Econometrica, Vol. 73, No. 3 (May, 2005), 771–836

ESTIMATING SEMIPARAMETRIC ARCH(∞) MODELS BY KERNEL SMOOTHING METHODS[1]

BY O. LINTON[2] AND E. MAMMEN[3]

We investigate a class of semiparametric ARCH(∞) models that includes as a special case the partially nonparametric (PNP) model introduced by Engle and Ng (1993) and which allows for both flexible dynamics and flexible function form with regard to the "news impact" function. We show that the functional part of the model satisfies a type II linear integral equation and give simple conditions under which there is a unique solution. We propose an estimation method that is based on kernel smoothing and profiled likelihood. We establish the distribution theory of the parametric components and the pointwise distribution of the nonparametric component of the model. We also discuss efficiency of both the parametric part and the nonparametric part. We investigate the performance of our procedures on simulated data and on a sample of S&P500 index returns. We find evidence of asymmetric news impact functions, consistent with the parametric analysis.

KEYWORDS: ARCH, inverse problem, kernel estimation, news impact curve, nonparametric regression, profile likelihood; semiparametric estimation, volatility.

1. INTRODUCTION

STOCHASTIC VOLATILITY MODELS are of considerable current interest in empirical finance following the seminal work of Engle (1982). Perhaps the most popular version of this is Bollerslev's (1986) GARCH(1, 1) model in which the conditional variance σ_t^2 of a martingale difference sequence y_t is

$$\sigma_t^2 = \beta\sigma_{t-1}^2 + \alpha + \gamma y_{t-1}^2. \tag{1}$$

This model has been extensively studied and generalized in various ways. See the review of Bollerslev, Engle, and Nelson (1994). This paper is about a particular class of nonparametric/semiparametric generalizations of (1). The motivation for this line of work is to increase the flexibility of the class of models we use and to learn from this the shape of the volatility function without restricting it a priori to have or not have certain shapes.

The nonparametric ARCH literature apparently begins with Pagan and Schwert (1990) and Pagan and Hong (1991). They consider the case where $\sigma_t^2 = \sigma^2(y_{t-1})$, where $\sigma(\cdot)$ is a smooth but unknown function, and the multilag

[1] We would like to thank Costas Meghir, two referees, Xiaohong Chen, Paul Doukhan, Wolfgang Härdle, Christian Huse, Dennis Kristensen, Jens Nielsen, and Eric Renault for helpful comments.

[2] Research was supported by the Economic and Social Science Research Council of the United Kingdom.

[3] Supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Humboldt-Universität zu Berlin and Project MA1026/6-2.

version $\sigma_t^2 = \sigma^2(y_{t-1}, y_{t-2}, \ldots, y_{t-d})$. Härdle and Tsybakov (1997) applied local linear fit to estimate the volatility function together with the mean function and derived their joint asymptotic properties. The multivariate extension is given in Härdle, Tsybakov, and Yang (1996). Masry and Tjøstheim (1995) also estimate nonparametric ARCH models using the Nadaraya–Watson kernel estimator. In practice, it is necessary to include many lagged variables. The problem with this is that nonparametric estimation of a multidimensional regression surface suffers from the well-known "curse of dimensionality": the optimal rate of convergence decreases with dimensionality d; see Stone (1980). In addition, it is hard to describe, interpret, and understand the estimated regression surface when the dimension is more than two. Furthermore, even for large d this model greatly restricts the dynamics for the variance process since it effectively corresponds to an ARCH(d) model, which is known in the parametric case not to capture the dynamics well. In particular, if the conditional variance is highly persistent, the nonparametric estimator of the conditional variance will provide a poor approximation, as reported in Perron (1998). So not only does this model not capture adequately the time series properties of many data sets, but the statistical properties of the estimators can be poor and the resulting estimators hard to interpret.

Additive models offer a flexible but parsimonious alternative to nonparametric models, and have been used in many contexts, see Hastie and Tibshirani (1990). Suppose that $\sigma_t^2 = c_v + \sum_{j=1}^d \sigma_j^2(y_{t-j})$. The best achievable rate of convergence for estimates of $\sigma_j^2(\cdot)$ is that of one-dimensional nonparametric regression; see Stone (1985). Yang, Härdle, and Nielsen (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is additive, but the volatility is multiplicative: $\sigma_t^2 = c_v \prod_{j=1}^d \sigma_j^2(y_{t-j})$. Their estimation strategy is based on the method of partial means/marginal integration using local linear fits as a pilot smoother. Kim and Linton (2004) generalize this model to allow for arbitrary (but known) transformations, i.e., $G(\sigma_t^2) = c_v + \sum_{j=1}^d \sigma_j^2(y_{t-j})$, where $G(\cdot)$ is a known function like log or level. Horowitz (2001) has analyzed the model where $G(\cdot)$ is also unknown, but his results were only in a cross-sectional setting. These separable models deal with the curse of dimensionality, but still do not capture the persistence of volatility and specifically they do not nest the favorite GARCH$(1, 1)$ process.

This paper analyzes a class of semiparametric ARCH models that has both general functional form aspects and flexible dynamics. A special case of our model is the Engle and Ng (1993) PNP model where $\sigma_t^2 = \beta\sigma_{t-1}^2 + m(y_{t-j})$, where $m(\cdot)$ is a smooth but unknown function. Our semiparametric model nests the simple GARCH$(1, 1)$ model but permits more general functional form: it allows for an asymmetric leverage effect and as much dynamics as GARCH$(1, 1)$. A major issue we solve is how to estimate the function $m(\cdot)$ by kernel methods. Our estimation approach is to derive population moment conditions for the nonparametric part and then solve them with empirical counterparts. The moment conditions we obtain are linear type II Fredholm integral equations, and so they fall in the class of inverse problems reviewed in Carrasco, Florens, and Renault (2003). These equations have been extensively studied in the applied mathematics literature; see, for example, Tricomi (1957). They also arise a lot in economic theory; see Stokey and Lucas (1989). The solution of these equations in our case only requires the computation of two-dimensional smoothing operations and one-dimensional integration, and so is attractive computationally. From a statistical perspective, there has been some recent work on this class of estimation problems. Starting with Friedman and Stuetzle (1981), in Breiman and Friedman (1985) and Hastie and Tibshirani (1990) these methods have been investigated in the context of additive nonparametric regression and related models, where the estimating equations are usually of type II. Recently, Opsomer and Ruppert (1997) and Mammen, Linton, and Nielsen (1999) have provided a pointwise distribution theory for this specific class of problems. Newey and Powell (2003) studied nonparametric simultaneous equations and obtained an estimation equation that was a linear integral equation also, except that it is the more difficult type I. They establish the uniform consistency of their estimator; see also Darolles, Florens, and Renault (2002). Hall and Horowitz (2003) establish the optimal rate for estimation in this problem and propose two estimators that achieve this rate. Neither paper provides pointwise distribution theory. Our estimation methods and proof technique are purely applicable to the type II situation, which is nevertheless quite common elsewhere in economics. For example, Berry and Pakes (2002) derive estimators for a class of semiparametric dynamic models used in industrial organization applications, and which solve type II equations similar to ours.

Our paper goes significantly beyond the existing literature in two respects. First, the integral operator does not necessarily have norm less than 1 so that the iterative solution method of successive approximations is not feasible. This also affects the way we derive the asymptotic properties, and we cannot directly apply the results of Mammen, Linton, and Nielsen (1999) here. Second, we have also finite-dimensional parameters and their estimation is of interest in itself. We establish the consistency and pointwise asymptotic normality of our estimates of the parameter and of the function. We establish the semiparametric efficiency bound for a Gaussian special case and show that our parameter estimator achieves this bound. We also discuss the efficiency question regarding the nonparametric component and conclude that a likelihood-based version of our estimator cannot be improved on without additional structure. We investigate the practical performance of our method on simulated data and present the result of an application to S&P500 data. The empirical results indicate some asymmetry and nonlinearity in the news impact curve.

Our model is introduced in the next section. In Section 3 we present our estimators. In Section 4 we give the asymptotic properties. In Section 5 we discuss an extension of our basic setting that accommodates a richer variety of tail behavior. Section 6 reports some numerical results and Section 7 concludes.

2. THE MODEL AND ITS PROPERTIES

We shall suppose throughout that the process $\{y_t\}_{t=-\infty}^{\infty}$ is stationary with finite fourth moment. We concentrate most of our attention on the case where there is no mean process, although we later discuss the extension to allow for some mean dynamics. Define the volatility process model

$$\sigma_t^2(\theta, m) = \mu_t + \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j}), \tag{2}$$

where $\mu_t \in \mathbb{R}$, $\theta \in \Theta \subset \mathbb{R}^p$, and $m \in \mathcal{M}$, where $\mathcal{M} = \{m\colon \text{measurable}\}$. The coefficients $\psi_j(\theta)$ satisfy at least $\psi_j(\theta) \geq 0$ and $\sum_{j=1}^{\infty} \psi_j(\theta) < \infty$ for all $\theta \in \Theta$. The true parameters θ_0 and the true function $m_0(\cdot)$ are unknown and to be estimated from a finite sample $\{y_1, \ldots, y_T\}$. The process μ_t can be allowed to depend on covariates and unknown parameters, but at this stage it assumed to be known. In much of the sequel it can be put equal to zero without any loss of generality. It will become important below when we consider more restrictive choices of $\mathcal{M}$. Robinson (1991) is perhaps the first study of ARCH(∞) models, although he restricted attention to the quadratic m case.

Following Drost and Nijman (1993), we can give three interpretations to (2). The *strong* form ARCH(∞) process arises when

$$\frac{y_t}{\sigma_t} = \varepsilon_t \tag{3}$$

is i.i.d with mean 0 and variance 1, where $\sigma_t^2 = \sigma_t^2(\theta_0, m_0)$. The *semistrong* form arises when

$$E(y_t|\mathcal{F}_{t-1}) = 0 \quad \text{and} \quad E(y_t^2|\mathcal{F}_{t-1}) \equiv \sigma_t^2, \tag{4}$$

where $\mathcal{F}_{t-1}$ is the sigma field generated by the entire past history of the y process. Finally, there is a *weak* form in which σ_t^2 is defined as the projection on a certain subspace. Specifically, let θ_0, m_0 be defined as the minimizers of the population least squares criterion function

$$S(\theta, m) = E\left[\left\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})\right\}^2\right] \tag{5}$$

and let $\sigma_t^2 = \sum_{j=1}^{\infty} \psi_j(\theta_0) m_0(y_{t-j})$. The criterion (5) is well defined only when $E(y_t^4) < \infty$.

In the special case that $\psi_j(\theta) = \theta^{j-1}$, with $0 < \theta < 1$, we can rewrite (2) as a difference equation in the unobserved variance

$$\sigma_t^2 = \theta \sigma_{t-1}^2 + m(y_{t-1}) \qquad (t = 1, 2, \ldots), \tag{6}$$

and this is consistent with a stationary GARCH(1, 1) structure for the unobserved variance when $m(y) = \alpha + \gamma y^2$ for some parameters α, γ. It also includes other parametric models as special cases: the Glosten, Jegannathan, and Runkle (1993) model, taking $m(y) = \alpha + \gamma y^2 + \delta y^2 \mathbb{1}(y < 0)$, the Engle (1990) asymmetric model, taking $m(y) = \alpha + \gamma(y + \delta)^2$, and the Engle and Bollerslev (1986) model, taking $m(y) = \alpha + \gamma|y|^\delta$.

The function $m(\cdot)$ is the "news impact function," and it determines the way in which the volatility is affected by shocks to y. Our model allows for general news impact functions including both symmetric and asymmetric functions, and so accommodates the leverage effect (Nelson (1991)). The parameter θ, through the coefficients $\psi_j(\theta)$, determines the persistence of the process, and we in principle allow for quite general coefficient values. A general class of coefficients can be obtained from the expansion of autoregressive moving average (ARMA) lag polynomials, as in Nelson (1991).

Our model generalizes the model considered in Carroll, Mammen, and Härdle (2002) in which $\sigma_t^2 = \sum_{j=1}^{\tau} \theta_0^{j-1} m_0(y_{t-j})$ for some finite τ. Their estimation strategy was quite different from ours: they relied on an initial estimator of a τ-dimensional surface and then marginal integration (Linton and Nielsen (1995)) to improve the rate of convergence. This method is likely to work poorly when τ is very large. Also, their theory requires the smoothness of m to increase with τ. Indeed, a contribution of our paper is to provide an estimation method for θ_0 and $m(\cdot)$ that just relies on one-dimensional smoothing operations, but is also amenable to theoretical analysis. Some other papers can be considered precursors to this one. First, Gouriéroux and Monfort (1992) introduced the qualitative threshold ARCH (QTARCH) which allowed quite flexible patterns of conditional mean and variance through step functions, although their analysis was purely parametric. Engle and Ng (1993) analyzed precisely the semistrong model (2) with $\psi_j(\theta) = \theta^{j-1}$ and called it partially nonparametric or PNP for short. They proposed an estimation strategy based on piecewise linear splines. Finally, we should mention some work by Audrino and Bühlmann (2001): their model is that $\sigma_t^2 = \Lambda(y_{t-1}, \sigma_{t-1}^2)$ for some smooth but unknown function $\Lambda(\cdot)$, and includes the PNP model as a special case. However, although they proposed an estimation algorithm, they did not establish the distribution theory of their estimator.

In the next subsection we discuss a characterization of the model that generates our estimation strategy. If m were known, it would be straightforward to estimate θ from some likelihood or least squares criterion. The main issue is how to estimate $m(\cdot)$ even when θ is known. The kernel method likes to express the function of interest as a conditional expectation or density of a small number of observable variables, but this is not directly possible here because m is only implicitly defined. However, we are able to show that m can be expressed in terms of all the bivariate joint densities of (y_t, y_{t-j}), $j = \pm 1, \ldots$, i.e., this collection of bivariate densities forms a set of sufficient statistics for our model. We use this relationship to generate our estimator.

2.1. *Linear Characterization*

Suppose for pedagogic purposes that the semistrong process defined in (4) holds, and for simplicity define $\widetilde{y}_t^2 = y_t^2 - \mu_t$. Take marginal expectations for any $j \geq 1$,

$$E(\widetilde{y}_t^2 | y_{t-j} = y) = \psi_j(\theta_0) m(y) + \sum_{k \neq j}^{\infty} \psi_k(\theta_0) E[m(y_{t-k}) | y_{t-j} = y]. \tag{7}$$

For each such j the above equation implicitly defines $m(\cdot)$. This is really a moment condition in the functional parameter $m(\cdot)$ for each j, and can be used as an estimating equation. As in the parametric method of moments case, it can pay to combine the estimating equations in terms of efficiency. Specifically, we take the linear combination of these moment conditions,

$$\begin{aligned} &\sum_{j=1}^{\infty} \psi_j(\theta_0) E(\widetilde{y}_t^2 | y_{t-j} = y) \\ &\quad = \sum_{j=1}^{\infty} \psi_j^2(\theta_0) m(y) + \sum_{j=1}^{\infty} \psi_j(\theta_0) \sum_{k \neq j}^{\infty} \psi_k(\theta_0) E[m(y_{t-k}) | y_{t-j} = y], \end{aligned} \tag{8}$$

which yields another implicit equation in $m(\cdot)$.

This equation arises as the first order condition from the least squares definition of σ_t^2, given in (5), as we now discuss. We can assume that the quantities $\theta_0, m_0(\cdot)$ are the unique minimizers of (5) over $\Theta \times \mathcal{M}$ by the definition of conditional expectation, see Drost and Nijman (1993). Furthermore, the minimizer of (5) satisfies a first-order condition and in the Appendix we show that this first-order condition is precisely (8). In fact, if we minimize (5) with respect to $m \in \mathcal{M}$ for any $\theta \in \Theta$ and let m_θ denote this minimizer, then m_θ satisfies (8) with θ_0 replaced by θ. Note that we are treating μ_t as a known quantity.

We next rewrite (8) (for general θ) in a more convenient form. Let p_0 denote the marginal density of y and let $p_{j,l}$ denote the joint density of y_j, y_l. Define

$$\mathcal{H}_\theta(y, x) = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta) \frac{p_{0,j}(y, x)}{p_0(y) p_0(x)}, \tag{9}$$

$$m_\theta^*(y) = \sum_{j=1}^{\infty} \psi_j^\dagger(\theta) g_j(y), \tag{10}$$

where $\psi_j^\dagger(\theta) = \psi_j(\theta) / \sum_{l=1}^{\infty} \psi_l^2(\theta)$ and $\psi_j^*(\theta) = \sum_{k \neq 0} \psi_{j+k}(\theta) \psi_j(\theta) / \sum_{l=1}^{\infty} \psi_l^2(\theta)$, while $g_j(y) = E(\widetilde{y}_t^2 | y_{t-j} = y)$ for $j \geq 1$. Then the function $m_\theta(\cdot)$ satisfies

$$m_\theta(y) = m_\theta^*(y) + \int \mathcal{H}_\theta(y, x) m_\theta(x) p_0(x)\, dx \tag{11}$$

for each $\theta \in \Theta$ (this equation is equivalent to (8) for all $\theta \in \Theta$). The operator $\mathcal{H}_j(y,x) = p_{0,j}(y,x)/p_0(y)p_0(x)$ is well studied in the statistics literature (see Bickel, Klaassen, Ritov, and Wellner (1993, p. 440)); our operator $\mathcal{H}_\theta$ is just a weighted sum of such operators, where the weights are declining to zero rapidly. In additive nonparametric regression, the corresponding integral operator is an unweighted sum of operators like $\mathcal{H}_j(y,x)$ over the finite number of dimensions (see Hastie and Tibshirani (1990) and Mammen, Linton, and Nielsen (1999)). Although the operators $\mathcal{H}_j$ are not self-adjoint without an additional assumption of time reversibility, it can easily be seen that $\mathcal{H}_\theta$ is self-adjoint in $L_2(p_0)$ due to the two-sided summation.[4]

Our estimation procedure will be based on plugging estimates $\widehat{m}_\theta^*$ and $\widehat{\mathcal{H}}_\theta$ of m_θ^* or $\mathcal{H}_\theta$, respectively, into (11) and then solving for $\widehat{m}_\theta$. The estimates $\widehat{m}_\theta^*$ and $\widehat{\mathcal{H}}_\theta$ will be constructed by plugging estimates of $p_{0,j}$, p_0, and g_j into (10) and (9). Nonparametric estimates of these functions only work accurately for arguments not too large. We do not want to enter into a discussion of tail behavior of nonparametric estimates at this point. For this reason we change our minimization problem (5), or rather restrict the parameter sets further. We consider minimization of (5) over all $\theta \in \Theta$ and $m \in \mathcal{M}_c$, where now $\mathcal{M}_c$ is the class of all bounded measurable functions that vanish outside $[-c, c]$, where c is some fixed constant (this makes $\sigma_t^2 = \mu_t$ whenever $y_{t-j} \notin [-c, c]$ for all j). Let us denote these minimizers by θ_c and m_c. Furthermore, denote the minimizer of (5) for fixed θ over $m \in \mathcal{M}_c$ by $m_{\theta,c}$. Then θ_c and m_c minimize $E[\{\widetilde{y}_t^2 - \sum_{j=1}^\infty \psi_j(\theta)m(y_{t-j})\}^2]$ over $\Theta \times \mathcal{M}_c$ and $m_{\theta,c}$ minimizes $E[\{\widetilde{y}_t^2 - \sum_{j=1}^\infty \psi_j(\theta)m_\theta(y_{t-j})\}^2]$ over $\mathcal{M}_c$. For now we adopt a fixed truncation where c and μ_t are constant and known, but return to this in Section 5. Then $m_{\theta,c}$ satisfies $m_{\theta,c}(y) = m_\theta^*(y) + \int_{-c}^c \mathcal{H}_\theta(y,x)m_{\theta,c}(x)p_0(x)\,dx$ for $|y| \le c$ and vanishes for $|y| > c$. For simplicity but in abuse of notation we omit the subindex c of $m_{\theta,c}$ and we write

$$m_\theta = m_\theta^* + \mathcal{H}_\theta m_\theta. \tag{12}$$

For each $\theta \in \Theta$, $\mathcal{H}_\theta$ is a self-adjoint linear operator on the Hilbert space of functions m that are defined on $[-c, c]$ with norm $\|m\|_2^2 = \int_{-c}^c m(x)^2 p_0(x)\,dx$ and (12) is a linear integral equation of the second kind. There are some general results providing sufficient conditions under which such integral equations have a unique solution. See Darolles, Florens, and Renault (2002) for a discussion on existence and uniqueness for the more general class of type I equations.

We assume the following high level condition:

[4] Specifically, with $\langle f, g\rangle = \int f(x)g(x)p_0(x)\,dx$ denoting the usual inner product in $L_2(p_0)$, we have $\langle g, \mathcal{H}_\theta m\rangle = -\sum\sum_{j\neq k} \psi_j(\theta)\psi_k(\theta)E[g(y_{t-j})E[m(y_{t-k})|y_{t-j}]] = -\sum\sum_{j\neq k} \psi_j(\theta)\psi_k(\theta)E[g(y_{t-j})m(y_{t-k})] = \langle \mathcal{H}_\theta g, m\rangle$ because the double sum is symmetric in j, k. The definition of adjoint operator can be found in Bickel, Klaassen, Ritov, and Wellner (1993, p. 416).

ASSUMPTION A1: *The operator $\mathcal{H}_\theta(x, y)$ is Hilbert–Schmidt uniformly over θ, i.e.*, $\sup_{\theta\in\Theta}\int_{-c}^{c}\int_{-c}^{c}\mathcal{H}_\theta(x, y)^2 p_0(x)p_0(y)\,dx\,dy < \infty$.

A sufficient condition for Assumption A1 is that the joint densities $p_{0,j}(y, x)$ are uniformly bounded for $j \neq 0$ and $|x|, |y| \le c$, and that the density $p_0(x)$ is bounded away from 0 for $|x| \le c$.

Under Assumption A1, for each $\theta \in \Theta$, $\mathcal{H}_\theta$ is a self-adjoint bounded compact linear operator on the Hilbert space of functions $L_2(p_0)$, and therefore has a countable number of eigenvalues: $\infty > |\lambda_{\theta,1}| \ge |\lambda_{\theta,2}| \ge \cdots$, with $\sup_{\theta\in\Theta}\sum_{j=1}^{\infty}\lambda_{\theta,j}^2 < \infty$.

ASSUMPTION A2: *There exist no $\theta \in \Theta$ and $m \in \mathcal{M}_c$ with $\|m\|_2 = 1$ such that $\sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j}) = 0$ with probability* 1.

This condition rules out a certain "concurvity" in the stochastic process. That is, the data cannot be functionally related in this particular way. It is a natural generalization to our situation of the condition that the regressors be not linearly related in a linear regression. A special case of this condition was used in Weiss (1986) and Kristensen and Rahbek (2003) for identification in parametric ARCH models, see also the arguments used in Lumsdaine (1996, Lemma 5) and Robinson and Zaffaroni (2002, Lemma 9).

ASSUMPTION A3: *The operator $\mathcal{H}_\theta$ fulfills the following continuity condition for $\theta, \theta' \in \Theta$: $\sup_{\|m\|_2\le 1}\|\mathcal{H}_\theta m - \mathcal{H}_{\theta'}m\|_2 \to 0$ for $\|\theta - \theta'\| \to 0$.*

This condition is straightforward to verify. We now argue that because of Assumptions A2 and A3, for a constant $0 < \gamma < 1$,

$$\sup_{\theta\in\Theta}\lambda_{\theta,1} < \gamma. \tag{13}$$

To prove this note that for $\theta \in \Theta$ and $m \in \mathcal{M}_c$ with $\|m\|_2 = 1$,

$$\begin{aligned}
0 &< E\left[\left(\sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j})\right)^2\right]\\
&= \chi_\theta\int_{-c}^{c} m^2(x)p_0(x)\,dx\\
&\quad + \chi_\theta\int_{-c}^{c}\int_{-c}^{c} m(x)m(y)\sum_{|k|\ge 1}\psi_k^*(\theta)p_{0,k}(x, y)\,dx\,dy\\
&= \chi_\theta\int_{-c}^{c} m^2(x)p_0(x)\,dx - \chi_\theta\int_{-c}^{c} m(x)\mathcal{H}_\theta m(x)p_0(x)\,dx,
\end{aligned}$$

where $\chi_\theta = \sum_{j=1}^{\infty} \psi_j^2(\theta)$ is a positive constant depending on θ. For eigenfunctions $m \in \mathcal{M}_c$ of $\mathcal{H}_\theta$ with eigenvalue λ this shows that $\int m^2(x)p_0(x)\,dx - \lambda \int m^2(x)p_0(x)\,dx > 0$. Therefore $\lambda_{\theta,j} < 1$ for $\theta \in \Theta$ and $j \geq 1$. Now, because of Assumption A3 and compactness of Θ, this implies (13).

From (13) we get that $I - \mathcal{H}_\theta$ has eigenvalues bounded from below by $1 - \gamma > 0$. Therefore $I - \mathcal{H}_\theta$ is strictly positive definite and hence invertible, and $(I - \mathcal{H}_\theta)^{-1}$ has only positive eigenvalues that are bounded by $(1-\gamma)^{-1}$:

$$\text{(14)} \qquad \sup_{\theta\in\Theta, m\in\mathcal{M}_c, \|m\|_2=1} \|(I - \mathcal{H}_\theta)^{-1} m\|_2 \leq (1-\gamma)^{-1}.$$

Therefore, we can directly solve the integral equation (12) and write

$$\text{(15)} \qquad m_\theta = (I - \mathcal{H}_\theta)^{-1} m_\theta^*$$

for each $\theta \in \Theta$. The representation (15) is fundamental to our estimation strategy, as it yields identification of m_θ.

We next discuss a further property that leads to an iterative solution method rather than a direct inversion. If it holds that $|\lambda_{\theta,1}| < 1$, then $m_\theta = \sum_{j=0}^{\infty} \mathcal{H}_\theta^j m_\theta^*$. In this case the sequence of successive approximations $m_\theta^{[n]} = m_\theta^* + \mathcal{H}_\theta m_\theta^{[n-1]}$, $n = 1, 2, \ldots,$ converges in norm geometrically fast to m_θ from any starting point. This sort of property has been established in other related problems—see Hastie and Tibshirani (1990) for discussion—and is the basis of most estimation algorithms in this area. Unfortunately, the conditions that guarantee convergence of the successive approximations method are not likely to be satisfied here even in the special case that $\psi_j(\theta) = \theta^{j-1}$. The reason is that the unit function is always an eigenfunction of $\mathcal{H}_\theta$ with eigenvalue determined by $-\sum_{j=\pm1}^{\pm\infty} \theta^{|j|} 1 = \lambda_\theta \cdot 1$, which implies that $\lambda_\theta = -2\theta/(1-\theta)$. This is less than 1 in absolute value only when $\theta < 1/3$. This implies that we will not be able to use directly the particularly convenient method of successive approximations (i.e., backfitting) for estimation: however, with some modifications it can be applied; see Linton and Mammen (2003).

2.2. *Likelihood Characterization*

In this section we provide an alternative characterization of m_θ, θ in terms of the Gaussian likelihood. We use this characterization later to define the semiparametric efficiency bound for estimating θ in the presence of unknown m. This characterization is also important for robustness reasons, since it does not require fourth moments on y_t.

Suppose that $m_0(\cdot)$, θ_0 are defined as the minimizers of the criterion function

$$\text{(16)} \qquad \ell(\theta, m) = E\left[\log \sigma_t^2(\theta, m) + \frac{y_t^2}{\sigma_t^2(\theta, m)}\right]$$

with respect to both $\theta, m(\cdot)$, where $\sigma_t^2(\theta, m) = \mu_t + \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})$. Notice that this criterion is well defined in many cases where the quadratic loss function is not.

