1987) and also as a minimum distance method. We will refer to the method using the former terminology, although the latter can also be used.

To explain this, first fix x. Then ignoring bias and other technical details, it is generally the case that for some constants $c_n \to 0$ and functions $w_j(x)$, the functions $\{\hat{m}_j(x)\}_{j=1}^{J}$ form a set of nearly independent, nearly normal random variables: $\hat{m}_j(x) \sim \text{Normal}\{\beta_0^{j-1} m_1(x), c_n/w_j(x)\}$. If one only had this single fixed x, then the unknowns would be β_0 and $m_1(x)$, and they could be estimated by minimizing in β and $m_1(x)$

$$\sum_{j=1}^{J} w_j(x)\{\hat{m}_j(x) - \beta^{j-1} m_1(x)\}^2. \tag{7}$$

That this is a minimum distance method is clear. It is also a classical functional errors-in-variables estimate with "responses" $\{\hat{m}_j(x)\}_{j=2}^{J}$, "true predictor" $m_1(x)$, and "error-prone predictor" $\hat{m}_1(x)$.

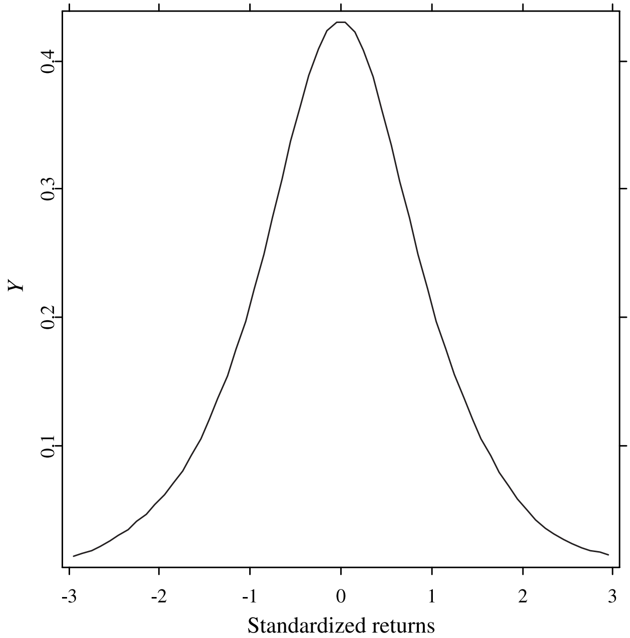

FIGURE 2. Kernel density estimate of the data from Figure 1 with bandwidth $h = 0.8$.

Continuing with fixed x, for a given β the minimizer of (7) is $G\{\hat{m}_1(x), \ldots, \hat{m}_J(x), \beta\}$, where

$$G\{m_1(x), \ldots, m_J(x), \beta\} = \sum_{j=1}^{J} w_j(x) m_j(x) \beta^{j-1} / \sum_{j=1}^{J} w_j(x) \beta^{2j-2}.$$

Note that $G\{m_1(x), \ldots, m_J(x), \beta_0\} = m_1(x)$ because $m_j(x) = \beta_0^{j-1} m_1(x)$.

Hence, if we worked only with a fixed x, β_0 would be estimated to minimize

$$\sum_{j=1}^{J} w_j(x)[\hat{m}_j(x) - \beta^{j-1} G\{\hat{m}_1(x), \ldots, \hat{m}_J(x), \beta\}]^2. \quad \textbf{(8)}$$

Summing (8) over all the data suggests that we estimate β by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{J} w_j(X_i)[\hat{m}_j(X_i) - \beta^{j-1} G\{\hat{m}_1(X_i), \ldots, \hat{m}_J(X_i), \beta\}]^2. \quad \textbf{(9)}$$

2.2. A Method Based on Least Squares

A special case is illuminating and suggests a second method. Suppose that $J = 2$ and that the variances of $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are asymptotically the same, so that we can set $w_1(\cdot) = w_2(\cdot) \equiv 1$. Then (9) reduces to minimizing in β

$$(1+\beta^2)^{-1}\sum_{i=1}^{n}\{\hat{m}_2(X_i)-\beta\hat{m}_1(X_i)\}^2. \tag{10}$$

The leading term $(1+\beta^2)^{-1}$ in (10) plays an important role in the usual errors-in-variables problem, but here we have a different situation, because the "errors in the variables" are small asymptotically as a result of the fact that the "error-prone predictor" $\hat{m}_1(x)$ has an error that is asymptotically small. This suggests that one might find a reasonable estimate if one simply removes this leading term and minimizes instead

$$\sum_{i=1}^{n}\{\hat{m}_2(X_i)-\beta\hat{m}_1(X_i)\}^2. \tag{11}$$

Although there are numerical differences between minimizing (10) or (11) (generally, the latter yields larger estimates for β_0), it can be shown that asymptotically the two lead to the same distribution for $\hat{\beta}$.

The method (11) can be obtained alternatively by replacing $G(\cdot)$ in (9) by $\hat{m}_1(\cdot)$. Thus the least squares method minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{J} w_j(X_i)\{\hat{m}_j(X_i)-\beta^{j-1}\hat{m}_1(X_i)\}^2. \tag{12}$$

For $J \geq 3$, (12) leads to an estimator that is asymptotically different from the solution to (9). We explore the differences numerically in Section 3.

2.3. Alternatives

There is a host of possible alternative methods based on the errors-in-variables connection (Amemiya and Fuller, 1988; Carroll, Ruppert, and Stefanski, 1995; Cook and Stefanski, 1995).

Alternatively, the method (12) can be looked upon as regressing $\hat{m}_j(\cdot)$ for $j \geq 2$ on $\hat{m}_1(\cdot)$. This could be expanded to doing all possible regressions of $\hat{m}_j(\cdot)$ on $\hat{m}_k(\cdot)$ for $j > k$.

Finally, a referee has suggested that (9) be placed by

$$\sum_{i=J+1}^{n}\left[Y_i-\sum_{j=1}^{J}\beta^{j-1}G\{\hat{m}_1(X_{i-j}),\ldots,\hat{m}_J(X_{i-j}),\beta\}\right]^2.$$

We have not explored these alternatives, although our methods of argument can in principle be used to obtain limit distributions for them.

3. THE EXAMPLE REVISITED

We will first describe the integration estimate used in this paper. For simplicity of notation, this will be done only for the case $J = 2$. The integration estimate has been introduced in Tjøstheim and Auestad (1994) and Linton and Nielsen (1995) for the estimation of additive nonparametric components $m_j(\cdot)$ in an additive model.

An alternative to the integration estimator is backfitting, which is described by Hastie and Tibshirani (1990), Mammen et al. (1999), Opsomer (2000), and Opsomer and Ruppert (1997), among others. Our theory applies only for estimates that allow a higher order stochastic expansion stated in Section 4. Although we are only able to check this expansion for integration estimators, we conjecture that for large J application of backfitting leads to more reliable estimates. This conjecture is motivated by the fact that the integration estimate works for many additive components only under additional higher order smoothness conditions on the additive components.

The first step in defining the integration estimate uses a full-dimensional local linear fit $\hat{m}_{LL}$ of the data; i.e., the preliminary estimate $\hat{m}_{LL}$ is defined as θ_0 where the vector $\theta = (\theta_0, \theta_1, \theta_2)^T$ is defined by

$$\sum_{i=3}^{n} K_{h_1}(X_{i-1} - x_1)\, K_{h_2}(X_{i-2} - x_2)[Y_i - \theta^T \xi_i(x)]\xi_i(x) = 0. \tag{13}$$

Here $\xi_i(x)$ denotes the vector $(1, (X_{i-1} - x_1)/h_1), (X_{i-2} - x_2)/h_2)^T$. The kernel $K(\cdot)$ is a symmetric density function. The integration estimate $\hat{m}_1^I$ of m_1 is defined as

$$\hat{m}_1^I(x_1) = \widetilde{m}_1^I(x_1) - \sum_{i=1}^{n} w(X_i)\widetilde{m}_1^I(X_i) \Big/ \sum_{i=1}^{n} w(X_i). \tag{14}$$

Here, w is a weight function. The estimate $\widetilde{m}_1^I$ is achieved by summing out an argument of the full-dimensional estimate $\hat{m}_{LL}$,

$$\widetilde{m}_1^I(x_1) = n^{-1} \sum_{i=1}^{n} w(X_i)\hat{m}_{LL}(x_1, X_i) \Big/ n^{-1} \sum_{i=1}^{n} w(X_i). \tag{15}$$

The estimate $\widetilde{m}_2^I$ is achieved by summing out the other argument of the full-dimensional estimate $\hat{m}_{LL}$.

Thus, the first step of the procedure is to fit an additive model, without assumed links on the components m_j, given by $Y_t = X_t^2 = \sum_{j=1}^{J} m_j(X_{t-j}) + \epsilon_t$. For our data set of DEM/USD exchange rates we chose $J = 5$ lags. Figure 3 shows the integration estimates of the additive components. The integration estimate was calculated by fitting the full-dimensional estimate on a grid of 26^5 points. The fits show a shape that is reasonably consistent with the damping implied by (6).

