ESTIMATING SEMIPARAMETRIC ARCH (∞) MODELS BY KERNEL SMOOTHING METHODS*

by

Oliver Linton
London School of Economics and Political Science

Enno Mammen
Universität Heidelberg

Contents:

Discussion Paper
No.EM/03/453
May 2003

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020 7955 6698

* We would like to thank Xiaohong Chen and Wolfgang Härdle for helpful discussions. Linton's research was supported by the National Science Foundation, the Economic and Social Research Council of the United Kingdom, and the Danish Social Science Research Council. Mammen's research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 'Quantifikation and Simulation Ökonimischer Prozesse', Humboldt Universität zu Berlin, and Project M1026/6-2.

Abstract

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 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 and nonparametric part. We investigate the performance of our procedures on simulated data and on a sample of S&P500 daily returns. We find some evidence of asymmetric news impact functions in the data.

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

JEL Nos.: C13**,** C14, G12.

Contact addresses:
Professor Oliver Linton, Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: o.linton@lse.ac.uk

Professor Enno Mammen, Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany. Email: enno@statlab.uni-heidelberg.de

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

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

This model has been extensively studied and many of its properties are now known, see Bollerslev, Engle, and Nelson (1994). Usually this model is estimated by Gaussian Quasi-Likelihood. In the last fifteen years there have many additional parametric volatility models studied in the literature. All these models are nonlinear, which poses difficulties both in computation and in deriving useful tools for statistical inference. Parametric models are also prone to misspecification especially in this context because of the lack of any theoretical guidance as to the correct functional form. Semiparametric models can provide greater flexibility and robustness to functional form misspecification, see Powell (1994).

Engle and Gonzalez-Rivera (1989) 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) proved that a kernel version of their procedure was 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 and construct an estimator that achieves the bound in large samples. This line of research is about refinements to existing consistent procedures.

More recently attention has focused on functional form issues in the conditional variance function itself. This literature 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 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 (1998). Masry and Tjøstheim (1995) also estimate nonparametric ARCH models using the Nadaraya-Watson kernel estimator. Fan and Yao (1998) have discussed efficiency issues in the model (2)

$$y_t = m(y_{t-1}) + \sigma(y_{t-1})\varepsilon_t, \tag{2}$$

where $m(\cdot)$ and $\sigma(\cdot)$ are smooth but unknown functions, and ε_t is a martingale difference sequence with unit conditional variance. In practice, including only one lag is unlikely to capture all the dynamics, and we must extend this model to include more lagged variables. The problem with this generalization is that nonparametric estimation of multi-dimension regression surface suffers from the well-known "curse of dimensionality": the optimal [Stone (1986)] rate of convergence decreases with dimensionality d. For example, under twice differentiability of $m(\cdot)$ and $\sigma(\cdot)$, the optimal rate is $T^{-2/(4+d)}$ for whatever d, which gets rapidly worse with dimension. In addition, it is hard to describe, interpret and understand the estimated regression surface when the dimension is more than two. Furthermore, 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 non-parametric estimator of the conditional variance will provide a poor approximation, as reported by Perron (1998). So not only does this model not capture adequately the time series properties of many datasets, 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 applications. A direct extension is to assume that the volatility [and perhaps the mean too] is additive, i.e.,

$$\sigma_t^2 = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j}). \tag{3}$$

Estimation in additive models has been studied in Hastie and Tibshirani (1990), Linton and Nielsen (1995) and Tjøstheim and Auestad (1994). Previous nonparametric approaches have considered only finite order ARCH(p) processes, see for example Pagan and Hong (1990), Masry and Tjøstheim (1997), and Carroll, Mammen, and Härdle (2002). The best achievable rate of convergence for estimates of $\sigma_j^2(.)$ is that of one-dimensional nonparametric regression. Yang, Härdle, and Nielsen (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is again additive, but the volatility is multiplicative:

$$\sigma_t^2 = c_v \prod_{j=1}^{d} \sigma_j^2(y_{t-j}). \tag{4}$$

To estimate (4) they applied the method of marginal integration using local linear fits as a pilot smoother, and derived the asymptotic normal distribution of the component estimates; they converge at the one-dimensional rate. The closed form of the bias and variance are also given. Kim and Linton

(2002) generalize this model to allow for arbitrary [but known] transformations, i.e.,

$$G(\sigma_t^2) = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j}), \tag{5}$$

where $G(.)$ is known function like log or level. In Xia, Tong, Li, and Zhu (2002) there is a discussion of index models of the form

$$\sigma_t^2 = \sigma^2\left(\sum_{j=1}^{d} \beta_j y_{t-j}^2\right), \tag{6}$$

where $\sigma^2(.)$ is an unknown function. Models (3)-(6) deal with the curse of dimensionality but still do not capture the persistence of volatility, and specifically they do not nest the favourite GARCH(1,1) process.

This paper analyses a class of semiparametric ARCH models that generalizes the Engle and Ng (1993) model and has both general functional form aspects and flexible dynamics. Specifically, our model nests the simple GARCH(1,1) model but permits more general functional form. It contains both finite dimensional parameters that control the dependence and a single unknown scalar function that determines the shape of the news impact curve. This model allows for an asymmetric leverage effect, and as much dynamics as GARCH(1,1). 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 integral equations, which have been extensively studied in the applied mathematics literature, see for example Tricomi (1955). The solution of these equations only requires the computation of two-dimensional smoothing operations, 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), Buja, Hastie, and Tibshirani (1989), and Hastie and Tibshirani (1990) these methods have been investigated in the context of additive nonparametric regression and related models. Recently, Opsomer and Ruppert (1997) and Mammen, Linton, and Nielsen (1999) have provided a distribution theory for this specific class of problems. Newey and Powell (1989,2003) 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. Hall and Horowitz (2003) establish the optimal rate for estimation in this problem and propose two estimators that achieve this rate. Neither paper provides distribution theory. Our estimation methods and proof technique is purely applicable to the type II situation, which is nevertheless quite common. Our paper goes significantly beyond the existing literature in two respects. First, the

integral operator does not necessarily have norm less than one so that the iterative solution method of successive approximations is not feasible. This also affects the way we derive the asymptotic properties, and we can't 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 and show that our parameter estimator achieves this bund. We also discuss the efficiency question regarding the nonparametric component and conclude that a likelihood-based version of our estimator can't 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 daily 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. Section 5 reports some numerical results and section 6 concludes.

2 The Model and its Properties

We shall suppose 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

$$\sigma_t^2(\theta, m) = \mu + \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j}), \tag{7}$$

where $\mu \in \mathbb{R}, \theta \in \Theta \subset \mathbb{R}^p$ and $m \in \mathcal{M}$, where $\mathcal{M} = \{m\text{: measurable}\}$. At this stage, the constant μ can be put equal to zero without any loss of generality. It will become important below when we will consider more restrictive choices of $\mathcal{M}$. Here, 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 $m_0(.)$ are unknown and to be estimated from a finite sample $\{y_1, \ldots, y_T\}$. Following Drost and Nijman (1993), we can give three interpretations to (7). The *strong* form ARCH(∞) process arises when

$$\frac{y_t}{\sigma_t} = \varepsilon_t \tag{8}$$

is i.i.d with mean zero and variance one, where $\sigma_t^2 = \sigma_t^2(\theta_0, m_0)$. The *semi-strong* form arises when

$$E(y_t \,|\mathcal{F}_{t-1}) = 0 \text{ and } E(y_t^2 \,|\mathcal{F}_{t-1}) \equiv \sigma_t^2, \tag{9}$$

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 following population least squares criterion function

$$S(\theta, m) = E\left[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j})\}^2\right]. \tag{10}$$

This criterion is well defined when $E(y_t^4) < \infty$.

In the special case that $\psi_j(\theta) = \theta^{j-1}$ we can rewrite model (7) as a difference equation in the unobserved variance

$$\sigma_t^2 = \theta_0 \sigma_{t-1}^2 + m(y_{t-1}),\ t = 1, 2, \ldots, \tag{11}$$

and this model is consistent with a stationary GARCH(1,1) structure for the unobserved variance when the further restriction is satisfied:

$$m(y) = \gamma y^2 + \alpha$$

for some parameters α, γ. More generally, m is the 'news impact function' and determines the way in which the volatility is affected by shocks to y, while the parameter θ, through the coefficients $\psi_j(\theta)$, determines the persistence. Our model allows for general "news impact functions" including both symmetric and asymmetric functions, and so accommodates the leverage effect [Nelson (1991)].

Our model generalizes the model considered in Carroll, Mammen, and Härdle (2001) 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. 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 this model (7) 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.[1] Finally, we should mention some work by Audrino and Bühlmann (2001): their 'model'

[1] Wu and Ziao (2002) investigate this model too, but they used data on the implied volatility from option prices, which means they can estimate the function m by standard partial linear regression.

includes ours as a special case.[2] However, although they proposed an estimation algorithm, they did not establish even consistency of the 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 $m(.)$ even when θ is known. The kernel method likes to express the function of interest as a conditional expectation 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 form a set of sufficient statistics for our model.[3] We use this relationship to generate our estimator.