Minimizing (16) with respect to m for each given θ yields the first-order condition, which is a nonlinear integral equation in m:

$$\sum_{j=1}^{\infty} \psi_j(\theta) E\big[\sigma_t^{-4}(\theta, m)\{y_t^2 - \sigma_t^2(\theta, m)\}|y_{t-j} = y\big] = 0. \tag{17}$$

This equation is difficult to work with from the point of view of statistical analysis because of the nonlinearity; see Horowitz and Mammen (2002). We consider instead a linearized version of this equation. Suppose that we have some initial approximation to σ_t^2. Then linearizing (17) about σ_t^2, we obtain the linear integral equation

$$\begin{aligned} &\overline{m}_\theta = \overline{m}_\theta^* + \overline{\mathcal{H}}_\theta \overline{m}_\theta; \\ &\overline{m}_\theta^* = \frac{\sum_{j=1}^{\infty} \psi_j(\theta) g_j^a(y)}{\sum_{j=1}^{\infty} \psi_j^2(\theta) g_j^b(y)}, \\ &\mathcal{H}_\theta(x, y) = \frac{-\sum_{j=1}^{\infty} \sum_{l=1, l\neq j}^{\infty} \psi_j(\theta) \psi_l(\theta) g_{l,j}^c(x, y) \frac{p_{0,l-j}(x,y)}{p_0(y) p_0(x)}}{\sum_{j=1}^{\infty} \psi_j^2(\theta) g_j^b(y)}. \end{aligned} \tag{18}$$

Here, $g_j^a(y) = E[\sigma_t^{-4} y_t^2 | y_{t-j} = y] = E[\sigma_t^{-2} | y_{t-j} = y]$, $g_j^b(y) = E[\sigma_t^{-4} | y_{t-j} = y]$, and $g_{l,j}^c(x, y) = E[\sigma_t^{-4} | y_{t-l} = x, y_{t-j} = y]$. This is a second kind linear integral equation in $\overline{m}_\theta(\cdot)$ but with a different intercept and operator from (12). See Hastie and Tibshirani (1990, Section 6.5) for a similar calculation. Under our assumptions, see B4, the weighted operator satisfies Assumptions A1 and A3 also. For a proof of Assumption A3 note that $0 < E[\sigma_t^{-4} \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})]^2$.

Note that in general $\overline{m}_\theta$ differs from m_θ, since they are defined as minimizers of different criteria. However, for the strong and semistrong versions of our model we get $\overline{m}_{\theta_0} = m_{\theta_0}$.

3. ESTIMATION

We shall construct estimates of θ and m from a sample $\{y_1, \dots, y_T\}$. We proceed in four steps. First, for each given θ we compute estimates of m_θ^* and $\mathcal{H}_\theta$, and then estimate m_θ by solving an empirical version of the integral equation (12). We then estimate θ by minimizing a profile least squares criterion. We then use the estimated parameter to give an estimator of $m(\cdot)$. Finally, we use our consistent estimators to define likelihood-based estimators that improve efficiency under some conditions. In particular, we solve an empirical version of the linearized likelihood implied integral equation (18) and then minimize a negative quasi-likelihood criterion to update the parameter estimate. In Section 3.1 we discuss how to compute m_θ^* and $\mathcal{H}_\theta$, while in Section 3.2 we state our estimation algorithm; in Section 3.3 we give further details about solving integral equations of this type.

3.1. *Our Estimators of m_θ^* and $\mathcal{H}_\theta$*

We now define local polynomial-based estimates $\widehat{m}_\theta^*$ of m_θ^* and kernel density estimates $\widehat{\mathcal{H}}_\theta$ of $\mathcal{H}_\theta$, respectively. Local linear estimation is a popular approach for estimating various conditional expectations with nice properties (see Fan (1992)). Define the estimator $\widehat{g}_j(y) = \widehat{a}_0$, where $(\widehat{a}_0, \dots, \widehat{a}_p)$ are the minimizers of the weighted sums of squares criterion

$$\sum_{t:1\le t-j\le T} \{y_t^2 - \mu_t - a_0 - a_1(y_{t-j} - y) - \cdots - a_p(y_{t-j} - y)^p\}^2 \\ \times K_h(y_{t-j} - y)$$

with respect to $(a_0, \dots, a_p)$, where K is a symmetric probability density function, h is a positive bandwidth, and $K_h(\cdot) = K(\cdot/h)/h$. We can allow $h = h_T(y)$, but for notational and theoretical simplicity we shall drop the dependence on y. Our theoretical properties are stated for the case $p = 1$, but the theory easily extends; in practice other choices may have some advantages.

Select a truncation sequence τ_T with $1 < \tau_T < T$ and compute $\widehat{m}_\theta^*(y) = \sum_{j=1}^{\tau_T} \psi_j^\dagger(\theta)\widehat{g}_j(y)$ for any $|y| \le c$. To estimate $\mathcal{H}_\theta$ we take the scheme

$$\widehat{\mathcal{H}}_\theta(y, x) = -\sum_{j=\pm 1}^{\pm\tau_T} \psi_j^*(\theta)\frac{\widehat{p}_{0,j}(y, x)}{\widehat{p}_0(y)\widehat{p}_0(x)},$$

$$\widehat{p}_{0,j}(y, x) = \frac{1}{T - |j|}\sum_{t:1\le t-j\le T} K_h(y - y_t)K_h(x - y_{t+j}) \quad \text{and}$$

$$\widehat{p}_0(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - y_t).$$

The action of the empirical operator is defined as $\widehat{\mathcal{H}}_\theta m = \int_{-c}^{c} \widehat{\mathcal{H}}_\theta(y, x)m(x) \times \widehat{p}_0(x)\,dx$. For each $\theta \in \Theta$, $\widehat{\mathcal{H}}_\theta$ is a self-adjoint linear operator on the Hilbert space of functions m that are defined on $[-c, c]$ with norm $\|m\|_2^2 = \int_{-c}^{c} m(x)^2 \times \widehat{p}_0(x)\,dx$.

Suppose that the sequence $\{\widehat{\sigma}_t^2, t = 1, \dots, T\}$ and θ are given. Then define $\widehat{g}_j^a(\cdot)$ to be the local linear smooth of $\widehat{\sigma}_t^{-4}\widetilde{y}_t^2$ on y_{t-j}, let $\widehat{g}_j^b(\cdot)$ be the local linear smooth of $\widehat{\sigma}_t^{-4}$ on y_{t-j}, and let $\widehat{g}_{l,j}^c(\cdot)$ be the bivariate local linear smooth of $\widehat{\sigma}_t^{-4}$ on (y_{t-l}, y_{t-j}). Then define

$$\widehat{\widetilde{m}}_\theta^*(y) = \frac{\sum_{j=1}^{\tau_T} \psi_j(\theta)\widehat{g}_j^a(y)}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta)\widehat{g}_j^b(y)},$$

$$\widehat{\widetilde{\mathcal{H}}}_\theta(x,y) = \frac{-\sum_{j=1}^{\tau_T}\sum_{l=1,l\neq j}^{\tau_T} \psi_j(\theta)\psi_l(\theta)\widehat{g}_{l,j}^c(x,y)\frac{\widehat{p}_{0,l-j}(x,y)}{\widehat{p}_0(x)\widehat{p}_0(y)}}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta)\widehat{g}_j^b(y)}.$$

3.2. *Our Estimators of θ and m*

Here we give a formal definition of our estimators.

STEP 1: Define $\widehat{m}_\theta(\cdot)$ as any sequence of random functions defined on $[-c, c]$ that approximately solves $\widehat{m}_\theta = \widehat{m}_\theta^* + \widehat{\mathcal{H}}_\theta \widehat{m}_\theta$. Specifically, we shall assume that $\widehat{m}_\theta$ is any sequence of functions that satisfies

$$\sup_{\theta\in\Theta, y\in[-c,c]} |(I - \widehat{\mathcal{H}}_\theta)\widehat{m}_\theta(y) - \widehat{m}_\theta^*(y)| = o_p(T^{-1/2}). \tag{19}$$

This step is the most difficult and requires a number of choices. In practice, we solve the integral equation on a finite grid of points, which reduces it to a large linear system.

STEP 2: Choose $\widehat{\theta} \in \Theta$ to be any sequence such that

$$\widehat{S}_T(\widehat{\theta}) \leq \arg\min_{\theta\in\Theta} \widehat{S}_T(\theta) + o_p(T^{-1/2}), \quad \text{where}$$

$$\widehat{S}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \{y_t^2 - \widehat{\sigma}_t^2(\theta)\}^2,$$

where $\widehat{\sigma}_1^2(\theta) = T^{-1}\sum_{t=1}^T y_t^2$ and $\widehat{\sigma}_t^2(\theta) = \max\{\mu_t + \sum_{j=1}^{\min\{t-1,\tau_T\}} \psi_j(\theta)\widehat{m}_\theta(y_{t-j}), \epsilon\}$, $t = 2, \ldots, T$. Here, ϵ is a small nonnegative number introduced to ensure that $\widehat{\sigma}_t^2(\theta) \geq 0$.[5] When θ is scalar this optimization can be done by grid search. Otherwise it may be desirable to use some derivative-based optimization algorithm like Newton–Raphson or its variants, which would require analytical or numerical derivatives of $\widehat{S}_T(\theta)$.

STEP 3: Define for any $y \in [-c, c]$ and $t \geq 2$,

$$\widehat{m}(y) = \widehat{m}_{\widehat{\theta}}(y),$$

[5] Note that in small samples we can find $\widehat{m}_\theta(y) < 0$ for some y, even if $\widehat{m}_\theta^*(y) > 0$ for all y and $\widehat{\mathcal{H}}_\theta(x,y) > 0$ for all x, y. One can replace $\widehat{m}_\theta(y)$ by a trimmed version to ensure its positivity.

$$\widehat{\sigma}_t^2 = \max\left\{\mu_t + \sum_{j=1}^{\min\{t-1,\tau_T\}} \psi_j(\theta)\widehat{m}_\theta(y_{t-j}), \epsilon\right\},$$

and $\widehat{\sigma}_1^2(\theta) = T^{-1}\sum_{t=1}^T y_t^2$. The estimates $(\widehat{m}(\cdot), \widehat{\theta})$ are our proposal for the weak version of our model. For the semistrong and strong version of the model the following updates of the estimate may yield improvements.

STEP 4: Given $(\widehat{\theta}, \widehat{m}(\cdot))$. Compute $\widehat{\widehat{m}}_\theta^*$ and $\widehat{\widehat{\mathcal{H}}}_\theta$ using the sequence $\{\widehat{\sigma}_t^2, t = 1, \dots, T\}$ defined in Step 3. Then solve the linear integral equation

$$\widetilde{m}_\theta = \widehat{\widehat{m}}_\theta^* + \widehat{\widehat{\mathcal{H}}}_\theta \widetilde{m}_\theta \tag{20}$$

for the estimator $\widetilde{m}_\theta$ and let $\widetilde{\sigma}_t^2(\theta) = \max\{\mu_t + \sum_{j=1}^{\tau T}\psi_j(\theta)\widetilde{m}_\theta(y_{t-j}), \epsilon\}$, $t = 2, \dots, T$, for each θ. Define $\widetilde{\theta} \in \Theta$ to be any sequence such that

$$\widetilde{\ell}_T(\widetilde{\theta}) \le \arg\min_{\theta\in\Theta} \widetilde{\ell}_T(\theta) + o_p(T^{-1/2}), \quad \text{where}$$

$$\widetilde{\ell}_T(\theta) = \frac{1}{T}\sum_{t=1}^T \log\widetilde{\sigma}_t^2(\theta) + \frac{y_t^2}{\widetilde{\sigma}_t^2(\theta)}.$$

To avoid a global search we suppose that $\widetilde{\theta}$ is the location of the local minimum of $\widetilde{\ell}_T(\theta)$ with smallest distance to $\widehat{\theta}$. Let $\widetilde{m}(y) = \widetilde{m}_{\widetilde{\theta}}(y)$ and $\widetilde{\sigma}_t^2 = \max\{\mu_t + \sum_{j=1}^{\tau T}\psi_j(\widetilde{\theta})\widetilde{m}(y_{t-j}), \epsilon\}$, $t = 2, \dots, T$.

These calculations may be iterated for numerical improvements. Step 4 can be interpreted as a version of Fisher scoring, discussed in Hastie and Tibshirani (1990, Section 6.2).

3.3. *Solution of Integral Equations*

There are many approaches to computing the solutions of integral equations. Rust (2000) gives a nice discussion about solution methods for a more general class of problems, with emphasis on the high-dimensional state. The two issues are how to approximate the integral in $\widehat{\mathcal{H}}_\theta m$ and how to solve the resulting linear system.

For any integrable function f on $[-c, c]$ define $J(f) = \int_{-c}^c f(t)\,dt$. Let $\{t_{j,n}, j = 1, \dots, n\}$ be some grid of points in $[-c, c]$ and let $w_{j,n}$ be some weights with n a chosen integer. A valid integration rule would satisfy $J_n(f) \to J(f)$ as $n \to \infty$, where $J_n(f) = \sum_{j=1}^n w_{j,n} f(t_{j,n})$. Simpson's rule and Gaussian quadrature both satisfy this for smooth f. Now approximate (19) by

$$\widehat{m}_\theta(x) = \widehat{m}_\theta^*(x) + \sum_{j=1}^n w_{j,n}\widehat{\mathcal{H}}_\theta(x, t_{j,n})\widehat{m}_\theta(t_{j,n})\widehat{p}_0(t_{j,n}). \tag{21}$$

In solvability, this is equivalent to the linear system (Atkinson (1976))

$$\widehat{m}_\theta(t_{i,n}) = \widehat{m}^*_\theta(t_{i,n}) + \sum_{j=1}^{n} w_{j,n}\widehat{\mathcal{H}}_\theta(t_{i,n}, t_{j,n})\widehat{m}_\theta(t_{j,n})\widehat{p}_0(t_{j,n}) \qquad (i = 1, \dots, n). \tag{22}$$

To each solution of (22) there is a unique corresponding solution of (21) with which it agrees at the node points. The solution of the system (22) converges in $L_2(\widehat{p})$ to the solution of (19) as $n \to \infty$, at a rate determined partly by the smoothness of $\widehat{\mathcal{H}}_\theta$. The linear system (22) can be written in matrix notation

$$(I_n - \widehat{\mathbf{H}}_\theta)\widehat{\mathbf{m}}_\theta = \widehat{\mathbf{m}}^*_\theta, \tag{23}$$

where I_n is the $n \times n$ identity, $\widehat{\mathbf{m}}_\theta = (\widehat{m}_\theta(t_{1,n}), \dots, \widehat{m}_\theta(t_{n,n}))^\top$, and $\widehat{\mathbf{m}}^*_\theta = (\widehat{m}^*_\theta(t_{1,n}), \dots, \widehat{m}^*_\theta(t_{n,n}))^\top$, while

$$\widehat{\mathbf{H}}_\theta = -\left[w_{j,n} \sum_{\ell=\pm 1}^{\pm\tau_T} \psi^*_\ell(\theta) \frac{\widehat{p}_{0,\ell}(t_{i,n}, t_{j,n})}{\widehat{p}_0(t_{i,n})} \right]_{i,j=1}^{n}$$

is an $n \times n$ matrix. We then find the solution values $\widehat{\mathbf{m}}_\theta = (\widehat{m}_\theta(t_{1,n}), \dots, \widehat{m}_\theta(t_{n,n}))^\top$ to this system (23). Note that once we have found $\widehat{m}_\theta(t_{j,n})$, $j = 1, \dots, n$, we can substitute back into (21) to obtain $\widehat{m}_\theta(x)$ for any $x \in [-c, c]$, which is called Nyström interpolation. More sophisticated methods also involve adaptive selection of the grid size n and the weighting scheme $\{w_{j,n}, t_{j,n}\}$.

There are two main classes of methods for solving large linear systems: direct methods, including Cholesky decomposition or straight inversion, and iterative methods. Direct methods work fine so long as n is only moderate, say up to $n = 1000$; we have used direct computation of $\widehat{\mathbf{m}}_\theta = (I_n - \widehat{\mathbf{H}}_\theta)^{-1}\widehat{\mathbf{m}}^*_\theta$ in our numerical work below. For larger grid sizes, iterative methods are indispensable. In Linton and Mammen (2003) we describe various iterative approaches.

4. ASYMPTOTIC PROPERTIES

4.1. *Regularity Conditions*

We will discuss properties of the estimates $\widehat{m}_\theta$ and $\widehat{\theta}$ first under the *weak* form model where we do not assume that (4) holds but where θ_0, m_0 are defined as the minimizers of the least squares criterion function (5). Asymptotics for $\widehat{m} = \widehat{m}_{\widehat{\theta}}$ and for the likelihood corrected estimates $\widetilde{m}$ and $\widetilde{\theta}$ will be discussed under the more restrictive setting that (4) holds. Note that as usual our regularity conditions are not necessary, only sufficient, and our method is expected to work well under more general circumstances.

Define $\eta_{j,t} = y_{t+j}^2 - E(y_{t+j}^2|y_t)$ and $\zeta_{j,t}(\theta) = m_\theta(y_{t+j}) - E[m_\theta(y_{t+j})|y_t]$, and let

$$(24) \qquad \eta_{\theta,t}^1 = \sum_{j=1}^{\infty} \psi_j^\dagger(\theta)\eta_{j,t} \quad \text{and} \quad \eta_{\theta,t}^2 = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta)\zeta_{j,t}(\theta),$$

where $\psi_j^\dagger(\theta)$, $\psi_j^*(\theta)$ were defined below (10). Let $\alpha(k)$ be the strong mixing coefficient of $\{y_t\}$ defined as $\alpha(k) \equiv \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_k^\infty} |P(A\cap B) - P(A)P(B)|$, where $\mathcal{F}_b^a$ is the sigma algebra of events generated by $\{y_t\}_a^b$.

B1: *The process* $\{y_t\}_{t=-\infty}^{\infty}$ *is stationary with absolutely continuous density* p_0, *and* α*-mixing with a mixing coefficient* $\alpha(k)$ *such that for some* $C \geq 0$ *and some large* s_0, $\alpha(k) \leq Ck^{-s_0}$.

B2: *The expectation* $E(|y_t|^{2\rho}) < \infty$ *for some* $\rho > 2$.

B3: *The kernel function is a symmetric probability density function with bounded support such that for some constant* C, $|K(u) - K(v)| \leq C|u - v|$. *Define* $\mu_j(K) = \int u^j K(u)\,du$ *and* $\nu_j(K) = \int u^j K^2(u)\,du$.

B4: *The function* m *together with the densities* (*marginal and joint*) $m(\cdot)$, $p_0(\cdot)$, *and* $p_{0,j}(\cdot)$ *are continuous and twice continuously differentiable over* $[-c, c]$, *and are uniformly bounded. The density* $p_0(\cdot)$ *is bounded away from zero on* $[-c, c]$, *i.e.*, $\inf_{-c\leq w\leq c} p_0(w) > 0$. *Furthermore, for a constant* $c_\sigma > 0$ *we have that a.s.*

$$(25) \qquad \sigma_t^2 > c_\sigma.$$

B5: *The density function* λ *of* $(\eta_{\theta,0}^1, \eta_{\theta,0}^2)$ *is Lipschitz continuous on its domain.*

B6: *The joint densities* $\lambda_{0,j}$, $j = 1, 2, \ldots,$ *of* $((\eta_{\theta,0}^1, \eta_{\theta,0}^2), (\eta_{\theta,j}^1, \eta_{\theta,j}^2))$ *are uniformly bounded.*

B7: *The parameter space* Θ *is a compact subset of* $\mathbb{R}^p$ *and the value* θ_0 *is an interior point of* Θ. *Also*, *Assumption* A2 *holds and for any* $\epsilon > 0$, $\inf_{\|\theta-\theta_0\|>\epsilon} S(\theta, m_\theta) > S(\theta_0, m_{\theta_0})$.

B8: *The truncation sequence* τ_T *satisfies* $\tau_T = C\log T$ *for some constant* C.

B9: *The bandwidth sequence* $h(T)$ *satisfies* $h(T) = \gamma(T)T^{-1/5}$ *with* $\gamma(T)$ *bounded away from zero and infinity.*

B10: *The coefficients satisfy* $\sup_{\theta\in\Theta, k=0,1,2} \|\partial^k\psi_j(\theta)/\partial\theta^k\| \leq C\overline{\psi}^j$ *for some* $\overline{\psi} < 1$ *and some finite constant* C, *while* $\inf_{\theta\in\Theta}\sum_{j=1}^{\infty}\psi_j^2(\theta) > 0$.

The following assumption will be used when we make asymptotics under the assumption of (4).

B11: *The semistrong model assumption* (4) *holds*, *so that the variables* $\eta_t = y_t^2 - \sigma_t^2$ *form a stationary ergodic martingale difference sequence with respect to* $\mathcal{F}_{t-1}$. *Let* $\varepsilon_t = y_t/\sigma_t$ *and* $u_t = (y_t^2 - \sigma_t^2)/\sigma_t^2$, *which are also both stationary ergodic martingale difference sequences.*

Note that B1–B11 imply Assumption A1–A3. Condition B1 is quite weak, although the value of s_0 can be quite large depending on the value of ρ given in B2. Carrasco and Chen (2002) provide some general conditions for a class of strong GARCH$(1,1)$-type processes to be strongly stationary, to have finite ρ moments, and to be exponentially β-mixing (which implies α-mixing); these conditions involve restrictions on the function m_0 and on the distribution of the innovations, in addition to restrictions on the parameters of the process. Masry and Tjøstheim (1995, Lemma 3.1) also provides conditions on finite order but "nonparametric" processes that imply geometric strong mixing.[6] We will make use of the mixing property to apply the exponential inequality of Bosq (1998) and to establish a central limit theorem for $\widehat{m}_\theta$ in the weak form case. In this weak form case we cannot apply martingale limit theory. We need to apply a central limit theorem to (local) averages of the processes $\eta^1_{\theta,t}$ and $\eta^2_{\theta,t}$ defined in (24). These processes need not be mixing but are near epoch dependent processes on the α-mixing bases y_t^2 or $m_\theta(y_t)$ (see Hansen (1991) for discussion) with exponentially declining weights under our conditions on $\psi_j(\theta)$; we apply a central limit theorem (CLT) due to Lu (2001) for such processes.

The moment condition B2 on y_t may appear quite strong: it is common practice now in the parametric literature to not assume any moments for y_t but to make assumptions on the rescaled error $\varepsilon_t = y_t/\sigma_t$; see Lee and Hansen (1994). This is because in many financial data sets there is evidence that the tails preclude fourth moments from existing. Note however that although we assume more than four moments in B2 and in defining (5), the moment conditions (7) and (8) are well defined under only second moments, and so some results like consistency will hold under less moments. Indeed, the results for likelihood-based estimators only require this condition because it provides a consistent initial estimator; if one is willing to assume the existence of a consistent estimator (with some rate) like in Horowitz and Mammen (2002), the distribution theory should follow through without moments on y. Bollerslev (1986) showed that in the strong GARCH$(1,1)$ model with $\varepsilon_t \sim N(0,1)$, it is necessary and sufficient for $E(y_t^4) < \infty$ that $2\gamma^2 + (\gamma+\beta)^2 < 1$. Thus only limited dynamics (β, γ) are consistent with fourth moments in this model. Because we have freed up the shape of m, this problem does not arise in our model. In principle, any value of the dynamic parameter θ is consistent with ρ moments existing provided the tails of m increase only slowly.

Conditions B3 and B4 are quite standard assumptions in the nonparametric regression literature. Under the assumption of (4), the bound (25) follows if we assume that $\inf_{-c\le w\le c} m(w) > -\sup_{t,\theta} \mu_t / \sum_{j=1}^{\infty} \psi_j(\theta)$.

Conditions B5 and B6 are used to apply the central limit theorem of Lu (2001) for NED processes over an α-mixing base.

[6] These include restrictions on the tail of the conditional moments, for example, that $\lim_{\|(y_1,\dots,y_d)\|^2\to\infty} \operatorname{var}(y_t|y_{t-1}=y_1,\dots,y_{t-d}=y_d)/\|(y_1,\dots,y_d)\|^2 \le c < 0$.

In B7 we explicitly assume the identification of the parametric part. We make this high level assumption for three reasons. First, we need identification in the weak ARCH(∞) case, and this seems like a natural assumption to make in view of our definition of the process through (5). Second, we allow the coefficients $\psi_j(\theta)$ to depend on θ in a complicated way. Third, the mapping $\theta \mapsto m_\theta$ may be quite complicated to analyze. Hannan (1973) used high level conditions (cf. his condition (4)) similar to ours. In special parametric ARCH models it has been possible to work from more primitive conditions: see Lee and Hansen (1994) and Lumsdaine (1996) for the GARCH(1, 1) model, and Robinson and Zaffaroni (2002) for a parametric ARCH(∞) model.

The distribution theory for parametric GARCH(1, 1) models has only recently been established. Lumsdaine (1996) established the consistency and asymptotic normality of the quasi-maximum likelihood estimator in a strong form model, while Lee and Hansen (1994) established the same results but for semistrong form case, i.e., they allowed for martingale difference errors. Both authors make use of ergodicity in their consistency proof and martingale central limit theorems in the asymptotic normality. The distribution theory for weak form GARCH processes has not yet been worked out, to our knowledge.

The truncation rate assumed in B8 can be weakened at the expense of more detailed argumentation. In B9 we are anticipating a rate of convergence of $T^{-2/5}$ for $\widehat{m}_\theta$, which is consistent with second-order smoothness on the data distribution. Assumption B10 is used for a variety of arguments; it can be weakened in some cases, but again at some cost. It is consistent with the GARCH case where $\psi_j(\theta) = \theta^{j-1}$ and $\partial^k \psi_j(\theta)/\partial\theta^k = (j-1)\cdots(j-k)\theta^{j-k-1}$.

The assumption we made in Section 2.1 about the fixed truncation c can also be weakened to allow $c = c(T) \to \infty$ as $T \to \infty$, and we discuss this issue below.

4.2. *Properties of* $\widehat{m}_\theta$ *and* $\widehat{\theta}$

We establish the properties of $\widehat{m}_\theta$ for all $\theta \in \Theta$ under the weak form assumption. Specifically, we do not require that (3) holds, but define m_θ as the minimizer of (5) over $\mathcal{M}_c$.

Define the functions $\beta_\theta^j(y)$, $j = 1, 2$, as solutions to the integral equations $\beta_\theta^j = \beta_\theta^{*,j}(y) + \mathcal{H}_\theta \beta_\theta^j$, in which (with $\nabla_2 = (\partial^2/\partial x^2) + 2(\partial^2/\partial x\,\partial y) + \partial^2/\partial y^2$) $\beta_\theta^{*,1}(y) = \frac{\partial^2}{\partial y^2} m_\theta^*(y)$ and

$$\beta_\theta^{*,2}(y) = \sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta) \left\{ E(m_\theta(y_{t+j}) | y_t = y) \frac{p_0''(y)}{p_0(y)} - \int [\nabla_2 p_{0,j}(y, x)] \frac{m_\theta(x)}{p_0(y)}\, dx \right\}.$$

Then define $\mu_\theta(y) = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta) E[m_\theta(y_{t+j})|y_t = y]$ and

$$\omega_\theta(y) = \frac{\nu_0(K)}{p_0(y)}\{\mathrm{var}[\eta_{\theta,t}^1 + \eta_{\theta,t}^2] + \mu_\theta^2(y)\} \quad \text{and}$$

$$b_\theta(y) = \frac{1}{2}\mu_2(K)[\beta_\theta^1(y) + \beta_\theta^2(y)],$$

where $\eta_{\theta,t}^j$, $j = 1, 2$, were defined in (24). We prove the following theorem in the Appendix.