Next we fitted the nonparametric GARCH model (4) with $m_j = \beta^{j-1} g$. For the estimation of β we used our method from errors in variables (see Section 2.1) and our least squares method (see Section 2.2). The resulting estimates were 0.892 and 0.779 for the integration estimate, respectively. There

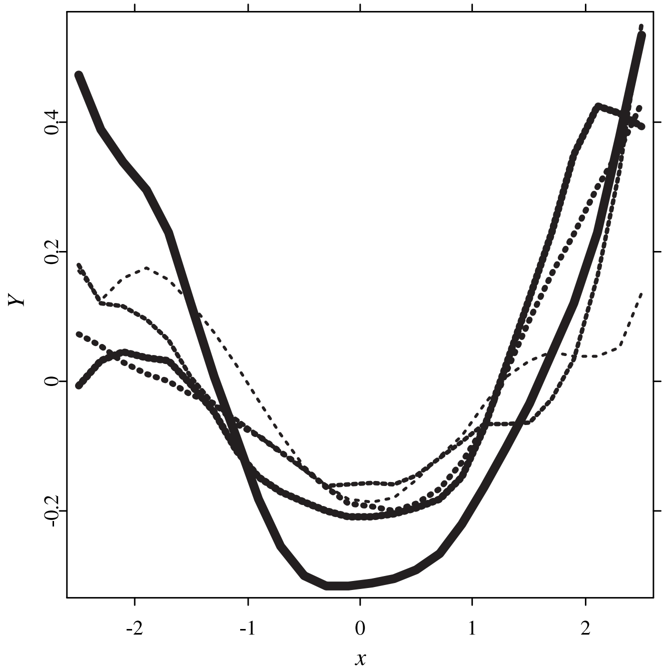

FIGURE 3. Integration estimates of the additive components m_j in the model $X_t^2 = \sum_{j=1}^{J} m_j(X_{t-j}) + \epsilon_t, j = 5$ (dashed line), 4 (small dots), 3 (large dots), 2 (black), 1 (thick black). Data as in Figure 1 (standardized returns).

are some differences between these estimates, largely along the lines of what one would expect from Figure 3. One would expect from these figures that the least squares estimate of β_0 would be smaller than the errors-in-variables/minimum distance estimate, because the latter compares $j = 2,3,4,5$ to a weighted average of $j = 1,\dots,5$, which is closer to the results for $j = 2,3,4,5$ than to the result for $j = 1$.

The estimates $\hat{m}_1,\dots,\hat{m}_J$ and $\hat{\beta}$ can be used to construct an estimate of m_1 that takes into account that the additive components are linked. This can be done by using the averaged estimate

$$\hat{m}_1^*(x) = \sum_{j=1}^{J} \hat{c}_j \hat{\beta}^{-(j-1)} \hat{m}_j(x) \Big/ \sum_{j=1}^{J} \hat{c}_j, \tag{16}$$

where $\hat{c}_j = \hat{\beta}^{2(j-1)}$ (see also (19) and the discussion following Theorem 4 in the next section). Here $\hat{\beta}$ denotes our estimate based on the method from errors in variables/minimum distance or the least squares method, respectively. Figures 4 and 5 show plots of the estimates $\hat{m}_1^*$ and $\hat{m}_j^* = \hat{\beta}^{j-1}\hat{m}_1^*$. When using

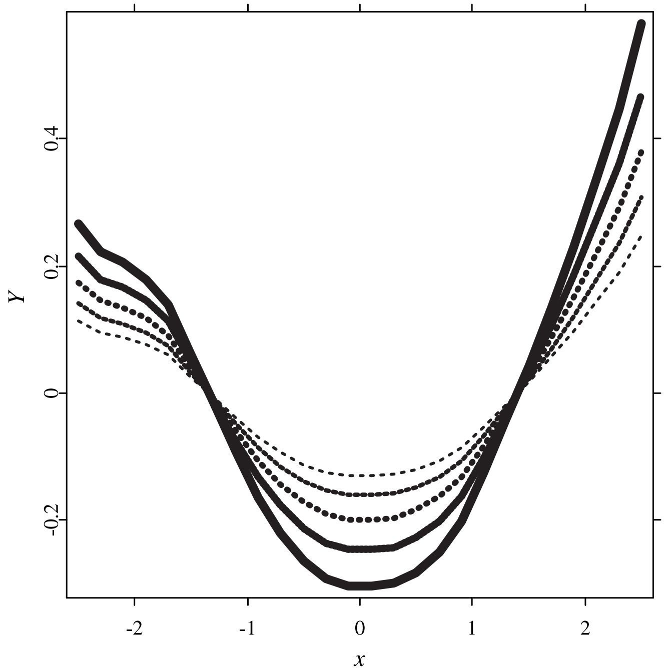

FIGURE 4. Plots of the estimates $\hat{m}_j^* = \hat{\beta}^{j-1}\hat{m}_1^*$ (see (16)), $j = 5$ (dashed line), 4 (small dots), 3 (large dots), 2 (black), 1 (thick black). (Estimation of the nonparametric components by the integration estimate and of β by the method from errors in variables.) Data as in Figure 1 (standardized returns).

backfitting, the results change only slightly. The plots differ slightly for the different methods. In particular use of the method from errors in variables leads to more asymmetric news impact functions.

Figure 6 shows how the nonparametric estimates depend on the chosen number J of lags. It compares the backfitting estimates of $\hat{m}_1^*$ for different numbers of lags ($J = 5, 10, 20, 30$). We here used the backfitting estimate because for $J \geq 10$ calculation of the integration estimate leads to major problems. It requires calculation of a J-dimensional smoother on a J-dimensional grid. This is computationally nearly infeasible. Furthermore, for a stable 10-dimensional smoother a larger sample size than 10,000 is required. In the calculations, β was estimated by the method from errors in variables. The estimated values are 0.789 ($J = 5$), 0.799 ($J = 10$), 0.809 ($J = 20$), and 0.788 ($J = 30$). The nonparametric estimates (besides small differences of the estimate for $J = 5$) are nearly indistinguishable. Because $J = \infty$ is the nonparametric GARCH model (2) and (3), we conclude that in this data example model (4) approximates the nonparametric GARCH model (2) and (3) reasonably well.

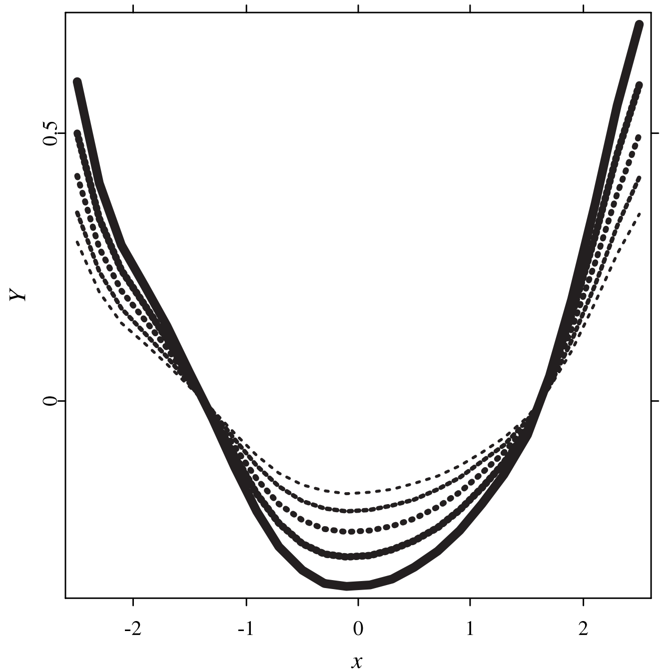

FIGURE 5. Plots of the estimates $\hat{m}_j^* = \hat{\beta}^{j-1}\hat{m}_1^*$ (see (16)), $j = 5$ (dashed line), 4 (small dots), 3 (large dots), 2 (black), 1 (thick black). (Estimation of the nonparametric components by the integration estimate and of β by the least squares method.) Data as in Figure 1 (standardized returns).

Of course, this analysis should be taken as illustrative, because it focuses on short-term dependencies. For modeling of long-range dependencies such as daily or weekly dependence, more complicated models may be needed. Furthermore, the plots only show minor deviations from symmetry. So we conjecture that for our data set the GARCH(1,1) model could not be rejected. However, a more careful analysis of the deviations requires the development of test procedures. A natural approach would be to base the test on the nonparametric estimate $\hat{m}_1^*$. For a test of the null hypothesis of a GARCH(1,1) model it could be compared with $\hat{E}_0\hat{m}_1^*$ where $\hat{E}_0\hat{m}_1^*$ is a bootstrap estimate of the expectation of $\hat{m}_1^*$ under the null hypothesis. Furthermore, a test of our model (4) could be based on the comparison of the fits $\hat{\beta}^{-(j-1)}\hat{m}_j(x)$. A rigorous treatment of such test procedures is out of the scope of this paper.

The next section discusses asymptotics of these estimates in an autoregression model.