2.1 Linear Characterization

Suppose for pedagogic purposes that the semi-strong process defined in (9) holds. Take marginal expectations for any $j \geq 1$

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

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

$$\begin{aligned} \sum_{j=1}^{\infty} \psi_j(\theta_0) E(y_t^2 \,|y_{t-j} = y) \;\; &= \;\; \mu \sum_{j=1}^{\infty} \psi_j(\theta_0) + \sum_{j=1}^{\infty} \psi_j^2(\theta_0) m(y) \\ &\quad + \sum_{j=1}^{\infty} \psi_j(\theta_0) \sum_{k \neq j}^{\infty} \psi_k(\theta_0) E[m(y_{t-k}) | y_{t-j} = y], \end{aligned} \tag{12}$$

which is another equation in $m(.)$. This equation arises as the first order condition from the least squares definition of σ_t^2, given in (10) as we now discuss. The quantities $\theta_0, m_0(.)$ are the unique minimizers of (10) over $\Theta \times \mathcal{M}$ by the definition of conditional expectation. Furthermore, the

[2] Their model is that $\sigma_t^2 = \Lambda(y_{t-1}, \sigma_{t-1}^2)$ for some smooth but unknown function $\Lambda(.)$.

Hafner (1998) and Carroll et al. (2002) have found evidence in support of the restriction that the news impact curve is similar across lags, which is implicit in our model.

[3] Hong and Li (2003) has recently proposed basing a test on a similar reduced class of distributions.

minimizer of (10) satisfies a first order condition and in the appendix we show that this first order condition is precisely (12). In fact, this equation also holds for any $\theta \in \Theta$ provided we replace m_0 by m_θ. Note that we are treating μ as a known quantity.

We next rewrite (12) for any given 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

$$\mathcal{H}_\theta(y,x) = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta) \frac{p_{0,j}(y,x)}{p_0(y)p_0(x)}, \tag{13}$$

$$m_\theta^*(y) = \sum_{j=1}^{\infty} \psi_j^\dagger(\theta) \left[g_j(y) - \mu \right], \tag{14}$$

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 $g_j(y) = E(y_0^2 | y_{-j} = y)$ for $j \geq 1$. Then the function $m_\theta(.)$ satisfies

$$m_\theta(y) = m_\theta^*(y) + \int \mathcal{H}_\theta(y,x) m_\theta(x) p_0(x) dx \tag{15}$$

for each $\theta \in \Theta$ [this equation is equivalent to (12) for all $\theta \in \Theta$]. The operator

$$\mathcal{H}_j(y,x) = \frac{p_{0,j}(y,x)}{p_0(y)p_0(x)}$$

is well studied in the literature [see Bickel et al. (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 the backfitting estimation of additive nonparametric regression, the corresponding integral operator is an unweighted sum of such kernels over the finite number of dimensions [see Mammen, Linton, and Nielsen (1999)]. In the fully independent case with $\psi_j(\theta) = \theta^{j-1}$ we have $\mathcal{H}_\theta(y,x) = -2\theta/(1-\theta)$.

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 (15) 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 (14) and (13). Nonparametric estimates of these functions only work accurately for arguments not too large. We do not want to enter into a discussion of tail behaviour of nonparametric estimates. For this reason we change our minimization problem (10), or rather restrict the parameter sets further. We consider minimization of (10) 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 $\sigma_t^2 = \mu$ 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 (10) for fixed θ over $m \in \mathcal{M}_c$ by $m_{\theta,c}$. Then θ_c and m_c minimize $E[\{y_t^2 - \mu - \sum_{j=1}^\infty \psi_j(\theta) m(y_{t-j})\}^2]$ over $\Theta \times \mathcal{M}_c$ and

$m_{\theta,c}$ minimizes $E[\{y_t^2 - \mu - \sum_{j=1}^{\infty} \psi_j(\theta) m_\theta(y_{t-j})\}^2]$ over $\mathcal{M}_c$. Estimates of m_0 can be constructed by estimating m_c and letting c converge to infinity.[4] In practice one might get better estimates if m_0 we fit nonparametrically inside $[-c, c]$ and parametrically outside this interval. In particular, μ could be fitted by a semiparametric approach. Motivated by traditional parametric GARCH models a more sophisticated parametric estimate for the tails would be a quadratic fit. We don't enter these discussions here and let c and μ be constant in the following. By the same arguments as above we get that $m_{\theta,c}$ 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 $|y| \leq c$ and vanishes for $|y| > c$. For simplicity but in abuse of notation we omit the subindex c of $m_{\theta,c}$ and we write

$$m_\theta = m_\theta^* + \mathcal{H}_\theta m_\theta. \tag{16}$$

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 (16) 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. Specifically, provided the Fredholm determinant of $\mathcal{H}_\theta$ is non-zero then there exists a unique solution given by the ratio of two infinite series in Fredholm determinants, see Tricomi (1957). See also Darolles, Florens, and Renault (2002) for a nice discussion on existence and uniqueness for type I equations.

We assume the following high level condition:

ASSUMPTION A1. The operator $\mathcal{H}_\theta(x, y)$ is Hilbert-Schmidt uniformly over θ, i.e.,

$$\sup_{\theta \in \Theta} \int \int \mathcal{H}_\theta(x, y)^2 p_0(x) p_0(y) dx dy < \infty.$$

A sufficient condition for A1 is that the joint densities $p_{0,j}(y, x)$ are uniformly bounded for $j \neq 0$ and $|x|, |y| \leq c$ and that the density $p_0(x)$ is bounded away from 0 for $|x| \leq c$. This condition can also be satisfied in certain unbounded cases. For example, when the process is stationary Gaussian provided that $\sup_{\theta \in \Theta} \sum_{j=1}^{\infty} \psi_j(\theta) < \infty$.

Under assumption A1, for each $\theta \in \Theta$, $\mathcal{H}_\theta$ is a self-adjoint bounded linear operator on the Hilbert space of functions $L_2(p_0)$. Also this condition implies that $\mathcal{H}_\theta$ is a compact operator and therefore has a countable number of eigenvalues[5]:

$$\infty > |\lambda_{\theta,1}| \geq |\lambda_{\theta,2}| \geq ....,$$

[4] It can be shown that $\lim_{c \to \infty} m_{\theta,c} = m_\theta$ in various ways.

[5] These are real numbers for which there exists functions $g_{\theta,j}(.)$ such that $\mathcal{H}_\theta g_{\theta,j} = \lambda_{\theta,j} g_{\theta,j}$.

with

$$\sup_{\theta\in\Theta}\sum_{j=1}^{\infty}\lambda_{\theta,j}^2<\infty. \tag{17}$$

ASSUMPTION A2. There exist no $\theta\in\Theta$ and $m\in\mathcal{M}_c$ with $\|m\|_2=1$ such that

$$\sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j})=0 \qquad \text{a.s.}$$

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.

ASSUMPTION A3. The operator $\mathcal{H}_\theta$ fulfills the following continuity condition for $\theta,\theta'\in\Theta$:

$$\sup_{\|m\|_2\le 1}\|\mathcal{H}_\theta m-\mathcal{H}_{\theta'}m\|_2\to 0 \qquad \text{for } |\theta-\theta'|\to 0.$$

This condition is straightforward to verify.

We now argue that because of (A2) and (A3) for a constant $0<\gamma<1$

$$\sup_{\theta\in\Theta}\lambda_{\theta,1}<\gamma. \tag{18}$$

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

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

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

$$\int m^2(x)p_0(x)dx-\lambda\int m^2(x)p_0(x)dx>0.$$

Therefore $\lambda_{\theta,j}<1$ for $\theta\in\Theta$ and $j\ge 1$. Now, because of (A3) and compactness of Θ, this implies (18).

From (18) we get that $I-\mathcal{H}_\theta$ has eigenvalues bounded from below by $1-\gamma>0$. Therefore $I-\mathcal{H}_\theta$ is invertible and $(I-\mathcal{H}_\theta)^{-1}$ has only positive eigenvalues that are bounded by $(1-\gamma)^{-1}$:

$$\sup_{\theta\in\Theta, m\in\mathcal{M}_c, \|m\|_2=1} \left\|(I-\mathcal{H}_\theta)^{-1}m\right\|_2 \leq (1-\gamma)^{-1}. \tag{19}$$

Therefore, we can directly solve the integral equation and write

$$m_\theta = (I-\mathcal{H}_\theta)^{-1}m^*_\theta \tag{20}$$

for each $\theta\in\Theta$.