THEOREM 1: *Suppose that* B1–B10 *hold. Then for each* $\theta \in \Theta$ *and* $y \in [-c, c]$,

$$\sqrt{Th}[\widehat{m}_\theta(y) - m_\theta(y) - h^2 b_\theta(y)] \Longrightarrow N(0, \omega_\theta(y)), \tag{26}$$

and $\widehat{m}_\theta(y)$ *and* $\widehat{m}_\theta(y')$ *are asymptotically independent when* $y \neq y'$. *Furthermore*,

$$\sup_{\theta\in\Theta, |y|\le c} |\widehat{m}_\theta(y) - m_\theta(y)| = o_p(T^{-1/4}), \tag{27}$$

$$\sup_{\theta\in\Theta, \tau_T\le t\le T} |\widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta)| = o_p(T^{-1/4}), \tag{28}$$

$$\sup_{\theta\in\Theta, \tau_T\le t\le T} \left| \frac{\partial\widehat{\sigma}_t^2}{\partial\theta}(\theta) - \frac{\partial\sigma_t^2}{\partial\theta}(\theta)\right| = o_p(T^{-1/4}). \tag{29}$$

Both the bias and the variance in this result are quite complicated even though a local linear smoother has been used in estimating g_j.

From Theorem 1 we obtain the properties of $\widehat{\theta}$ by an application of the asymptotic theory for semiparametric profiled estimators; see Severini and Wong (1992) and Newey (1994). This requires a uniform expansion for $\widehat{m}_\theta(y)$ and for the derivatives (with respect to θ) of $\widehat{m}_\theta(y)$.

THEOREM 2: *Suppose that* B1–B10 *hold except that in* B2 *we require* $\rho > 4$. *Then* $\sqrt{T}(\widehat{\theta} - \theta_0) = O_p(1)$.

In Theorem 2 we require stronger moment conditions for the root-T consistency of $\widehat{\theta}$ than for the $\sqrt{Th}$ consistency of $\widehat{m}_\theta(y)$. By using the quasi-likelihood criterion these moment conditions can be reduced to $\rho > 2$. These results can be applied to get the asymptotic distribution of $\widehat{m} = \widehat{m}_{\widehat{\theta}}$. Define

$$\omega(y) = \frac{\nu_0(K)\sum_{j=1}^{\infty}\psi_j^2(\theta_0)E[(y_t^2 - \sigma_t^2)^2|y_{t-j} = y]}{p_0(y)[\sum_{j=1}^{\infty}\psi_j^2(\theta_0)]^2}, \tag{30}$$

$$b(y) = \mu_2(K)\left\{\frac{1}{2}m''(y) + (I - \mathcal{H}_\theta)^{-1}\left[\frac{p_0'}{p_0}\frac{\partial}{\partial y}(\mathcal{H}_\theta m)\right](y)\right\}.$$

THEOREM 3: *Suppose that* B1–B10 *hold and that* $\widehat{\theta}$ *is an arbitrary estimate* (*possibly different from the above definition*) *with* $\sqrt{T}(\widehat{\theta}-\theta_0)=O_p(1)$. *Then for* $y \in [-c, c]$,

$$\sqrt{Th}[\widehat{m}_{\widehat{\theta}}(y)-\widehat{m}_{\theta_0}(y)]=o_p(1) \tag{31}$$

and $\widehat{m}_{\widehat{\theta}}(y)$ *and* $\widehat{m}_{\widehat{\theta}}(y')$ *are asymptotically independent when* $y \neq y'$. *Under the additional assumption of* B11 *we get that*

$$\sqrt{Th}[\widehat{m}_{\widehat{\theta}}(y)-m_{\theta_0}(y)-h^2b(y)] \Longrightarrow N(0,\omega(y)). \tag{32}$$

The asymptotic variance has contributions from the estimation of m^* and from the estimation of $\mathcal{H}_\theta$ which combine to give a nice simple formula. The bias of $\widehat{m}$ is rather complicated and it contains a term that depends on the density p_0 of y_t. We now introduce a modification of $\widehat{m}$ that has a simpler bias expansion. For $\theta \in \Theta$ the modified estimate $\widehat{m}_\theta^{\mathrm{mod}}$ is defined as any (approximate) solution of $\widehat{m}_\theta^{\mathrm{mod}}=\widehat{m}_\theta^*+\widehat{\mathcal{H}}_\theta^{\mathrm{mod}}\widehat{m}_\theta^{\mathrm{mod}}$, where the operator $\widehat{\mathcal{H}}^{\mathrm{mod}}$ is defined by use of modified kernel density estimates

$$\widehat{\mathcal{H}}_\theta^{\mathrm{mod}}(y,x)=-\sum_{j=\pm1}^{\pm\tau_T}\psi_j^*(\theta)\frac{\widehat{p}_{0,j}^{\mathrm{mod}}(y,x)}{\widehat{p}_0^{\mathrm{mod}}(y)\widehat{p}_0(x)},$$

$$\widehat{p}_{0,j}^{\mathrm{mod}}(y,x)=\widehat{p}_{0,j}(x,y) + \frac{\widehat{p}_0'(x)}{\widehat{p}_0(x)}\frac{1}{T-|j|}\sum_t (y_t-y)K_h(y_t-y)K_h(y_{t+j}-x),$$

$$\widehat{p}_0^{\mathrm{mod}}(x)=\widehat{p}_0(x)+\frac{\widehat{p}_0'(x)}{\widehat{p}_0(x)}\frac{1}{T}\sum_{t=1}^{T}(y_t-y)K_h(y_t-y).$$

In the definition of the modified kernel density estimates $\widehat{p}_0'$ could be replaced by another estimate of the derivative of p_0 that is uniformly consistent on $[-c, c]$, e.g., $T^{-1}\sum_{t=1}^{T}(y_t-y)K_h(y_t-y)/[h^2\mu_2(K)]$. The asymptotic distribution of the modified estimate is stated in the next theorem.

THEOREM 4: *Suppose that* B1–B11 *hold and that* $\widehat{\theta}$ *is an estimate as in Theorem* 3. *Then for* $y \in [-c, c]$, $\sqrt{Th}[\widehat{m}_{\widehat{\theta}}^{\mathrm{mod}}(y)-m_{\theta_0}(y)-h^2b^{\mathrm{mod}}(y)] \Longrightarrow N(0,\omega(y))$, *where* $\omega(y)$ *is defined as in Theorem* 3 *and where* $b^{\mathrm{mod}}(y)=\mu_2(K)m''(y)/2$.

This bias has a particularly appealing form since it is the bias that would result were $m(\cdot)$ a one-dimensional regression function and the estimator a local linear kernel smoother. Hence, this estimator is design adaptive (Fan (1992)).

4.3. *Properties of $\widetilde{m}$ and $\widetilde{\theta}$*

We now assume that $\widehat{\theta}$ is consistent and so we can confine ourselves to working in a small neighborhood of θ_0, and our results will be stated only for such θ. We shall now assume that (4) holds, so that the variables $\eta_t = y_t^2 - \sigma_t^2$ form a martingale difference sequence with respect to $\mathcal{F}_{t-1}$. Let $\varepsilon_t = y_t/\sigma_t$ and $u_t = (y_t^2 - \sigma_t^2)/\sigma_t^2 = \varepsilon_t^2 - 1$, which are also both martingale difference sequences by assumption.

Define

$$\omega^{\mathrm{eff}}(y) = \frac{1}{p_0(y)} \frac{\nu_0(K)\sum_{j=1}^{\infty}\psi_j^2(\theta_0)E(\sigma_t^{-4}u_t^2|y_{t-j}=y)}{[\sum_{j=1}^{\infty}\psi_j^2(\theta_0)E(\sigma_t^{-4}|y_{t-j}=y)]^2}.$$

Note that $\omega^{\mathrm{eff}}(y)$ can exist even when the fourth moments of y_t do not exist.

THEOREM 5: *Suppose that* B1–B11 *hold. Then, for some bounded continuous function* $b^{\mathrm{eff}}(y)$ *we have* $\sqrt{Th}[\widetilde{m}_{\widehat{\theta}}(y) - m_{\widehat{\theta}}(y) - h^2 b^{\mathrm{eff}}(y)] \Longrightarrow N(0, \omega^{\mathrm{eff}}(y))$.

The next theorem gives the asymptotic distribution of $\widetilde{\theta}$. Define the "least favorable" process $\overline{\sigma}_t^2(\theta) = \mu + \sum_{j=1}^{\infty}\psi_j(\theta)\overline{m}_\theta(y_{t-j})$, where $\overline{m}_\theta(\cdot)$ was defined below (18). Define also

$$\mathcal{J} = E\left(\sigma_t^{-4}\frac{\partial\overline{\sigma}_t^2}{\partial\theta}\frac{\partial\overline{\sigma}_t^2}{\partial\theta^{\top}}(\theta_0)\right) \quad \text{and} \quad \mathcal{I} = \mathrm{var}\left[\sigma_t^{-2}u_t\frac{\partial\overline{\sigma}_t^2}{\partial\theta}(\theta_0)\right].$$

THEOREM 6: *Suppose that* B1–B11 *hold. Then* $\sqrt{T}(\widetilde{\theta} - \theta_0) \Longrightarrow N(0, \mathcal{J}^{-1} \times \mathcal{I}\mathcal{J}^{-1})$.

The result permits inference robust to higher-order moment variation and distributional shape. Consistent standard errors can be obtained by the formula

$$\widehat{\mathcal{J}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\sigma}_t^{-4}\frac{\partial\widehat{\sigma}_t^2}{\partial\theta}\frac{\partial\widehat{\sigma}_t^2}{\partial\theta^{\top}}(\widehat{\theta}) \quad \text{and} \quad \widehat{\mathcal{I}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\sigma}_t^{-4}\frac{\partial\widehat{\sigma}_t^2}{\partial\theta}\frac{\partial\widehat{\sigma}_t^2}{\partial\theta^{\top}}(\widehat{\theta})\widehat{u}_t^2,$$

where hats denote estimated quantities. We show in the next section that when the rescaled errors are Gaussian, the semiparametric efficiency bound for θ is $2\mathcal{J}^{-1}$, and that our estimator achieves this bound.

4.4. *Semiparametric Efficiency*

We next investigate the semiparametric efficiency question, confining our attention to the *strong form* model where ε_t is i.i.d. and in fact standard normal. Our approach to this is heuristic, but is founded on the work of

Bickel, Klaassen, Ritov, and Wellner (1993) and Newey (1990) for i.i.d. data. There has been some previous work on semiparametric efficiency in related semiparametric ARCH models. Engle and González-Rivera (1991) considered a semiparametric model with a standard GARCH$(1,1)$ specification for the conditional variance, but allowed the error distribution to be of unknown functional form. They suggested a semiparametric estimator of the variance parameters based on splines. Linton (1993) examined the Engle and González-Rivera (1991) model and proved that a kernel version of their procedure was semiparametrically efficient and even adaptive in the ARCH(p) model when the error distribution was symmetric about zero. Drost and Klaassen (1997) extended this work to consider GARCH structures and asymmetric distributions: they compute the semiparametric efficiency bound for a general class of models.

We will represent our semiparametric model by $\mathbf{P}_{\theta,m} = \{P_{\theta,m}\}$, where $P_{\theta,m}$ is the probability distribution of the process with parameters $\theta, m(\cdot)$. Now suppose that m is a known function but θ is unknown, in which case we have a specific parametric model, denoted $\mathbf{P}_\theta = \{P_\theta\}$, where $\mathbf{P}_\theta \subset \mathbf{P}_{\theta,m}$. The log-likelihood function is proportional to $\ell_T(\theta) = \frac{1}{2}\sum_{t=1}^T \log s_t^2(\theta) + y_t^2/s_t^2(\theta)$, where $s_t^2(\theta) = \sum_{j=1}^\infty \psi_j(\theta) m(y_{t-j})$. The score function with respect to θ is

$$\begin{aligned}\frac{\partial \ell_T(\theta)}{\partial\theta} &= -\frac{1}{2}\sum_{t=1}^T u_t(\theta)\frac{\partial \log s_t^2(\theta)}{\partial\theta}\\ &= -\frac{1}{2}\sum_{t=1}^T u_t(\theta)\frac{1}{s_t^2(\theta)}\sum_{j=1}^\infty \dot{\psi}_j(\theta) m(y_{t-j}),\end{aligned}$$

where $u_t(\theta) = (y_t^2/s_t^2(\theta) - 1)$ and $\dot{\psi}_j(\theta) = \partial\psi_j(\theta)/\partial\theta$. The Cramer–Rao lower bound in the model P_θ is then $\mathcal{I}_{\theta\theta}^{-1} = 2(E[[\frac{\partial \log \sigma_t^2}{\partial\theta}\frac{\partial \log \sigma_t^2}{\partial\theta^\top}]])^{-1}$, since $E(u_t^2) = 2$.

Suppose that we parameterize m by a scalar η and write m_η, so that we have a parametric model $\mathbf{P}_{\theta,\eta} = \{P_{\theta,\eta}\}$, where $\mathbf{P}_{\theta,\eta} \subset \mathbf{P}_{\theta,m}$. For simplicity we just assume temporarily that θ is also a scalar. The score with respect to η is

$$\begin{aligned}\frac{\partial \ell_T(\theta,\eta)}{\partial\eta} &= -\frac{1}{2}\sum_{t=1}^T u_t(\theta,\eta)\frac{\partial \log \sigma_t^2(\theta,\eta)}{\partial\eta}\\ &= -\frac{1}{2}\sum_{t=1}^T u_t(\theta,\eta)\frac{1}{\sigma_t^2(\theta,\eta)}\sum_{j=1}^\infty \psi_j(\theta)\frac{\partial m_\eta(y_{t-j})}{\partial\eta}.\end{aligned}$$

The efficient score function $\partial\ell_T^*(\theta,\eta)/\partial\theta$ is the projection of $\partial\ell_T(\theta,\eta)/\partial\theta$ onto the orthocomplement of span$[\partial\ell_T(\theta,\eta)/\partial\eta]$ in $\mathbf{P}_{\theta,\eta}$, where span$[\cdot]$ denotes the linear subspace generated by the given element. It follows that $\partial\ell_T^*(\theta,\eta)/\partial\theta$ is a linear combination of $\partial\ell_T(\theta,\eta)/\partial\theta$ and $\partial\ell_T(\theta,\eta)/\partial\eta$

and has variance (called the efficient information) less than the variance of $\partial \ell_T(\theta, \eta)/\partial\theta$; this reflects the cost of estimating the nuisance parameter.

Now consider the semiparametric model $\mathbf{P}_{\theta,m}$. We compute the efficient score functions for all such parameterizations of m, and find the worst such case. Because of the definition of the process σ_t^2, the set of all possible score functions with respect to parameters of m at the true parameters θ_0 is

$$\mathcal{S}_m = \left\{ \sum_{t=1}^{T} u_t \frac{1}{\sigma_t^2} \sum_{j=1}^{\infty} \psi_j(\theta_0) g(y_{t-j}) : g \text{ measurable} \right\}.$$

To find the efficient score function in the semiparametric model, we find the projection of $\partial \ell_T(\theta_0, m)/\partial\theta$ onto the orthocomplement of $\mathcal{S}_m$. We seek a function g_0 that minimizes

$$E\left[\left\{ \frac{\partial \log s_t^2}{\partial\theta} - \frac{1}{s_t^2} \sum_{j=1}^{\infty} \psi_j(\theta_0) g(y_{t-j}) \right\}^2 \right] \tag{33}$$

over all measurable g. This minimization problem is similar to that which m_{θ_0} solves. We show that g_0 satisfies the linear integral equation (see the Appendix for details)

$$g_0 = g^* + \overline{\mathcal{H}}_{\theta_0} g_0, \tag{34}$$

where the operator $\overline{\mathcal{H}}_\theta$ was defined below (18), while

$$g^*(y) = \sum_{j=1}^{\infty} \psi_j(\theta_0) E\left[\frac{1}{s_t^4} \frac{\partial s_t^2}{\partial\theta} \middle| y_{t-j} = y \right] \Big/ \sum_{j=1}^{\infty} \psi_j^2(\theta_0) E[s_t^{-4} | y_{t-j} = y].$$

Note that the integral equation (34) is similar to (18) except that the intercept function g^* is different from $m_{\theta_0}^*$; it has solution $g_0 = (I - \overline{\mathcal{H}}_{\theta_0})^{-1} g^*$. The implied predictor of $\partial \log s_t^2/\partial\theta$ in (33) is $s_t^{-2} \sum_{j=1}^{\infty} \psi_j(\theta_0) g_0(y_{t-j})$, which we denote by $E_m(\partial \log s_t^2/\partial\theta)$. The efficient score function in the semiparametric model is thus

$$\begin{aligned} \frac{\partial \ell_T^*(\theta_0, m)}{\partial\theta} &= \frac{1}{2} \sum_{t=1}^{T} u_t \left[\frac{\partial \log s_t^2}{\partial\theta} - E_m\left(\frac{\partial \log s_t^2}{\partial\theta} \right) \right] \\ &= \frac{1}{2} \sum_{t=1}^{T} u_t \frac{1}{s_t^2} \sum_{j=1}^{\infty} \big[\dot{\psi}_j(\theta_0)(I - \overline{\mathcal{H}}_{\theta_0})^{-1} m_{\theta_0}^* \\ &\qquad - \psi_j(\theta_0)(I - \overline{\mathcal{H}}_{\theta_0})^{-1} g^* \big](y_{t-j}) \end{aligned}$$

$$= \frac{1}{2}\sum_{t=1}^{T} u_t \frac{1}{s_t^2} \sum_{j=1}^{\infty} \big[(I - \overline{\mathcal{H}}_{\theta_0})^{-1} \\ \times \{ \dot{\psi}_j(\theta_0)\overline{m}_{\theta_0}^* - \psi_j(\theta_0) g^* \} \big](y_{t-j}).$$

By construction $\partial \ell_T^*(\theta_0, m)/\partial\theta$ is orthogonal to any element of $\mathcal{S}_m$. The semiparametric efficiency information bound is $\mathcal{I}_{\theta\theta}^* = \mathrm{var}[\partial \ell_T^*(\theta_0, m)/\partial\theta]$. It follows that any regular estimator of θ in this semiparametric model has asymptotic variance not less than $\mathcal{I}_{\theta\theta}^{*-1}$. This bound is clearly larger than in the parametric submodel where m is known. It can be easily checked that $\partial \log \overline{\sigma}_t^2/\partial\theta = \partial \log s_t^2/\partial\theta - E_m(\partial \log s_t^2/\partial\theta)$ from which it follows that our estimator achieves the bound.

An alternative justification for our claims comes from working with the least favorable parametric submodel of $\mathbf{P}_{\theta,m}$, which is $\{P_{\theta,\eta} : m_\eta = m_0 + \eta g_0, \eta \in \mathbb{R}, \theta \in \mathbb{R}^p\}$, where g_0 is defined in (34). For this parametric model the asymptotic (efficient) information for θ is precisely $\mathcal{I}_{\theta\theta}^*$. Since our estimator, which does not use this parametric structure, has the asymptotic variance $\mathcal{I}_{\theta\theta}^{*-1}$, it must be semiparametrically efficient.

We have taken a constructive approach to finding the information bound and we acknowledge that more work is needed to make this rigorous. Perhaps this could be done along the lines of Drost and Klaassen (1997).

4.5. *Nonparametric Efficiency*

Here, we discuss the issue about efficiency of the nonparametric estimators. Our discussion is confined to a special case of the *strong* model. In this case,

$$\omega^{\mathrm{eff}}(y) = \frac{1}{p_0(y)} \frac{(\kappa_4 + 2)\nu_0(K)}{\sum_{j=1}^{\infty} \psi_j^2(\theta_0) E(\sigma_j^{-4} | y_0 = y)},$$

where κ_4 is the excess kurtosis of ε_t.

Our discussion is heuristic and is confined to the comparison of asymptotic variances. This type of analysis has been carried out before in many separable nonparametric models; see Linton (2000). The general idea is to set out a standard of efficiency against which to measure a given procedure along with a strategy for achieving efficiency. Horowitz and Mammen (2002) apply this in generalized additive models. In our model, there are some novel features due to the presence of the infinite number of lags.

Horowitz, Klemelä, and Mammen (2002) establish the minimax superiority of a local linear backfitting estimator in an additive nonparametric regression model.

We first compare the asymptotic variance of $\widehat{m}_{\widehat{\theta}}$ and $\widehat{m}_{\widehat{\theta}}^{\mathrm{mod}}$ with the variance of an infeasible estimator that is based on certain least squares criteria. Define for each $j = 1, 2, \ldots,$

$$S_j(\lambda) = \frac{1}{Th}\sum_t K\left(\frac{y - y_{t-j}}{h}\right)[y_t^2 - \sigma_{t;j}^2(\lambda)]^2, \tag{35}$$

where $\sigma_{t;j}^2(\lambda) = \sum_{k=1,k\neq j}^{\tau_T} \psi_k(\theta)m(y_{t-k}) + \psi_j(\theta)\lambda$, and let $\widetilde{m}_j(y) = \widetilde{\lambda}_j = \arg\max_\lambda S_j(\lambda)$. This least squares estimator is infeasible since it requires knowledge of m at $\{y_{t-k}, k \neq j\}$ points. We suppose without loss of generality that $\psi_j(\theta) > 0$ for each j. It can then be shown that

$$\sqrt{Th}[\widetilde{m}_j(y) - m(y) - h^2 b_j(y)]$$
$$\Longrightarrow N\left(0, \frac{(\kappa_4 + 2)\nu_0(K)E((y_t^2 - \sigma_t^2)^2 | y_{t-j} = y)}{\psi_j^2(\theta)p_0(y)}\right)$$

for all $j = 1, 2, \ldots$ with some bounded continuous bias functions $b_j(\cdot)$. Furthermore, $\widetilde{m}_j(y), \widetilde{m}_k(y)$ with $j \neq k$ are asymptotically independent. Now define a class of estimators $\{\sum_j w_j\widetilde{m}_j : \sum_j w_j = 1\}$, each of which will satisfy a similar central limit theorem. The optimal (according to variance) linear combination of these least squares estimators satisfies

$$\sqrt{Th}[\widetilde{m}_{\mathrm{opt}}(y) - m(y) - h^2 b(y)] \tag{36}$$
$$\Longrightarrow N\left(0, \frac{\nu_0(K)}{p_0(y)\sum_{j=1}^{\infty}\psi_j^2(\theta)[\mathrm{E}((y_t^2 - \sigma_t^2)^2 | y_{t-j} = y)]^{-1}}\right)$$

with some bias function $b(y)$. See Xiao, Linton, Carroll, and Mammen (2003). This is the best that one could do by this strategy; the question is, does our estimator achieve the same efficiency?

Define $s_j(y) = E(\sigma_t^4 u_t^2 | y_{t-j} = y)$. By the Cauchy–Schwarz inequality, $1 = \sum_{j=1}^{\infty}\alpha_j = \sum_{j=1}^{\infty}\alpha_j^{1/2}s_j^{1/2}(y)\alpha_j^{1/2}s_j^{-1/2}(y) \leq \sum_{j=1}^{\infty}\alpha_j s_j(y)\sum_{j=1}^{\infty}\alpha_j s_j^{-1}(y)$, where $\alpha_j = \psi_j^2(\theta)/\sum_{j=1}^{\infty}\psi_j^2(\theta)$, which implies that $\sum_{j=1}^{\infty}\psi_j^2(\theta)s_j(y)/(\sum_{j=1}^{\infty}\psi_j^2(\theta))^2 \geq 1/\sum_{j=1}^{\infty}\psi_j^2(\theta)s_j^{-1}(y)$ with equality only when $s_j(y)$ does not depend on j. So our estimator with variance (30) would achieve the asymptotic efficiency bound (36) in the case of constant conditional variances $s_j(y)$. It is generally inefficient when $s_j(y)$ are not constant. Because our estimator is motivated by an unweighted least squares criterion, it could not be expected that it corrects for heteroscedasticity. We next turn to the likelihood criterion, which takes account of the heteroscedasticity in a natural way.

Define analogously to (35) the (infeasible) local likelihoods

$$\ell_j(\lambda) = \frac{1}{Th}\sum_t K\left(\frac{y - y_{t-j}}{h}\right)\left[\log\sigma_{t;j}^2(\lambda) + \frac{y_t^2}{\sigma_{t;j}^2(\lambda)}\right]$$

and let $\widetilde{m}_j^{\text{lik}}(y) = \widetilde{\lambda}_j = \arg\max_\lambda \ell_j(\lambda)$. It can be shown that

$$\sqrt{Th}[\widetilde{m}_j^{\text{lik}}(y) - m(y)] \Longrightarrow N\left(0, \frac{(\kappa_4 + 2)\nu_0(K)}{\psi_j^2(\theta)p_0(y)E(\sigma_t^{-4}|y_{t-j} = y)}\right)$$

for each j, and again $\widetilde{m}_j^{\text{lik}}(y), \widetilde{m}_k^{\text{lik}}(y)$ with $j \neq k$ are asymptotically independent. As before this suggests that any single $\widetilde{m}_j^{\text{lik}}(y)$ is inefficient and can be improved on by taking linear combinations. It can be shown that the optimal linear combination of $\widetilde{m}_j^{\text{lik}}(y)$ has asymptotic variance $(\kappa_4 + 2)\nu_0(K)/(p_0(y\sum_j \psi_j^2(\theta)E(\sigma_t^{-4}|y_{t-j} = y))$. This is precisely the variance achieved by our estimator $\widetilde{m}_{\widehat{\theta}}(y)$. In other words, our likelihood-based estimator $\widetilde{m}_{\widehat{\theta}}(y)$ appears to be as efficient as it can be, at least under Gaussianity.

5. MODELING THE TAILS

In this section we discuss how to select c and μ_t. A simple method is just to set c at some quantile of the empirical distribution of the data and let $\mu_t = 0$. This works well when c is taken pretty large and when the tails are not so influential. This is the sort of trimming that one finds in much work in econometrics. It may, however, be preferable in some cases to allow for a more sophisticated tail model. We propose below some more refined methods and then give theoretical results about one of them.

5.1. *Estimation Method*

We consider fits of the news impact curve that are of the following form. For $|y| \leq c$ the fit is a nonparametric smoother $\widehat{m}$ and for the tails $|y| > c$ it is chosen as a parametric fit $\mu(y; \widehat{\xi})$. Here $\mu(y; \xi)$ is a parametric model depending on a vector of unknown parameters ξ. We also write $\mu^c(y; \xi)$ for the function that is equal to $\mu(y; \xi)$ for $|y| > c$ and vanishes for $|y| \leq c$. Then we have the estimate of the volatility,

$$\widehat{\sigma}_t^2 = \sigma_t^2(\widehat{\theta}, \widehat{\xi}, \widehat{m}),$$

where $\sigma_t^2(\theta, \xi, m) = \sum_{j=1}^{\tau} \psi_j(\theta)[m(y_{t-j}) + \mu^c(y_{t-j}; \xi)]$ with parametric estimates $\widehat{\theta}$ and $\widehat{\xi}$ and a smoothing estimate $\widehat{m}$ that vanishes for $|y| \geq c$. This generalizes our approach where we have chosen $\mu(y; \xi) \equiv 0$. Parametric specifications include $\mu(y; \xi) = \xi_1 + \xi_2 y^2$, which effectively imposes that the news impact curve is quadratic in the tails (as $y \to \pm\infty$). The Engle and Ng (1993) procedure assumed a linear tail (indeed, they assumed piecewise linear everywhere).

The estimation strategy for this case is pretty much the same as in Section 3. For given θ and ξ one can estimate $m_{\theta,\psi}$ on $[-c, c]$ by putting now

$\tilde{y}_t^2 = y_t^2 - \mu_t(\theta, \xi)$. To estimate θ and ξ one maximizes the profiled least squares criterion or the profiled likelihood with respect to this larger parameter vector. In practice one has to choose c. One could treat c as an unknown parameter and try to estimate it or one can select it on a pragmatic basis by setting it to a high empirical quantile.

The resulting estimate $\widehat{m}(y) + \mu^c(y; \widehat{\xi})$ of the news impact curve is discontinuous at the point c. This can be repaired by calculating the estimates for a continuum (i.e., for a large number) of values of c and by taking an average of these estimates. Another possibility would be to calculate the estimate for two values of c and to smoothly change from the first estimate to the second one when going to $\pm\infty$. Some heuristics that support the second modification of our estimate will be given below.