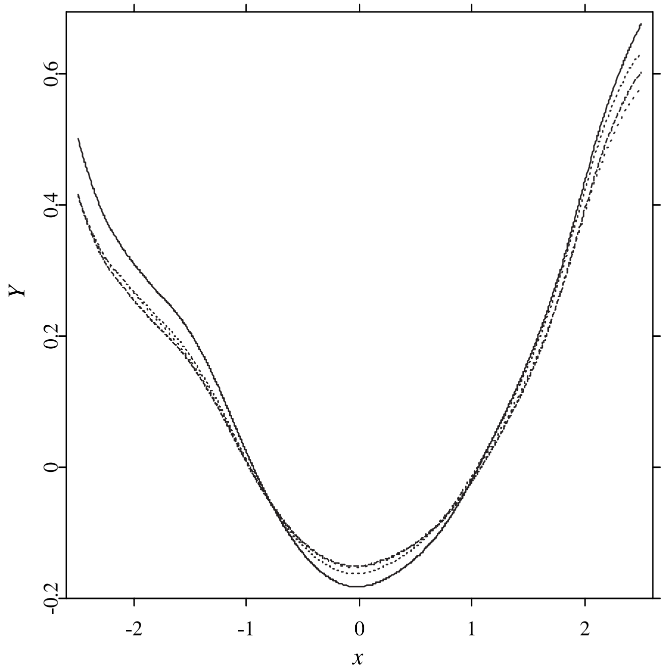

FIGURE 6. Backfitting estimates of $\hat{m}_1^*$ for different numbers of lags. Orienting at the right upper corner, the highest to lowest values are associated with $J = 5,10,30,20$, respectively. Data as in Figure 1 (standardized returns).

4. ASYMPTOTICS FOR AUTOREGRESSION

We suppose that a stationary time series $X_0,\ldots,X_n$ is observed and that $E(X_{J+1}^2|X_J,\ldots,X_1) = \alpha + m_1(X_J) + \beta m_1(X_{J-1}) + \cdots + \beta^{J-1}m_1(X_1)$, where for a weight function w the function m_1 satisfies $E\,w(X_i)m_1(X_i) = 0$. The density of $(X_J,\ldots,X_1)$ is denoted by p, the marginal density of X_i by p_X. For this setup, all the results will be based on the validity of the expansion

$$\hat{m}_j(x) = m_j(x) + \left(\frac{1}{2}\right)h^2\beta^{j-1}r(x) + n^{-1}\sum_{i=J+1}^{n} K_h(X_{i-j}-x)u_j(\mathbf{X}_i,x)\epsilon_i + n^{-1}\sum_{i=J+1}^{n} v_j(\mathbf{X}_i,x)\epsilon_i + n^{-1}\sum_{i=J+1}^{n} t_j(\mathbf{X}_i) + o_P(n^{-1/2}), \qquad \textbf{(17)}$$

where $\mathbf{X}_i$ is the vector $(X_{i-1},\ldots,X_{i-J})$ and $\epsilon_i = Y_i - \alpha - m_1(X_{i-1}) - \beta m_1(X_{i-2}) - \cdots - \beta^{J-1}m_1(X_{i-J})$ with $Y_i = X_i^2$.

First we study the validity of (17) for the integration estimate. Our first result shows that $\hat{m}_j^I$ satisfy (17). For simplicity of presentation, this is only done for the case $J = 2$. The result can be easily generalized to larger J for higher order kernels at the cost of additional smoothness conditions of m_1 (see the discussion after Theorem 1, which follows). These additional conditions reflect the well-known fact that the integration estimator does not perform well for large dimensions. The main reason is that the integration estimate uses a full-dimensional smoother in a pilot step. An attractive alternative is the backfitting estimate. In the backfitting algorithm only one-dimensional smoothing is used, and thus the curse of dimensionality is effectively circumvented. We conjecture that an expansion of the form (17) holds (uniformly) for the backfitting estimate. In Mammen et al. (1999), for a version of the backfitting estimate $\hat{m}_j^{BACK}$, a stochastic expansion has been given. Applied to our setup this expansion is

$$\hat{m}_j^{BACK}(x) = m_j(x) + \left(\frac{1}{2}\right)h^2\beta^{j-1}r(x) + n^{-1}\sum_{i=J+1}^{n} K_h(X_{i-j} - x)u_j(\mathbf{X}_i, x)\epsilon_i + O_P(n^{-1/2}), \tag{18}$$

with an appropriate choice of u_j and where $r(x)$ is as for $\hat{m}_j^I$ (see Theorem 1). The $O_P(n^{-1/2})$ term in (18) can be explicitly given by an infinite series (see Mammen et al., 1999). However, it seems to be complicated to show that this term has the form of the terms in (17). Hence we can only conjecture that the results to follow hold for the backfitting estimate. For another recent asymptotic treatment of (another version of) the backfitting estimate, see Opsomer (2000) and Opsomer and Ruppert (1997).

THEOREM 1. *Suppose $J = 2$. Under the regularity conditions (A) listed in the Appendix, and if the bandwidths fulfill that $h_1 = h_2$, $h_2^4 = o(n^{-1/2})$, and $(\log n)^2[\sqrt{n}h_1h_2]^{-1} \to 0$, and if p has two continuous partial derivatives and m_1 has four continuous derivatives, then the estimate $\hat{m}_1^I$ has a stochastic expansion (17) with $h = h_1 = h_2$ and*

$$r(x) = m_1''(x) - \int w m_1'' p_X \left[\int w p_X\right]^{-1},$$

$$u_1(\mathbf{X}_i, x) = w(X_{i-2})p_X(X_{i-2})p(x, X_{i-2})^{-1},$$

$$v_1(\mathbf{X}_i, x) = -w(X_{i-1})w(X_{i-2})p_X(X_{i-1})p_X(X_{i-2})p(X_{i-1}, X_{i-2})^{-1}\left[\int w p_X\right]^{-1},$$

$$t_1(\mathbf{X}_i) = [w(X_{i-1})m_1(X_{i-1})]\left[\int w p_X\right]^{-1}.$$

Here, $\mathbf{X}_i$ denotes the vector $(X_{i-1}, X_{i-2})^T$. For $\hat{m}_2^I$ the expansion (17) holds with the same $r(x)$, with the same $v_2(\mathbf{X}_i, x) = v_1(\mathbf{X}_i, x)$, and with $t_2(\mathbf{X}_i) = \beta t_1(\mathbf{X}_i)$ and also with $u_2(\mathbf{X}_i, x) = w(X_{i-1})p_X(X_{i-1})p(X_{i-1}, x)^{-1}$. In both cases, *the expansion (17) holds uniformly for $x \in B$. The set B is introduced in Assumption A(iii) in the Appendix.*

The statement of Theorem 1 can be easily generalized to $J > 2$ at the cost of a rather involved notation. Then higher order local polynomials should be used to meet the conditions that then become necessary for the orders of the bandwidths. For the pilot estimate $\widetilde{m}^I$ we propose to use a local linear fit (with bandwidth h, say) for the estimated component (m_j, say) and to use local polynomial fits of degree $2l - 1$ (for an $l \geq 1$) with common bandwidth (g, say) for the other $J - 1$ "nuisance" components m_l with $l \neq j$. For the resulting estimator $\hat{m}_j^I$ it is possible to verify a stochastic expansion of the form (17). The following additional assumptions are needed. The density p has $l + 1$ continuous derivatives, m_1 has $2l + 2$ continuous derivatives, and the bandwidths fulfill $h = o(n^{-1/8}), g = o(n^{-1/[4(l+1)]})$, and $(\log n)^2 [\sqrt{n} h g^{J-1}]^{-1} \to 0$.

The assumptions A(i)–A(viii) are given in the Appendix. We note that the major assumption is A(i), namely, that $X_1, X_2, \ldots$ is a geometrically strongly mixing stationary process. Assumption A(vi) could be replaced by a moment condition. This can be done by using the truncation techniques in Masry (1996a, 1996b). However, then stronger conditions on the rates of bandwidths have to be introduced.

We now describe the main results. Suppose that we have estimates of m_j that fulfill the expansion (17). These estimates can be used to construct an estimate of β. Asymptotics for this estimate are given in the next two theorems. The first theorem describes least squares estimation of β (see Section 2.2).

Make the definitions

$$\begin{aligned}\mathcal{H}_1(\mathbf{X},\beta) = \sum_{j=1}^{J} (j-1)\beta^{j-2} \Bigg\{ & [v_j^*(\mathbf{X}) - \beta^{j-1} v_1^*(\mathbf{X})] \\ & + \sum_{k=1}^{J} [w_j(X_j) m_1(X_j) p_X(X_j) u_j(\mathbf{X}, X_j) \\ & - \beta^{j-1} w_j(X_1) m_1(X_1) p_X(X_1) u_1(\mathbf{X}, X_1)] \Bigg\};\end{aligned}$$

$$v_j^*(\mathbf{x}) = E[w_j(X_i) m_1(X_i) v_j(\mathbf{x}, X_i)];$$

$$\mathcal{H}_2(\mathbf{X},\beta) = \sum_{j=1}^{J} (j-1)\beta^{j-2} \{ s_j^* [t_j(\mathbf{X}) - \beta^{j-1} t_1(\mathbf{X})]\};$$

$$s_j^* = E[w_j(X_i) m_1(X_i)];$$

$$D_{LS} = \sum_{j=1}^{J} \{(j-1)\beta_0^{j-2}\}^2 E\{w_j(X_i) m_1^2(X_i)\}.$$

THEOREM 2. *Suppose that the regularity conditions in the Appendix hold and for some estimates $\hat{m}_j$ assume that they fulfill (17) uniformly for $x \in B$ with*

h of order $n^{-1/5}$, where r, u_j, v_j, and t_j are bounded functions with $E\, t_j(\mathbf{X}_i) = 0$ and $\sup_{x,z}|(\partial^3/(\partial x)^3)u_j(z,x)| < \infty$. Then the estimate $n^{1/2}(\hat\beta_{LS} - \beta_0)$ has an asymptotic normal distribution with mean 0 and variance $D_{LS}^{-2}\sum_{k\in\mathbb{Z}}\operatorname{cov}(U_0,U_k)$ where $U_k = \mathcal{H}_1(\mathbf{X}_k,\beta_0)\epsilon_k + \mathcal{H}_2(\mathbf{X}_k,\beta_0)$.