The representation (20) is fundamental to our estimation strategy. We next discuss a further property that leads to an iterative solution method rather than a direct inversion. If it holds that

$$|\lambda_{\theta,1}|<1, \text{ then } m_\theta = \sum_{j=0}^{\infty}\mathcal{H}_\theta^j m^*_\theta.$$

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 to the truth from any starting point. This sort of property has been established in other related problems, see Mammen, Linton, and Nielsen (1999) and Linton, Mammen, Nielsen, and Tanggaard (2001), and is the basis of most estimation algorithms in this area.[6] 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 one in absolute value only when $\theta<1/3$. This implies that we will not be able to use directly the particular convenient method of successive approximations [aka backfitting] for estimation. However, one can apply this solution method after first transforming the integral equation. Define

$$\begin{aligned}\nu &= \min\{j:|\lambda_j|<1\}\\ \pi_\nu &= L_2 \text{ projection onto } \text{span}(e_1,\ldots,e_{v-1}),\end{aligned}$$

where e_j is the eigenfunction corresponding to λ_j. Then

$$m = m^* + \mathcal{H}(I-\pi_\nu)m + \mathcal{H}\pi_\nu m,$$

[6] The standard "Hastie and Tibhsirani backfitting' approach to estimation here would be to substitute empirical versions in equation (12) and iteratively update. In this method you are estimating for example $E[m(y_{t-k})|y_{t-j}=y]$ rather than the operator that produces it, which makes it slightly different from the Mammen, Linton, and Nielsen (1999) 'Smooth backfitting approach'.

which is equivalent to

$$m = m^*_\pi + \mathcal{H}_\pi m, \tag{21}$$

where $m^*_\pi = (I - \mathcal{H}\pi_\nu)^{-1} m^*$ and $\mathcal{H}_\pi = (I - \mathcal{H}\pi_\nu)^{-1}\mathcal{H}(I - \pi_\nu)$. It is easy to check that $||\mathcal{H}_\pi|| < 1$, and so the method of successive approximations for example can be applied to the transformed equation.

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.

We now suppose that $m_0(.), \theta_0$ are defined as the minimizers of the criterion function

$$\ell(\theta, m) = E\left[\log \sigma_t^2(\theta, m) + \frac{y_t^2}{\sigma_t^2(\theta, m)}\right] \tag{22}$$

with respect to both $\theta, m(.)$, where $\sigma_t^2(\theta, m) = \mu + \sum_{j=1}^\infty \psi_j(\theta) m(y_{t-j})$. Notice that this criterion is well defined [i.e., the expectation is finite] in many cases where the quadratic loss function is not defined because say $E(y_t^4) = \infty$.

Minimizing (22) with respect to m for each given θ leads to the nonlinear integral equation for m

$$\sum_{j=1}^\infty \psi_j(\theta) E\left[\frac{1}{\sigma_t^2(\theta, m)} | y_{t-j} = y\right] = \sum_{j=1}^\infty \psi_j(\theta) E\left[\frac{y_t^2}{\sigma_t^4(\theta, m)} | y_{t-j} = y\right]. \tag{23}$$

This equation is difficult to work with from the point of view of statistical analysis. We consider instead a linearized version of this equation. Suppose that we have some initial value (or approximation) to σ_t^2, then linearizing (23) about σ_t^2, we obtain the linear integral equation

$$\overline{m}_\theta = \overline{m}^*_\theta + \mathcal{H}^w_\theta \overline{m}_\theta, \tag{24}$$

$$\overline{m}^*_\theta = \frac{\sum_{j=1}^\infty \psi_j(\theta) E\left[\sigma_t^{-4} y_t^2 | y_{t-j} = y\right]}{\sum_{j=1}^\infty \psi_j^2(\theta) E\left[\sigma_t^{-4} | y_{t-j} = y\right]} \quad ; \quad \mathcal{H}^w_\theta(x, y) = -\sum_{\substack{j=1 \\ l \neq j}}^\infty \sum_{l=1}^\infty \psi_j(\theta)\psi_l(\theta) w_{j,l}(x, y) \frac{p_{0,l-j}(x, y)}{p_0(y)}$$

$$w_{j,l}(x, y) = \frac{E[\sigma_t^{-4} | y_{t-l} = x, y_{t-j} = y]}{\sum_{j=1}^\infty \psi_j^2(\theta) E[\sigma_t^{-4} | y_{t-j} = y]}.$$

This is a second kind linear integral equation in $\overline{m}_\theta(.)$ but with a different intercept and operator from (16). See Hastie and Tibshirani (1990, Section 6.5) for a similar calculation. Under our assumptions,

see B4 below, the weighted operator satisfies assumptions A1 and A3 also. For a proof of A3 note that

$$0 < E\left[\sigma_t^{-4}\sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j})\right]^2.$$

Note that in generally $\overline{m}_\theta$ differs from m_θ . They are defined as minimizers for different functionals. However, for the strong and semistrong versions of our model we get $\overline{m}_{\theta_0} = m_{\theta_0}$. We also write $\overline{\sigma}_t^2(\theta) = \mu + \sum_{j=1}^{\infty}\psi_j(\theta)\overline{m}_\theta(y_{t-j})$. Compare with $\sigma_t^2(\theta) = \mu + \sum_{j=1}^{\infty}\psi_j(\theta)m_\theta(y_{t-j})$.

2.3 Efficiency Bound for θ

We now turn to a discussion about some properties of θ. Specifically, we discuss the semiparametric efficiency bound for estimation of θ in the strong ARCH model when m is unknown in the case where y_t/σ_t is iid normal. This discussion is indirectly related to the characterizations of m_θ that we have obtained.

Suppose that m is a known function but the parameter θ is unknown, i.e., we have a specific parametric model. The log likelihood function is proportional to

$$\ell_T(\theta) = \frac{1}{2}\sum_{t=1}^{T}\log s_t^2(\theta) + \frac{y_t^2}{s_t^2(\theta)}, \text{ where } s_t^2(\theta) = \sum_{j=1}^{\infty}\psi_j(\theta)m(y_{t-j}).$$

The score function with respect to θ is

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

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 here (when m is known) is

$$\mathcal{I}_{\theta\theta}^{-1} = 2\left(E\left[\left(\frac{\partial\log\sigma_t^2}{\partial\theta}\right)^2\right]\right)^{-1},$$

since $E(u_t^2) = 2$. See Bollerslev and Wooldridge (1992).

Now suppose that we parameterize m by η and write m_η, so that we have a parametric model with parameters (θ, η). The score with respect to η is

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

The efficient score function [see Bickel et al. (1993, pp)] is the projection of $\partial \ell_T(\theta,\eta)/\partial\theta$ onto the orthocomplement of span$[\partial \ell_T(\theta,\eta)/\partial\eta]$; this is a linear combination of $\partial \ell_T(\theta,\eta)/\partial\theta, \partial \ell_T(\theta,\eta)/\partial\eta$ and has variance less than $\partial \ell_T(\theta,\eta)/\partial\theta$ reflecting the cost of the nuisance parameter. Now consider the semiparametric case. We have to compute the efficient score functions for all such parameterizations of m. Because of the definition of the process σ_t^2 the set of possible score functions with respect to m is

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

where we have evaluated at the true parameters. To find the efficient score function in the semiparametric model we have to find the projection of $\partial \ell_T(\theta,\eta)/\partial\theta$ onto the orthocomplement of $\mathcal{S}_m$. We seek a function g_0 that minimizes

$$E\left[\left\{ \frac{1}{s_t^2}\frac{\partial s_t^2}{\partial\theta} - \frac{1}{s_t^2}\sum_{j=1}^{\infty}\psi_j(\theta) g(y_{t-j}) \right\}^2 \right] \tag{25}$$

over all measurable g. This minimization problem is similar to that which m_θ solves. In particular, we can show that g_0 satisfies the linear integral equation (see appendix for details)

$$g_0 = g^* + \mathcal{H}_{\theta_0}^w g_0, \tag{26}$$

where the operator $\mathcal{H}_{\theta_0}^w$ was defined below (24), while

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

Note that the integral equation (26) is similar to (24) except that the intercept function g^* is different from m_θ^*.

Denote the implied least squares predictor in (25) as

$$\frac{1}{s_t^2}\sum_{j=1}^{\infty}\psi_j(\theta) g_0(y_{t-j}) = \frac{1}{s_t^2}\sum_{j=1}^{\infty}\psi_j(\theta)(I-\mathcal{H}_{\theta_0}^w)^{-1} g^*(y_{t-j}) \equiv \mathcal{P}_m \frac{\partial \log s_t^2}{\partial\theta}. \tag{27}$$

We assume that this predictor of $\partial \log s_t^2/\partial\theta$ is imperfect, in the sense that the residual variance in

(25) is positive. The efficient score function in the semiparametric model is thus

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

By construction this score function is orthogonal to any element of $\mathcal{S}_m$. The semiparametric efficiency bound is

$$\mathcal{I}^{*-1}_{\theta\theta} = 2\left(E\left[\left(\frac{\partial \log s_t^2}{\partial \theta} - \mathcal{P}_m \frac{\partial \log s_t^2}{\partial \theta}\right)^2\right]\right)^{-1}, \tag{28}$$

i.e., any regular estimator of θ in this semiparametric model has asymptotic variance not less than $\mathcal{I}^{*-1}_{\theta\theta}$. This bound is clearly larger than in the case where m is known. We will construct an estimator that achieves this semiparametric efficiency bound. It can be easily checked that

$$\frac{\partial \log \overline{\sigma}_t^2}{\partial \theta} = \frac{\partial \log s_t^2}{\partial \theta} - \mathcal{P}_m \frac{\partial \log s_t^2}{\partial \theta}.$$

3 Estimation

We shall construct estimates of θ and m from a sample $\{y_1, \ldots, y_T\}$. We proceed in four steps. First, for each given θ we estimate m_θ by solving an empirical version of the integral equation (16). We then estimate θ by maximizing a profile least squares criterion. We then use the estimated parameter to give an estimator of m. The last step consists in solving an empirical version of the linearized likelihood implied integral equation (24), and doing a two step quasi-Newton method to update the parameter estimate.