We also consider a slightly different approach to explicit trimming based on variable bandwidths. In this method one computes the standard estimators of Section 3.2, but uses a variable bandwidth $h_T(y)$ in the local polynomial estimators of $g_j(\cdot)$. If $h_T(y) \to \infty$ as $|y| \to c$, then the local polynomial estimates of g_j and hence m^* become global polynomials for all y with $|y| \geq c$. Likewise the estimated operator $\mathcal{H}$ becomes proportional to the identity operator in the tails. Therefore, we can expect the estimated m to be polynomial in the tails. This method therefore achieves a similar objective to the explicit trimming approach we described above, but it has a nice advantage: provided $h_T^{-1}(y)$ is continuous in y, the resulting estimator of m will also be continuous. We use this method in the simulations below.

5.2. *Asymptotic Properties*

In this section we discuss the properties of the trimming-based estimators. We focus on the case that $c \to \infty$. The strategy is to analyze the corresponding population problem for given c and then to let $c \to \infty$. We will discuss this for the weak form specification of our ARCH(∞) model.

We still suppose that y_t is a stationary process. Define for fixed θ the function $m_{\theta,\infty}$ as minimizer of $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})\}^2]$. The best fit in the trimmed model, with fixed c, is given by $m_{\theta,c}(y) + \mu^c(y; \xi_{\theta,c})$, where $m = m_{\theta,c}(y)$ and $\xi = \xi_{\theta,c}$ minimize $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta)[m(y_{t-j}) + \mu^c(y_{t-j}; \xi)]\}^2]$ over all functions m with support $[-c, c]$ and over all parameters ξ. As above it can be checked that for fixed θ the best fit $m_{\theta,c}$ is uniquely determined by the integral equations

$$m_{\theta,c}(y) = m_\theta^*(y) + \int_{|x|\geq c} \mathcal{H}_\theta(y, x)\mu^c(y; \xi_{\theta,c}) p_0(x)\, dx$$
$$+ \int_{-c}^{c} \mathcal{H}_\theta(y, x) m_{\theta,c}(x) p_0(x)\, dx$$

for $|y| \leq c$, if $c < \infty$, and $m_{\theta,\infty}(y) = m_\theta^*(y) + \int_{-\infty}^{\infty} \mathcal{H}_\theta(y, x) m_{\theta,\infty}(x) p_0(x)\, dx$ for all y. The model bias, caused by trimming, is equal to $m_{\theta,\infty}(y) - m_{\theta,c}(y)$. The bias will depend on the choice of c and on how well $m_{\theta,\infty}$ is approximated by $\mu^c(y;\xi_{\theta,c})$ in the tails. The following theorem gives an estimate for the bias. For the theorem we need the following assumptions that are slightly stronger than Assumptions A1–A3.

C1: *It holds that*

$$\sup_{j\neq 0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{p_{0,j}(y,x)^2}{p_0(x)p_0(y)}\,dx\,dy<\infty,$$

$$\sup_{j\neq 0}\int_{-\infty}^{\infty}\int_{|x|\geq c}\frac{p_{0,j}(y,x)^2}{p_0(x)p_0(y)}\,dx\,dy\to 0\quad \textit{for}\quad c\to\infty.$$

C2: *There exist no* $\theta\in\Theta$ *and no function* m *with* $\int_{-\infty}^{\infty}m^2(x)p_0(x)\,dx=1$ *such that* $\sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j})=0$ *with probability* 1.

C3: *It holds that* $\sup_{\theta\in\Theta}\sum_{j\geq 1}\psi_j(\theta)<\infty$, $\inf_{\theta\in\Theta}\sum_{j\geq 1}\psi_j(\theta)^2>0$, *and* $\sup_{\theta,\theta^*:\,\|\theta-\theta^*\|\to 0}\sum_{j\geq 1}|\psi_j(\theta)-\psi_j(\theta^*)|\to 0$.

In particular, conditions C1 and C3 imply that $\sup_{\theta\in\Theta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{H}_\theta(y,x)^2\times p_0(x)p_0(y)\,dx\,dy<\infty$. Note that this is stronger than Assumption A1, where the integral only runs over $[-c,c]$.

THEOREM 7: *Suppose that* C1–C3 *hold. Then for some constants* $C_1,C_2>0$ (*not depending on* c) *it holds that*

$$\int_{-c}^{c}[m_{\theta,c}-m_{\theta,\infty}](x)^2p_0(x)\,dx\leq C_1\Delta_\theta(c)^2,\tag{37}$$

$$|m_{\theta,c}(y)-m_{\theta,\infty}(y)|\leq C_2\rho_\theta(y)\Delta_\theta(c)\quad \textit{for}\quad |y|\leq c\tag{38}$$

with $\Delta_\theta(c)^2=\int_{|x|\geq c}[m_{\theta,\infty}(x)-\mu^c(x;\xi_{\theta,c})]^2p_0(x)\,dx$ *and* $\rho_\theta(y)^2=\int_{-\infty}^{\infty}\mathcal{H}_\theta(y,x)^2p_0(x)\,dx$.

Condition C1 can be checked for transformed Gaussian processes $y_t=G(x_t)$, where x_t is a stationary Gaussian process with variance 1 and autocorrelation function $r(j)$. Then it holds with a constant C that $\rho_\theta(y)^2\leq[(1-\delta)(1+\delta)]^{-1/2}\exp[\delta(1+\delta)^{-1}G^{-1}(y)^2]$ with $\delta=\sup_{j\geq 1}r(j)^2$. Then $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{H}_\theta(y,x)^2\times p_0(x)p_0(y)\,dx\,dy=E[\rho_\theta(y_t)^2]<\infty$. This discussion shows that C1 is fulfilled also for heavy tailed processes. Condition C1 only puts a condition on the dependence structure of the transformed process $x_t=G^{-1}(y_t)$. It requires that the conditional correlation of x_t and x_s (given $|x_t|$ is larger than c) is bounded away from 1 or does not converge to 1 too fast (for $c\to\infty$).

In this example it holds that $\rho_\theta(c)\to\infty$ for $c\to\infty$. Thus Theorem 7 does not imply that $m_{\theta,c}(c)$ approximates $m_{\theta,\infty}(c)$ if $\Delta_\theta(c)$ converges to zero too slowly. This suggests to use $m_{\theta,c}(y)$ as an approximation of $m_{\theta,\infty}(y)$ only for $|y| \ll c$. For fixed y, $m_{\theta,\infty}(y)$ can be always approximated by $m_{\theta,c}(y)$ with $c \to \infty$. This heuristics also supports the use of the second proposal of smooth trimming that we had discussed above.

We conjecture that condition C1 also holds for (strong form) GARCH(1, 1) processes. This conjecture is supported by Theorem 2.3 of Mikosch and Stărică (2000). This theorem gives expansions for tail probabilities of (y_t, y_{t+j}) and of y_t, and it suggests the approximations $p_{0,j}(y, x) \approx (x^2 + y^2)^{-\kappa-2} f[(x, y)(x^2 + y^2)^{-1/2}]$ and $p_0(x) \approx Cx^{-\kappa-1}$ for $|x|$ and $|y|$ large. Here, C is a constant and f is a function on the sphere. The constant κ is determined by the equation $E(\beta + \gamma\varepsilon_t^2)^{\kappa/2} = 1$. Plugging these approximations into the integral $\int\int p_{0,j}^2(x, y) p_0(x)^{-1} p_0(y)^{-1}\, dx\, dy$ results in a finite integral. This suggests that C1 holds. For $\rho_\theta^2(c)$ we get an approximation that is of order c^{-2}. It follows that in this case $\rho_\theta^2(c) \to 0$ for $c \to \infty$.

We next provide results for a linear parametrization $\mu(y, \xi) = \xi^\top \nu(y)$, where ν is a vector of known functions. For doing so we need slightly stronger conditions. In particular, smoothness conditions for the densities and regression functions have to been stated for the whole real line.

C4: *The trimming threshold $c = c_T$ converges to ∞ for $T \to \infty$ with $c_T \le CT^\gamma$ for constants $C, \gamma > 0$.*

C5: *It holds that $E|y_t|^{2\rho} < \infty$ for a constant $\rho > 5/2$.*

Condition C5 is slightly stronger than B2. Note that in the following theorem we show a faster uniform rate of convergence.

C6: *The function m together with the densities p_0 and $p_{0,j}$ are twice differentiable on $(-\infty, \infty)$. The functions p_0 and $p_{0,j}$ and their derivatives are uniformly bounded on $(-\infty, \infty)$. For p_0 it holds that $p_0(x) \ge C'T^{-\gamma'}$ for $|x| \le c_T$ for some positive constants C', γ'. The function $m(x)$ and its derivatives are bounded for $|x| \le c_T$ by $C''T^{\gamma''}$ for some positive constants C'', γ''. Furthermore, for a constant $c_w > 0$ we have that $\sigma_t^2 \ge c_w$ (a.s.).*

This condition replaces the old condition B4.

C7: *It holds that $E|\nu_j(x)|^\rho < \infty$ for a constant $\rho > 5/2$. Furthermore it holds that the minimal eigenvalue of the matrix $\int_{|x|\ge c} \nu(x)\nu(x)^\top p_0(x)\, dx$ is larger than $C'''T^{-\gamma'''}$ for $C''', \gamma''' > 0$.*

Conditions C4–C7 hold for the strong GARCH(1, 1) process under some conditions on the dynamic parameters and the moments of the innovation. We now state our result that gives similar results as Theorem 1. We do not consider asymptotic normality at a fixed point because now we are estimating a function in a growing interval.

THEOREM 8: *Choose* $\delta > 0$ *and suppose that* C1–C7, B1, B3, *and* B5–B10 *hold with* $\gamma, \dots, \gamma''' > 0$ *small enough* (*depending on* δ). *Then there exist estimators* $\widehat{m}_{\theta,c}(\cdot)$, $\widehat{\xi}_{\theta,c}$ *such that* $\sup_{\theta\in\Theta,|y|\le c_T} |\widehat{m}_{\theta,c}(y) - m_{\theta,c}(y)| = O_P(T^{-2/5+\delta})$ *and* $\sup_{\theta\in\Theta} \|\widehat{\xi}_{\theta,c} - \xi_{\theta,c}\| = O_P(T^{-2/5+\delta})$.

As above, we define θ_0 as the parameter that minimizes $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) \times m_{\theta,\infty}(y_{t-j})\}^2]$. Then θ_0 and $m_0 = m_{\theta_0}$ minimize $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})\}^2]$. From Theorems 7 and 8 we get the following result about the accuracy in estimating the target function m_0.

THEOREM 9: *Make the assumptions of Theorem* 8 *and assume additionally that* $\Delta_{\theta_0}(c_T) = O(T^{-2/5+\delta})$. *Put* $\widehat{m}(y)$ *equal to* $m_{\widehat{\theta},c_T}(y)$ *for* $|y| \le c_T$ *and equal to* $\xi^T_{\widehat{\theta},c_T} \nu(y)$ *for* $|y| \ge c_T$. *Here* $\widehat{\theta}$ *is an estimate with* $\widehat{\theta} - \theta_0 = O_P(T^{-2/5+\delta})$. *Then it holds that* $\int_{-\infty}^{\infty} [\widehat{m}(y) - m_0(y)]^2 p_0(y)\, dy = O_P(T^{-4/5+2\delta})$ *and* $\widehat{m}(y) - m_0(y) = O_P(T^{-2/5+\delta})$ *for a fixed* y.

A possible candidate for the estimate $\widehat{\theta}$ is the profile least squares estimate.

6. NUMERICAL RESULTS

6.1. *Bandwidth Choice and Lag Truncation*

One approach is to choose the bandwidth h to minimize the asymptotic mean squared error of $\widehat{m}$ derived above. This requires estimation of the second derivatives of m and other quantities, so may not work well in practice. Instead we develop a rule of thumb bandwidth using the pointwise mean squared error implied by Theorem 4 when the process is a strong GARCH(1, 1), although we will use this bandwidth more widely. In this case the bias function is just $b^{\mathrm{mod}}(y) = \mu_2(K)\gamma$. In the variance term $\omega(y)$ we replace $E((y_t^2 - \sigma_t^2)^2 | y_{t-j} = y)$ by the unconditional average $\widehat{m}_4 = T^{-1} \sum_{t=1}^{T} (y_t^2 - \widehat{\sigma}_t^2)^2$, where $\widehat{\sigma}_t^2$ are estimated from a preliminary GARCH(1, 1) fit, as are $\widehat{\gamma}$ and $\widehat{\theta}$. It may be desirable to replace $\widehat{m}_4$ by a more robust measure like the median of $(y_t^2 - \widehat{\sigma}_t^2)^2$. Then the pointwise mean squared error optimal bandwidth for the least squares estimator can be approximated by

$$\widehat{h}_{\mathrm{ROT}}(y) = \left[\frac{(1 - \widehat{\theta}^2)\nu_0(K)\widehat{m}_4}{4\mu_2^2(K)\widehat{\gamma}^2 \widehat{p}_0(y)} \right]^{1/5} T^{-1/5}. \tag{39}$$

For the likelihood-based estimator the same formula applies but with $\widehat{m}_4$ replaced by $\widehat{m}_4^*$, where $\widehat{m}_4^* = 1/(T^{-1} \sum_{t=1}^{T} \widehat{\sigma}_t^{-4})$ where $\widehat{\sigma}_t^{-2}$ is the GARCH volatility estimator. These bandwidths are defined on $[-c, c]$. This approach gives moderate increase of bandwidth in the tails; one can magnify the increase in bandwidth by the following method. Replace $\widehat{\gamma}^2$ in (39) by $\widehat{\gamma}^2 \pi(y)$, where π is a function that decreases to zero rapidly after some threshold c_0. Specifically, $\pi(y) = 1$ for all y with $|y| \leq c_0 < c$, while $\pi(y) \to 0$ as $|y| \to c$. Note that as $|y| \to c$, the bandwidth increases to infinity and so the local polynomial estimate of $E(y_t^2|y_{t-j} = y)$ becomes a global polynomial; therefore this estimation strategy forces $\widehat{m}(y)$ to have the same polynomial shape in the tails.

For the truncation parameter τ_T, we have chosen τ to make $\sup_{\theta\in\Theta} \sum_{j=\tau+1}^{\infty} \psi_j(\theta) < \epsilon$ for some small prespecified tolerance level ϵ. One can also use some formal model selection technique but at computational cost.

6.2. *Simulated Data*

We report the results of a small simulation experiment. There are several papers that provide simulation evidence on the finite sample performance of GARCH quasimaximum likelihood (and related) estimators (QMLE). A major issue in these studies is the reliability of the results and their robustness to alternative implementations. This is acknowledged in most of the studies we examined: e.g., Lumsdaine (1995) and Fiorentini, Calzolari, and Pannatoni (1996). Nonlinear estimators in nonconvex optimization problems can have a variety of problems. To some extent this is a problem with the nature of large scale simulations rather than with the estimator itself—when one runs 10,000 replications of a procedure one is restricted to a relatively crude implementation, whereas for a single data set one can modify the procedure as required for that particular sample. However, there are also studies that report finding significantly different results for a given single data set using different commercial software; see Brooks, Burke, and Persand (2001) and McCullough and Renfro (1999).

The focus of our study is on the news impact curve $m(\cdot)$. In an earlier version of this paper, Linton and Mammen (2003), we report results for a design where θ was estimated along with $m(y)$ by choosing θ from a grid of 100 points on $(0, 1)$. The estimates of θ were quite well behaved even for relatively small sample sizes. In parametric applications one often finds estimates of θ to be strongly significant. The results we report here address the issue of how well the nonparametric estimate of the news impact curve m performs in comparison with a parametric method in a situation that is favorable to the parametric method. Specifically, we shall assume that θ is known in both procedures.

We consider two sets of experiments. In the first case (model 1) we generated data from (6), where $y_t = \varepsilon_t\sigma_t$ and ε_t is standard normal, with $\theta = 0.45$, $\gamma = 0.35$, $\alpha = 0.20$. These are the parameter values chosen in Fiorentini et al. (1996). In the second case (model 2) we consider y_t, ε_t as above and $\sigma_t^2 = \theta\sigma_{t-1}^2 + \alpha + \gamma y_{t-1}^2 + \delta y_{t-1}^2 \mathbb{1}(y_{t-1} < 0)$ with $\theta = 0.9$, $\gamma = 0.06$, $\delta = 0.03$ and α as before. For model 1, $E(|y_t|^8) < \infty$ and so both least squares and likelihood estimates of the parameters are consistent and asymptotically normal, while for model 2 we have $E(|y_t|^{4+\epsilon}) < \infty$ for some small $\epsilon > 0$ but $E(|y_t|^8) = \infty$. Although model 1 is far from the sort of model one encounters with daily stock return data, it is not a bad match for standardized monthly data. Model 2 is more realistic for daily data and poses a challenge for the least squares methods because of the approximate violation of our regularity conditions. We consider $T \in \{200, 400, 800\}$.

We investigate both least squares estimators $\widehat{m}(y)$ and likelihood estimators $\widetilde{m}(y)$. In each case the intercept functions were estimated with local constant, local linear, and local quadratic smoothers with a Gaussian kernel. We chose throughout $n = 200$ grid points equally spaced in quantile space.[7] We estimate on the entire sample range of the data,[8] but use the variable bandwidth method (39) with the downweighting described directly afterward with $\pi(y) = \exp(-(|y| - 2)^2)$ for $|y| > 2$. Although the estimates of m sometimes take negative values, we do not trim them.

To compare the performance of the nonparametric estimators we need a benchmark. Our benchmark is the asymptotic variance that would apply to a GARCH maximum likelihood estimator (MLE) (assuming θ is known). This avoids the tricky implementation issues associated with these estimators as discussed above. It has to be noted that this sets a very high standard, since it is an infeasible estimator. The GARCH MLE of the news impact curve is $\widehat{m}_{\text{Lik}}(y) = \widehat{\alpha} + \widehat{\gamma} y^2$, where $(\widehat{\alpha}, \widehat{\gamma})$ are the MLEs of (α, γ). The asymptotic variance of $\widehat{m}_{\text{Lik}}(y)$ is $v_{\text{Lik}}(y) = (\sigma_{\alpha\alpha} + \sigma_{\gamma\gamma} y^4 + 2\sigma_{\alpha\gamma} y^2)/T$, where $\sigma_{\alpha\alpha}$, $\sigma_{\gamma\gamma}$, and $\sigma_{\alpha\gamma}$ are the corresponding asymptotic variances and covariances of the parameter estimates. We compute $v_{\text{Lik}}(y)$ by simulation to three decimal place accuracy.

We present in Table I the bias and standard deviation of the local constant, local linear, and local quadratic implementations of $\widehat{m}(y)$ and $\widetilde{m}(y)$ along with the (asymptotic) MLE at the 1%, 10%, 25%, 50%, 75%, 90%, and 99% quantiles of the distribution of y_t. We summarize the main findings for model 1 as follows: 1. The results for all implementations seem to improve with sample size, with some exceptions regarding the biases in the extreme tails. 2. The performance is much better in the center of the news distribution, but this is also true with the parametric estimator. 3. The MLE, $\widehat{m}_{\text{Lik}}(y)$, performs better according to mean squared error. However, this advantage decreases relatively with sample size, due to the large small sample component in the performance of the nonparametric estimators.[9] 4. The local likelihood estimator $\widetilde{m}$ generally performs much better than the least squares estimator $\widehat{m}$ according to mean squared error, regardless of whether a local constant, local linear, or local quadratic smoother is used, except in the tails where it can perform worse. 5. The local constant implementation generally works better in terms of mean squared error than the local linear or local quadratic implementations of $\widehat{m}$.

[7] That is, the grid points $t_{j,n}$ are chosen to be the j/n sample quantile, where $j = 0, \dots, n$.

[8] This means that we take c to be the maximum value of y_t and $-c$ to be the minimum value of y_t.

[9] Note that the comparable results in Fiorentini et al. (1996) for an implementation of the MLE show slightly worse performance due to small sample issues.

TABLE I

FINDINGS FOR PARAMETER VALUES IN FIORENTINI ET AL. (1996), $\sigma_t^2 = 0.2 + 0.45\sigma_{t-1}^2 + 0.35y_{t-1}^2$[a]

	Quantile	1%		10%		25%		50%		75%		90%		99%	
	n	Bias	Std	Bias	Std	Bias	Std	Bias	Std	Bias	Std	Bias	Std	Bias	Std
$\widehat{m}$															
Quadratic	200	0.598	3.712	0.089	0.363	−0.086	0.210	−0.122	0.213	−0.079	0.212	0.039	0.307	0.462	2.054
	400	0.619	2.524	0.062	0.243	−0.092	0.147	−0.124	0.142	−0.085	0.146	0.025	0.222	0.575	1.559
	800	0.608	1.750	0.029	0.170	−0.091	0.100	−0.117	0.092	−0.090	0.100	0.013	0.168	0.540	1.100
Linear	200	−2.349	3.453	0.152	0.422	−0.013	0.138	−0.057	0.127	−0.012	0.133	0.105	0.357	−0.046	1.749
	400	−1.396	2.213	0.127	0.267	−0.024	0.100	−0.063	0.084	−0.022	0.096	0.104	0.240	0.371	1.288
	800	−0.398	1.435	0.088	0.168	−0.028	0.068	−0.057	0.054	−0.025	0.069	0.087	0.176	0.528	1.011
Constant	200	−3.332	2.955	−0.035	0.222	0.026	0.104	0.042	0.077	0.031	0.096	−0.035	0.194	−1.100	1.221
	400	−2.804	2.004	−0.016	0.159	0.018	0.078	0.025	0.054	0.019	0.077	−0.017	0.149	−0.927	0.782
	800	−2.218	1.325	−0.014	0.113	0.009	0.055	0.014	0.041	0.011	0.058	−0.005	0.118	−0.683	0.540
$\widetilde{m}$															
Quadratic	200	−0.746	20.000	−0.109	0.278	0.025	0.082	0.073	0.055	0.027	0.063	−0.096	0.186	−0.460	1.960
	400	−0.445	3.159	−0.102	0.145	0.015	0.048	0.058	0.036	0.017	0.041	−0.091	0.103	−0.344	1.034
	800	−0.441	3.143	−0.092	0.060	0.009	0.030	0.046	0.025	0.010	0.028	−0.083	0.058	−0.339	0.683
Linear	200	−0.562	12.504	−0.092	0.223	0.034	0.074	0.079	0.052	0.036	0.067	−0.081	0.200	−0.422	2.232
	400	−0.404	3.527	−0.072	0.173	0.032	0.060	0.068	0.036	0.032	0.045	−0.066	0.107	−0.221	1.127
	800	−0.047	1.765	−0.057	0.065	0.028	0.030	0.058	0.025	0.030	0.031	−0.046	0.065	−0.200	0.773
Constant	200	−1.022	10.612	−0.154	0.137	0.029	0.066	0.089	0.056	0.033	0.059	−0.130	0.112	−0.730	1.814
	400	−0.951	3.040	−0.146	0.092	0.029	0.047	0.081	0.039	0.028	0.042	−0.130	0.073	−0.744	0.663
	800	−0.921	1.508	−0.144	0.042	0.023	0.027	0.075	0.026	0.025	0.028	−0.127	0.042	−0.845	0.458
$\widehat{m}_{\text{Lik}}$															
MLE	200		0.534		0.095		0.036		0.048		0.036		0.096		0.539
	400		0.378		0.067		0.025		0.035		0.025		0.068		0.381
	800		0.267		0.047		0.018		0.024		0.018		0.047		0.267

[a] The quantiles are of the distribution of y_t. Bias and std denote bias and standard deviation, respectively. MLE results are taken from the simulated asymptotic distribution, hence there is no bias.

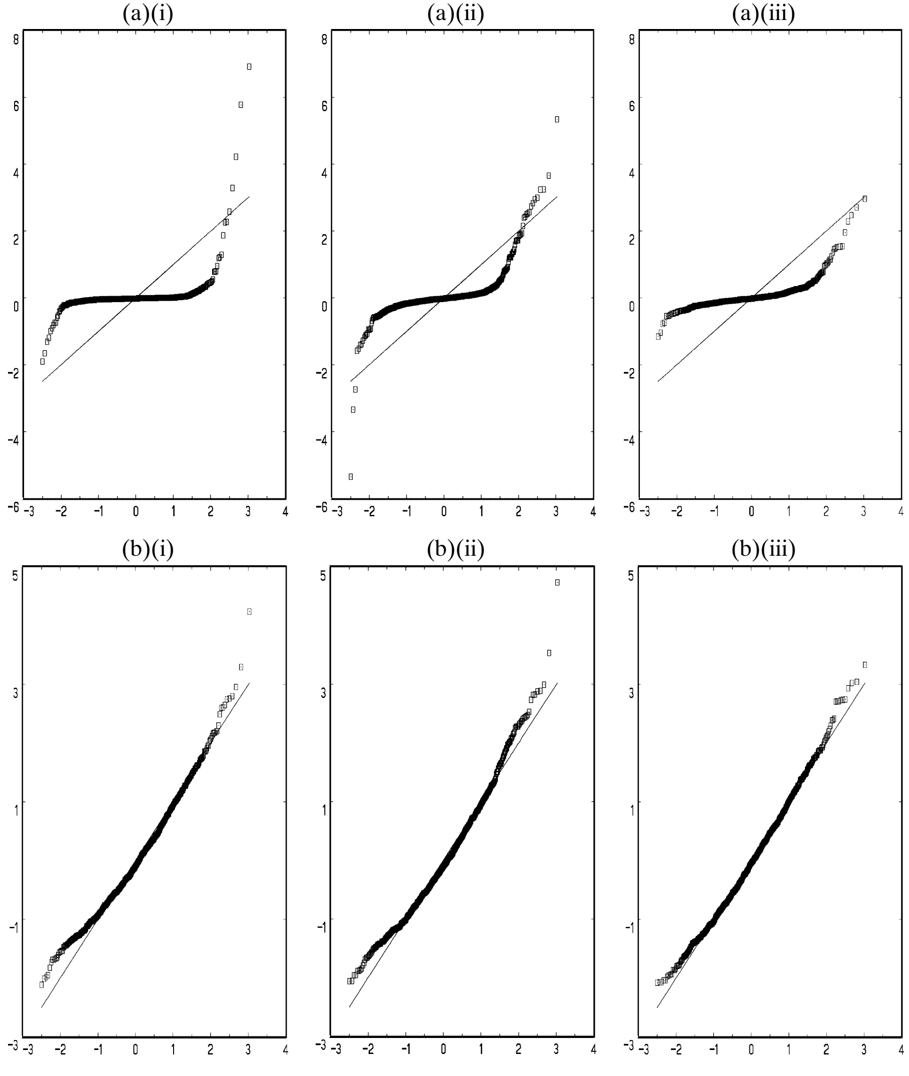
For $\widetilde{m}$, the local quadratic method seems to work best in the center of the distribution, while the local linear method works better in the tails. The local constant method tends to do better in the small sample sizes.

The poor performance in the tails can perhaps be explained by the fact that the population moments of $\widehat{m}$ and $\widetilde{m}$ are not guaranteed to exist; robust estimates of the scale of $\widehat{m}$ and $\widetilde{m}$ give dramatically smaller numbers out in the tail. For example, the local quadratic likelihood estimator at the 1% quantile has for $n = 200$ a standard deviation of 20.00 across simulated samples, but the robust scale estimate of interquartile range/1.35 is only 0.497. The median bias is also somewhat smaller than the mean bias, where this is very large in absolute value. This suggests that the poor performance is driven by a few "rogue" data sets that perhaps require the special treatment that could be given to a unique data set but not across simulations. In Figure 1 we show the q–q plot of the distribution of the centered estimators $\widetilde{m}$ for the 0.01 and 0.50 quantiles. Clearly, in the tails convergence to the normal distribution is slow in comparison with the median.