The next theorem gives the asymptotic distribution of the errors-in-variables estimate of β (see Section 2.1). For the statement of the theorem we need the following additional definitions:

$$s(\beta,x) = \left[\sum_{j=1}^{J} w_j(x)\beta^{2j-2}\right]^{-2};$$

$$R_j(x,\beta) = \sum_{\ell=1}^{J} w_\ell(x)\{(2\ell-2)\beta^{2\ell-3}m_j(x) - (\ell+j-2)\beta^{\ell+j-3}m_\ell(x)\};$$

$$\mathcal{M}(\mathbf{x},\beta) = E\left[\sum_{j,\ell=1}^{J} w_j(X_i)s(\beta,X_i)R_j(X_i,\beta)w_\ell(X_i) \times \{\beta^{2\ell-2}v_j(\mathbf{x},X_i) - \beta^{\ell+j-2}v_\ell(\mathbf{x},X_k)\}\right];$$

$$\mathcal{T}_1(j,\ell,\beta,\mathbf{x}) = \beta^{2\ell-2}w_j(x_j)w_\ell(x_j)s(\beta,x_j)R_j(x_j,\beta)p_X(x_j)u_j(\mathbf{x},x_j);$$

$$\mathcal{T}_2(j,\ell,\beta,\mathbf{x}) = \beta^{\ell+j-2}w_j(x_\ell)w_\ell(x_\ell)s(\beta,x_\ell)R_j(x_\ell,\beta)p_X(x_\ell)u_\ell(\mathbf{x},x_\ell);$$

$$\mathcal{D}_{j,\ell}(\beta) = E\{w_j(X_i)s(\beta,X_i)R_j(X_i,\beta)w_\ell(X_i)\};$$

$$\mathcal{G}(\mathbf{X},\beta) = \sum_{j,\ell=1}^{J}\mathcal{D}_{j,\ell}(\beta)\{\beta^{2\ell-2}t_j(\mathbf{X}) - \beta^{\ell+j-2}t_\ell(\mathbf{X})\};$$

$$\mathcal{H}_3(\mathbf{x},\beta) = \sum_{j,\ell=1}^{J}\{\mathcal{T}_1(j,\ell,\beta,\mathbf{x}) - \mathcal{T}_2(j,\ell,\beta,\mathbf{x})\};$$

$$D_{EIV} = \sum_{j=1}^{J} Ew_j(X_i)s(\beta_0,X_i)m_1^2(X_i)\left\{\sum_{\ell}^{J} w_\ell(X_i)\beta^{2\ell+j-4}(\ell-j)\right\}^2.$$

THEOREM 3. *Suppose that the assumptions of Theorem 2 hold for some estimates $\hat m_j$. Then the estimate $n^{1/2}(\hat\beta_{EIV} - \beta_0)$ has an asymptotic normal distribution with mean 0 and variance $D_{LS}^{-2}\sum_{k\in\mathbb{Z}}\operatorname{cov}(V_0,V_k)$, where $V_k = \{\mathcal{H}_3(\mathbf{X}_k,\beta_0) + \mathcal{M}(\beta_0,\mathbf{X}_k)\}\epsilon_k + \mathcal{G}(\mathbf{X}_k,\beta_0)$.*

Under our model assumption that $m_j = \beta^{j-1}m_1$ an improved estimate $\hat m_1^*$ of m_1 can be constructed by using the estimates $\hat m_1,\dots,\hat m_J$ and an estimate $\hat\beta$ of β. This can be done, e.g., by putting

$$\hat m_1^*(x) = \sum_{j=1}^{J}\hat c_j\hat\beta^{-(j-1)}\hat m_j(x)\Big/\sum_{j=1}^{J}\hat c_j, \qquad \textbf{(19)}$$

where $\hat c_j$ are some data adaptive choices of weights. The next theorem gives the asymptotic distribution $\hat m_1^*(x)$.

THEOREM 4. *Suppose the conditions of the Appendix hold and assume that* $\hat{m}_j$ *are estimates with*

$$\hat{m}_j(x) = m_j(x) + \left(\frac{1}{2}\right) h^2 \beta^{j-1} r(x) + n^{-1} \sum_{i=J+1}^{n} K_h(X_{i-j} - x) u_j(\mathbf{X}_i, x)\epsilon_i + o_P(n^{-2/5}), \tag{20}$$

where u_j *is a function with* $\sup_z \sup_{|x-y|<\delta} |u_j(z,y) - u_j(z,x| \to 0$ *for* $\delta \to 0$ *and where* h *is of order* $n^{-1/5}$. *Furthermore suppose that* $\hat{\beta}$ *is an estimate of* β *with* $\hat{\beta} = \beta + o_P(n^{-2/5})$ *and that for some constants* c_j *it holds that* $\hat{c}_j = c_j + o_P(n^{-2/5})$. *Then* $n^{2/5}[\hat{m}_1^*(x) - m(x)]$ *has an asymptotic normal distribution with mean* $(\frac{1}{2}) n^{2/5} h^2 r(x)$ *and variance* $n^{1/5} h^{-1} \nu(K) f(x) \times \sum_{j=1}^{J} c_j^2 \beta^{-2(j-1)} s_j^2(x) / (\sum_{j=1}^{J} c_j)^2$, *where* $\nu(K) = \int K^2(u)\, du$ *and* $s_j^2(x) = E\{\epsilon_{J+1}^2 u_j^2(X_J, \ldots, X_1, x) | X_{J-j+1} = x\}$. *The variance is minimized by a choice* $\hat{c}_j$ *with* $\hat{c}_j = c\beta^{2(j-1)} s_j^{-2}(x) + o_P(n^{-2/5})$, *where* c *is some constant. In this case* $n^{2/5}\{\hat{m}_1^*(x) - m(x)\}$ *has an asymptotic variance* $n^{1/5} h^{-1} \nu(K) f(x) / \sum_{j=1}^{J} \beta^{2(j-1)} s_j^{-2}(x)$.

The asymptotic variance of $n^{2/5}[\hat{\beta}^{-(j-1)} \hat{m}_j(x) - m(x)]$ is equal to $n^{1/5} h^{-1} \nu(K) f(x) \beta^{-2(j-1)} s_j^2(x)$ (see the proof of Theorem 4). Clearly, the asymptotic variance of $n^{2/5}[\hat{m}_1^*(x) - m(x)]$ is strictly smaller for all j for an asymptotically optimal choice of $\hat{c}_j$. Typically, application of asymptotically optimal weights requires estimation of $s_j^2(x)$. However, if the weight function w is chosen as indicator function of an interval $[-c, c]$ with c large enough we conjecture that for the backfitting estimate, $s_j^2(x)$ does not depend strongly on j. This motivates in these cases the choice $\hat{c}_j = \hat{\beta}^{2(j-1)}$ that leads to a nearly minimal asymptotic variance of $\hat{m}_1^*(x)$ for all x.

It can be easily seen that the asymptotic result of Theorem 4 applies under the conditions of Theorems 2 and 3 for the choices $\hat{\beta} = \hat{\beta}_{EIV}$ and $\hat{\beta} = \hat{\beta}_{LS}$.

5. SIMULATION

In a small simulation study we generated time series from model (4) with $J_{model} = \infty$, $J_{model} = 5$, and $J_{model} = 2$. Remember that for $J_{model} = \infty$ model (4) coincides with the nonparametric GARCH model (2) and (3). For these time series we fitted the model (4) with $J_{fit} = 5$ and $J_{fit} = 2$. So our simulation includes the case of model misspecification. In particular, we studied how estimates from model (4) with $J_{fit} = 5$ and $J_{fit} = 2$ perform for the nonparametric GARCH model (2) and (3) (i.e., $J_{model} = \infty$). We have run 200 simulations. The simulations were done using the statistical software Voyager (see Sawitzki, 1996). The sample size was 5000. As function $g = m_1$ we chose

$$g(x) = 1 - 0.9 \exp(-2x^2).$$

The variable ξ_t has a standard normal distribution (and thus Y_t/σ_t^2 a χ^2 distribution). The parameter β was equal to 0.7. We used the marginal integration estimate with quartic kernel and with bandwidth 0.4 for the estimated compo-