3.1 Our Estimators of m^*_θ and $\mathcal{H}_\theta$

We now define local linear 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, 1993)). Define the estimators $(\widehat{g}_j(y), \widehat{g}'_j(y))$ of $(g_j(y), g'_j(y))$ as

the minimizers of the weighted sums of squares criterion

$$(\widehat{g}_j(y), \widehat{g}'_j(y)) = \arg\min_{\alpha,\beta} \sum_t \left\{ y_t^2 - \mu - \alpha - \beta(y_{t-j} - y) \right\}^2 K_h\left(y_{t-j} - y\right), \tag{29}$$

where K is a symmetric probability density function, h is a positive bandwidth, and $K_h(.) = K(./h)/h$. The summation is over all t such that $1 \leq t - j \leq T$.

First select a truncation sequence τ_T with $1 < \tau_T < T$. Then compute

$$\widehat{m}^*_\theta(y) = \begin{cases} (1-\theta^2) \sum_{j=1}^{\tau_T} \psi_j(\theta) \widehat{g}_j(y) & \text{if } |y| \leq c \\ 0 & \text{else.} \end{cases}$$

To estimate $\mathcal{H}_\theta$ we take the following scheme

$$\begin{aligned} \widehat{\mathcal{H}}_\theta(y,x) &= -\sum_{|j|=1}^{\tau_T} \psi_j^*(\theta) \frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)\widehat{p}_0(x)}, \qquad (30) \\ \widehat{p}_{0,j}(y,x) &= \frac{1}{T-|j|} \sum_{t=\max\{1,j\}}^{\min\{T,T+j\}} K_h(y - y_t) K_h(x - y_{t+j}), \\ \widehat{p}_0(x) &= \frac{1}{T} \sum_{t=1}^{T} K_h(x - y_t). \end{aligned}$$

We define

$$\widehat{\mathcal{H}}_\theta m = \int_{-c}^{c} \widehat{\mathcal{H}}_\theta(y,x) m(x) \widehat{p}_0(x) dx. \tag{31}$$

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 $\|m\|_2^2 = \int_{-c}^{c} m(x)^2 \widehat{p}_0(x) dx$. Note that when $\theta = 0$, the operator $\widehat{\mathcal{H}}_\theta(y,x) = 0$ and $\widehat{m}_\theta$ is the corresponding kernel regression smoother.

Suppose that the sequence $\{\widehat{\sigma}_t^2,\ t = 1, \ldots, T\}$ and θ are given. Then define $\widehat{g}_j^a(\cdot)$ to be the local linear smooth of $\widehat{\sigma}_t^{-4} y_t^2$ 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}), with the population quantities defined correspondingly. Then define

$$\widehat{\overline{m}}^*_\theta = \frac{\sum_{j=1}^{\tau_T} \psi_j(\theta) \widehat{g}_j^a(y)}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta) \widehat{g}_j^b(y)} \quad ; \quad \widehat{\mathcal{H}}_\theta^{\widehat{w}_\theta}(x,y) = -\sum_{j=1}^{\tau_T} \sum_{\substack{l=1 \\ l \neq j}}^{\tau_T} \psi_j(\theta)\psi_l(\theta) \widehat{w}_{j,l}^\theta(x,y) \frac{\widehat{p}_{0,l-j}(x,y)}{\widehat{p}_0(x)\widehat{p}_0(y)},$$

where

$$\widehat{w}_{j,l}^\theta(x,y) = \frac{\widehat{g}_{l,j}^c(x,y)}{\sum_{j=1}^{\tau_T} \psi_j^2(\theta) \widehat{g}_j^b(y)}.$$

3.2 Our Estimators of θ and m

STEP 1. Define $\widehat{m}_\theta$ as any solution of the equation

$$\widehat{m}_\theta = \widehat{m}^*_\theta + \widehat{\mathcal{H}}_\theta \widehat{m}_\theta. \tag{32}$$

This step is the most difficult and requires a number of choices. In practice, we are going to solve the integral equation on a grid of points, which reduces it to a large linear system. As the number of grid points increases the approximation becomes arbitrarily good. See below for more discussion.

STEP 2. We next choose $\theta \in \Theta$ to minimize the following criterion

$$\widehat{S}_T(\theta) = \frac{1}{T} \sum_{t=\tau_T+1}^{T} \left\{ y_t^2 - \widehat{\sigma}_t^2(\theta) \right\}^2, \text{ where } \widehat{\sigma}_t^2(\theta) = \sum_{j=1}^{\tau_T} \psi_j(\theta) \widehat{m}_\theta(y_{t-j}).$$

When θ is one dimensional this optimization can be done by grid search since θ is a scalar and lies in a compact set.

STEP 3. Define for any $y \in [-c, c]$ and $t \geq \tau_T + 1$:

$$\widehat{m}(y) = \widehat{m}_{\widehat{\theta}}(y) \text{ and } \widehat{\sigma}_t^2 = \sum_{j=1}^{\tau_T} \psi_j(\widehat{\theta}) \widehat{m}(y_{t-j}).$$

The estimates $(\widehat{m}(y), \widehat{\theta})$ 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 are proposed.

STEP 4. Given $(\widehat{\theta}, \widehat{m}(.))$. Compute $\widehat{\overline{m}}^*_\theta$ and $\widehat{\mathcal{H}}^{\widehat{w}_\theta}_\theta$ using the sequence $\{\widehat{\sigma}_t^2,\ t = 1, \ldots, T\}$ defined above, then solve the linear integral equation

$$\widetilde{m}_\theta = \widehat{\overline{m}}^*_\theta + \widehat{\mathcal{H}}^{\widehat{w}_\theta}_\theta \widetilde{m}_\theta \tag{33}$$

for the estimator $\widetilde{m}_\theta$, and let $\widetilde{\sigma}_t^2(\theta) = \sum_{j=1}^{\tau_T} \psi_j(\theta)\ \widetilde{m}_\theta(y_{t-j})$ for each θ. Next define $\widetilde{\theta}$ as any minimizer of

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

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 $\widehat{\theta}$. Compute $\widetilde{m}(y) = \widetilde{m}_{\widetilde{\theta}}(y)$ and $\widetilde{\sigma}_t^2 = \sum_{j=1}^{\tau_T} \psi_j(\widetilde{\theta}) \widetilde{m}(y_{t-j})$.

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

3.2.1 Solution of Integral Equations

There are many approaches to computing the solutions, see for example Riesz and Nagy (1990). Rust (2000) gives a nice discussion about solution methods for a more general class of problems. The two issues are: how to approximate the integral in (31), 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 $\{t_{j,n}, j = 1, \ldots, n\}$ be some grid of points in $[-c, c]$ and $w_{j,n}$ 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 $J_n(f) = \sum_{j=1}^{n} w_{j,n} f(t_{j,n})$. For example, Simpson's rule or Gaussian Quadrature both satisfy this. Now approximate (32) by

$$\widehat{m}_\theta(x) = \widehat{m}^*_\theta(x) + \sum_{j=1}^{n} w_{j,n} \widehat{\mathcal{H}}_\theta(x, t_{j,n}) \widehat{m}_\theta(t_{j,n}) \widehat{p}_0(t_{j,n}). \tag{34}$$

This is equivalent to the linear system [Atkinson (1976)]

$$\widehat{m}_\theta(t_{i,n}) = \widehat{m}^*_\theta(t_{i,n}) + \sum_{j=1}^{n} w_{j,n} \widehat{\mathcal{H}}_\theta(t_{i,n}, t_{j,n}) \widehat{m}_\theta(t_{j,n}) \widehat{p}_0(t_{j,n}), \ i = 1, \ldots, n. \tag{35}$$

To each solution of equation (35) there is a unique corresponding solution of (34) with which it agrees at the node points. Under smoothness conditions on $\widehat{\mathcal{H}}_\theta$, the solution of the system (35) converges in L_2 to the solution of (32) as $n \to \infty$, and at a geometric rate. The linear system can be written in matrix notation

$$(I_n - \widehat{B}_\theta)\widehat{m}_\theta = \widehat{m}^*_\theta,$$

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

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

is an $n \times n$ matrix. We then find the solution values $\widehat{m}_\theta = (\widehat{m}_\theta(y_1), \ldots, \widehat{m}_\theta(y_n))'$ to this system. Note that once we have found $\widehat{m}_\theta(t_{j,n})$, $j = 1, \ldots, n$, we can substitute back into (34) to obtain $\widehat{m}_\theta(x)$ for any $x \in [-c, c]$. More sophisticated methods also involve selection of the grid size n and scheme.