In the interests of space we report briefly here on our results for model 2, where $\sigma_t^2 = 0.2 + 0.90\sigma_{t-1}^2 + 0.06y_{t-1}^2 + 0.03y_{t-1}^2\mathbb{1}(y_{t-1} < 0)$. The least squares methods then exhibit poor mean squared performance relative to the benchmark—although the standard deviations decrease with sample size they do so slowly and from a high level, while the biases remain large. By contrast the likelihood methods generally perform reasonably well. At the median, the benchmark MLE has asymptotic standard deviations of 0.126, 0.089, and 0.063 for sample sizes 200, 400, and 800. By contrast, the local quadratic likelihood method has standard deviations 0.268, 0.192, and 0.145 with biases 0.118, 0.065, and 0.038. At the 1% quantile, the benchmark MLE has asymptotic standard deviations of 1.737, 1.228, and 0.868, whereas the local constant likelihood method has standard deviations 6.474, 3.761, and 2.727 with biases 2.303, 1.200, and 0.390. The performance of all estimators is better in an absolute sense in the right tail, where the news impact curve is smaller, than in the left tail.

6.3. *Investigation of the News Impact Curve in S&P500 Index Returns: 1955–2002*

We next provide a study of the news impact curve on various stock return series. The purpose here is to discover the relationship between past return shocks and conditional volatility. We investigate a sample of daily returns on the S&P500 from 1955 to 2002, a total of 11,893 observations, a sample of weekly returns with 2464 observations, and a monthly sample with 570 observations. In an earlier version of this paper we concentrated on the daily data, while here we give more results for the weekly estimation. Table II gives the unconditional cumulants: the fourth cumulant is quite large for the daily data and suggests that the fourth moment may not exist, whereas the fourth cumulants and "Hill plots" for the weekly data point to much lighter tails.


(a)(i) (a)(ii) (a)(iii)

(b)(i) (b)(ii) (b)(iii)

FIGURE 1.—The q–q plots of local constant $\widetilde{m}$ from model 1. Panels (a) show results for y quantile equal to 0.01, while (b) show results for y quantile equal to 0.50: (i) corresponds to $n = 200$, (ii) $n = 400$, (iii) $n = 800$.

In Figure 2 we show nonparametric estimates of the first four conditional cumulants, i.e., $E(y_t|y_{t-k})$, $\text{var}(y_t|y_{t-k})$, $\text{skew}(y_t|y_{t-k})$, and $\text{kurt}(y_t|y_{t-k})$ for the weekly data. These are computed using local linear smoothers and a rule of thumb bandwidth from Fan and Gijbels (1996). We show the curves for $k = 1, 2, \ldots, 10$.

TABLE II

CUMULANTS BY FREQUENCY[a]

	Daily	Weekly	Monthly
Mean ($\times 100$)	0.029	0.141	0.606
Std ($\times 100$)	0.038	0.200	0.903
Skewness	−1.546	−0.375	−0.589
Excess kurtosis	43.334	6.521	5.588
Minimum	−25.422	−6.577	−5.984
Maximum	9.623	6.534	3.450

[a] Descriptive statistics for the returns on the S&P500 index for the period 1955–2002 for three different data frequencies. Minimum and maximum are measured in standard deviations and from the mean.

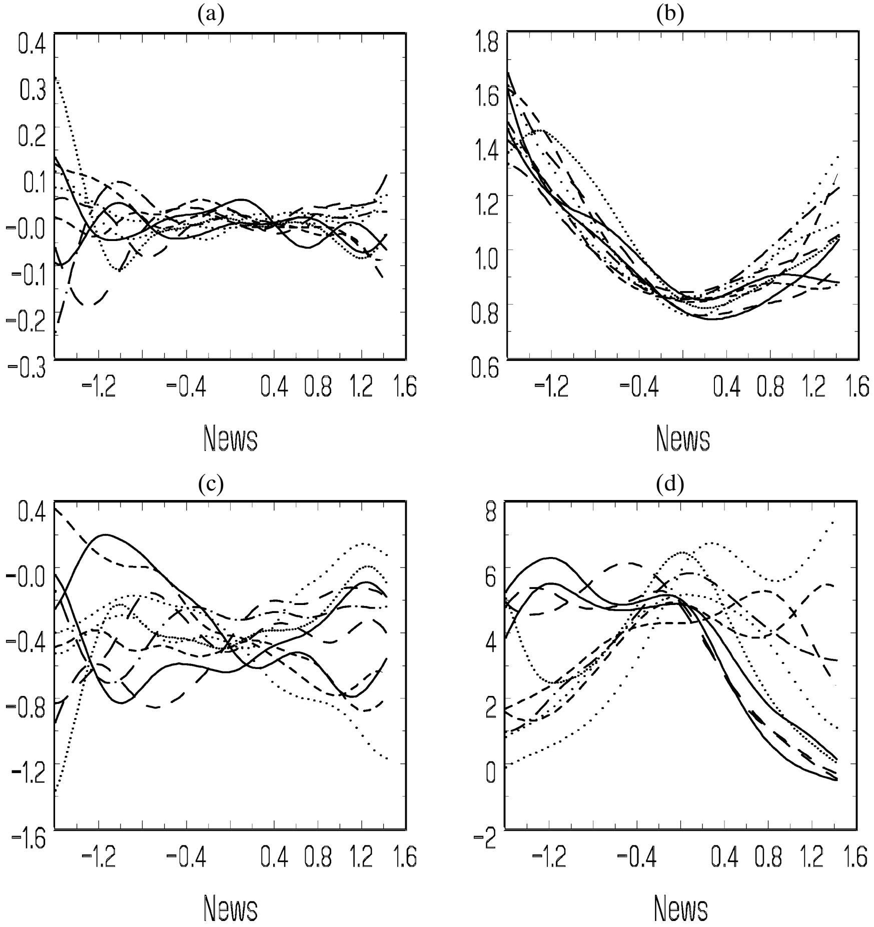


FIGURE 2.—Conditional cumulants of weekly S&P500 returns for lags $k = 1, \ldots, 10$: (a) mean; (b) variance; (c) skewness; (d) kurtosis.

There does not appear to be much common structure in the conditional mean. The conditional variances are significantly different from constants and appear to have similar asymmetric U shapes. The conditional skewness and kurtosis are large in absolute value and have a variety of different shapes, with no common pattern, although the skewness is mostly negative and the kurtosis is mostly positive. Given the heavy tails in the data, these curves may not be significant relative to the sampling variability. Similar patterns are observed in the daily and monthly data. Although there is no necessary relationship between these marginal curves and the corresponding joint cumulants $E(y_t|\mathcal{F}_{t-1})$, $\mathrm{var}(y_t|\mathcal{F}_{t-1})$, $\mathrm{skew}(y_t|\mathcal{F}_{t-1})$, and $\mathrm{kurt}(y_t|\mathcal{F}_{t-1})$, this is suggestive of a common structure similar to what is imposed in our model.

Following Engle and Ng (1993) we fitted regressions on seasonal dummies, but, unlike them, found little significant effects. In Table III we report the results of estimating the Glosten, Jegannathan, and Runkle (1993) model (which we call GJR) parametric fits on these standardized series. All parameters appear significant and there is quite strong evidence of asymmetry at all frequencies.

We next applied our methods. We fitted an AR(2) process to the data and then worked with the standardized residual series. We computed our estimators using $\tau = 50$ for the daily data and $\tau = 25$ for the weekly and monthly data, where the dynamic coefficients were $\psi_j(\theta) = \theta^{j-1}$ with $\theta \in (0, 1)$. We estimated the function m on the entire range of the data using local constant, local linear, and local quadratic smoothers with variable bandwidth selected

TABLE III

PARAMETRIC ESTIMATION[a]

	Daily	Weekly	Monthly
ρ_1	0.138788	0.007065	0.14661
	(0.009524)	(0.022000)	(0.045131)
ρ_2	−0.01906	0.051815	−0.018694
	(0.009449)	(0.022044)	(0.045083)
$\alpha(\times 1{,}000)$	0.0000721	0.00130	0.862000
	(0.0000064)	(0.000242)	(0.249000)
θ	0.920489	0.850348	0.442481
	(0.002243)	(0.015580)	(0.176365)
γ	0.034018	0.047885	−0.076662
	(0.002613)	(0.013504)	(0.042047)
δ	0.078782	0.140013	0.266916
	(0.003302)	(0.020349)	(0.094669)

[a] Standard errors are given in parentheses. These estimates are for the raw data series and refer to the AR(2)–GJR–GARCH(1, 1) model

$$y_t = c + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t \sigma_t,$$

$$\sigma_t^2 = \alpha + \theta\sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 \mathbb{1}(y_{t-1} < 0).$$

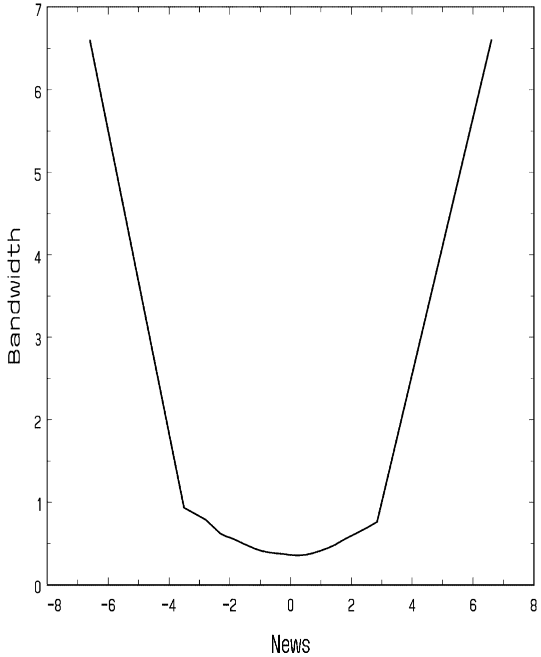


FIGURE 3.—The bandwidth $h_T(y)$ used in the computation of $\widehat{m}_\theta(y)$ for the weekly S&P500 returns data.

by the rule of thumb (39) with the tail modification. Specifically, we chose $\pi(y) = \exp(-(|y| - 3)^2)$ for $|y| > 3$. In Figure 3 we plot the bandwidth used for the computation of $\widehat{m}_\theta$ for weekly data as a function of y. For comparison, the Silverman rule of thumb bandwidth is 0.224, which is smaller than our bandwidth ever is.

In Figure 4 we show our two semiparametric news impact curve estimates, local quadratic $\widehat{m}_\theta$ and $\widetilde{m}_\theta$, along with the parametric alternative $\widehat{m}_{\text{GJR}}$ for the three data frequencies using the GJR dynamic parameters, denoted $\widehat{\theta}_{\text{GJR}}$, which are taken from Table III. Our graphs show the curves on the interval defined by the 0.01–0.99 quantiles along with the standard errors for the three estimates. The main conclusions are the following: 1. There is evidence of asymmetry for daily, weekly, and monthly frequencies. 2. The least squares estimators $\widehat{m}_\theta$ show the greatest growth on the negative side, while typically likelihood estimator $\widetilde{m}_\theta$ is a bit closer to the parametric curve. 3. The minima of the curves in all cases occur on the positive side for each frequency. 4. The monthly $\widehat{m}_{\text{GJR}}$ is monotonic decreasing on this range, which is unexpected.[10] 5. The daily standard errors obey the anticipated ordering: the largest

[10] We have recomputed the parametric estimates using Eviews, Gauss, and Matlab, but in all cases find qualitatively similar results.

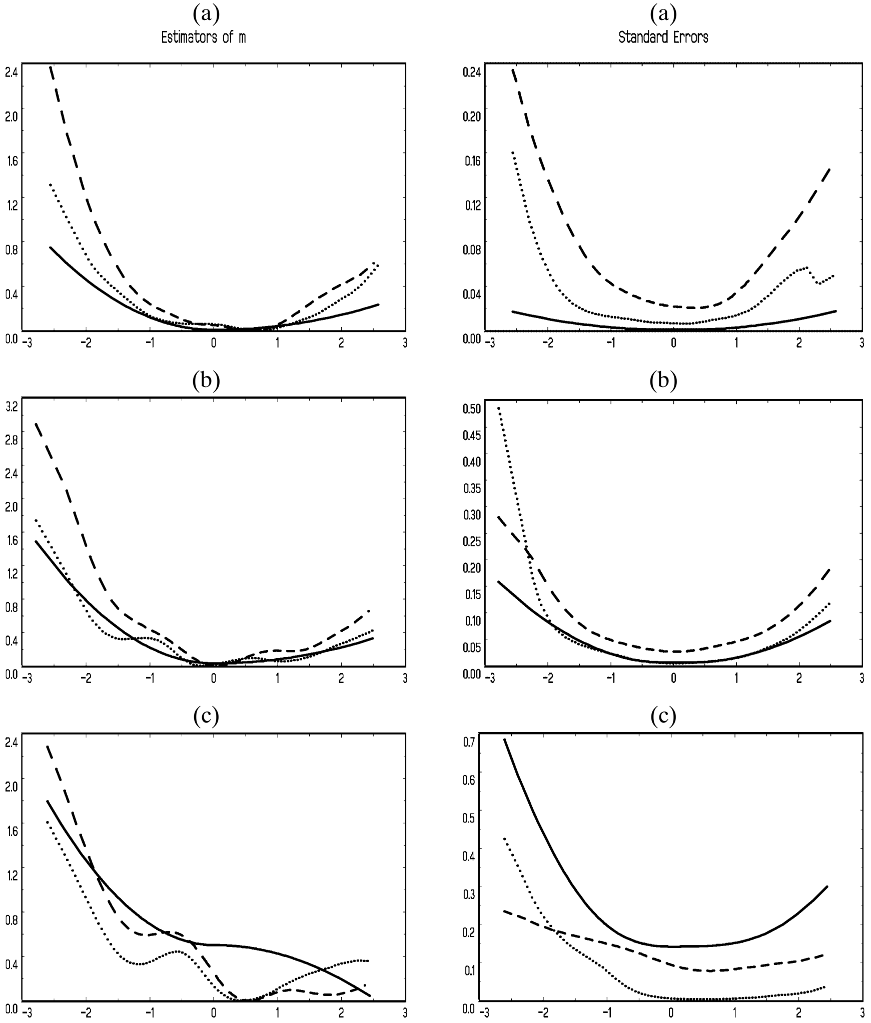


FIGURE 4.—Estimated news impact curves for (a) daily, (b) weekly, and (c) monthly S&P500 returns along with standard errors in right panel. Solid line is $\widehat{m}_{\mathrm{GJR}}$; dashed line is $\widehat{m}_{\widehat{\theta}_{\mathrm{GJR}}}$; dotted line is $\widetilde{m}_{\widehat{\theta}_{\mathrm{GJR}}}$.

are for $\widehat{m}_\theta$ and the smallest are for $\widehat{m}_{\mathrm{GJR}}$. Both semiparametric standard errors increase rapidly when $|y| > 2$. The weekly standard errors follow a similar pattern except that the standard errors for $\widetilde{m}_\theta$ are larger than those for $\widehat{m}_\theta$ when $y < -2.5$. 6. The magnitudes of the standard errors are such that there are significant differences between the news impact curves at various points. 7. The monthly standard errors seem a bit erratic: the parametric ones are too large and those for $\widetilde{m}_\theta$ seem way too small when $y > 0$. 8. The tail part of the estima-

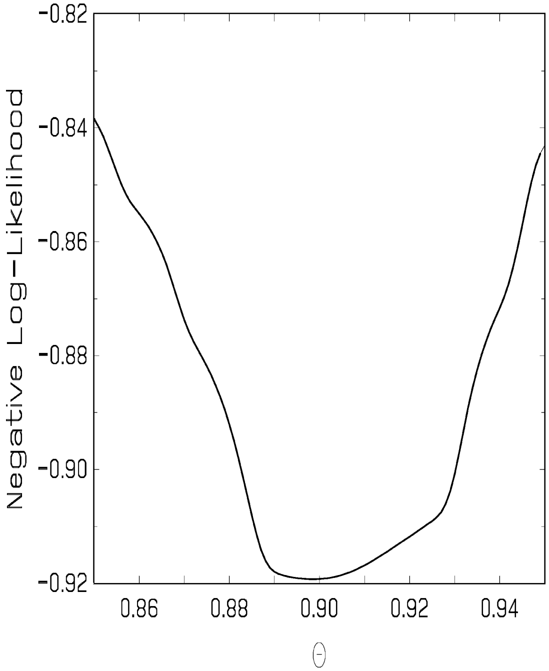


FIGURE 5.—S&P500 weekly data. Negative of log-likelihood as a function of θ.

tion, which is not shown in the graphics, reveals quite substantial differences between $\widehat{m}_{\mathrm{GJR}}$, $\widehat{m}_\theta$, and $\widetilde{m}_\theta$. This is especially so in the daily data where there is a single isolated observation at -25 standard deviations (the 1987 crash) and this forces big differences in the tail functions. Engle and Ng (1993) found similar results.

We next estimated the full semiparametric model on the weekly data. We took the dynamic coefficients to be $\psi_j(\theta) = \theta^{j-1}$, where θ was selected by a grid search on $(0, 1)$ with width 0.001; we computed $\widehat{m}_\theta$ and $\widetilde{m}_\theta$ as described above. In Figure 5 we report the negative likelihood function on the range $[0.85, 0.95]$. The global minimum is at $\widehat{\theta} = 0.899$, which is slightly larger than the value estimated in the GJR QMLE. The likelihood is a bit flat near the minimum and, consequently, the standard error of $\widetilde{\theta}$ is quite large at 0.036, more than twice the standard error of the parametric estimator. The news impact curves are similar in shapes to those reported above and are omitted for brevity.

Finally, we looked at some diagnostics based on the standardized residuals $\widehat{\varepsilon}_t = y_t/\widehat{\sigma}_t$ and $\widetilde{\varepsilon}_t = y_t/\widetilde{\sigma}_t$ from the full semiparametrically estimated model, weekly data. The conditional variances of these series show much less evidence of systematic shapes; the skewnesses and kurtosises are smaller in absolute value. We report in Figure 6 the correlogram for the squared residuals

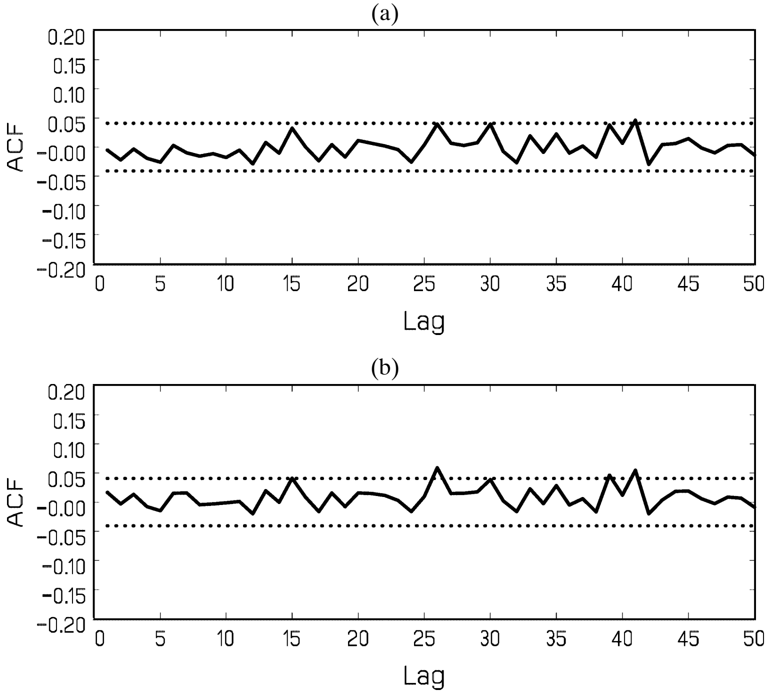


FIGURE 6.—Residual diagnostics of estimated semiparametric model for weekly S&P500 data. The estimated ACF along with 95% Bartlett intervals is shown (a) for $\widehat{\varepsilon}_t^2$ and (b) for $\widetilde{\varepsilon}_t^2$.

$\widehat{\varepsilon}_t^2$ and $\widetilde{\varepsilon}_t^2$ along with the Bartlett interval $\pm 1.96/\sqrt{T}$.[11] There is very little evidence of autocorrelation in either series.

Our application has confirmed some of the findings of Engle and Ng (1993), namely the asymmetric news impact curve, on the S&P500 data set. We acknowledge that we are not able to give a definitive statement of the shape of the news impact curve out in the tails, but our asymptotic theory better reflects this uncertainty than the theory for parametric models, which is overly precise. Thus we are able to provide a better idea of what *we know we do not know*.

7. CONCLUSIONS AND EXTENSIONS

Although we have relied on the least squares criterion to obtain consistency, in practice one can avoid least squares estimations altogether and just apply an iterated version of the likelihood-based method. We expect that the distribu-

[11] It should be noted that these confidence intervals do not take account of the additional variation induced by the various estimations; taking account of this estimation error would widen the confidence intervals considerably.

tion theory for such a method is the same as the distribution of our two-step version of this procedure. This is to be expected from results of Mammen, Linton, and Nielsen (1999) and Linton (2000) in other contexts.

Other estimation methods can be used here like series expansion or splines. However, although one can obtain the distribution theory for parameters θ and rates for estimators of m in that case, the pointwise distribution theory for the nonparametric part is elusive. Furthermore, such methods may be inefficient in the sense of Section 4.4. One might want to combine the series expansion method with a likelihood iteration, an approach taken in Horowitz and Mammen (2002). However, one would still need either to apply our estimation method and theory or to develop a theory for combining an increasing number of Horowitz and Mammen (2002) estimators.

In the working paper version we discuss some extensions of our method to allow a model for the mean and some transformation models like logarithmic variance, and present some ideas about "integrated" versions of our model.

Dept. of Economics, London School of Economics, Houghton Street, London WC2A 2AE, U.K.; lintono@lse.ac.uk

and

Dept. of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany; emammen@rumms.uni-heidelberg.de.

Manuscript received September, 2003; final revision received November, 2004.

APPENDIX A: PROOFS OF (8) AND (34)

In the sequel we take $\mu_t = 0$ without loss of generality.

PROOF OF (8): It is convenient to break the joint optimization problem down into two separate problems: first, for each $\theta \in \Theta$, let m_θ be the function that minimizes (5) with respect to $m \in \mathcal{M}$; second, let θ_* be the parameter that minimizes the profiled criterion $E[y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m_\theta(y_{t-j})]^2$ with respect to $\theta \in \Theta$. It follows that $\theta_0 = \theta_*$ and $m_0 = m_{\theta_0}$. We next find the first-order conditions for this sequential population optimization problem. We write $m = m_0 + \epsilon \cdot f$ for any function f, differentiate with respect to ϵ, and, setting $\epsilon = 0$, we obtain the first-order condition $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m_0(y_{t-j})\}\{\sum_{l=1}^{\infty} \psi_l(\theta) f(y_{t-l})\}] = 0$, which can be rewritten as

$$\text{(40)} \qquad \sum_{j=1}^{\infty} \psi_j(\theta) E[y_0^2 f(y_{-j})] - \sum_{j=1}^{\infty} \sum_{l=1, j\neq l}^{\infty} \psi_j(\theta)\psi_l(\theta) E[m_0(y_{-j}) f(y_{-l})]$$
$$= \sum_{j=1}^{\infty} \psi_j^2(\theta) E[m_0(y_{-j}) f(y_{-j})]$$

for all f. Taking $f(\cdot) = \delta_y(\cdot)$, where $\delta_y(\cdot)$ is the Dirac delta function, we have

$$\begin{aligned} E[y_0^2 f(y_{-j})] &= \int E[y_0^2 | y_{-j} = y'] f(y') p_0(y')\, dy' \\ &= \int E[y_0^2 | y_{-j} = y'] \delta_y(y') p_0(y')\, dy' \\ &= E[y_0^2 | y_{-j} = y] p_0(y), \end{aligned}$$

while $E[m_0(y_{-j}) f(y_{-j})] = \int m_0(y')\delta_y(y') p_0(y')\, dy' = m_0(y) p_0(y)$. Finally, $E[m_0(y_{-j}) f(y_{-l})] = E[E[m_0(y_{-j})|y_{-l}] f(y_{-l})] = \int E[m_0(y_{-j})|y_{-l} = y']\delta_y(y') \times p_0(y')\, dy' = E[m_0(y_{-j})|y_{-l} = y] p_0(y)$. The next step is to change the variables in the double sum. Note that $E[m_0(y_{-j})|y_{-l} = y] = E[m_0(y_0)|y_{j-l} = y]$ by stationarity. Let $t = j - l$. Then for any function $c(\cdot)$ defined on the integers,

$$\begin{aligned} \sum_{j=1}^{\infty} \sum_{l=1, j\neq l}^{\infty} \psi_j(\theta)\psi_l(\theta) c(j-l) &= \sum_{t=\pm 1}^{\infty} \sum_{l=1}^{\infty} \psi_{t+l}(\theta)\psi_l(\theta) c(t) \\ &= \sum_{t=\pm 1}^{\infty} \left(\sum_{l=1}^{\infty} \psi_{t+l}(\theta)\psi_l(\theta) \right) c(t). \end{aligned}$$

Therefore, dividing through by $p_0(y)$ and $\sum_{j=1}^{\infty} \psi_j^2(\theta)$, (40) can be written $\sum_{j=1}^{\infty} \psi_j^{\dagger}(\theta) E(y_0^2 | y_{-j} = y) - \sum_{j=\pm 1}^{\pm\infty} \psi_t^*(\theta) E(m_0(y_0)|y_j = y) = m_0(y)$, which is the stated answer. *Q.E.D.*

PROOF OF (34): We write $g = g_0 + \epsilon \cdot f$ for any function f, differentiate with respect to ϵ, and, setting $\epsilon = 0$, we obtain the first-order condition

$$E\left[\left\{ \frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta} - \frac{1}{\sigma_t^2}\sum_{j=1}^{\infty} \psi_j g_0(y_{t-j}) \right\} \frac{1}{\sigma_t^2} \sum_{l=1}^{\infty} \psi_l f(y_{t-l}) \right] = 0,$$

which can be rewritten

$$\begin{aligned} 0 = \sum_{l=1}^{\infty} \psi_l E\left[\sigma_t^{-4} \frac{\partial \sigma_t^2}{\partial \theta} \middle| y_{t-l} = y \right] - g_0(y) \sum_{j=1}^{\infty} \psi_j^2 E[\sigma_t^{-4} | y_{t-j} = y] \\ - \sum_{j=1}^{\infty} \sum_{l=1, j\neq l}^{\infty} \psi_j \psi_l E[\sigma_t^{-4} g_0(y_{t-j}) | y_{t-l} = y]. \end{aligned}$$

Now use the law of iterated expectations to write $E[\sigma_t^{-4} g_0(y_{t-j})|y_{t-l} = y] = E[E[\sigma_t^{-4}|y_{t-j}, y_{t-l}] g_0(y_{t-j}) | y_{t-l} = y]$. Then

$$E[\sigma_t^{-4} g_0(y_{t-j})|y_{t-l} = y] = \int q_{j,l}(x, y) \frac{p_{0,j-l}(x, y)}{p_0(y) p_0(x)} g_0(x) p_0(x)\, dx,$$

where $q_{j,l}(y,x) = E[\sigma_t^{-4}|y_{t-j} = x, y_{t-l} = y]$. The result follows. *Q.E.D.*

APPENDIX B: PROOFS OF THEOREMS

PROOF OF THEOREM 1: We first outline the approach to obtaining the asymptotic properties of $\widehat{m}_\theta(\cdot)$ for any $\theta \in \Theta$. We give some high level Assumptions A4–A6 under which we have an expansion for $\widehat{m}_\theta - m_\theta$ in terms of $\widehat{m}^*_\theta - m^*_\theta$ and $\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta$. Both terms will contribute a bias and a stochastic term to the expansion. We then verify the Assumptions A4–A6 and verify the central limit theorem.

ASSUMPTION A4: *Suppose that for a sequence* $\delta_T \to 0$, $\sup_{\theta\in\Theta, \|m\|_2=1, |x|\le c} |\widehat{\mathcal{H}}_\theta \times m(x) - \mathcal{H}_\theta m(x)| = o_p(\delta_T)$.