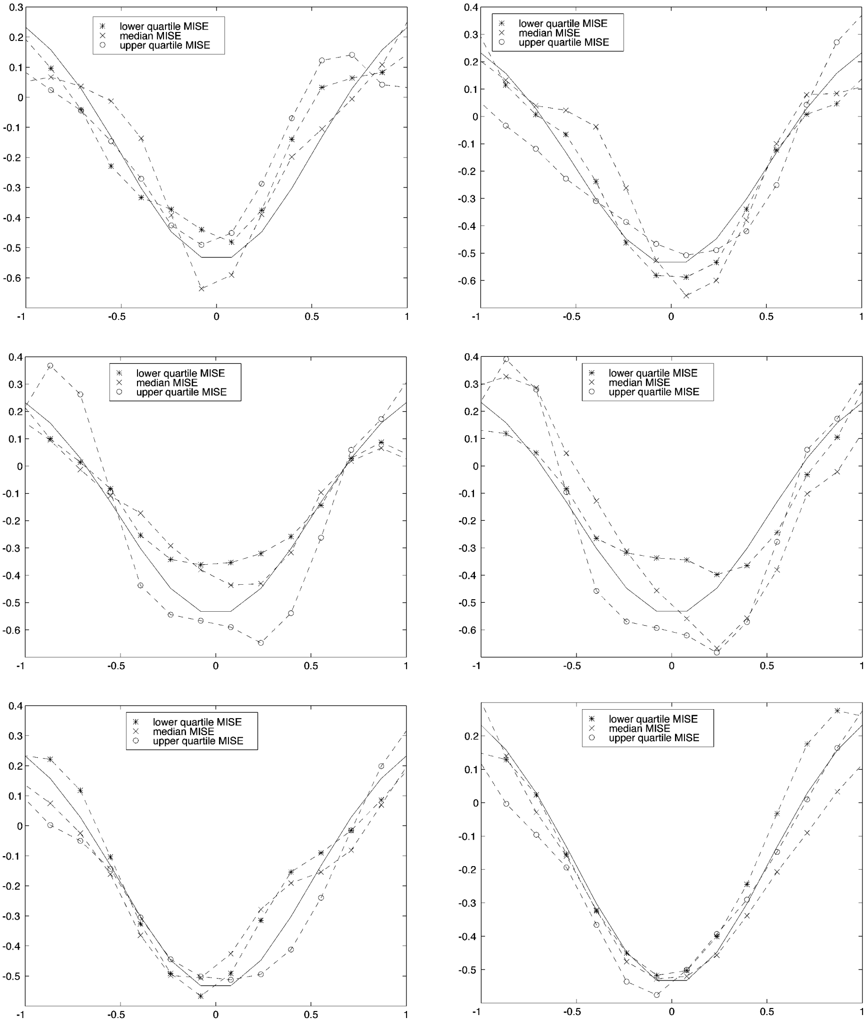

FIGURE 7. Simulated estimates of the function g. Left plots show estimates with method from errors in variables; right plots are made with the method based on least squares. The plots show three estimates from a simulation with 200 replications. The MISEs (mean integrated squared errors) of the plotted estimates are the 25%, 50%, and 75% quantiles among the simulated values, respectively. The black line shows the underlying curve g. The first two plots show the results for $(J_{model}, J_{fit}) = (\infty, 2)$, the next two for $(\infty, 5)$, followed by $(5,2)$, $(5,5)$, and $(2,2)$. (*Figure continues on facing page*.)

nent and bandwidth 0.9 for the other components. The integration estimate was calculated on a grid in $[-1.5, 1.5]^{J_{fit}}$.

For each resulting estimate of m_1 we have calculated its MISE (mean integrated squared error). In Figure 7 we show the estimates with lower quar-

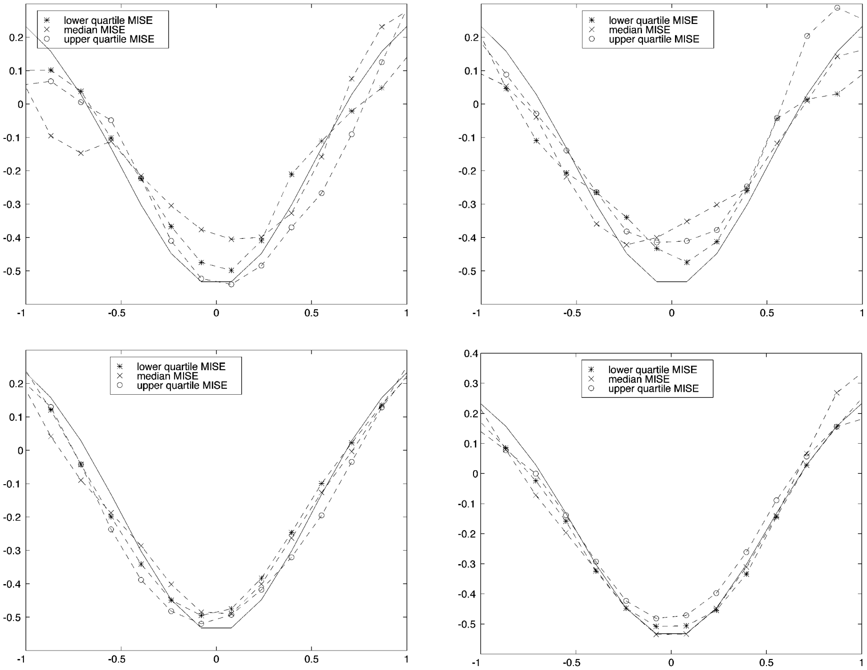

FIGURE 7. *Continued.*

tile, median, and upper quartile performance. The estimates reflect well the shape of the underlying function g. This also holds in case of misspecification of the model. In particular, for $J_{fit} = 5$ the estimates show much fluctuation. This is related to the fact that the "actual sample size" N_{act} is relatively small in that case. We denote by N_{act} the number of observations that lie in $[-1.5,1.5]^{J_{fit}}$. We have that the average of N_{act} in our simulations is equal to 2566 ($J_{model} = \infty, J_{fit} = 2$), 1050 ($J_{model} = \infty, J_{fit} = 5$), 2987 ($J_{model} = 5$, $J_{fit} = 2$), 1519 ($J_{model} = 5, J_{fit} = 5$), and 4041 ($J_{model} = 2, J_{fit} = 2$).

6. DISCUSSION

The key feature of our model (5) and (6) is that of an additive model with parametrically linked components. We have illustrated the use of the model in a financial time series context and obtained asymptotic results for autoregression.

The methods are conceptually simple. One first uses standard additive model techniques to obtain estimates of the components, and then estimates the linking parameter β_0 by combining the components. It is surprising and pleasing that standard additive model techniques can be used for the first stage without the need for the undersmoothing which often occurs in semiparametric modeling.

We have illustrated the use of two such combinations of the component estimates, one an intuitive least squares approach (Section 2.2) and one motivated by errors-in-variables/minimum distance considerations (Section 2.1). At least in principle one would conjecture that the basic idea of estimating β_0 should generalize to such things as generalized linear models. Obtaining asymptotic distributions for such generalizations is likely to be challenging.

In this paper we present results only for methods using the integration estimate. This estimate has drawbacks for high dimensions. The backfitting estimator is a popular competitor that more efficiently circumvents the curse of dimensionality. So it would be interesting to study how the backfitting estimate works as a pilot estimate in our procedure. Recently some work has been done on the asymptotics of backfitting. Unfortunately this work does not easily allow to derivation of second-order properties that are necessary to study backfitting for our setup.

An interesting generalization of the model (4) would be to allow for the addition of other parametrically linked terms of the form $\theta^T Z_t$ based on covariates Z_t. In the context of the example, these covariates might include information about previous market behavior, e.g., yesterday's volatility. Again, although the ideas may seem straightforward, actually obtaining asymptotic results may well prove to be difficult.

REFERENCES

Amemiya, Y. & W.A. Fuller (1988) Estimation for the nonlinear functional relationship. *Annals of Statistics* 16, 147–160.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.

Bollerslev, T., R.F. Engle, & D.B. Nelson (1994) ARCH models. In R.F. Engle & D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, pp. 2959–3038. Amsterdam: Elsevier Science.

Bosq, D. (1996) *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. New York: Springer.

Bossaerts, P., W. Härdle, & C. Hafner (1996) Foreign exchange rates have surprising volatility. In P. Robinson (ed.), *Athens Conference on Applied Probability and Time Series,* vol. 2, pp. 55–72. Lecture Notes in Statistics 115. New York: Springer.

Box, G.E.P. & G.M. Jenkins (1976) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Carroll, R.J., D. Ruppert, & L.A. Stefanski (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Cook, J. & L.A. Stefanski (1995) A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association* 89, 1314–1328.

Dacorogna, M.M., U.A. Müller, R.J. Nagler, R.B. Olsen, & O.V. Pictet (1993) A geographical model for daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance* 12, 413–426.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.

Fan, J., W. Härdle, & E. Mammen (1998) Direct estimation of low dimensional components in additive models. *Annals of Statistics* 26, 943–971.

Fuller, W.A. (1987) *Measurement Error Models*. New York: Wiley.

Goodhart, C. & L. Figliuoli (1991) Every minute counts in financial markets. *Journal of International Money and Finance* 10, 23–52.