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 $n = 1000$, and so long as we do not require too much accuracy in the computation of θ. For larger problems, iterative methods are indispensable. We next describe the sort of iterative approaches that we have tried.

When $\widehat{\mathcal{H}}_\theta$ has operator norm strictly smaller than one, one can directly apply a version of the Backfitting/Gauss-Seidel/Successive approximation method of Hastie and Tibshirani (1990) or Mammen, Linton, and Nielsen (1999). However, as we pointed out already the operator $\widehat{\mathcal{H}}_\theta$ only satisfies this condition for a small subset of θ values. Instead, it is necessary to modify the algorithm along the line discussed in (21), see also Hastie and Tibshirani (1990, Section 5.2). Below we describe a simple version of this, called 'convergent splitting' in the numerical analysis literature. We factorize

$$I_n - \widehat{B}_\theta = C_\theta - R_\theta \tag{36}$$

where the matrices C_θ and R_θ are chosen to satisfy

$$\rho(C_\theta^{-1} R_\theta) < 1,$$

where ρ denotes spectral radius.[7] Then from starting value $\widehat{m}_\theta^{[0]}$, compute the iteration

$$\widehat{m}_\theta^{[r+1]} = C_\theta^{-1} R_\theta \widehat{m}_\theta^{[r]} + C_\theta^{-1} \widehat{m}_\theta^*$$

until convergence. In practice, we have chosen $C_\theta = 2\theta/(1-\theta) I_n$ and found good results. The iteration is continued until some convergence criterion is satisfied. For example, one can stop when

$$||\widehat{m}_\theta^{[r+1]} - \widehat{m}_\theta^{[r]}|| < \epsilon \text{ or } ||(I_n - \widehat{B}_\theta)\widehat{m}_\theta - \widehat{m}_\theta^*|| < \epsilon$$

for some small ϵ. Here, $||x|| = (x'x)^{1/2}$ is the Euclidean norm on vectors in $\mathbb{R}^n$.

4 Asymptotic Properties

4.1 Regularity Conditions

We will discuss properties of the estimates $\widehat{m}_\theta$ and $\widehat{\theta}$ under the *weak* form where we do not assume that (9) holds but where θ_0, m_0 are defined as the minimizers of the least squares criterion function (10). 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 (9) holds.

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

$$\eta_{\theta,t}^1 = \sum_{j=1}^{\infty} \psi_j^\dagger(\theta)\eta_{j,t} \text{ and } \eta_{\theta,t}^2 = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta)\zeta_{j,t}(\theta). \tag{37}$$

[7] The spectral radius of a square symmetric matrix is the largest (in absolute value) eigenvalue.

Let $\alpha(k)$ be the strong mixing coefficient of $\{y_t\}$ defined as

$$\alpha(k) \equiv \sup_{A \in \mathcal{F}_{-\infty}^{0},\, B \in \mathcal{F}_{k}^{\infty}} |P(A \cap B) - P(A)\,P(B)|, \tag{38}$$

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 and alpha 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 $E\left(|y_t|^{2\rho}\right) < \infty$ *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.* $p_0(\cdot)$ *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.*

$$\sigma_t^2 > c_\sigma. \tag{39}$$

B5 *The conditional distribution functions of* $\eta_{\theta,t}^1$ *and* $\eta_{\theta,t}^2$ *given* $y_t = u$ *are continuous at the point* y.

B6 $E\left[\left(\eta_{\theta,1}^j\right)^2 + \left(\eta_{\theta,t}^j\right)^2 |y_1 = u_1, y_t = u_2\right]$ *is uniformly bounded for* $j = 1, 2$, $t \geq 1$ *and* u_1 *and* u_2 *in neighborhoods of* y.

B7 *The parameter* θ *is contained in the compact set* $\Theta \subset \mathbb{R}^p$. *Also, 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* $\sup_{\theta\in\Theta,k=0,1,2}|\partial^k\psi_j(\theta)/\partial\theta^k| \leq j^p\overline{\psi}^j$ *for some finite* p *and* $\overline{\psi}<1$, *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 (9).

B11 *The semistrong model assumption (9) 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* $\varepsilon_t = y_t/\sigma_t$, $u_t = (y_t^2-\sigma_t^2)/\sigma_t^2$, *which are also both martingale difference sequences by assumption.*

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 such processes to be strongly stationary and $\beta-$mixing; these conditions involve restrictions on the function m_0 and the distribution of the innovations, in addition to restrictions on θ_0. Conditions B3,B4 are quite standard assumptions in the nonparametric regression literature. Under the assumption of (9), the bound (39) follows if we assume that $\inf_{-c\leq w\leq c}m(w) > -\mu/\sum_{j=1}^{\infty}\psi_j$. Conditions B5, B6 are used to apply the central limit theorem of Masry and Fan (1997) and can be replaced by more primitive conditions. Assumption B7 is for the identification of the parametric part. Following Hannan (1973) it is usual to impose these high level conditions [c.f. his condition (4)]. 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 $\partial^k\psi_j(\theta)/\partial\theta^k = (j-1)\cdots(j-k)\theta^{j-k-1}$.

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 (8) holds, but define m_θ as the minimizer of (10) over $\mathcal{M}_c$.

Define the functions $\beta_\theta^j(y)$ as solutions to the integral equations

$$\beta_\theta^j = \beta_\theta^{*,j}(y) + \mathcal{H}_\theta\beta_\theta^j,\ j=1,2,$$

in which:

$$\beta_\theta^{*,1}(y) = \frac{\partial^2}{\partial y^2}m_\theta^*(y),$$

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

where $\nabla_2 = (\partial^2/\partial x^2) + 2(\partial^2/\partial x \partial y) + \partial^2/\partial y^2$ is the Laplacian operator. Then define $\mu_\theta(y) = -\sum_{j=\pm 1}^{\pm\infty} \psi_j^*(\theta) E[m_\theta(y_{t+j})|y_t = y]$, and

$$\omega_\theta(y) = \frac{\nu_0(K)}{p_0(y)} \left\{ \text{var}[\eta_{\theta,t}^1 + \eta_{\theta,t}^2] + \mu_\theta^2(y) \right\} \tag{40}$$

$$b_\theta(y) = \frac{1}{2}\mu_2(K) \left[\beta_\theta^1(y) + \beta_\theta^2(y) \right], \tag{41}$$

where $\eta_{\theta,t}^j$, $j = 1, 2$ were defined in (37). We prove the following theorem in the appendix.

THEOREM 1. *Suppose that B1-B9 hold. Then for each $\theta \in \Theta$ and $y \in [-c, c]$*

$$\sqrt{Th} \left[\widehat{m}_\theta(y) - m_\theta(y) - h^2 b_\theta(y) \right] \Longrightarrow N\left(0, \omega_\theta(y)\right). \tag{42}$$

Furthermore,

$$\sup_{\theta \in \Theta, |y| \le c} |\widehat{m}_\theta(y) - m_\theta(y)| = o_p(T^{-1/4}), \tag{43}$$

$$\sup_{\theta \in \Theta, \tau_T \le t \le T} \left| \widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta) \right| = o_p(T^{-1/4}), \tag{44}$$

$$\sup_{\theta \in \Theta, \tau_T \le t \le T} \left| \frac{\partial \widehat{\sigma}_t^2}{\partial \theta}(\theta) - \frac{\partial \sigma_t^2}{\partial \theta}(\theta) \right| = o_p(T^{-1/4}), \tag{45}$$

From this result we obtain the properties of $\widehat{\theta}$ by an application of the 'standard' asymptotic theory for semiparametric estimators [see for example Bickel, Klaassen, Ritov, and Wellner (1993)]; this requires a uniform expansion for $\widehat{m}_\theta(y)$ and some similar properties on the derivatives (with respect to θ) of $\widehat{m}_\theta(y)$.

THEOREM 2. *Suppose that B1-B10 hold. Then*

$$\sqrt{T}(\widehat{\theta} - \theta_0) = O_p(1). \tag{46}$$

These results can be applied to get the asymptotic distribution of $\widehat{m} = \widehat{m}_{\widehat{\theta}}$.