In particular, Assumption A4 gives that $\sup_{\theta\in\Theta,\|m\|_2=1} \|[\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta]m\|_2 = o_p(\delta_T)$. We now show by virtue of Assumption A4 that $(I - \widehat{\mathcal{H}}_\theta)$ is invertible for all $\theta \in \Theta$, with probability tending to 1, and it holds that (see also (14))

$$\sup_{\theta\in\Theta, \|m\|_2=1, |y|\le c} \left|[(I - \widehat{\mathcal{H}}_\theta)^{-1} - (I - \mathcal{H}_\theta)^{-1}]m(y)\right| = o_p(\delta_T). \tag{41}$$

In particular, $\sup_{\theta\in\Theta,\|m\|_2=1} \|[(I - \widehat{\mathcal{H}}_\theta)^{-1} - (I - \mathcal{H}_\theta)^{-1}]m\|_2 = o_p(\delta_T)$. For a proof of claim (41) note that for $m \in \mathcal{M}_c$,

$$m = (I - \widehat{\mathcal{H}}_\theta)^{-1}(I - \mathcal{H}_\theta)^{-1} \sum_{j=0}^{\infty} [(\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)(I - \mathcal{H}_\theta)^{-1}]^j m$$

because of

$$\begin{aligned} \sum_{j=0}^{\infty} [(\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)(I - \mathcal{H}_\theta)^{-1}]^j &= [I - (\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)(I - \mathcal{H}_\theta)^{-1}]^{-1} \\ &= [(I - \widehat{\mathcal{H}}_\theta)(I - \mathcal{H}_\theta)^{-1}]^{-1}. \end{aligned}$$

This gives

$$(I - \widehat{\mathcal{H}}_\theta)^{-1}m - (I - \mathcal{H}_\theta)^{-1}m = \sum_{j=0}^{\infty} [(\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)(I - \mathcal{H}_\theta)^{-1}]^j m.$$

We suppose that $\widehat{m}^*_\theta(y)$ has an asymptotic expansion where the components have certain properties.

ASSUMPTION A5: *Suppose that with* δ_T *as in Assumption* A4, $\widehat{m}^*_\theta(y) - m^*_\theta(y) = \widehat{m}^{*,B}_\theta(y) + \widehat{m}^{*,C}_\theta(y) + \widehat{m}^{*,D}_\theta(y)$, *where* $\widehat{m}^{*,B}_\theta$, $\widehat{m}^{*,C}_\theta$, *and* $\widehat{m}^{*,D}_\theta$ *satisfy*

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^{*,B}(y)| = O_p(T^{-2/5}) \quad \textit{with } \widehat{m}^{*,B} \textit{ deterministic}, \tag{42}$$

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^{*,C}_\theta(y)| = O_p(T^{-2/5}\delta_T^{-1}), \tag{43}$$

$$\sup_{\theta\in\Theta,|y|\le c} |\mathcal{H}_\theta(I-\mathcal{H}_\theta)^{-1}\widehat{m}^{*,C}_\theta(y)| = o_p(T^{-2/5}), \tag{44}$$

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^{*,D}_\theta(y)| = o_p(T^{-2/5}). \tag{45}$$

Here, $\widehat{m}^{*,B}_\theta$ is the bias term, $\widehat{m}^{*,C}_\theta$ is the stochastic term, and $\widehat{m}^{*,D}_\theta$ is the remainder term. For local linear estimates of $g_j(y)$ it follows that under standard smoothness conditions, (42), (43), and (45) hold. The argument is complicated by the fact that $\widehat{m}^*_\theta$ depends on a large number of $g_j(y)$'s, although it effectively behaves like a single smoother. The intuition behind (44) is based on the fact that an integral operator applies averaging to a local smoother and transforms it into a global average, thereby reducing its variance.

Define now for $j = B, C, D$ the terms $\widehat{m}^j_\theta$ as solutions to the integral equations $\widehat{m}^j_\theta = \widehat{m}^{*,j}_\theta + \widehat{\mathcal{H}}_\theta\, \widehat{m}^j_\theta$ and $\widehat{m}^A_\theta$ implicitly from writing the solution $m_\theta + \widehat{m}^A_\theta$ to the integral equation

$$(m_\theta + \widehat{m}^A_\theta) = m^*_\theta + \widehat{\mathcal{H}}_\theta(m_\theta + \widehat{m}^A_\theta). \tag{46}$$

The existence and uniqueness of $\widehat{m}^j_\theta$ follows from the invertibility of the operator $I - \widehat{\mathcal{H}}_\theta$ (at least with probability tending to 1). It now follows that $\widehat{m}_\theta = m_\theta + \widehat{m}^A_\theta + \widehat{m}^B_\theta + \widehat{m}^C_\theta + \widehat{m}^D_\theta$ by linearity of the operator $(I - \widehat{\mathcal{H}}_\theta)^{-1}$. Note that $\widehat{m}^j_\theta = (I - \widehat{\mathcal{H}}_\theta)^{-1}\widehat{m}^{*,j}_\theta$ for $j = B, C, D$, while $m_\theta + \widehat{m}^A_\theta = (I - \widehat{\mathcal{H}}_\theta)^{-1} m^*_\theta$. Define also m^B_θ as the solution to the equation

$$m^B_\theta = \widehat{m}^{*,B}_\theta + \mathcal{H}_\theta m^B_\theta. \tag{47}$$

We now claim that under Assumptions A1–A5,

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^B_\theta(y) - m^B_\theta(y)| = o_p(T^{-2/5}), \tag{48}$$

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^C_\theta(y) - \widehat{m}^{*,C}_\theta(y)| = o_p(T^{-2/5}), \tag{49}$$

$$\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}^D_\theta(y)| = o_p(T^{-2/5}). \tag{50}$$

Here, claims (48) and (50) immediately follow from (14) and (41). For (49) note that because of (43)–(44), (41), and Assumption A4, $\sup_{\theta\in\Theta,|y|\le c}|\widehat{\mathcal{H}}_\theta(I-\widehat{\mathcal{H}}_\theta)^{-1}\widehat{m}_\theta^{*,C}(y)|=o_p(T^{-2/5})$. So we arrive at the expansion of $\widehat{m}_\theta$:

$$\sup_{\theta\in\Theta,|y|\le c}\left|\widehat{m}_\theta(y)-m_\theta(y)-\widehat{m}_\theta^A(y)-m_\theta^B(y)-\widehat{m}_\theta^{*,C}(y)\right|=o_p(T^{-2/5}).$$

This gives an approximation to $\widehat{m}_\theta(y)-m_\theta(y)$ in terms of the expansion of $\widehat{m}_\theta^*$, the population operator $\mathcal{H}_\theta$, and the quantity $\widehat{m}_\theta^A(y)$. This latter quantity depends on the random operator $\widehat{\mathcal{H}}_\theta$.

Next we approximate the quantity $\widehat{m}_\theta^A(y)$ by simpler terms. By subtracting $m_\theta=m_\theta^*+\mathcal{H}_\theta m_\theta$ from (46) we get $\widehat{m}_\theta^A=(\widehat{\mathcal{H}}_\theta-\mathcal{H}_\theta)m_\theta+\widehat{\mathcal{H}}_\theta\widehat{m}_\theta^A$. We next write $\widehat{\mathcal{H}}_\theta$ as a sum of terms with convenient properties.

ASSUMPTION A6: *Suppose that for δ_T as in Assumption* A4, $(\widehat{\mathcal{H}}_\theta-\mathcal{H}_\theta)\times m_\theta(y)=\widehat{m}_\theta^{*,E}(y)+\widehat{m}_\theta^{*,F}(y)+\widehat{m}_\theta^{*,G}(y)$, *where* $\widehat{m}_\theta^{*,E},\widehat{m}_\theta^{*F}$, *and* $\widehat{m}_\theta^{*,G}$ *satisfy* $\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}^{*,E}(y)|=O_p(T^{-2/5})$ *with $\widehat{m}^{*,E}$ deterministic*, $\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}_\theta^{*,F}(y)|=O_p(T^{-2/5}\delta_T^{-1})$, $\sup_{\theta\in\Theta,|y|\le c}|\mathcal{H}_\theta(I-\mathcal{H}_\theta)^{-1}\widehat{m}_\theta^{*,F}(y)|=o_p(T^{-2/5})$, *and* $\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}_\theta^{*,G}(y)|=o_p(T^{-2/5})$.

Again, $\widehat{m}_\theta^{*,E}$ is a bias term, $\widehat{m}_\theta^{*F}$ is a stochastic term, and $\widehat{m}_\theta^{*,G}$ is a remainder term. For kernel density estimates of $\widehat{\mathcal{H}}_\theta$ under standard smoothness conditions, the expansion in Assumption A6 follows from similar arguments to those given for Assumption A5. Define for $j=E,F,G$ the terms $\widehat{m}_\theta^j$ as the unique solutions to the equations $\widehat{m}_\theta^j=\widehat{m}_\theta^{*,j}+\widehat{\mathcal{H}}_\theta\widehat{m}_\theta^j$. It now follows that $\widehat{m}_\theta^A$ can be decomposed into $\widehat{m}_\theta^A=\widehat{m}_\theta^E+\widehat{m}_\theta^F+\widehat{m}_\theta^G$. Define m_θ^E as the solution to the second kind linear integral equation

$$m_\theta^E=\widehat{m}_\theta^{*,E}+\mathcal{H}_\theta m_\theta^E. \tag{51}$$

As above we get that

$$\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}_\theta^E(y)-m_\theta^E(y)|=o_p(T^{-2/5}),$$

$$\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}_\theta^F(y)-\widehat{m}_\theta^{*,F}(y)|=o_p(T^{-2/5}),\quad\text{and}$$

$$\sup_{\theta\in\Theta,|y|\le c}|\widehat{m}_\theta^G(y)|=o_p(T^{-2/5}).$$

We summarize our discussion in the following proposition.

PROPOSITION 1: *Suppose that Assumptions* A1–A6 *hold for some estimators $\widehat{m}_\theta^*$ and $\widehat{\mathcal{H}}_\theta$. Define $\widehat{m}_\theta$ as any solution of $\widehat{m}_\theta=\widehat{m}_\theta^*+\widehat{\mathcal{H}}_\theta\widehat{m}_\theta$. Then the following* *expansion holds for* $\widehat{m}_\theta$:

$$\sup_{\theta\in\Theta,|y|\le c}\left|\widehat{m}_\theta(y)-m_\theta(y)-m_\theta^B(y)-m_\theta^E(y)-\widehat{m}_\theta^{*,C}(y)-\widehat{m}_\theta^{*,F}(y)\right| \tag{52}$$
$$=o_p(T^{-2/5}).$$

The terms m_θ^B *and* m_θ^E *have been defined in* (47) *and* (51), *respectively.*

Equation (52) gives a uniform expansion for $\widehat{m}_\theta(y)-m_\theta(y)$ in terms of a deterministic expression $m_\theta^B(y)+m_\theta^E(y)$ and a random variable $\widehat{m}_\theta^{*,C}(y)+\widehat{m}_\theta^{*,F}(y)$ that is explicitly defined. We have hitherto just made high level assumptions on $\widehat{m}_\theta^*$ and the operator $\widehat{\mathcal{H}}_\theta$ in Assumptions A4–A6, so our result applies to any smoothing method that satisfies these conditions. It remains to prove that Assumptions A4–A6 hold under our primitive conditions B1–B7 and that a central limit theorem (and uniform convergence) applies to $\widehat{m}_\theta^{*,C}(y)+\widehat{m}_\theta^{*,F}(y)$.

PROOF OF HIGH LEVEL CONDITIONS A1, A3–A6, AND CLT: We first define the concept of near epoch dependence (NED) for stationary processes, which we will use in the sequel.

DEFINITION: The stationary process $\{x_t\}$ is said to be stable (NED) in L_2-norm on the stationary α-mixing process $\{z_t\}$ if there exist measurable functions g_m such that, as $m\to\infty$, $v(m)=E[|x_t-g_m(z_{t-1},\dots,z_{t-m})|^2]\to 0$.

This definition provides a sufficient condition for the more general NED definition in say Andrews (1995). The process $\sigma_t^2(\theta)=\sum_{j=1}^\infty\psi_j(\theta)m(y_{t-j})$ is stable in L_2-norm on the process $\{m(y_t)\}$ (which is α-mixing) and the stable numbers satisfy $v(m)\le\exp(-cm)$ for some constant $c>0$. Likewise, the process $\eta_{\theta,t}^1$ is geometrically stable on $\{y_t^2\}$ and $\eta_{\theta,t}^2$ is geometrically stable on $\{m_\theta(y_t)\}$. We use this property below.

Assumptions A1 and A3 follow immediately from our conditions on the parameter space and density functions. We assumed Assumptions A2 in B7.

We verify Assumptions A4–A6 with the choice $\delta_T=T^{-3/10+\xi}$ with $\xi>0$ small enough. This rate is arbitrarily close to the rate of convergence of two-dimensional nonparametric density or regression estimators. We will verify Assumptions A5 and A6 with

$$\widehat{m}_\theta^{*,B}(y)=\frac{h^2}{2}\mu_2(K)\times\beta_\theta^1(y),\widehat{m}_\theta^{*,C}(y)$$
$$=\frac{1}{Tp_0(y)}\sum_{t=1}^{T-\tau_T}K_h(y_t-y)\eta_{\theta,t}^1,\widehat{m}_\theta^{*,E}(y)$$
$$=\frac{h^2}{2}\mu_2(K)\times\beta_\theta^2(y),$$

$$\widehat{m}_\theta^{*,F}(y) = \frac{1}{Tp_0(y)} \sum_{t=1}^{T-\tau_T} K_h(y_t - y)\eta_{\theta,t}^2$$
$$+ \frac{1}{T}\sum_{t=1}^{T-\tau_T} \frac{\mu_\theta(y)}{p_0(y)}[K_h(y_t - y) - EK_h(y_t - y)],$$

where $\eta_{\theta,t}^1 = \sum_{j=1}^\infty \psi_j^\dagger(\theta)\eta_{j,t}$ and $\eta_{\theta,t}^2 = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^2(\theta)\zeta_{j,t}(\theta)$, while $\eta_{j,t} = y_{t+j}^2 - E(y_{t+j}^2|y_t)$ and $\zeta_{j,t}(\theta) = m_\theta(y_{t+j}) - E[m_\theta(y_{t+j})|y_t]$.

PROOF OF ASSUMPTION A4: It suffices to show that

$$\sup_{x|,|y|\le c, 1\le j\le \tau_T} |\widehat{p}_{0,j}(x,y) - p_{0,j}(x,y)| = o_p(\delta_T), \tag{53}$$

$$\sup_{|x|\le c} |\widehat{p}_0(x) - p_0(x)| = o_p(\delta_T). \tag{54}$$

Note that by assumption B4 the density p_0 is bounded from below on $|x| \le c$. For the proof of (53) we make use of an exponential inequality. Using Theorem 1.3 in Bosq (1998) one gets

$$\Pr\left(\left|T^{3/10-\xi}[\widehat{p}_{0,j}(x,y) - E\widehat{p}_{0,j}(x,y)]\right| \ge C\right)$$
$$\le \Pr\left(\left|T^{3/10}\sum_{t=1}^{T-j} K_h(y_t - x)K_h(y_{t+j} - y)\right.\right.$$
$$\left.\left. - EK_h(y_t - x)K_h(y_{t+j} - y)\right| \ge \frac{T}{2}T^\xi\right)$$
$$\le 4\exp\left(-\frac{T^{2\xi}}{32v^2(q)}q\right) + 22(1 + 8T^{-\xi}b)^{1/2}q\alpha\left(\left[\frac{T}{2q}\right] - j\right),$$

where $[x]$ denotes the largest integer smaller or equal to x, and where $q = T^\beta$ with $\frac{7}{10} < \beta < 1$, $j^2 \le T^{1-\beta}$, $b = CT^{7/10}$ for a constant C, $v^2(q) = 8(q^2/T^2)\sigma^2(q) + \frac{b}{4}T^\xi$, and $\sigma^2(q) = E[\sum_{t=1}^{[T/2q]+1} K_h(y_t - x)K_h(y_{t+j} - y) - EK_h(y_t - x)K_h(y_{t+j} - y)]^2$. The variance $\sigma^2(q)$ can be bounded by use of Corollary 1.1. in Bosq (1998). This gives $\sigma^2(q) \le C'T^{2-\beta+(2/5)\gamma}$ for $0 < \gamma < 1$ with a constant C' depending on γ. This gives with constants $C_1, C_2, \ldots > 0$ for $|x|, |y| \le c$, $1 \le j \le \tau_T$,

$$\Pr\left(\left|T^{3/10}[\widehat{p}_{0,j}(x,y) - E\widehat{p}_{0,j}(x,y)]\right| \ge T^\xi\right)$$
$$\le C_1\exp(-C_2T^{C_3}) + C_4T^{C_5}\alpha(T^{C_6}).$$

Define $z = (x,y)$ and let $V_j(z) = \widehat{p}_{0,j}(z) - E\widehat{p}_{0,j}(z)$. Let $B(z_1, \epsilon_T), \ldots, B(z_Q, \epsilon_T)$ be a cover of $\{|x| \le c, |y| \le c\}$, where $B(z_q, \epsilon)$ is a ball centered at z_q

of radius ϵ, while $Q(T)$ is a sufficiently large integer and $Q(T) = 2c^2/\epsilon_T$. By the triangle inequality,

$$\begin{aligned}
&\Pr\Biggl[\sup_{\substack{|x|\le c,|y|\le c\\ 1\le j\le \tau}} |V_j(z)| \ge 2c\delta_T\Biggr]\\
&\quad\le \Pr\Biggl[\max_{\substack{1\le j\le\tau\\ 1\le q\le Q}} |V_j(z_q)| > c\delta_T\Biggr]\\
&\qquad + \Pr\Biggl[\max_{\substack{1\le j\le\tau\\ 1\le q\le Q}} \sup_{z\in B(z_q,\epsilon_T)} |V_j(z_q) - V_j(z)| > c\delta_T\Biggr]
\end{aligned}$$

for any constant c. By the Bonferroni and exponential inequalities,

$$\begin{aligned}
\Pr\Biggl[\max_{\substack{1\le j\le\tau\\ 1\le q\le Q}} |V_j(z_q)| > c\delta_T\Biggr] &\le \sum_{j=1}^{\tau}\sum_{q=1}^{Q} \Pr\bigl[|V_j(z_q)| > c\delta_T\bigr]\\
&\le Q\tau[C_1\exp(-C_2T^{C_3}) + C_4T^{C_5}\alpha(T^{C_6})]\\
&= o(1),
\end{aligned}$$

provided s_0 in B1 is chosen large enough. By the Lipschitz continuity of K, $|K_h(y_t - x) - K_h(y_t - x_q)| \le \overline{K}|x - x_q|/h$, where $\overline{K}$ is finite, and so $T^{3/10-\xi} \times |V_j(z_q) - V_j(z)| \le T^{3/10-\xi}(1/h^2)[c_1|x - x_q| + c_2|y - y_q|] \le c\epsilon_T T^{7/10-\xi}$ for some constants c_1, c_2. This bound is independent of j and uniform over z, so that provided $\epsilon_T T^{7/10-\xi} \to 0$, this term is $o(1)$. This requires that $Q(T)/T^{7/10-\xi} \to \infty$.

We have given the detailed proof of (53) because similar arguments are used in the sequel. Equation (54) follows by the same type of argument. *Q.E.D.*

PROOF OF ASSUMPTION A5: Claim (42) immediately follows from assumption B4. For the proof of (45) we use the usual variance + bias + remainder term decomposition of the local linear estimates $\widehat{g}_j$ as in Masry (1996). Write $M(y) = p_0(y)\,\mathrm{diag}(1, \mu_2(K))$ and

$$M_{Tj}(y) = \frac{1}{Th}\sum_{t=1}^{T} K\biggl(\frac{y - y_{t-j}}{h}\biggr)\begin{bmatrix} 1 & (\frac{y-y_{t-j}}{h}) \\ (\frac{y-y_{t-j}}{h}) & (\frac{y-y_{t-j}}{h})^2 \end{bmatrix}.$$

Then $\widehat{g}_j(y) - g_j(y) = \widehat{B}_{jy} + \widehat{V}_{jy}$, where $\widehat{B}_{jy} = e_1' M_{Tj}^{-1}(y) B_{Tj}(y)$, and $B_{Tj}(y)$ is a vector $B_{Tj}(y) = [B_{Tj,0}(y), B_{Tj,1}(y)]^\top$, where

$$B_{Tj,l}(y) = \frac{1}{Th}\sum_{t=1}^{T}\biggl(\frac{y - y_{t-j}}{h}\biggr)^l K\biggl(\frac{y - y_{t-j}}{h}\biggr)\varDelta_{tj}(y),$$

where $\Delta_{tj}(y) = g_j(y_{t-j}) - g_j'(y)(y_{t-j} - y) = g_j''(y^*_{t,j})(y_{t-j} - y)^2/2$ for some intermediate point $y^*_{t,j}$. The variance effect is $\widehat{V}_{jy} = e_1' M_{Tj}^{-1}(y) U_{Tj}(y)$. The stochastic term $U_{Tj}(y)$ is $U_{Tj}(y) = [U_{Tj,0}(y), U_{Tj,1}(y)]^\top$, where

$$U_{Tj,l}(y) = \frac{1}{Th}\sum_{t=1}^{T}\left(\frac{y - y_{t-j}}{h}\right)^{l} K\left(\frac{y - y_{t-j}}{h}\right)\eta_{j,t-j}.$$

We have $\widehat{m}^*_\theta(y) - m^*_\theta(y) = \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)[\widehat{g}_j(y) - g_j(y)] - \sum_{j=\tau+1}^{\infty}\psi_j^\dagger(\theta) g_j(y)$, where $\sup_{\theta\in\Theta}\sup_{|y|\le c}|\sum_{j=\tau+1}^{\infty}\psi_j^\dagger(\theta) g_j(y)| \le c'\sum_{j=\tau+1}^{\infty}\overline{\psi}^{j-1}/\inf_{\theta\in\Theta}\sum_{j=1}^{\infty}\psi_j^2(\theta)$ for some finite constant c', and $\sum_{j=\tau+1}^{\infty}\overline{\psi}^{j-1} \le \overline{\psi}^{\tau}/(1-\overline{\psi}) = o(T^{-1/2})$. Therefore,

$$\widehat{m}^*_\theta(y) - m^*_\theta(y) = \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)\widehat{V}_{jy} + \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)\widehat{B}_{jy} + o_p(T^{-1/2}).$$

We then use the fact that $\sup_{|y|\le c, 1\le j\le \tau_T}\|M_{T,j}(y) - M(y)\| = o_p(1)$, which follows by the same reasoning as for (53) and (54). Defining V_{jy} and B_{jy} as $\widehat{V}_{jy}$ and $\widehat{B}_{jy}$ with $M_{Tj}(y)$ replaced by $M(y)$, we have

$$\begin{aligned}\widehat{m}^*_\theta(y) - m^*_\theta(y) &= \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)V_{jy} + \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)B_{jy}\\ &\qquad + R_{T1}(y,\theta) + R_{T2}(y,\theta) + o_p(T^{-1/2}),\end{aligned}$$

where $R_{T1}(y,\theta) = \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)[\widehat{V}_{jy} - V_{jy}]$ and $R_{T2}(y,\theta) = \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)[\widehat{B}_{jy} - B_{jy}]$. We have

$$\begin{aligned}\sum_{j=1}^{\tau}\psi_j^\dagger(\theta)V_{jy} &= \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)\sum_{t=\tau_T+1}^{T}\frac{K_h(y - y_{t-j})\eta_{j,t-j}}{Tp_0(y)}\\ &= \sum_{j=1}^{\tau}\psi_j^\dagger(\theta)\sum_{s=1}^{T-\tau_T}\frac{K_h(y - y_s)\eta_{j,s}}{Tp_0(y)}\\ &= \sum_{s=1}^{T-\tau_T}\frac{K_h(y - y_s)\sum_{j=1}^{\tau}\psi_j^\dagger(\theta)\eta_{j,s}}{p_0(y)T}\\ &= \sum_{t=1}^{T-\tau_T}\frac{K_h(y_t - y)\eta^1_{\theta,t}}{Tp_0(y)} + \sum_{t=1}^{T-\tau_T}K_h(y_t - y)\sum_{j=\tau+1}^{\infty}\frac{\psi_j^\dagger(\theta)\eta_{j,t}}{Tp_0(y)}\end{aligned}$$

by changing variable $t \mapsto t - j = s$ and interchanging summation.

We show that

$$\sup_{|y|\le c,\theta\in\Theta} |R_{T1}(y,\theta)| = o_p(T^{-2/5}), \tag{55}$$

$$\sup_{|y|\le c,\theta\in\Theta} |R_{T2}(y,\theta)| = o_p(T^{-2/5}), \tag{56}$$

$$\sup_{|y|\le c,\theta\in\Theta} \left| \frac{1}{Tp_0(y)} \sum_{t=1}^{T-\tau_T} K_h(y_t-y) \sum_{j=\tau+1}^{\infty} \psi_j^{\dagger}(\theta)\eta_{j,t} \right| = o_p(T^{-2/5}). \tag{57}$$

It follows that

$$\begin{aligned} &\widehat{m}_\theta^*(y) - m_\theta^*(y) \\ &\quad = \frac{1}{Tp_0(y)} \sum_{t=1}^{T-\tau_T} K_h(y_t-y)\eta_{\theta,t}^1 + \sum_{j=1}^{\tau} \psi_j^{\dagger}(\theta) B_{jy} + o_p(T^{-2/5}). \end{aligned}$$

We establish next (57). Define $T_n = T^{-1}\sum_{t=1}^{T-\tau_T} K_h(y_t-y)\sum_{j=\tau+1}^{\infty}\psi_j^{\dagger}(\theta)\eta_{j,t}/p_0(y)$. First note that $E(T_n)=0$ and

$$\begin{aligned} \operatorname{var}(T_n) &= \frac{1}{T^2h^2p_0^2(y)} \sum_{t=1}^{T-\tau_T}\sum_{s=1}^{T-\tau_T}\sum_{j=\tau+1}^{\infty}\sum_{l=\tau+1}^{\infty} \psi_j^{\dagger}(\theta)\psi_l^{\dagger}(\theta)E[K_tK_s\eta_{j,t}\eta_{l,s}] \\ &\le C\frac{1}{Th^{2(1-1/\rho)}p_0^2(y)} \\ &\quad \times \sum_{j=\tau+1}^{\infty}\sum_{l=\tau+1}^{\infty} \psi_j^{\dagger}(\theta)\psi_l^{\dagger}(\theta)\sum_{s=1}^{\infty}\alpha(j-(s+l))^{1-1/2\rho} \\ &\le C'\frac{1}{Th^{2(1-1/\rho)}}\overline{\psi}^{2(\tau+1)} = o(T^{-1}h^{-1}) \end{aligned}$$

by Davydov's inequality, the mixing condition B1, and the decay conditions B10. Here, $K_t = K((y_t-y)/h)$ and C, C' are generic finite constants. This establishes the pointwise rate of T_n. The uniformity of the bound in (57) can be achieved by application of the exponential inequality in Theorem 1.3 of Bosq (1998) used also in the proof of (53). The proofs of (55) and (56) are similar.

For the proof of (43) we apply this exponential inequality to bound

$$\Pr\left(\left|T^{2/5}\sum_{t=1}^{T} K_h(y_t-y)\frac{\widetilde{\eta}_{\theta,t}}{p_0(y)}\right| \ge \frac{T}{2}T^{3/10+\xi}\right),$$

where $\widetilde{\eta}_{\theta,t} = \sum_{j=1}^{\tau_T} \psi_j^{\dagger}(\theta)[\min\{y_{t+j}^2, T^{1/\rho}\} - E(\min\{y_{t+j}^2, T^{1/\rho}\}|y_t)]$. The truncated random variables $\widetilde{\eta}_{\theta,t}$ can be replaced by $\eta_{\theta,t}$ using the fact that $1 - \Pr(y_t^2 \le T^{1/\rho}$ for $1 \le t \le T) \le T \Pr(y_t^2 > T^{1/\rho}) \le E[y_t^{2\rho} \mathbb{1}(y_t^2 > T^{1/\rho})] \to 0$.