Gouriéroux, C. (1997) *ARCH Models and Financial Applications*. Heidelberg: Springer.
Gouriéroux, C. & A. Monfort (1992) Qualitative threshold ARCH models. *Journal of Econometrics* 52, 159–199.
Hafner, C.M. (1998) *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*. Heidelberg: Physica.
Härdle, W. & A. Tsybakov (1997) Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics* 81, 223–242.
Härdle, W., A. Tsybakov, & L. Yang (1998) Nonparametric vector autoregression. *Journal of Statistical Planning and Inference* 68, 221–245.
Hastie, T.J. & R. Tibshirani (1990), *Generalized Additive Models*. London: Chapman and Hall.
Linton, O.B. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–473.
Linton, O.B. & J.P. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–101.
Mammen, E., O.B. Linton, & J. Nielsen (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* 27, 1443–1490.
Masry, E. (1996a) Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications* 65, 81–101.
Masry, E. (1996b) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17, 571–599.
Nelson, D.B. (1991) Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347–370.
Opsomer, J.D. (2000) Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* 73, 166–179.
Opsomer, J.D. & D. Ruppert (1997) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* 25, 186–211.
Rabemananjara, R. & J.M. Zakoian (1993) Threshold ARCH models and asymmetries in volatility. *Journal of Applied Econometrics* 8, 31–49.
Sawitzki, G. (1996) Extensible statistical software: On a Voyage to Oberon. *Journal of Computational and Graphical Statistics* 5, 263–283.
Tjøstheim, D. & B.H. Auestad (1994) Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* 89, 1398–1409.
Zakoian, J.M. (1994) Threshold heteroskedastic functions. *Journal of Economic Dynamics and Control* 18, 931–955.

APPENDIX

Assumptions: Condition A.

(i) $X_1, X_2, \ldots$ is a stationary process that is geometrically strongly mixing, i.e., $\alpha(k) \le c_0 \rho^k$ for some constants c_0 and $0 < \rho < 1$.
(ii) For all q there exists a constant c_q such that for all indices $i_1, \ldots, i_q$ the density of $(X_{i_1}, \ldots, X_{i_q})$ is bounded by c_q.
(iii) The weight function w has a continuous derivative and a bounded support B.
(iv) The density p of $(X_J, \ldots, X_1)$ is continuously differentiable, and on B^J it is bounded away from 0.
(v) The regression function m_1 is two times continuously differentiable.
(vi) The variables ϵ_i have a finite Laplace transform, i.e., $E \exp(\gamma \epsilon_i) < \infty$ for $|\gamma|$ small enough.

(vii) The conditional variance $\sigma^2(x) = E(\epsilon_i^2|\mathbf{X}_i = x)$ is continuous.

(viii) The kernel K is a symmetric probability density with compact support. Without loss of generality we assume that $\int u^2 K(u)\,du = 1$.

Proof of Theorem 1. Note that $\hat{m}^{LL}(x) = e_1^T\{n^{-1}\sum_{i=1}^n A_i(x)\xi_i(x)\xi_i(x)^T\}^{-1} \times n^{-1}\sum_{i=1}^n A_i(x)Y_i\xi_i(x)$, where $A_i(x) = K_{h_1}(X_{i-1} - x_1)K_{h_2}(X_{i-2} - x_2)$ and where $e_1^T = (1,0,0)$. Note that $E(Y_i|X_{i-1}, X_{i-2}) = \alpha + m_1(X_{i-1}) + \beta m_1(X_{i-2})$. Put $\epsilon_i = Y_i - \alpha - m_1(X_{i-1}) - \beta m_1(X_{i-2})$, $m(x) = \alpha + m_1(x_1) + \beta m_1(x_2)$, and $\hat{r}_i(x) = m(X_{i-1}, X_{i-2}) - m(x) - (X_{i-1} - x_1)m_1'(x_1) - (X_{i-2} - x_2)\beta m_1'(x_2)$. We can write $\hat{m}^{LL}(x) - m(x) = e_1^T\{n^{-1}\sum_{i=1}^n A_i(x)\xi_i(x)\xi_i(x)^T\}^{-1} n^{-1}\sum_{i=1}^n A_i(x)\{\epsilon_i + \hat{r}_i(x)\}\xi_i(x)$.

Now note that

$$
\begin{aligned}
n^{-1}\sum_{i=1}^n A_i(x)\xi_i(x)\xi_i(x)^T &= p(x)\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + h_1 p^{(1,0)}(x)\begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix}\\
&\quad + h_2 p^{(0,1)}(x)\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix} + O_P\left(h_1^2 + h_2^2 + \frac{\log n}{\sqrt{nh_1h_2}}\right)\\
&= p(x)I + h_1 p^{(1,0)}(x)I_1 + h_2 p^{(0,1)}(x)I_2\\
&\quad + O_P\left(h_1^2 + h_2^2 + \frac{\log n}{\sqrt{nh_1h_2}}\right). \qquad \textbf{(A.1)}
\end{aligned}
$$

Here $p^{(0,1)}$ and $p^{(1,0)}$ denote the partial derivative of p with respect to x_1 or x_2, respectively. The expansion in (A.1) holds uniformly for $x \in B \times B$. For a proof of (A.1) one proceeds as in Bosq (1996) where uniform rates are shown for kernel density estimates of strongly mixing observations by using exponential inequalities for mixing sequences (see Bosq, 1996, Theorem 2.2).

Equation (A.1) shows

$$
\begin{aligned}
&\left[n^{-1}\sum_{i=1}^n A_i(x)\xi_i(x)\xi_i(x)^T\right]^{-1}\\
&\quad = p(x)^{-1}\Big\{I - h_1 p^{(1,0)}(x)p(x)^{-1}I_1 - h_2 p^{(0,1)}(x)p(x)^{-1}I_2\\
&\qquad + O_P\left(h_1^2 + h_2^2 + \frac{\log n}{\sqrt{nh_1h_2}}\right)\Big\}. \qquad \textbf{(A.2)}
\end{aligned}
$$

Similarly as (A.1) one shows uniformly for $x \in B \times B$

$$
n^{-1}\sum_{i=1}^n A_i(x)\epsilon_i\xi_i(x) = O_P\left(\frac{\log n}{\sqrt{nh_1h_2}}\right) \qquad \textbf{(A.3)}
$$

and

$$
n^{-1}\sum_{i=1}^n A_i(x)\hat{r}_i(x)\xi_i(x) = O_P(h_1^2 + h_2^2). \qquad \textbf{(A.4)}
$$

Equations (A.2)–(A.4) give uniformly for $x \in B \times B$:

$$\hat{m}^{LL}(x) - m(x) = e_1^T p(x)^{-1}\Big\{I - h_1 p^{(1,0)}(x)p(x)^{-1}I_1 - h_2 p^{(0,1)}(x)p(x)^{-1}I_2 + O_P\Big(h_1^2 + h_2^2 + \frac{\log n}{\sqrt{nh_1h_2}}\Big)\Big\}n^{-1} \times \sum_{i=1}^n A_i(x)[\epsilon_i + \hat{r}_i(x)]\xi_i(x) + o_P(n^{-1/2}). \qquad \textbf{(A.5)}$$

We use now that uniformly in $x \in B \times B$:

$$n^{-1}\sum_{i=1}^n A_i(x)\hat{r}_i(x)\xi_i(x) = (1/2)h_1^2 m_1''(x_1)e_1 + (1/2)h_2^2\beta m_1''(x_2)e_1 + o_P(n^{-1/2}). \qquad \textbf{(A.6)}$$

For a proof of (A.6) one uses the fact that m_1 is four times continuously differentiable, and for the treatment of the resulting sum of strongly mixing summands one proceeds as in the proof of (A.1).

We now treat the estimate

$$\widetilde{m}_1^I(x_1) = n^{-1}\sum_{i=1}^n w(X_{i-2})\hat{m}^{LL}(x, X_{i-2}) \Big/ \Big[n^{-1}\sum_{i=1}^n w(X_{i-2})\Big].$$

Note now that for all q with $\gamma_1 > 0$ arbitrarily small and $\gamma_2 > 0$ arbitrarily large:

$$\sup_{x_1 \in B} E\Big[h_1 n^{-1}\sum_{j=1}^n w(X_{j-2})\frac{p^{(1,0)}(x_1, X_{j-2})}{p(x_1, X_{j-2})^2} n^{-1}\sum_{i=1}^n A_i(x_1, X_{j-2})\epsilon_i^*\Big]^{2q} = O(h_1^{2q}(nh_1)^{-q}n^{\gamma_1 q}), \qquad \textbf{(A.7)}$$

where $\epsilon_i^* = \epsilon_i \mathbf{1}[|\epsilon_i| \le \gamma_2 \log n] - E\{\epsilon_i \mathbf{1}[|\epsilon_i| \le \gamma_2 \log n]\}$. Claim (A.7) can be shown by application of Davydov's inequality (see Bosq, 1996 Corollary 1.1). For this purpose one writes the left-hand side of (A.7) as $\sum_{j_1,\dots,j_{2q}}\sum_{i_1,\dots,i_{2q}} h_1^{2q} n^{-4q} EZ_{i_1,j_1}\cdot\ldots\cdot Z_{i_{2q},j_{2q}}$, where $Z_{i,j} = w(X_{j-2})p^{(1,0)}(x_1, X_{j-2})\, A_i(x_1, X_{j-2})\, \epsilon_i^*/p(x_1, X_{j-2})^2$. Davydov's inequality and our mixing conditions imply that for arbitrarily large C there exist constants C' and C'' such that

$$|EZ_{i_1,j_1}\cdot\ldots\cdot Z_{i_{2q},j_{2q}}| \le C'' n^{-C}, \qquad \textbf{(A.8)}$$

for all indices $i_1,\dots,j_{2q}$ such that there exists a $1 \le l \le 2q$ with $|i_l - i_k| \ge C' \log n$ for all $k \ne l$ and $|i_l - j_k| \ge C' \log n$ for all k. For the proof of (A.8) one makes use of A(ii)–A(iv) and of the fact that $|\epsilon_i^*| \le 2\gamma_2 \log n$. Claim (A.7) follows by a bound on the remaining terms.