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 $y \in [-c, c]$*

$$\sqrt{Th} \left[\widehat{m}_{\widehat{\theta}}(y) - \widehat{m}_{\theta_0}(y) \right] = o_p(1). \tag{47}$$

Under the additional assumption of B11 we get that

$$\sqrt{Th}\left[\widehat{m}_{\widehat{\theta}}(y) - m_{\theta_0}(y) - h^2 b(y)\right] \Longrightarrow N\left(0, \omega(y)\right), \tag{48}$$

where

$$\omega(y) = \frac{\nu_0(K)}{p_0(y)\left[\sum_{j=1}^{\infty}\psi_j^2(\theta_0)\right]^2}\sum_{j=1}^{\infty}\psi_j^2(\theta_0)E\left[\sigma_t^4 u_t^2 | y_{t-j} = y\right] \tag{49}$$

and

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

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 $\widehat{m}_\theta^{mod}$ is defined as any solution of

$$\widehat{m}_\theta^{mod} = \widehat{m}_\theta^* + \widehat{\mathcal{H}}_\theta^{mod}\widehat{m}_\theta^{mod},$$

where the operator $\widehat{\mathcal{H}}^{mod}$ is defined by use of modified kernel density estimates

$$\begin{aligned}
\widehat{\mathcal{H}}_\theta^{mod}(y,x) &= -\sum_{j=\pm 1}^{\pm\tau_T}\psi_j^*(\theta)\frac{\widehat{p}_{0,j}^{mod}(y,x)}{\widehat{p}_0^{mod}(y)\widehat{p}_0(x)}, \\
\widehat{p}_{0,j}^{mod}(y,x) &= \widehat{p}_{0,j}(x,y) + \frac{\widehat{p}_0'(x)}{\widehat{p}_0(x)}\frac{1}{T-|j|}\sum_t (y_t - y)K_h(y_t - y)K_h(y_{t+j} - x), \\
\widehat{p}_0^{mod}(x) &= \widehat{p}_0(x) + \frac{\widehat{p}_0'(x)}{\widehat{p}_0(x)}\frac{1}{T}\sum_{t=1}^{T}(y_t - y)K_h(y_t - y).
\end{aligned}$$

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. $\frac{1}{T}\sum_{t=1}^{T}(y_t - y)K_h(y_t - y)/[h^2\mu_2(K)]$. 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 $y \in [-c,c]$*

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

where $\omega(y)$ is defined as in Theorem 3 and where

$$b^{mod}(y) = \frac{1}{2}\mu_2(K)m''(y).$$

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 (9) 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 $\varepsilon_t = y_t/\sigma_t$, $u_t = (y_t^2 - \sigma_t^2)/\sigma_t^2$, which are also both martingale difference sequences by assumption. We suppose for simplicity that ε_t has a time invariant kurtosis κ_4.

B12 *The variables ε_t have a time invariant conditional kurtosis κ_4*

$$\mathrm{E}\left[u_t^2 | y_{t-j} = y\right] = \kappa_4 + 2.$$

Define

$$\omega^{eff}(y) = \frac{\nu_0(K)(\kappa_4 + 2)}{p_0(y) \sum_{j=1}^{\infty} \psi_j^2(\theta) E\left(\sigma_t^{-4} | y_{t-j} = y\right)} \tag{50}$$

THEOREM 5. *Suppose that B1-B12 hold. For some bounded continuous function $b^{eff}(y)$ we have*

$$\sqrt{Th}\left[\widetilde{m}_{\widehat{\theta}}(y) - m_{\widehat{\theta}}(y) - h^2 b^{eff}(y)\right] \Longrightarrow N\left(0, \omega^{eff}(y)\right).$$

The next theorem discuss the asymptotic distribution of $\widetilde{\theta}$.

THEOREM 6. *Suppose that B1-B12 hold. Then*

$$\sqrt{T}(\widetilde{\theta} - \theta_0) \Longrightarrow N(0, V), \quad \text{where } V = (\kappa_4 + 2)\left(E\left[\left(\frac{\partial \log s_t^2}{\partial \theta} - \mathcal{P}_m \frac{\partial \log s_t^2}{\partial \theta}\right)^2\right]\right)^{-1}.$$

Thus when the errors are Gaussian $\widetilde{\theta}$ achieves the semiparametric efficiency bound. When B12 does not hold, asymptotic normality can still be shown but the limiting distribution has a more complicated sandwich form. Standard errors robust to departures from B12 can be constructed from the representation (94) given in the appendix.

4.4 Nonparametric Efficiency

Here, we discuss the issue about efficiency of the nonparametric estimators. Our discussion is heuristic and is confined to the *semistrong* case and to comparison of asymptotic variances. This type of analysis has been carried out before in many separable models, see Linton (1996,2000); it sets out a standard of efficiency and a strategy for achieving it and hence improving on the given method.

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.

We first compare the asymptotic variance of $\widehat{m}_{\widehat{\theta}}$ and $\widehat{m}_{\widehat{\theta}}^{mod}$ with the variance of an infeasible estimator that is based on a certain least squares criterion. Let

$$S_j(\lambda) = \frac{1}{Th} \sum_t K\left(\frac{y - y_{t-j}}{h}\right) \left[y_t^2 - \sigma_{t;j}^2(\lambda)\right]^2, \tag{51}$$

where $\sigma_{t;j}^2(\lambda) = \sum_{\substack{k=1 \\ k \neq j}}^{\tau_T} \psi_k(\theta) m(y_{t-k}) + \psi_j(\theta)\lambda$, and define $\widetilde{m}_j(y) = \widetilde{\lambda}_j = \arg\max_\lambda S_j(\lambda)$. This least squares estimator is infeasible since it requires knowledge of m at $\{y_{t-k},\, k \neq j\}$ points. It can easily be shown that

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

for all $j = 1, 2, \ldots$ with some appropriately chosen bias terms b_j. Now define a class of such 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

$$\sqrt{Th}[\widetilde{m}_{opt}(y) - m(y) - h^2 b(y)] \Longrightarrow N\left(0, \frac{\nu_0(K)}{p_0(y) \sum_{j=1}^\infty \psi_j^2(\theta) \left[\mathrm{E}\left(\sigma_t^4 u_t^2 | y_{t-j} = y\right)\right]^{-1}}\right)$$

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

Define $s_j(y) = E\left(\sigma_t^4 u_t^2 | y_{t-j} = y\right)$. 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

$$\left(\sum_{j=1}^\infty \psi_j^2(\theta)\right)^{-2} \sum_{j=1}^\infty \psi_j^2(\theta) s_j(y) \geq \frac{1}{\sum_{j=1}^\infty \psi_j^2(\theta) s_j^{-1}(y)}$$

with equality only when $s_j(y)$ does not depend on j. So our estimate with variance (49) would achieve the asymptotic efficiency bound in case of constant conditional variances $s_j(y)$. It is inefficient in case of heteroscedasticity. Because our estimator is motivated by an unweighted least squares criterion it could not been expected that it corrects for heteroscedasticity. The asymptotic

efficiency of the estimator for homoscedasticity supports the power of our approach. For the case of heteroscedasticity we conjecture that one could improve the efficiency of our estimator along the lines of Linton (1996,2000), but we do not pursue this because the likelihood based procedure can be even more efficient.

Define analogously to (51) the (infeasible) local likelihoods

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

and let $\widetilde{m}_j(y) = \widetilde{\lambda}_j = \arg\max_\lambda \ell_j(\lambda)$. The properties of $\widetilde{m}_j(y)$ are easy to find. We have

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

Thus the optimal linear combination of $\widetilde{m}_j(y)$ has asymptotic variance

$$\frac{(\kappa_4+2)\nu_0(K)}{p_0(y)}\frac{1}{\sum_j \psi_j^2(\theta)E\left(\sigma_t^{-4}|y_{t-j}=y\right)}.$$

This is precisely the variance achieved by our weighted smooth backfitting estimator. In other words our estimator $\widetilde{m}_{\widehat{\theta}}(y)$ appears to be as efficient as it can be.[8]

4.5 Some Practical Issues

There remain some choices to be determined including the truncation parameter τ_T and the bandwidth or bandwidths used in smoothing. For the truncation parameter τ_T, in practice we use various selection criteria such as AIC and BIC. If the true model has a finite τ, the order selection based on the BIC criterion is consistent and thus might be preferred. However, if the true model is not of finite order, AIC may be preferred since it leads to asymptotically efficient choice of optimal order in the class of some projected infinite order processes. Define

$$RSS_T(\tau) = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T}\{y_t^2 - \widehat{\sigma}_t^2\}^2$$

[8] Note that

$$\sum_{j=1}^{\infty}\psi_j^2(\theta)\frac{1}{E\left(\sigma_t^4|y_{t-j}=y\right)} \le \sum_{j=1}^{\infty}\psi_j^2(\theta)E\left(\sigma_t^{-4}|y_{t-j}=y\right)$$

by the Cauchy-Schwarz inequality. It follows that the likelihood based estimator is superior to the least squares one according to asymptotic variance.

to be the residual sum of squares. Then the Akaike Information, Bayesian Information, and Hannan-Quinn model selection criteria are

$$\begin{aligned} AIC &= \log RSS_T(\tau) + \frac{2\tau}{T} \\ BIC &= \log RSS_T(\tau) + \frac{\tau \log T}{T} \\ HQ &= \log RSS_T(\tau) + \frac{2\tau \log\log T}{T}. \end{aligned}$$

For the bandwidth h, one objective is to choose h to minimize the integrated mean squared error of $\widehat{m}$ derived above. This can be done using simulation methods, but requires estimation of second derivatives of m and other quantities, so may not work well in practice. Instead we develop a rule of thumb bandwidth using the mean squared error implied by Theorem 4. If we take as pilot model that the process is GARCH(1,1), then the bias function is just $b^{mod}(y) = \mu_2(K)\gamma$. We propose the following automatic bandwidth

$$h_{ROT} = \left[\frac{c(1-\widehat{\theta}^2)\nu_0(K)m_4}{\mu_2^2(K)\widehat{\gamma}} \right]^{1/5} T^{-1/5},$$

where m_4 is the sample fourth moment, and $\widehat{\gamma}$ is the estimated parameter from a GARCH(1,1) model. This has no more justification than the Silverman's rule of thumb, but at least does reflect some aspects of the problem.