It remains to check (44). Define the operator $\mathcal{L}_\theta(x, y)$ by $\mathcal{H}_\theta(I - \mathcal{H}_\theta)^{-1} \times m(x) = \int_{-c}^{c} \mathcal{L}_\theta(x, y) m(y) p_0(y)\, dy$. The $\mathcal{L}_\theta(x, y)$ can be constructed by use of the eigenfunctions $\{e_{\theta,j}\}_{j=1}^{\infty}$ of $\mathcal{H}_\theta$. Denote as above the corresponding eigenvalues by $\lambda_{\theta,j}$. Then

$$\mathcal{H}_\theta(x, y) = \sum_{j=1}^{\infty} \lambda_{\theta,j} e_{\theta,j}(x) e_{\theta,j}(y) \quad \text{and}$$

$$\mathcal{L}_\theta(x, y) = \sum_{j=1}^{\infty} \frac{\lambda_{\theta,j}}{1 - \lambda_{\theta,j}} e_{\theta,j}(x) e_{\theta,j}(y).$$

Note that for a constant $0 < \gamma < 1$ we have $\sup_{\theta \in \Theta, j \ge 1} \lambda_{\theta,j} < \gamma$. This shows that

$$\begin{aligned} &\int_{-c}^{c} \mathcal{L}_\theta^2(x, y) p_0(y) p_0(x)\, dx\, dy \\ &\quad = \sum_{j=1}^{\infty} \frac{\lambda_{\theta,j}^2}{(1 - \lambda_{\theta,j})^2} \le \frac{1}{(1 - \gamma)^2} \sum_{j=1}^{\infty} \lambda_{\theta,j}^2 < \infty. \end{aligned}$$

Furthermore, it can be checked that $\mathcal{L}_\theta(x, y)$ is continuous in θ, x, y. This follows from Assumption A3 and the continuity of $\mathcal{H}_\theta(x, y)$.

Therefore, we write $\mathcal{H}_\theta(I - \mathcal{H}_\theta)^{-1} \widehat{m}_\theta^{*,C}(x) = \frac{1}{T} \sum_{t=1}^{T} \nu_\theta(y_t, x) \eta_{\theta,t}^1$ with $\nu_\theta(z, x) = \int_{-c}^{c} \mathcal{L}_\theta(x, y)(1/p_0(y)) K_h(z - y)\, dy$. The function $\nu_\theta(z, x)$ is continuous in θ, z, x. Using this fact, claim (44) can be easily checked, e.g., again by application of the exponential inequality in Theorem 1.3 of Bosq (1998). *Q.E.D.*

Proof of Assumption A6: Write

$$\begin{aligned} &\int \widehat{\mathcal{H}}_\theta(y, x) m_\theta(x) \widehat{p}_0(x)\, dx - \int \mathcal{H}_\theta(y, x) m_\theta(x) p_0(x)\, dx \\ &\quad = -\sum_{j=\pm 1}^{\pm \tau_T} \psi_j^*(\theta) \int \left[\frac{\widehat{p}_{0,j}(y, x)}{\widehat{p}_0(y)} - \frac{p_{0,j}(y, x)}{p_0(y)} \right] m_\theta(x)\, dx \\ &\quad = -\sum_{j=\pm 1}^{\pm \tau_T} \psi_j^*(\theta) \int \left[\frac{\widehat{p}_{0,j}(y, x) - p_{0,j}(y, x)}{p_0(y)} \right] m_\theta(x)\, dx \\ &\qquad + \sum_{j=\pm 1}^{\pm \tau_T} \psi_j^*(\theta)(\widehat{p}_0(y) - p_0(y)) \int \left[\frac{p_{0,j}(y, x)}{p_0^2(y)} \right] m_\theta(x)\, dx \\ &\qquad + o_p(T^{-2/5}). \end{aligned}$$

Using this expansion one can show that $\widehat{m}_\theta^{*,G}(y) = (\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)m_\theta(y) - \widehat{m}_\theta^{*,E}(y) - \widehat{m}_\theta^{*,F}(y)$ is of order $o_p(T^{-2/5})$. The other conditions of Assumption A6 can be checked as in the proof of Assumption A5. *Q.E.D.*

PROOF OF CLT FOR $\widehat{m}_\theta^{*,C}(y) + \widehat{m}_\theta^{*,F}(y)$: This follows by an application of a central limit theorem for triangular arrays of NED processes along the lines of Lu (2001). The argument is first to replace, for example, $\eta_{\theta,t}^1$ by the logarithmic truncation $\eta_{\theta,t}^{1,\tau} = \sum_{j=1}^{\tau} \psi_j^\dagger(\theta)\eta_{j,t}$. Then divide the sum $\sum_{t=1}^T K_h(y_t - y)\eta_{\theta,t}^{1,\tau}$ into the usual Bernstein large block/small blocks. Then apply Davydov's inequality for random variables with finite p moments. Because of the exponential decline of the stability numbers $v(m)$, the CLT follows. This concludes the proof of (26). *Q.E.D.*

PROOFS OF (27) AND (28): The only additionality here is to show that $\sup_{\theta\in\Theta,|y|\le c} |\widehat{m}_\theta^{*,C}(y) + \widehat{m}_\theta^{*,F}(y)| = o_p(T^{-1/4})$. This follows by applying the exponential inequality again.

Finally,

$$\begin{aligned}
&\sup_{\theta,1\le t} |\widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta)| \\
&\quad \le \sup_\theta \sum_{j=1}^{\infty} \psi_j(\theta) \sup_{\theta,|y|\le c} |\widehat{m}_\theta(y) - m_\theta(y)| \\
&\quad\quad + \sup_\theta \sum_{j=\tau+1}^{\infty} \psi_j(\theta) \sup_{|y|\le c} m_\theta(y) + \frac{\tau_T}{T}\sum_{t=1}^{T} y_t^2 - E[y_t^2] \\
&\quad \le \frac{1}{1-\overline{\psi}}\Big[\sup_{\theta,|y|\le c} |\widehat{m}_\theta(y) - m_\theta(y)| + \overline{\psi}^{\tau+1}\sup_{|y|\le c} m_\theta(y)\Big] + O_p(\tau_T T^{-1/2}) \\
&\quad = o_p(T^{-1/4})
\end{aligned}$$

by the summability conditions on $\psi_j(\theta)$, the boundedness of $m_\theta(y)$ on $[-c, c]$, and the uniform convergence result (27). *Q.E.D.*

PROOF OF THEOREM 2: Consistency. We apply some general results for semiparametric estimators. Write $S_T(\theta) = T^{-1}\sum_{t=1}^T \{y_t^2 - \sigma_t^2(\theta)\}^2$, where $\sigma_t^2(\theta) = \sum_{j=1}^\infty \psi_j(\theta)m_\theta(y_{t-j})$, and let $S(\theta) = ES_T(\theta)$. We show that $S_T(\theta) - S(\theta) = o_p(1)$ by applying a law of large numbers for near epoch dependent functions of mixing processes. Let $\overline{m}(y) = \sup_{\theta\in\Theta} m_\theta(y)$ and $\overset{\circ}{\overline{m}}_\ell(y) = \sup_{\theta\in\Theta} |\partial m_\theta(y)/\partial\theta_\ell|$, which are bounded continuous functions on $[-c, c]$. It follows that $\sup_{\theta\in\Theta} \sigma_t^2(\theta) \le C\sum_{j=1}^\infty \overline{\psi}^{j-1}\overline{m}(y_{t-j})$ and $\sup_{\theta\in\Theta} |\partial\sigma_t^2(\theta)/\partial\theta_\ell| \le C\sum_{j=1}^\infty \overline{\psi}^{j-1}(\overline{m}(y_{t-j}) + \overset{\circ}{\overline{m}}_\ell(y))$, which are both bounded processes. Therefore, the law of large numbers can be made uniform over $\theta \in \Theta$. In conclusion we have

$$\sup_{\theta\in\Theta} |S_T(\theta) - S(\theta)| = o_p(1). \tag{58}$$

Then, letting $\eta_t(\theta) = y_t^2 - \sigma_t^2(\theta)$, we have for each $\theta \in \Theta$,

$$\begin{aligned}|\widehat{S}_T(\theta) - S_T(\theta)| &\le \frac{2}{T}\sum_{t=1}^{T} |\eta_t(\theta)| \max_{1\le t\le T} |\widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta)| \\ &\quad + \Big[\max_{1\le t\le T} |\widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta)|\Big]^2 + \frac{1}{T}\sum_{t=1}^{\tau T} \eta_t^2(\theta) \\ &= o_p(1)\end{aligned}$$

because of (28). In fact, this order is uniform in θ and we have

$$\sup_{\theta\in\Theta} |\widehat{S}_T(\theta) - S_T(\theta)| \xrightarrow{p} 0. \tag{59}$$

Therefore, by (58) and (59) we have $\sup_{\theta\in\Theta} |\widehat{S}_T(\theta) - S(\theta)| = o_p(1)$. By assumption B7, $S(\theta)$ is uniquely minimized at $\theta = \theta_0$, which then implies consistency of $\widehat{\theta}$.

Root-N Consistency. Consider the derivatives

$$\frac{\partial \widehat{S}_T(\theta)}{\partial \theta} = -\frac{2}{T}\sum_{t=1}^{T} \widehat{\eta}_t(\theta)\frac{\partial \widehat{\sigma}_t^2(\theta)}{\partial \theta} \quad \text{and}$$

$$\frac{\partial^2 \widehat{S}_T(\theta)}{\partial \theta\, \partial \theta^\top} = \frac{2}{T}\sum_{t=1}^{T} \frac{\partial \widehat{\sigma}_t^2(\theta)}{\partial \theta}\frac{\partial \widehat{\sigma}_t^2(\theta)}{\partial \theta^\top} - \widehat{\eta}_t(\theta)\frac{\partial^2 \widehat{\sigma}_t^2(\theta)}{\partial \theta\, \partial \theta^\top},$$

where $\widehat{\eta}_t(\theta) = (y_t^2 - \widehat{\sigma}_t^2(\theta))$. We have shown that $\widehat{\theta} \to^p \theta_0$, where θ_0 is an interior point of Θ. We make a Taylor expansion about θ_0,

$$o_p(1) = \sqrt{T}\frac{\partial \widehat{S}_T(\widehat{\theta})}{\partial \theta} = \sqrt{T}\frac{\partial \widehat{S}_T(\theta_0)}{\partial \theta} + \frac{\partial^2 \widehat{S}_T(\overline{\theta})}{\partial \theta\, \partial \theta^\top}\sqrt{T}(\widehat{\theta} - \theta_0),$$

where $\overline{\theta}$ is an intermediate value. We then show that for all sequences $\epsilon_T \to 0$, we have for a constant $C > 0$,

$$\inf_{\|\theta-\theta_0\|\le\epsilon_T} \lambda_{\min}\left(\frac{\partial^2 \widehat{S}_T(\theta)}{\partial \theta\, \partial \theta^\top}\right) \ge C + o_p(1), \tag{60}$$

$$\sqrt{T}\frac{\partial \widehat{S}_T(\theta_0)}{\partial \theta} = O_p(1). \tag{61}$$

This implies the $\sqrt{T}$ consistency.

To establish the results (60) and (61) we use some expansions given in Lemma 1.

PROOF OF (60): By straightforward but tedious calculation we show that

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_T, 1\leq t\leq T}\left\|\frac{\partial^2\widehat{S}_T(\theta)}{\partial\theta\,\partial\theta^\top}-\frac{\partial^2 S_T(\theta)}{\partial\theta\,\partial\theta^\top}\right\|=o_p(1).$$

Specifically, it suffices to show that

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_T, 1\leq t\leq T}\left\|\frac{\partial^j\widehat{\sigma}_t^2(\theta)}{\partial\theta^j}-\frac{\partial^j \sigma_t^2(\theta)}{\partial\theta^j}\right\|=o_p(1),$$

$j=0,1,2$. For $j=0,1$ this follows from (28) and (29). For $j=2$ this follows by similar arguments using Lemma 1. Note also that by B4 for a constant $c>0$, $\inf_{\|\theta-\theta_0\|\leq\epsilon_T,1\leq t\leq T}\sigma_t^2(\theta)>c$. Furthermore,

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_T}\left\|\frac{\partial^2 S_T(\theta)}{\partial\theta\,\partial\theta^\top}-E\left[\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta^\top}\right]\right\|=o_p(1)$$

by standard arguments. Therefore, by the triangle inequality,

$$\sup_{\|\theta-\theta_0\|\leq\epsilon_T}\left\|\frac{\partial^2 \widehat{S}_T(\theta)}{\partial\theta\,\partial\theta^\top}-E\left[\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta^\top}\right]\right\|=o_p(1).$$

Q.E.D.

PROOF OF (61): Write

$$\frac{\partial\widehat{S}_T(\theta_0)}{\partial\theta}=-\frac{2}{T}\sum_{t=1}^{T}\left[y_t^2-\sigma_t^2(\theta_0)-[\widehat{\sigma}_t^2(\theta_0)-\sigma_t^2(\theta_0)]\right]$$
$$\times\left[\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}+\frac{\partial\widehat{\sigma}_t^2(\theta_0)}{\partial\theta}-\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right]$$

and let with $\eta_t=\eta_t(\theta_0)$, $\sqrt{T}E_T(\theta_0)=E_{T1}+E_{T2}$,

$$E_{T1}=-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta},$$

$$E_{T2}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[\widehat{\sigma}_t^2(\theta_0)-\sigma_t^2(\theta_0)]\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}$$
$$-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\left[\frac{\partial\widehat{\sigma}_t^2(\theta_0)}{\partial\theta}-\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right].$$

Then

$$\begin{aligned}
&\left|\sqrt{T}\frac{\partial\widehat{S}_T(\theta_0)}{\partial\theta}-\sqrt{T}E_T(\theta_0)\right|\\
&\quad\le\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[\widehat{\sigma}_t^2(\theta)-\sigma_t^2(\theta_0)]\left[\frac{\partial\widehat{\sigma}_t^2(\theta_0)}{\partial\theta}-\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right]\right|\\
&\quad\le\sqrt{T}\max_{1\le t\le T}|\widehat{\sigma}_t^2(\theta_0)-\sigma_t^2(\theta_0)|\times\max_{1\le t\le T}\left\|\frac{\partial\widehat{\sigma}_t^2(\theta_0)}{\partial\theta}-\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right\|=o_p(1)
\end{aligned}$$

by (28) and (29).

The term E_{T1} is asymptotically normal with mean zero and finite variance by the central limit theorem for (geometric) NED processes over an α-mixing base. Note that $E[\eta_t(\partial\sigma_t^2(\theta_0)/\partial\theta)]=0$ by definition of θ_0.

For the treatment of E_{T2} we now use that

$$\begin{aligned}
(62)\qquad E_{T2}=&\frac{h^2}{\sqrt{T}}\sum_{t=1}^{T}\left\{\sum_{j=1}^{\tau_T}\psi_j(\theta_0)b^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)+\eta_t\sum_{j=1}^{\tau_T}\psi_j'(\theta_0)b^0(y_{t-j})\right\}\\
&+\frac{h^2}{\sqrt{T}}\sum_{t=1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j(\theta_0)b^1(y_{t-j})\right\}\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\sum_{j=1}^{\tau_T}\psi_j(\theta_0)s^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)\right\}\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j'(\theta_0)s^0(y_{t-j})\right\}\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j(\theta_0)s^1(y_{t-j})\right\}+o_P(1),
\end{aligned}$$

where $b_\theta(y)=h^{-2}[m_\theta^B(y)+m_\theta^E(y)]$, $s_\theta(y)=(I-\mathcal{H}_\theta)^{-1}(m_\theta^{*,C}+m_\theta^{*,F})(y)$, $b^j(y)=\partial^j/(\partial\theta)^j b_{\theta_0}(y)$, and $s^j(y)=\partial^j/(\partial\theta)^j s_{\theta_0}(y)$. By tedious calculations it can be shown that the last three terms on the right-hand side of (62) are of order $o_P(1)$. For this purpose one has to plug in the definitions of s^0 and s^1 as local weighted sums of mixing mean zero variables. For the first two terms on the right-hand side of (62) note that b^0 and b^1 are deterministic functions. Furthermore, we will show that

$$(63)\qquad E\left[\sum_{j=1}^{\infty}\psi_j(\theta_0)b^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)+\eta_t\psi_j'(\theta_0)b^0(y_{t-j})\right]=0,$$

$$E\left[\eta_t \sum_{j=1}^{\infty} \psi_j(\theta_0) b^1(y_{t-j})\right] = 0. \tag{64}$$

Note that in (63) and (64) we have replaced the upper index of the sum by ∞. Thus, with (63) and (64) we see that the first two terms on the right-hand side of (62) are sums of variables with mean geometrically tending to zero. The sums are multiplied by factors $h^2 T^{-1/2}$. By using mixing properties it can be shown that these sums are of order $O_P(h^2) = o_p(1)$. It remains to check (63) and (64). By definition for each function g, $E[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta)\delta g(y_{t-j})\}^2]$ is minimized for $\delta = 0$. By taking derivatives with respect to δ we get that

$$E\left\{[y_t^2 - \sigma_t^2(\theta)] \sum_{j=1}^{\infty} \psi_j(\theta) g(y_{t-j})\right\} = 0. \tag{65}$$

With $g = b^0$ and θ_0 this gives (64). For the proof of (63) we now take the difference of (65) for θ and θ_0. This gives

$$\begin{aligned} &E[y_t^2 - \sigma_t^2(\theta_0)] \sum_{j=1}^{\infty} [\psi_j(\theta) - \psi_j(\theta_0)] g(y_{t-j}) \\ &\quad - E[\sigma_t^2(\theta) - \sigma_t^2(\theta_0)] \sum_{j=1}^{\infty} \psi_j(\theta) g(y_{t-j}) = 0. \end{aligned}$$

Taking derivatives with respect to θ gives $E[u_t \sum_{j=1}^{\infty} \psi_j'(\theta_0) g(y_{t-j}) - \partial\sigma_t^2/\partial\theta(\theta_0) \sum_{j=1}^{\infty} \psi_j(\theta_0) g(y_{t-j})] = 0$. With $g = b^0$ this gives (63). *Q.E.D.*

PROOFS OF THEOREMS 3 AND 4: We only give a proof of Theorem 3. Theorem 4 follows along the same lines. For a proof of (31) one shows that for $C > 0$, $\sup_{\|\theta-\theta_0\| \le CT^{-1/2}} |\widehat{m}_\theta(y) - \widehat{m}_{\theta_0}(y)| = o_P[(Th)^{-1/2}]$. This claim follows by using appropriate bounds on $\widehat{\mathcal{H}}_\theta - \widehat{\mathcal{H}}_{\theta_0}$ and $\widehat{m}_\theta^* - \widehat{m}_{\theta_0}^*$.

Because of (31), for a proof of (32) it suffices to show

$$\sqrt{Th}[\widehat{m}_{\theta_0}(y) - m_{\theta_0}(y) - h^2 b(y)] \Longrightarrow N(0, \omega(y)). \tag{66}$$

So it remains to show (66). Put $\widehat{p}_0^1(y) = T^{-1} \sum_{t=1}^{T} (y_t - y) K_h(y_t - y)$ and $\widehat{p}_0^2(y) = T^{-1} \sum_{t=1}^{T} (y_t - y)^2 K_h(y_t - y)$. Then, by using similar arguments as in the proof of Theorem 1, we have for $\gamma > 0$, $\sup_{|y| \le c} |\widehat{p}_0^1(y) - h^2 \mu_2(K) p_0'(y)| =$

$O_p(h^{1/2}T^{-1/2+\gamma} + h^3)$ and $\sup_{|y|\le c}|\widehat{p}_0^2(y) - h^2\mu_2(K)p_0(y)| = O_p(h^{3/2}T^{-1/2+\gamma} + h^3)$. Furthermore, $\sup_{|y|\le c}|\widehat{p}_0(y) - p_0(y)| = O_p(h^2 + h^{-1/2}T^{-1/2+\gamma})$.

These results can be applied to show that uniformly in $|y| \le c$ and $j \le \tau_T$,

$$\begin{aligned}
\hat{g}_j(y) = {} & \frac{1}{T}\sum_{t=1}^{T}\frac{K_h(y_{t-j}-y)\sigma_t^2 u_t}{p_0(y)(y)} \\
& + \frac{1}{T}\sum_{t=1}^{T}\frac{K_h(y_{t-j}-y)}{p_0(y)}\sum_{\ell=1}^{\infty}\psi_\ell(\theta_0)m(y_{t-\ell}) \\
& + \frac{\widehat{p}_0^1(y)^2}{\widehat{p}_0(y)^2\widehat{p}_0^2(y)}\frac{1}{T}\sum_{t=1}^{T}K_h(y_{t-j}-y)\sum_{\ell=1}^{\infty}\psi_\ell(\theta_0)m(y_{t-\ell}) \\
& - \frac{\widehat{p}_0^1(y)^2}{\widehat{p}_0(y)\widehat{p}_0^2(y)}\frac{1}{T}\sum_{t=1}^{T}(y_{t-j}-y)K_h(y_{t-j}-y)\sum_{\ell=1}^{\infty}\psi_\ell(\theta_0)m(y_{t-\ell}) \\
& + o_p(T^{-1/2}) \\
= {} & \frac{1}{T}\sum_{t=1}^{T}\frac{K_h(y_{t-j}-y)}{p_0(y)}\sigma_t^2 u_t \\
& + \frac{1}{T}\sum_{t=1}^{T}\frac{K_h(y_{t-j}-y)}{\widehat{p}_0(y)}\sum_{\ell=1}^{\infty}\psi_\ell(\theta_0)m(y_{t-\ell}) \\
& + h^2\Biggl\{\mu_2(K)\frac{p_0'(y)^2}{p_0(y)^3}\sum_{\ell=1,\ell\ne j}^{\infty}\psi_\ell(\theta_0)\int m(u)p_{j,\ell}(y,u)\,du \\
& \qquad - \mu_2(K)\frac{p_0'(y)}{p_0(y)^2}\sum_{\ell=1,\ell\ne j}^{\infty}\psi_\ell(\theta_0)\int m(u)\frac{\partial}{\partial y}p_{j,\ell}(y,u)\,du \\
& \qquad - \mu_2(K)\psi_j(\theta)\frac{p_0'(y)m'(y)}{p_0(y)}\Biggr\} + o_p(T^{-1/2}).
\end{aligned}$$

By plugging this into

$$\begin{aligned}
& \widehat{m}_{\theta_0}^*(y) - (I - \widehat{\mathcal{H}}_{\theta_0})m_0(y) \\
& \quad = \sum_{j=1}^{\tau_T}\psi_j^\dagger(\theta_0)\widehat{g}_j(y) - m_0(y) - \sum_{0<|j|<\tau_T}\psi_j^*(\theta_0)\int\frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)}m_0(x)\,dx,
\end{aligned}$$

we get $\widehat{m}^*_{\theta_0}(y) - (I - \widehat{\mathcal{H}}_{\theta_0})m_0(y) = S_1 + S_2 + S_3 + S_4 - m_0(y) + o_p(T^{-1/2})$, where

$$S_1 = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\frac{\psi_j^\dagger(\theta_0)K_h(y_{t-j}-y)\sigma_t^2 u_t}{p_0(y)},$$

$$S_2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty}\psi_j^\dagger(\theta_0)\psi_\ell(\theta_0)\frac{K_h(y_{t-j}-y)m_0(y_{t-\ell})}{\widehat{p}_0(y)},$$

$$S_3 = h^2\mu_2(K)\frac{p_0^1(y)}{p_0(y)}\left[\frac{\partial}{\partial y}(\mathcal{H}_{\theta_0}m_0(y) - m_0(y))\right],$$

$$S_4 = -\sum_{j\neq 0}\psi_j^*(\theta_0)\int\frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)}m(x)\,dx.$$

We have

$$\begin{aligned}
&S_2 + S_4 - m_0(y)\\
&\quad = \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau_T}^{j=1}\psi_j(\theta_0)\psi_j^\dagger(\theta_0)\frac{K_h(y_{t-j}-y)}{\widehat{p}_0(y)}[m_0(y_{t-j}) - m_0(y)]\\
&\qquad + \frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq 0}^{\tau_T}\psi_j^*(\theta_0)\frac{K_h(y_{t-j}-y)}{\widehat{p}_0(y)}m_0(y_{t-j})\\
&\qquad - \sum_{j\neq 0}^{\tau_T}\int\psi_j^*(\theta_0)\frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)}m_0(x)\,dx\\
&\quad = h^2\mu_2(K)\left[\frac{p_0'(y)m_0'(y)}{p_0(y)} + \frac{m_0''(y)}{2}\right]\\
&\qquad + \sum_{j\neq 0}\psi_j^*(\theta_0)\frac{1}{T}\sum_{t=1}^{T}\frac{K_h(y_t - y)}{\widehat{p}_0(y)}\\
&\qquad\qquad \times\left\{m_0(y_{t+j}) - \int K_h(y_{t+j} - x)m_0(x)\,dx\right\}\\
&\qquad + o_p(T^{-1/2})\\
&\quad = h^2\mu_2(K)\left[\frac{p_0'(y)m_0'(y)}{p_0(y)} + \frac{1}{2}m_0''(y)\right.\\
&\qquad\qquad \left. + \frac{1}{2}\sum_{j\neq 0}\frac{\psi_j^*(\theta_0)}{p_0(y)}\int m_0''(u)p_{0,j}(y,u)\,du\right] + o_p(T^{-1/2})
\end{aligned}$$

$$= h^2\mu_2(K)\left[\frac{p_0'(y)}{p_0(y)}m_0'(y) + \frac{1}{2}m_0''(y) - \frac{1}{2}\mathcal{H}_{\theta_0}m_0''(y)\right] + o_p(T^{-1/2}).$$

Therefore we get uniformly in $|y| \le c$,

$$\begin{aligned}
&\widehat{m}_{\theta_0}(y) - m_{\theta_0}(y)\\
&= (I - \widehat{\mathcal{H}}_{\theta_0})^{-1}[\widehat{m}^*_{\theta_0}(y) - (I - \widehat{\mathcal{H}}_{\theta_0})m_{\theta_0}(y)]\\
&= (I - \mathcal{H}_{\theta_0})^{-1}[\widehat{m}^*_{\theta_0}(y) - (I - \widehat{\mathcal{H}}_{\theta_0})m_{\theta_0}(y)] + o_p(T^{-1/2})\\
&= \frac{1}{T}\sum_{t=1}^{T}(I - \mathcal{H}_{\theta_0})^{-1}\left[\sum_{j=1}^{\tau_T}\psi_j^{\dagger}(\theta_0)\frac{K_h(y_{t-j} - y)}{p_0(y)}\right]\sigma_t^2 u_t\\
&\quad + h^2\mu_2(K)(I - \mathcal{H}_{\theta_0})^{-1}\\
&\quad \times \left\{\frac{p_0'(y)}{p_0(y)}\left[\frac{\partial}{\partial y}\mathcal{H}_{\theta_0}m_0(y) - m_0'(y) + \mathcal{H}_{\theta_0}m_0'(y)\right]\right.\\
&\qquad \left. + \frac{m_0''(y)}{2} - \frac{\mathcal{H}_{\theta_0}m_0''(y)}{2}\right\} + o_p(T^{-1/2})\\
&= \frac{1}{T}\sum_{t=1}^{T}K_t^*\sigma_t^2 u_t\\
&\quad + h^2\mu_2(K)\left\{\frac{m_0''(y)}{2} + (I - \mathcal{H}_{\theta_0})^{-1}\left[\frac{p_0'(y)}{p_0(y)}(\mathcal{H}_{\theta_0}m_0)\right](y)\right\}\\
&\quad + o_p(T^{-1/2})
\end{aligned}$$

with $K_t^* = \sum_{j=1}^{\tau_T}\psi_j^{\dagger}(\theta_0)K_h(y_{t-j} - y)/p_0(y)$. From this stochastic expansion we immediately get an expansion for the asymptotic bias. For the calculation of the asymptotic variance note that

$$\begin{aligned}
hEK_t^{*2} &= h\frac{1}{p_0^2(y)}\left\{\sum_{j\neq\ell}\psi_j^{\dagger}(\theta_0)\psi_\ell^{\dagger}(\theta_0)E\{K_h(y_{t-j} - y)K_h(y_{t-\ell} - y)\right.\\
&\qquad\qquad \left. \times E[\sigma_t^4 u_t^2|y_{t-j}, y_{t-\ell}]\}\right\}\\
&\qquad + \sum_{j=1}^{\infty}\psi_j^{\dagger}(\theta_0)^2 E\{K_h^2(y_{t-j} - y)E[\sigma_t^4 u_t^2|y_{t-j} = y]\}\\
&= \frac{1}{p_0(y)}\nu_0(K)\sum_{j=1}^{\infty}\psi_j^{\dagger}(\theta_0)^2 E(\sigma_t^4 u_t^2|y_{t-j} = y) + o(1)
\end{aligned}$$

$$= \frac{1}{p_0(y)} \left[\sum_{l=1}^{\infty} \psi_l(\theta_0)^2 \right]^{-1} \nu_0(K) \sum_{j=1}^{\infty} \psi_j(\theta_0)^2 E[\sigma_t^4 u_t^2 | y_{t-j} = y]$$
$$+ o(1). \qquad\qquad \textit{Q.E.D.}$$