With the help of (A.7) and using the fact that, for c large enough,

$$\Big|\frac{\partial}{\partial x_1} h_1 n^{-1}\sum_{j=1}^n w(X_{j-2})\frac{p^{(1,0)}(x_1, X_{j-2})}{p(x_1, X_{j-2})^2} n^{-1}\sum_{i=1}^n A_i(x_1, X_{j-2})\epsilon_i^*\Big| \le n^c \qquad \textbf{(A.9)}$$

we get that uniformly for $x_1 \in B$

$$h_1 n^{-1} \sum_{j=1}^{n} w(X_{j-2}) \frac{p^{(1,0)}(x_1, X_{j-2})}{p(x_1, X_{j-2})^2} n^{-1} \sum_{i=1}^{n} A_i(x_1, X_{j-2})\epsilon_i^* = o_P(n^{-1/2}). \qquad \textbf{(A.10)}$$

For a proof that (A.10) holds uniformly for $x_1 \in B$ one shows first that it holds uniformly for $x_1 \in B_n$ where B_n is a subset of B with $n^{c'}$ equidistant points and where c' is an arbitrary positive constant. Then the claim follows from (A.9) if c' has been chosen large enough.

Note now that for all C''' there exists γ_2 such that $\max_{1\le i\le n}|\epsilon_i| \le \gamma_2 \log n$ (with probability tending to 1) and $|E\{\epsilon_i \mathbf{1}[|\epsilon_i| \le \gamma_2 \log n]\}| \le n^{-C'''}$. Remember that ϵ_i has a finite Laplace transform (see A(vi)). Therefore (A.10) with γ_2 large enough implies

$$h_1 n^{-1} \sum_{j=1}^{n} w(X_{j-2}) \frac{p^{(1,0)}(x_1, X_{j-2})}{p(x_1, X_{j-2})^2} n^{-1} \sum_{i=1}^{n} A_i(x_1, X_{j-2})\epsilon_i = o_P(n^{-1/2}). \qquad \textbf{(A.11)}$$

Again, this expansion holds uniformly for $x_1 \in B$.

With (A.11) and with similar expansions for other terms we arrive at

$$\begin{aligned}\tilde{m}_1^I(x_1) &= \alpha + m_1(x_1) + \left(\frac{1}{2}\right) h_1^2 m_1''(x_1) + (1/2) h_2^2 \beta \int m_1'' wf \\ &\quad + \frac{1}{n^2} \sum_{i,j=1}^{n} w(X_{j-2}) p(x_1, X_{j-2})^{-1} A_i(x_1, X_{j-2})\epsilon_i + o_P(n^{-1/2}).\end{aligned}$$

This expansion holds uniformly for $x_1 \in B$. Again using Davydov's inequality one shows that

$$\begin{aligned}&E\left[n^{-2} \sum_{i,j=1}^{n} \left\{ w(X_{j-2}) \frac{K_{h_2}(X_{i-2} - X_{j-2})}{p(x_1, X_{j-2})^2} - w(X_{i-2}) \frac{f(X_{i-2})}{p(x_1, X_{i-2})^2} \right\} K_{h_1}(X_{i-1} - x_1)\epsilon_i^* \right]^{2q} \\ &\quad = O(h_2^{4q}(nh_1)^{-q} n^{\gamma q}).\end{aligned}$$

As before this shows that uniformly for $x_1 \in B$

$$\begin{aligned}&\frac{1}{n^2} \sum_{i,j=1}^{n} \left[w(X_{j-2}) \frac{K_{h_2}(X_{i-2} - X_{j-2})}{p(x_1, X_{j-2})^2} - w(X_{i-2}) \frac{f(X_{i-2})}{p(x_1, X_{i-2})^2} \right] K_{h_1}(X_{i-1} - x_1)\epsilon_i \\ &\quad = o_P(n^{-1/2}).\end{aligned}$$

This gives uniformly for $x_1 \in B$

$$\begin{aligned}\tilde{m}_1^I(x_1) &= \alpha + m_1(x_1) + \left(\frac{1}{2}\right) h_1^2 m_1''(x_1) + \left(\frac{1}{2}\right) h_2^2 \beta \int m_1'' wf \\ &\quad + n^{-1} \sum_{i=1}^{n} w(X_{i-2}) p(x_1, X_{i-2})^{-1} f(X_{i-2}) K_{h_1}(X_{i-1} - x_1)\epsilon_i + o_P(n^{-1/2}).\end{aligned}$$

The expansion for $\hat{m}_1^I$ stated in the theorem follows by some straightforward calculations. For the treatment of $\hat{m}_2^I$ one proceeds similarly. ■

Sketch of Proof of Theorem 2. We first verify the stochastic expansion

$$n^{1/2}(\hat{\beta}_{LS} - \beta_0) = D_{LS}^{-1} n^{-1/2} \sum_{i=1}^{n} \{\epsilon_i \mathcal{H}_1(\mathbf{X}_i, \beta_0) + \mathcal{H}_2(\mathbf{X}_i, \beta_0)\} + o_p(1). \quad \textbf{(A.12)}$$

Again, the main tools are exponential inequalities and Davydov's inequality for mixing sequences. For doing this one proceeds as in the proof of Theorem 1. We omit details and give a short sketch for the proof of (A.12).

By a Taylor series expansion,

$$\begin{aligned}
0 = {} & n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} w_j(X_i)\hat{m}_1(X_i)\{\hat{m}_j(X_i) - \beta_0^{j-1}\hat{m}_1(X_i)\} \\
& + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)(j-2)\beta_0^{j-3} w_j(X_i)\hat{m}_1(X_i) \\
& \times \{\hat{m}_j(X_i) - \beta_0^{j-1}\hat{m}_1(X_i)\} n^{1/2}(\hat{\beta}_{LS} - \beta_0) \\
& - n^{-1} \sum_{j=1}^{n} \sum_{j=1}^{J} (j-1)^2 \beta_0^{2j-4} w_j(X_i)\hat{m}_1^2(X_i) n^{1/2}(\hat{\beta}_{LS} - \beta_0) + o_p(1).
\end{aligned}$$

The middle term is easily seen to be $o_p(1)$. The last term is easily seen to be $D_{LS} n^{1/2} \times (\hat{\beta}_{LS} - \beta_0) + o_p(1)$, where $D_{LS} = \sum_{j=1}^{J}\{(j-1)\beta_0^{j-2}\}^2 E\{w_j(X_i)m_1^2(X_i)\}$. Finally, the first term has the same behavior as if the leading $\hat{m}_1(X_i)$ were the same as $m_1(X_i)$. Making this substitution and invoking (17), because $m_j^{(2)}(x) = \beta_0^{j-1} m_1^{(2)}(x)$, we find that the result is asymptotically equivalent to $R_1 + R_2$ where

$$\begin{aligned}
R_1 = {} & n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} w_j(X_i) m_1(X_i) \\
& \times \sum_{\ell=1}^{n} \epsilon_\ell \{K_h(X_{\ell-j} - X_i) u_j(\mathbf{X}_l, X_i) - \beta_0^{j-1} K_h(X_{\ell-1} - X_i) u_1(\mathbf{X}_l, X_i)\}, \\
R_2 = {} & n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} w_j(X_i) m_1(X_i) \\
& \times \sum_{\ell=1}^{n} [\epsilon_\ell \{v_j(\mathbf{X}_l, X_i) - \beta_0^{j-1} v_1(\mathbf{X}_l, X_i)\} + \{t_j(\mathbf{X}_l) - \beta_0^{j-1} t_1(\mathbf{X}_l)\}].
\end{aligned}$$

Interchanging the indices i and ℓ and using the fact that $K(\cdot)$ is symmetric, the term R_1 equals

$$\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \epsilon_i \sum_{j=1}^{J} (j-1)\beta_0^{j-2} n^{-1} \sum_{\ell=1}^{n} w_j(X_\ell) m_1(X_\ell) \\
& \times \{K_h(X_\ell - X_{i-j}) u_j(\mathbf{X}_i, X_\ell) - \beta_0^{j-1} K_h(X_\ell - X_{i1}) u_1(\mathbf{X}_i, X_{\ell k})\}.
\end{aligned}$$

Using standard kernel theory and $h^3 n^{1/2} \to 0$ one shows that the last summation has the limit

$$w_j(X_{i-j}) m_1(X_{i-j}) p_{X_1}(X_{i-j}) u_j(\mathbf{X}_i, X_{i-j}) - \beta_0^{j-1} w_j(X_{i-1}) m_1(X_{i-1}) p_X(X_{i-1}) u_1(\mathbf{X}_i, X_{i-1}).$$