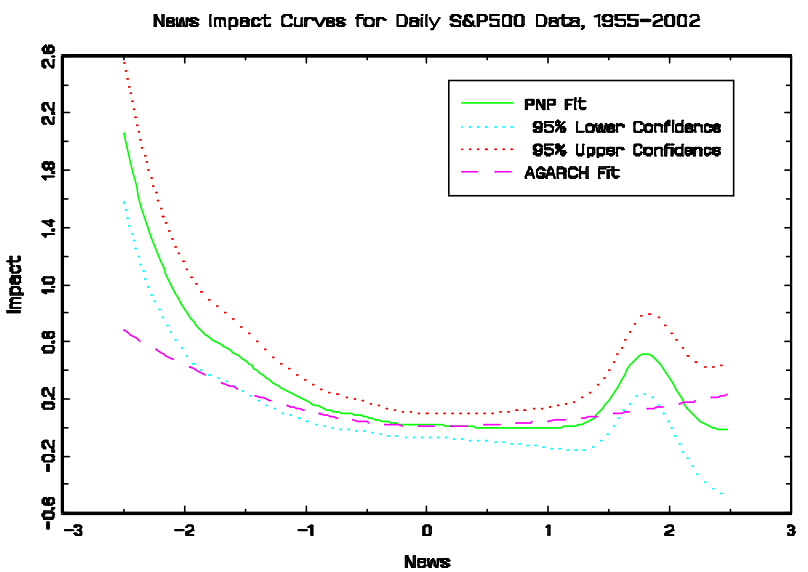
5 Numerical Results

5.1 Simulated Data

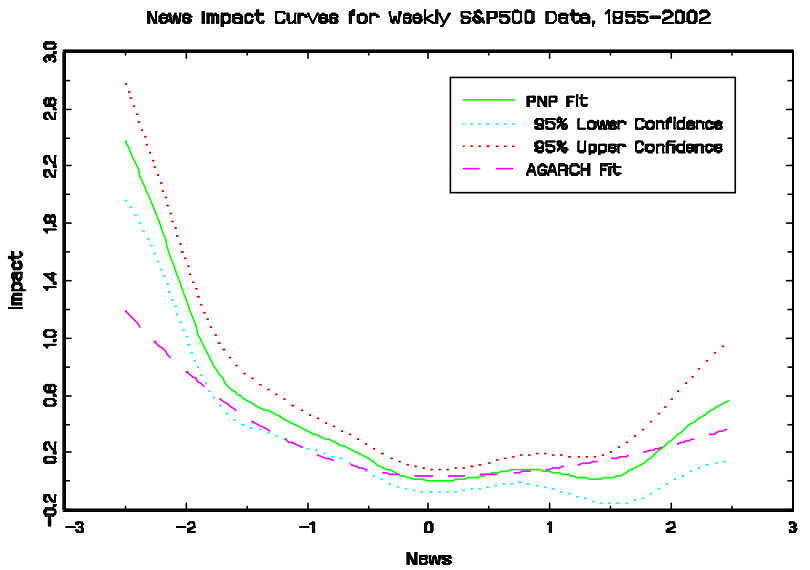
We report the results of a small simulation experiment on GARCH(1,1) data. Specifically, we generated data from (11), where $y_t = \varepsilon_t \sigma_t$ and ε_t is standard normal. We took the parameter values from a real dataset, in particular $\theta = 0.75$. We took sample sizes $T \in \{50, 100, 200\}$ and vary both the truncation parameter τ_T and the bandwidth h, or rather a bandwidth constant c_h that we multiply against h_{ROT}. The results are shown in Table 1. The performance of both $\widetilde{\theta}$ and $\widetilde{\sigma}_t^2$ improves with sample size.

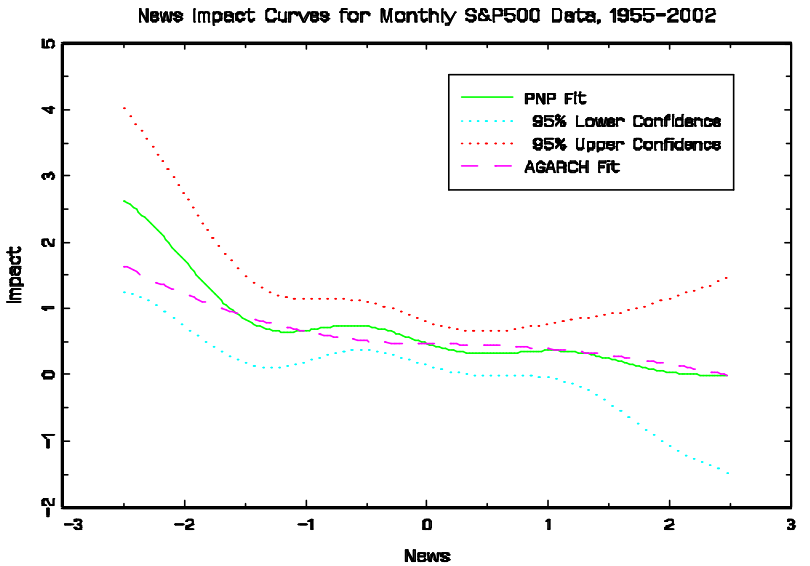
5.2 Investigation of the News Impact Curve in S&P500 Index Returns

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 samples of daily, weekly, and monthly returns on the S&P500 from 1955 to 2002, a total of 11,893, 2464, and 570 observations respectively. Following Engle and Ng (1993) we fitted regressions on seasonal dummies and lagged values, but, unlike them, found little significant effects other than the mean. Therefore, we work with the standardized return series. In table 3 we report the results of Asymmetric GARCH [AGARCH] parametric fits on these standardized series. There is quite strong evidence of asymmetry at all frequencies. We computed our estimators using $\tau = 50$ for daily data and $\tau = 20$ for weekly and monthly data. Our estimation was on the range [-2.5,2.5] and bandwidth selected by rule of thumb. In figures 1-3 we report the estimated news impact curve and its 95% confidence envelope along with the implied AGARCH curve for the three dataseries. The confidence intervals obviously widen at the edges but it is still clear that the news impact curve from the AGARCH fits deviate significantly from the nonparametric fits at least for the daily and weekly data. For the monthly data the AGARCH curve provides a reasonable fit.[9]



[9]Note that the confidence intervals get narrower for the larger sample sizes but not so much since the kurtosis in the daily data is very large.





6 Conclusions and Extensions

We have established the pointwise distribution theory of our least squares and likelihood-based nonparametric methods and have discussed the efficiency question. It is perhaps a weakness of our approach that we have relied on the least squares criterion to obtain consistency, as some may

be concerned about the existence of moments. However, in practice one can avoid least squares estimations altogether and just apply an iterated version of the likelihood based method. We expect that the distribution 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 to either apply our theory or to develop a theory for combining an increasing number of Horowitz and Mammen (2002) estimators.

In some datasets it may be important to allow some model for the mean of the process so that for example $y_t = \beta' x_t + \varepsilon_t \sigma_t$, where σ_t^2 is as in (11). In this case one has to apply our procedure to (parametric) residuals obtained by some preliminary estimation. This will certainly affect the parametric asymptotics, but should not affect the distributions for the nonparametric part.

We can also treat transformation models of the form

$$E(\chi(y_t;\lambda)\,|\mathcal{F}_{t-1}) = \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j}),$$

where χ is monotonic in y for each λ, for example the Box-Cox model $\chi(y;\lambda) = |y|^\lambda$, $\lambda \geq 0$. This would include the logarithmic and standard deviation specifications as well as many other cases. We can apply our estimation procedures to estimate the function m, for given λ, θ, and then choose λ, θ to maximize the implied profile likelihood. Under stronger conditions than this paper it is possible to identify both λ, θ. To construct the likelihood we would need to obtain σ_t. We can obtain the conditional variance process itself under some conditions. Suppose that $y_t = \sigma_t \varepsilon_t$ with ε_t iid. Then,

$$E(\chi(y_t;\lambda_0)\,|\mathcal{F}_{t-1}) = \int \chi(\sigma_t \varepsilon;\lambda_0) f_\varepsilon(\varepsilon) d\varepsilon = \Psi(\sigma_t),$$

where f_ε is the (known) density of ε. The function Ψ is monotonic and so $\sigma_t = \Psi^{-1}(\sum_{j=1}^{\infty} \psi_j(\theta)$ $m(y_{t-j}))$, which can then be plugged into an estimating equation for the parameters. In practice we would have to compute Ψ by numerical methods.