PROOFS OF THEOREMS 5 AND 6: The proof makes use of similar arguments as in Theorems 1–4. For this reason we only give a short outline. We first discuss $\widetilde{m}_{\theta_0}$. Below we will show that $\widetilde{\theta} - \theta_0 = O_P(T^{-1/2})$. This can be used to show that $\sup_{|y| \le c} |\widetilde{m}_{\theta_0}(y) - \widetilde{m}_{\widetilde{\theta}}(y)| = o_P(T^{-2/5})$. Thus, up to first order the asymptotics of both estimates coincide. We compare $\widetilde{m}_{\theta_0}$ with the following theoretical estimate $\widetilde{\overline{m}}_\theta$. This estimate is defined by the integral equation $\widetilde{\overline{m}}_\theta = \widetilde{\overline{m}}_\theta^* + \widetilde{\overline{\mathcal{H}}}_\theta \widetilde{\overline{m}}_\theta$, where

$$\widetilde{\overline{m}}_\theta^*(y) = \frac{\sum_{j=1}^{\tau_T} \psi_j(\theta) \widetilde{g}_j^a(y)}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta) \widetilde{g}_j^b(y)},$$

$$\widetilde{\overline{\mathcal{H}}}_\theta(x, y) = \frac{-\sum_{j=1}^{\tau_T} \sum_{l=1, l \ne j}^{\tau_T} \psi_j(\theta) \psi_l(\theta) \widetilde{g}_{l,j}^c(x, y) \frac{\widehat{p}_{0,l-j}(x,y)}{\widehat{p}_0(y)\widehat{p}_0(y)}}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta) \widetilde{g}_j^b(y)}.$$

Here $\widetilde{g}_j^a$ is the local linear smooth of $\sigma_t^{-4} y_t^2$ on y_{t-j}, $\widetilde{g}_j^b$ is the local linear fit of σ_t^{-4} on y_{t-j}, and $\widetilde{g}_{l,j}^c$ is the bivariate local linear fit of σ_t^{-4} on (y_{t-l}, y_{t-j}). Note that $\widetilde{g}_j^a, \widetilde{g}_j^b, \widetilde{g}_{l,j}^c$ are defined as $\widehat{g}_j^a, \widehat{g}_j^b, \widehat{g}_{l,j}^c$, but with $\widehat{\sigma}_t^2$ replaced by σ_t^2. Furthermore, $\widetilde{\overline{m}}_\theta$ is defined as $\widetilde{m}_\theta$ but with $\widehat{g}_j^a, \widehat{g}_j^b, \widehat{g}_{l,j}^c$ replaced by $\widetilde{g}_j^a, \widetilde{g}_j^b, \widetilde{g}_{l,j}^c$.

By tedious calculations one can verify for a constant $C > 0$ that there exists a bounded function b such that uniformly for $|y| \le c$, $\|\theta - \theta_0\| \le CT^{-1/2}$, $\widetilde{\overline{m}}_\theta(y) - \widetilde{m}_\theta(y) - h^2 b(y) = o_P(T^{-1/2})$. The bias term b is caused by bias terms of $\widehat{\sigma}_t^2 - \sigma_t^2$. So up to bias terms the asymptotics of $\widetilde{\overline{m}}_\theta(y)$ and $\widetilde{m}_\theta(y)$ coincide.

The estimate $\widetilde{\overline{m}}_{\theta_0}(y)$ can be treated as $\widehat{m}_{\theta_0}(y)$ in the proof of Theorem 3. As the stochastic term of $\widetilde{\overline{m}}_{\theta_0}(y)$ we get $T^{-1} \sum_{t=1}^{T} \overline{K}_t(y) \sigma_t^{-4}(y_t^2 - \sigma_t^2) = T^{-1} \sum_{t=1}^{T} \overline{K}_t(y) \sigma_t^{-2} u_t$, where $\overline{K}_t(y) = \sum_{j=1}^{\tau_T} \psi_j(\theta_0) K_h(y_{t-j} - y)/p_0(y) \times \sum_{j=1}^{\tau_T} \psi_j^2(\theta_0) E[\sigma_t^{-4} | y_{t-j} = y]$. Asymptotic normality of this term can be shown by use of central limit theorems as in the proof of Theorem 1. For the calculation of the asymptotic variance it can be easily checked that

$$hE[\overline{K}_t(y)^2 \sigma_t^{-4} u_t^2] = \frac{1}{p_0(y)} \frac{\nu_0(K) \sum_{j=1}^{\infty} \psi_j^2(\theta_0) E(\sigma_j^{-4} u_j^2 | y_0 = y)}{[\sum_{j=1}^{\infty} \psi_j^2(\theta_0) E(\sigma_j^{-4} | y_0 = y)]^2}$$
$$+ o(1),$$

from which the result follows. In the special case of homokurtosis, the numerator simplifies as stated.

Use of the above arguments gives the statement of Theorem 5. For the proof of Theorem 6 one shows

$$\frac{\partial \widetilde{l}}{\partial \theta}(\theta_0) = -\frac{1}{T}\sum_{t=1}^{T} \sigma_t^{-2} u_t \frac{\partial \overline{\sigma}_t^2}{\partial \theta}(\theta_0) + o_P(T^{-1/2}), \tag{67}$$

$$\frac{\partial^2 \widetilde{l}}{\partial \theta^2}(\theta) = -E\left[\sigma_t^{-4} \frac{\partial \overline{\sigma}_t^2}{\partial \theta} \frac{\partial \overline{\sigma}_t^2}{\partial \theta^\top}(\theta_0)\right] + o_P(1) \tag{68}$$

uniformly for $|\theta - \theta_0| < CT^{-1/2}$ for all $C > 0$. This shows that for $c_T \to \infty$ slowly enough, there exists a unique local minimizer $\widetilde{\theta}$ of $\widetilde{l}(\theta)$ in a $c_T T^{-1/2}$ neighborhood of θ_0 with

$$\widetilde{\theta} = \theta_0 - \left\{E\left[\sigma_t^{-4} \frac{\partial \overline{\sigma}_t^2}{\partial \theta} \frac{\partial \overline{\sigma}_t^2}{\partial \theta^\top}(\theta_0)\right]\right\}^{-1} T^{-1} \sum_{t=1}^{T} \sigma_t^{-2} u_t \frac{\partial \overline{\sigma}_t^2}{\partial \theta}(\theta_0) + o_P(T^{-1/2}).$$

This expansion can be used to show the desired asymptotic normal limit for $\widetilde{\theta}$. It remains to show (67) and (68). This can be done by using similar arguments as for the proof of (60) and (61). *Q.E.D.*

PROOF OF THEOREM 7: For $0 < c \le \infty$ we define the operator $\mathcal{H}_{\theta,c} m(y) = \int_{-c}^{c} \mathcal{H}_\theta(y,x) m(x) p_0(x)\,dx$. We write $\|m\|_{\infty,2}$ for the $L_2(p_0)$-norm $\|m\|_{\infty,2}^2 = \int_{-\infty}^{\infty} m(x)^2 p_0(x)\,dx$. For a linear operator $A{:}L_2(p_0) \to L_2(p_0)$ we write $\|A\|_{\infty,2} = \sup_{\|m\|_{\infty,2}\le 1} \|Am\|_{\infty,2}$. We have added the subindex ∞ to indicate that integration now runs from $-\infty$ to ∞. For Hilbert–Schmidt operators $A{:}L_2(p_0) \to L_2(p_0)$ we denote the maximal eigenvalue by $\lambda_{\max}(A)$. Using the same arguments as in Section 2.1 we get from C2 that for $0 < c \le \infty$, $\theta \in \Theta$, $\lambda_{\max}(\mathcal{H}_{\theta,c}) < 1$. With the help of C1 and C3 we conclude that there exist constants $c_* > 0$ and $0 < \gamma_* < 1$ with $\lambda_{\max}(\mathcal{H}_{\theta,c}) < \gamma_*$ for $c_* \le c \le \infty$, $\theta \in \Theta$. This implies that $\|(I - \mathcal{H}_{\theta,c})^{-1}\|_{\infty,2} \le (1 - \gamma_*)^{-1}$. By definition we have, with $\delta_\theta(y) = \mu^c(y; \xi_{\theta,c}) - m_{\theta,\infty}(y)$,

$$m_{\theta,c}(y) - m_{\theta,\infty}(y) = \mathcal{H}_{\theta,c}(m_{\theta,c} - m_{\theta,\infty})(y) + (\mathcal{H}_{\theta,\infty} - \mathcal{H}_{\theta,c})\delta_\theta(y). \tag{69}$$

This implies $\|m_{\theta,c} - m_{\theta,\infty}\|_{\infty,2} \le \|(I - \mathcal{H}_{\theta,c})^{-1}\|_{\infty,2} \|\mathcal{H}_{\theta,\infty} - \mathcal{H}_{\theta,c}\|_{\infty,2} \Delta(c)$. This shows claim (37). For the proof of claim (38) note that we get from (69) that $m_{\theta,c} - m_{\theta,\infty} = [I + \mathcal{H}_{\theta,c} + \mathcal{H}_{\theta,c}^2 (I - \mathcal{H}_{\theta,c})^{-1}][\mathcal{H}_{\theta,\infty} - \mathcal{H}_{\theta,c}]\delta_\theta$. *Q.E.D.*

PROOFS OF THEOREMS 8 AND 9: We first define explicitly the estimators $\widehat{m}_{\theta,c}$, $\widehat{\xi}_{\theta,c}$. To do this we obtain a population characterization of $m_{\theta,c}$, $\xi_{\theta,c}$. Write ν^c for the function vector that vanishes on $[-c, c]$ and is equal to ν outside of $[-c, c]$. The functions are then elements of a linear subspace L_2 of $L_2(p_0)$. This subspace consists of all functions m of $L_2(p_0)$ that fulfill

$m(y) = \xi^\top \nu(y)$ for $|y| \geq c$ for some parameter ξ. We also write $m = (m_c, \xi)$ for elements of L_2. As above we now consider the target function $(m_{\theta,c}, \xi_{\theta,c})$ that minimizes $E[\{y_t^2 - \sum_{j=1}^\infty \psi_j(\theta)[m_c(y_{t-j}) + \xi^\top \nu^c(y_{t-j})]\}^2]$ over all elements (m_c, ξ) of L_2. This tuple is uniquely determined by the linear operator equation $(m_{\theta,c}, \xi_{\theta,c}) = (m_{\theta,c}^*, \xi_{\theta,c}^*) + \mathcal{H}_{\theta,c}^*(m_{\theta,c}, \xi_{\theta,c})$. Here for $|y| \leq c$ the function $m_{\theta,c}^*(y)$ is defined as $m_\theta^*(y)$. The parameter $\xi_{\theta,c}^*$ is given by $\xi_{\theta,c}^* = E[y_t^2 \sum_{k=1}^\infty \psi_k^\dagger(\theta)\nu(y_{t-k})]$. The operator $\mathcal{H}_{\theta,c}^*$ is defined by $\mathcal{H}_{\theta,c}^*(m_c, \xi) = (m_c^H, \xi^H)$ with

$$\begin{aligned} m_c^H(y) &= \int_{|x|\leq c} \mathcal{H}_\theta(y,x) m_c(x) p_0(x) \\ &\qquad + \int_{|x|>c} \mathcal{H}_\theta(y,x)\xi^\top \nu(x) p_0(x)\,dx \quad \text{for} \quad |y| \leq c, \\ \xi^H &= \left[\int_{|x|\geq c} \nu(x)\nu(x)^\top p_0(x)\,dx\right]^{-1} \\ &\qquad \times \left[\int_{|x|\geq c, |y|\geq c} \nu(x)\nu(y)^\top \xi \mathcal{H}_\theta(y,x) p_0(x) p_0(y)\,dx\,dy \right. \\ &\qquad \left. + \int_{|x|\geq c, |y|\leq c} \nu(x)\mathcal{H}_\theta(x,y) m_c(x) p_0(x) p_0(y)\,dx\,dy\right]. \end{aligned}$$

Estimates of $(m_{\theta,c}, \xi_{\theta,c})$ are given by the solution $(\widehat{m}_{\theta,c}, \widehat{\xi}_{\theta,c})$ of the linear equation

$$(\widehat{m}_{\theta,c}, \widehat{\xi}_{\theta,c}) = (\widehat{m}_{\theta,c}^*, \widehat{\xi}_{\theta,c}^*) + \widehat{\mathcal{H}}_{\theta,c}^*(\widehat{m}_{\theta,c}, \widehat{\xi}_{\theta,c}). \tag{70}$$

Here $\widehat{m}_{\theta,c}^*$ is defined as in Section 2.1. The parameter ξ^* can be estimated by $\widehat{\xi}^* = \sum_{k=1}^{\tau_T} \psi_k^\dagger(\theta)(T-k)^{-1}\sum_{t=k+1}^T y_t^2 \nu(y_{t-k})$. The operator $\widehat{\mathcal{H}}_{\theta,c}^*$ is defined by $\widehat{\mathcal{H}}_{\theta,c}^*(m_c, \xi) = (m_c^{\widehat{H}}, \xi^{\widehat{H}})$ with

$$\begin{aligned} m_c^{\widehat{H}}(y) &= \int_{|x|\leq c} \widehat{\mathcal{H}}_\theta(y,x) m_c(x) p_0(x) + \xi^\top \widehat{\nu}^H(y) \quad \text{for} \quad |y| \leq c, \\ \xi^{\widehat{H}} &= \widehat{A}_\nu^{-1}\left[\widehat{B}_\nu \xi + \int_{|y|\leq c} \widehat{\nu}^H(y) m_c(y) p_0(y)\,dy\right]. \end{aligned}$$

Here $\widehat{A}_\nu = T^{-1}\sum_{t=1}^T \nu^c(y_t)\nu^c(y_t)^\top$ and $\widehat{B}_\nu = \sum_{1<|l|\leq\tau_T} \psi_l^*(\theta)(T-l)^{-1} \times \sum_{t=l+1}^T \nu(y_t)\nu(y_{t-l})^\top$. The function $\widehat{\nu}^H(y)$ is defined as $\widehat{\nu}^H(y) = \sum_{1<|l|\leq\tau_T} \psi_l^*(\theta) \times \widehat{r}_l(y)$, where $\widehat{r}_l(y)$ is a local linear fit of the conditional expectation $E[\nu(y_{t+l})| y_t = y]$. Equation (70) can be solved by first eliminating the unknown $\widehat{\xi}_{\theta,c}$. Then one has a linear integral equation with unknown $\widehat{m}_{\theta,c}$. The integral equation can be solved by the numerical methods discussed above. The random operator $\widehat{\mathcal{H}}^*_{\theta,c}$ can be discussed as the operator $\widehat{\mathcal{H}}_\theta$ in the proof of Theorem 1. This leads to quite analogous results for the estimates $\widehat{\xi}_{\theta,c}$ and $\widehat{m}_{\theta,c}$. The proofs of Theorems 8 and 9 follow now directly along the lines of Theorem 1. *Q.E.D.*

APPENDIX C: ADDITIONAL LEMMA

LEMMA 1: *We have for* $j = 0, 1, 2$,

$$\sup_{|y|\le c, \theta\in\Theta} \left| \frac{\partial^j}{\partial\theta^j} \left[\widehat{m}_\theta(y) - m^B_\theta(y) - m^E_\theta(y) - (I - \mathcal{H}_\theta)^{-1} (\widehat{m}^{*,C}_\theta + \widehat{m}^{*,F}_\theta)(y) \right] \right|$$
$$= o_p(T^{-1/2}).$$

PROOF OF LEMMA 1: For $j = 0$ the claim follows along the lines of the proof of Theorem 1. Note that in the expansions of the theorem, $(\widehat{m}^{*,C}_\theta + \widehat{m}^{*,F}_\theta)(y)$ is now replaced by $(I - \mathcal{H}_\theta)^{-1}(\widehat{m}^{*,C}_\theta + \widehat{m}^{*,F}_\theta)(y)$. The difference of these terms is of order $O_P(T^{-1/2})$. For the proof for $j = 1$ we make use of the integral equation for $\widehat{m}^1_\theta = \frac{\partial}{\partial\theta}\widehat{m}_\theta$, $\widehat{m}^1_\theta = \frac{\partial}{\partial\theta}\widehat{m}^*_\theta + [\frac{\partial}{\partial\theta}\widehat{\mathcal{H}}_\theta]\widehat{m}_\theta + \widehat{\mathcal{H}}_\theta\widehat{m}^1_\theta$. Thus with $\widehat{m}^{*,1}_\theta = \frac{\partial}{\partial\theta}\widehat{m}^*_\theta + [\frac{\partial}{\partial\theta}\widehat{\mathcal{H}}_\theta]\widehat{m}_\theta$, the derivative $\widehat{m}^1_\theta$ fulfills $\widehat{m}^1_\theta = \widehat{m}^{*,1}_\theta + \widehat{\mathcal{H}}_\theta\widehat{m}^1_\theta$. This is an integral equation with the same integral kernel $\widehat{\mathcal{H}}_\theta$ but with another intercept. An expansion for the solution can be achieved by the same approach as for $\widehat{m}$. Similarly, one proceeds for $j = 2$. These arguments use condition B10. *Q.E.D.*

REFERENCES

ANDREWS, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

ATKINSON, K. (1976): "An Automatic Program for Linear Fredholm Integral Equations of the Second Kind," *ACM Transactions on Mathematical Software*, 2, 154–171.

AUDRINO, F., AND P. BÜHLMANN (2001): "Tree-Structured GARCH Models," *Journal of the Royal Statistical Society*, 63, 727–744.

BERRY, S., AND A. PAKES (2002): "Two Estimators for the Parameters of Discrete Dynamic Games," Unpublished Manuscript, Yale University.

BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: The John Hopkins University Press.

BOLLERSLEV, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

BOLLERSLEV, T., R. F. ENGLE, AND D. B. NELSON (1994), "ARCH Models," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engle and D. L. McFadden. Amsterdam: Elsevier Science, 2959–3038.

BOSQ, D. (1998): *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*. Berlin: Springer-Verlag.

BREIMAN, L., AND J. H. FRIEDMAN (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation" (with Discussion), *Journal of the American Statistical Association*, 80, 580–619.

BROOKS, C., S. P. BURKE, AND G. PERSAND (2001): "Benchmarks and the Accuracy of GARCH Model Estimation," *International Journal of Forecasting*, 17, 45–56.

CARRASCO, M., AND X. CHEN (2002): "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39.

CARRASCO, M., J. P. FLORENS, AND E. RENAULT (2003): "Linear Inverse Problems in Structural Econometrics," in *Handbook of Econometrics*, Vol. 6 (forthcoming), ed. by J. J. Heckman and E. Leamer. Amsterdam: Elsevier Science.

CARROLL, R., E. MAMMEN, AND W. HÄRDLE (2002): "Estimation in an Additive Model when the Components Are Linked Parametrically," *Econometric Theory*, 18, 886–912.

DAROLLES, S., J. P. FLORENS, AND E. RENAULT (2002): "Nonparametric Instrumental Regression," Working Paper, GREMAQ, Toulouse.

DROST, F. C., AND C. A. J. KLAASSEN (1997): "Efficient Estimation in Semiparametric GARCH Models," *Journal of Econometrics*, 81, 193–221.

DROST, F. C., AND T. E. NIJMAN (1993): "Temporal Aggregation of GARCH Processes," *Econometrica*, 61, 909–927.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica*, 50, 987–1008.

——— (1990): "Discussion: Stock Market Volatility and the Crash of '87," *Review of Financial Studies*, 3, 103–106.

ENGLE, R. F., AND T. BOLLERSLEV (1986): "Modeling the Persistence of Conditional Variances," *Econometric Reviews*, 5, 1–50.

ENGLE, R. F., AND G. GONZÁLEZ-RIVERA (1991): "Semiparametric ARCH Models," *Journal of Business and Economic Statistics*, 9, 345–359.

ENGLE, R. F., AND V. K. NG (1993): "Measuring and Testing the Impact of News on Volatility," *The Journal of Finance*, 48, 1749–1778.

FAN, J. (1992): "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 82, 998–1004.

FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

FIORENTINI, G., G. CALZOLARI, AND L. PANNATONI (1996): "Analytic Derivatives and the Computation of GARCH Estimates," *Journal of Applied Econometrics*, 11, 399–417.

FRIEDMAN, J. H., AND W. STUETZLE (1981): "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

GLOSTEN, L. R., R. JAGANNATHAN, AND D. E. RUNKLE (1993): "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Returns on Stocks," *Journal of Finance*, 48, 1779–1801.

GOURIÉROUX, C., AND A. MONFORT (1992): "Qualitative Threshold ARCH Models," *Journal of Econometrics*, 52, 159–199.

HALL, P., AND J. L. HOROWITZ (2003): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," Unpublished Manuscript, Northwestern University.

HANNAN, E. J. (1973): "The Asymptotic Theory of Linear Time-Series Models," *Journal of Applied Probability*, 10, 130–145.

HANSEN, B. A. (1991): "GARCH(1, 1) Processes Are Near Epoch Dependent," *Economics Letters*, 36, 181–186.

HÄRDLE, W., AND A. B. TSYBAKOV, (1997): "Locally Polynomial Estimators of the Volatility Function," *Journal of Econometrics*, 81, 223–242.

HÄRDLE, W., A. B. TSYBAKOV, AND L. YANG, (1996): "Nonparametric Vector Autoregression," Discussion Paper SFB 373, Humbodt-Universität Berlin.

HASTIE, T., AND R. TIBSHIRANI (1990): *Generalized Additive Models*. London: Chapman & Hall.

HOROWITZ, J. L. (2001): "Nonparametric Estimation of a Generalized Additive Model with Unknown Link Function," *Econometrica*, 69, 499–513.

HOROWITZ, J. L., J. KLEMELÄ, AND E. MAMMEN (2002): "Optimal Estimation in Additive Regression," Unpublished Manuscript, Heidelberg University.

HOROWITZ, J. L., AND E. MAMMEN (2002): "Nonparametric Estimation of an Additive Model with a Link Function," Unpublished Manuscript, Northwestern University.

KIM, W., AND O. LINTON (2004): "A Local Instrumental Variable Estimation Method for Generalized Additive Volatility Models," *Econometric Theory*, 20, 1094–1139.

KRISTENSEN, D., AND A. RAHBEK (2003): "Asymptotics of the QMLE for a Class of ARCH(q) Models," Unpublished Manuscript, University of Copenhagen.

LEE, S., AND B. HANSEN (1994): "Asymptotic Theory for the GARCH(1, 1) Quasi-Maximum Likelihood Estimator," *Econometric Theory*, 10, 29–52.

LINTON, O. B. (1993): "Adaptive Estimation in ARCH Models," *Econometric Theory*, 9, 539–569.

——— (2000): "Efficient Estimation of Generalized Additive Nonparametric Regression Models," *Econometric Theory*, 16, 502–523.

LINTON, O. B., AND E. MAMMEN (2003): "Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods," Working Paper EM/2003/453, STICERD.

LINTON, O. B., AND J. P. NIELSEN (1995): "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–100.

LU, Z. (2001): "Kernel Density Estimation for Time Series under Generalized Conditions: Asymptotic Normality and Applications," *Annals of the Institute of Statistical Mathematics*, 53, 447–468.

LUMSDAINE, R. (1995): "Finite-Sample Properties of the Maximum Likelihood Estimator in GARCH(1, 1) and IGARCH(1, 1) Models: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 13, 1–10.

LUMSDAINE, R. L. (1996): "Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1, 1) and Covariance Stationary GARCH(1, 1) Models," *Econometrica*, 64, 575–596.

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

MASRY, E. (1996): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.

MASRY, E., AND D. TJØSTHEIM (1995): "Nonparametric Estimation and Identification of Nonlinear ARCH Time Series: Strong Convergence and Asymptotic Normality," *Econometric Theory*, 11, 258–289.

MCCULLOUGH, B. D., AND C. G. RENFRO (1999): "Benchmarks and Software Standards: A Case Study of GARCH Procedures," *Journal of Economic and Social Measurement*, 25, 59–71.

MIKOSCH, T., AND C. STĂRICĂ (2000): "Limit Theory for the Sample Autocorrelations and Extremes of a GARCH(1, 1) Process," *The Annals of Statistics*, 28, 1427–1451.

NELSON, D. B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347–370.

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

NEWEY, W. K., AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

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

PAGAN, A. R., AND Y. S. HONG (1991): "Nonparametric Estimation and the Risk Premium," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. Barnett, J. Powell, and G. E. Tauchen. Cambridge, U.K.: Cambridge University Press, 51–75.

PAGAN, A. R., AND G. W. SCHWERT (1990): "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics*, 45, 267–290.

PERRON, B. (1998): "A Monte Carlo Comparison of Non-Parametric Estimators of the Conditional Variance," Unpublished Manuscript, Université de Montréal.

ROBINSON, P. M. (1991): "Testing for Strong Serial Correlation and Dynamic Conditional Heteroskedasticity in Multiple Regression," *Journal of Econometrics*, 47, 67–84.

ROBINSON, P. M., AND P. ZAFFARONI (2002): "Pseudo-Maximum Likelihood Estimation of ARCH(∞) Models," Unpublished Manuscript, London School of Economics.

RUST, J. (1997): "Using Randomization to Break the Curse of Dimensionality," *Econometrica*, 65, 487–516.

——— (2000): "Nested Fixed Point Algorithm Documentation Manual. Version 6," Yale University.

SEVERINI, T. A., AND W. H. WONG (1992): "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

STOKEY, N. L., AND R. E. LUCAS (1989): *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press.

STONE, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

——— (1985): "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 685–705.

TRICOMI, F. G. (1957): *Integral Equations*. New York: Interscience.

WEISS, A. A. (1986): "Asymptotic Theory for ARCH Models: Estimation and Testing," *Econometric Theory*, 2, 107–131.

XIAO, Z., O. LINTON, R. CARROLL, AND E. MAMMEN (2003): "More Efficient Local Polynomial Estimation in Nonparametric Regression with Autocorrelated Errors," *Journal of the American Statistical Association*, 98, 980–992.

YANG, L., W. HÄRDLE, AND J. P. NIELSEN (1999): "Nonparametric Autoregression with Multiplicative Volatility and Additive Mean," *Journal of Time Series Analysis*, 20, 579–604.