Similarly, $R_2 = R_{21} + R_{22}$, where

$$R_{21} = n^{-1/2} \sum_{i=1}^{n} \epsilon_i n^{-1} \sum_{j=1}^{J} \sum_{\ell=1}^{n} (j-1)\beta_0^{j-2} w_j(X_\ell) m_1(X_\ell) \{v_j(\mathbf{X}_i, X_\ell) - \beta_0^{j-1} v_1(\mathbf{X}_i, X_\ell)\}$$

$$\approx n^{-1/2} \sum_{i=1}^{n} \epsilon_i n^{-1} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} \{v_j^*(\mathbf{X}_i) - \beta_0^{j-1} v_1^*(\mathbf{X}_i)\}$$

and

$$R_{22} = n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} w_j(X_i) m_1(X_i) \sum_{\ell=1}^{n} \{t_j(\mathbf{X}_l) - \beta_0^{j-1} t_1(\mathbf{X}_l)\}$$

$$\approx n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J} (j-1)\beta_0^{j-2} s_j^* \{t_j(\mathbf{X}_i) - \beta_0^{j-1} t_1(\mathbf{X}_i)\}.$$

This shows (A.12). Asymptotic normality of $n^{1/2}(\hat{\beta}_{LS} - \beta_0)$ follows from (A.12) by application of a central limit theorem for strongly mixing sequences (see, e.g., Bosq, 1996, Theorem 1.7). We have to verify that for some $\gamma > 2$ $E|\epsilon_i \mathcal{H}_1(\mathbf{X}_i, \beta_0) + \mathcal{H}_2(\mathbf{X}_i, \beta_0)|^\gamma < \infty$. This follows easily from A(vi) and from the assumption that the functions $w_j, m_1, v_j, f_j, u_j, t_j$ are bounded for $j = 1, \ldots, J$. ■

Sketch of Proof of Theorem 3. One first shows that the following stochastic expansion (A.13) holds:

$$n^{1/2}(\hat{\beta}_{EIV} - \beta_0) = D_{LS}^{-1} n^{-1/2} \sum_{i=1}^{n} \{\epsilon_i \{\mathcal{H}_3(\mathbf{X}_i, \beta_0) + \mathcal{M}(\mathbf{X}_i, \beta_0)\} + \mathcal{G}(\mathbf{X}_i, \beta_0)\} + o_p(1). \quad \textbf{(A.13)}$$

Then one proceeds as in the last proof. Again, the main tools are exponential inequalities and Davydov's inequality for mixing sequences. We omit such details and refer to the arguments in the proof of Theorem 1.

The proof of (A.13) is facilitated by noting that

$$m_j(x) - \beta^{j-1} G\{m_1(x), \ldots, m_J(x), \beta\}$$

$$= \left\{\sum_{j=1}^{J} w_j(x)\beta^{2j-2}\right\}^{-1} \sum_{k=1}^{J} \{w_k(x)\beta^{2k-2} m_j(x) - w_k(x)\beta^{k+j-2} m_k(x)\}$$

$$= \left\{\sum_{j=1}^{J} w_j(x)\beta^{2j-2}\right\}^{-1} \sum_{k=1}^{J} w_k(x)\{\beta^{2k-2} m_j(x) - \beta^{k+j-2} m_k(x)\}.$$

Hence, $\hat{\beta}_{EIV}$ is formed by minimizing

$$n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J} w_j(X_i) s(\beta, X_i) \left[\sum_{\ell=1}^{J} w_\ell(X_i)\{\beta^{2\ell-2} \hat{m}_j(X_i) - \beta^{\ell-j-2} \hat{m}_\ell(X_i)\}\right]^2.$$

Define $s_\beta(\beta,x) = (\partial/\partial\beta)s(\beta,x)$ and $s_{\beta\beta}(\beta,x) = (\partial^2/\partial\beta^2)s(\beta,x)$. Further define

$$Q_j(x,\beta,m_1,\dots,m_J) = \sum_{\ell=1}^{J} w_\ell(x)\{\beta^{2\ell-2}m_j(x) - \beta^{\ell+j-2}m_\ell(x)\};$$

$$R_j(x,\beta,m_1,\dots,m_J) = \sum_{\ell=1}^{J} w_\ell(x)\{(2\ell-2)\beta^{2\ell-3}m_j(x) - (\ell+j-2)\beta^{\ell+j-3}m_\ell(x)\}.$$

Then $\hat\beta_{EIV}$ minimizes $n^{-1/2}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta,X_i)Q_j^2(X_i,\beta,\hat m_1,\dots,\hat m_j)$ and hence necessarily solves

$$\begin{aligned}0 &= n^{-1/2}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s_\beta(\beta,X_i)Q_j^2(X_i,\beta,\hat m_1,\dots,\hat m_J)\\ &\quad + 2n^{-1/2}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta,X_i)Q_j(X_i,\beta,\hat m_1,\dots,\hat m_J)R_j(X_i,\beta,\hat m_1,\dots,\hat m_J).\end{aligned}$$

This is an estimating equation, and with an admitted lack of rigor we proceed to analyze it in a standard fashion. Indeed, $n^{-1/2}$ times its first derivative evaluated at $(\beta_0,m_1,\dots,m_J)$ is

$$\begin{aligned}&n^{-1}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s_{\beta\beta}(\beta_0,X_i)Q_j^2(X_i,\beta,m_1,\dots,m_J)\\ &\quad + 4n^{-1}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s_\beta(\beta_0X_i)Q_j(X_i,\beta_0,m_1,\dots,m_J)R_j(X_i,\beta_0,m_1,\dots,m_J)\\ &\quad + 2n^{-1}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta_0,X_i)R_j^2(X_i,\beta_0,m_1,\dots,m_J)\\ &\quad + 2n^{-1}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta_0,X_i)Q_j(X_i,\beta_0,m_1,\dots,m_J)(\partial/\partial\beta)R_j(X_i,\beta_0,m,\dots,m_J).\end{aligned}$$

However, note that $Q_j(x,\beta_0,m_1,\dots,m_J) = 0$, so that the first derivative of the estimating equation when normalized and evaluated of $(\beta_0,m_1,\dots,m_J)$ is

$$\begin{aligned}&2n^{-1}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta_0,X_i)R_j^2(X_i,\beta_0,m_1,\dots,m_J)\\ &\quad \xrightarrow{p} 2\sum_{j=1}^J Ew_j(X_i)s(\beta_0,X_i)m_1^2(X_i)\left\{\sum_{\ell}^{J} w_\ell(X_i)\beta^{2\ell+j-4}(\ell-j)\right\}^2 = 2D_{EIV}.\end{aligned}$$

It is immediately obvious that because $Q_j^2(x,\beta_0,\hat m_1,\dots,\hat m_J) = o_p(n^{-1/2})$, then $n^{1/2}(\hat\beta_{EIV}-\beta_0)$ is asymptotically equivalent to

$$D_{EIV}^{-1}n^{-1/2}\sum_{i=1}^n\sum_{j=1}^J w_j(X_i)s(\beta_0,X_i)Q_j(X_i,\beta_0,\hat m_1,\dots,\hat m_J)R_j(X_i,\beta_0,\hat m_1,\dots,\hat m_J).$$

Because $Q_j(x, \beta_0, m_1, \ldots, m_J) = 0$, $n^{1/2}(\hat{\beta}_{EIV} - \beta_0)$ is asymptotically equivalent to

$$D_{EIV}^{-1} n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J} w_j(X_i) s(\beta_0, X_i) R_j(X_i, \beta_0, m_1, \ldots, m_J)$$

$$\times \sum_{\ell}^{J} w_\ell(X_i)[\beta_0^{2\ell-2}\{\hat{m}_j(X_i) - m_j(X_i)\} - \beta_0^{\ell+j-2}\{\hat{m}_\ell(X_i) - m_\ell(X_i)\}]. \qquad \textbf{(A.14)}$$

The proof is completed by using expansion (A.7) and the fact that $m_j^{(2)}(x) = \beta_0^{j-1} m_1^{(2)}(x)$, in a manner similar to that of Theorem 2. After some algebra one arrives at (A.13). ■

Sketch of Proof of Theorem 4. Proceeding as in the proof of central limit theorems for kernel estimates of strongly mixing sequences (see, e.g., Bosq, 1996, Theorems 2.3 and 3.4) and using the stochastic expansion (20), one shows that the vector $n^{2/5}\{\hat{m}_1(x) - m_1(x), \ldots, \hat{m}_J(x) - \beta^{J-1} m_1(x)\}^T$ has a normal limit with mean $(\frac{1}{2}) n^{2/5} h^2 r(x)(1, \ldots, \beta^{J-1})^T$ and covariance matrix equal to a diagonal matrix with diagonal elements $n^{1/5} h^{-1} \nu(K) f(x)\, s_j^2(x)$. Note that $\hat{m}_j(x)$ and $\hat{m}_k(x)$ are asymptotically independent for $j \neq k$. The first statement of Theorem 4 follows from the fact that $\hat{m}_1^*(x) - \sum_{j=1}^J c_j \beta^{-(j-1)} \hat{m}_j(x) / \sum_{j=1}^J c_j = o_P(n^{-2/5})$. For the second statement of Theorem 4 note that $\sum_{j=1}^J c_j^2 \beta^{-2(j-1)} s_j^2(x) / \sum_{j=1}^J c_j^2$ is minimized by $c_j = \beta^{2(j-1)} s_j^{-2}(x)$. ■