A Appendix

PROOF OF (12). It is convenient to break the joint optimization problem down in to two separate problems: first, for each $\theta \in \Theta$ let m_θ be the function that minimizes (10) 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

$$E\left[\{y_t^2 - \sum_{j=1}^{\infty} \psi_j(\theta) m_0(y_{t-j})\}\{\sum_{l=1}^{\infty} \psi_l(\theta) f(y_{t-l})\}\right] = 0,$$

which can be rewritten as

$$\sum_{j=1}^{\infty} \psi_j(\theta) E\left[y_0^2 f(y_{-j})\right] - \sum_{\substack{j=1 \\ j \neq l}}^{\infty} \sum_{l=1}^{\infty} \psi_j(\theta)\psi_l(\theta) E\left[m_0(y_{-j}) f(y_{-l})\right] = \sum_{j=1}^{\infty} \psi_j^2(\theta) E\left[m_0(y_{-j}) f(y_{-j})\right] \quad (52)$$

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

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

while

$$E\left[m_0(y_{-j}) f(y_{-j})\right] = \int m_0(y') \delta_y(y') p_0(y') dy' = m_0(y) p_0(y).$$

Finally,

$$\begin{aligned} E\left[m_0(y_{-j}) f(y_{-l})\right] &= E\left[E[m_0(y_{-j}) | y_{-l}] f(y_{-l})\right] \\ &= \int E[m_0(y_{-j}) | y_{-l} = y'] \delta_y(y') p_0(y') dy' \\ &= E[m_0(y_{-j}) | y_{-l} = y] p_0(y). \end{aligned}$$

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 $c(.)$ defined on the integers:

$$\sum_{\substack{j=1 \\ j \neq l}}^{\infty} \sum_{l=1}^{\infty} \psi_j(\theta)\psi_l(\theta) c(j-l) = \sum_{t=\pm 1}^{\infty} \sum_{l=1}^{\infty} \psi_{t+l}(\theta)\psi_l(\theta) c(t) = \sum_{t=\pm 1}^{\infty} \left(\sum_{l=1}^{\infty} \psi_{t+l}(\theta)\psi_l(\theta) \right) c(t). \quad (53)$$

Therefore, dividing through by $p_0(y)$ and $\sum_{j=1}^{\infty}\psi_j^2(\theta)$, (52) can be written

$$\sum_{j=1}^{\infty}\psi_j^{\dagger}(\theta)E(y_0^2|y_{-j}=y)-\sum_{j=\pm 1}^{\pm\infty}\psi_t^{*}(\theta)E(m_0(y_0)|y_j=y)=m_0(y), \tag{54}$$

which is the stated answer. ■

PROOF OF (26). 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

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

Now use the law of iterated expectations to write

$$E\left[\sigma_t^{-4}g_0(y_{t-j})|y_{t-l}=y\right]=E\left[E[\sigma_t^{-4}|y_{t-j},y_{t-l}]g_0(y_{t-j})|y_{t-l}=y\right].$$

Then

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

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

A.1 Proof of Theorem 1

A.1.1 Outline of Asymptotic Approach

We first outline the approach to obtaining the asymptotic properties of $\widehat{m}_\theta(.)$ for any $\theta\in\Theta$. We give some high level conditions A4-A6 below 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 conditions A4-A6 and verify the central limit theorem.

ASSUMPTION A4. Suppose that for a sequence $\delta_T\to 0$:

$$\sup_{\theta\in\Theta,\|m\|_2=1,|x|\leq c}\left|\widehat{\mathcal{H}}_\theta m(x)-\mathcal{H}_\theta m(x)\right|=o_p(\delta_T).$$

In particular (A4) gives that

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

We now show that by virtue of (A4) that $(I-\widehat{\mathcal{H}}_\theta)$ is invertible for all $\theta\in\Theta$, with probability tending to one, and it holds that (see also (19))

$$\sup_{\theta\in\Theta,\|m\|_2=1,|y|\leq c}\left|\left[\left(I-\widehat{\mathcal{H}}_\theta\right)^{-1}-(I-\mathcal{H}_\theta)^{-1}\right]m(y)\right|=o_p(\delta_T). \tag{55}$$

In particular,

$$\sup_{\theta\in\Theta,\|m\|_2=1}\left\|\left[\left(I-\widehat{\mathcal{H}}_\theta\right)^{-1}-(I-\mathcal{H}_\theta)^{-1}\right]m\right\|_2=o_p(\delta_T). \tag{56}$$

For a proof of claim (55) note that for $m\in\mathcal{M}_c$

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

because of

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

This gives

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

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 (A4)

$$\widehat{m}^*_\theta(y)-m^*_\theta(y)=\widehat{m}^{*,B}_\theta(y)+\widehat{m}^{*C}_\theta(y)+\widehat{m}^{*,D}_\theta(y), \tag{57}$$

where $\widehat{m}_\theta^{*,B}, \widehat{m}_\theta^{*C}$, and $\widehat{m}_\theta^{*,D}$ satisfy:

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}^{*,B}(y)\right| = O_p(T^{-2/5}) \text{ with } \widehat{m}^{*,B} \text{ deterministic} \tag{58}$$

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^{*,C}(y)\right| = O_p\left(T^{-2/5}\delta_T^{-1}\right) \tag{59}$$

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\mathcal{H}_\theta\left(I-\mathcal{H}_\theta\right)^{-1}\widehat{m}_\theta^{*,C}(y)\right| = o_p(T^{-2/5}), \tag{60}$$

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^{*,D}(y)\right| = o_p(T^{-2/5}). \tag{61}$$

Here, $\widehat{m}_\theta^{*,B}$ is the bias term, $\widehat{m}_\theta^{*C}$ is the stochastic term and $\widehat{m}_\theta^{*,D}$ is the remainder term. For local linear estimates of $g_j(y)$ it follows that under standard smoothness conditions, (58)–(59), (61) 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 (60) 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 $\widehat{m}_\theta^j$ as solutions to the integral equations

$$\widehat{m}_\theta^j = \widehat{m}_\theta^{*,j} + \widehat{\mathcal{H}}_\theta\ \widehat{m}_\theta^j$$

and $\widehat{m}_\theta^A$ implicitly from writing the solution $m_\theta + \widehat{m}_\theta^A$ to the integral equation

$$\left(m_\theta + \widehat{m}_\theta^A\right) = m_\theta^* + \widehat{\mathcal{H}}_\theta\ \left(m_\theta + \widehat{m}_\theta^A\right). \tag{62}$$

The existence and uniqueness of $\widehat{m}_\theta^j$ follows from the invertibility of the operator $I - \widehat{\mathcal{H}}_\theta$ (at least with probability tending to one). It now follows that

$$\widehat{m}_\theta = m_\theta + \widehat{m}_\theta^A + \widehat{m}_\theta^B + \widehat{m}_\theta^C + \widehat{m}_\theta^D$$

by linearity of the operator $(I - \widehat{\mathcal{H}}_\theta)^{-1}$. Note that $\widehat{m}_\theta^j = (I - \widehat{\mathcal{H}}_\theta)^{-1}\widehat{m}_\theta^{*,j}$ for $j = B, C, D$, while $m_\theta + \widehat{m}_\theta^A = (I - \widehat{\mathcal{H}}_\theta)^{-1} m_\theta^*$. Define also m_θ^B as the solution to the equation

$$m_\theta^B = \widehat{m}_\theta^{*,B} + \mathcal{H}_\theta\ m_\theta^B. \tag{63}$$

We now claim that under (A1)–(A5):

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^B(y) - m_\theta^B(y)\right| = o_p(T^{-2/5}). \tag{64}$$

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^C(y) - \widehat{m}_\theta^{*,C}(y)\right| = o_p(T^{-2/5}) \tag{65}$$

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^D(y)\right| = o_p(T^{-2/5}) \tag{66}$$

Here, claims (64) and (66) immediately follow from (19) and (55). For (65) note that because of (59)–(60), (55) and (A4)

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

So we arrive at the following expansion of $\widehat{m}_\theta$.

$$\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta(y)-m_\theta(y)-\widehat{m}_\theta^A(y)-m_\theta^B(y)-\widehat{m}_\theta^{*,C}(y)\right|=o_p(T^{-2/5}). \tag{67}$$

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 (62) we get

$$\widehat{m}_\theta^A=\left(\widehat{\mathcal{H}}_\theta\ -\mathcal{H}_\theta\right)m_\theta+\widehat{\mathcal{H}}_\theta\ \widehat{m}_\theta^A. \tag{68}$$

We next write $\widehat{\mathcal{H}}_\theta$ as a sum of terms with convenient properties.

ASSUMPTION A6. Suppose that for δ_T as in (A4)

$$\left(\widehat{\mathcal{H}}_\theta-\mathcal{H}_\theta\right)m_\theta(y)=\widehat{m}_\theta^{*,E}(y)+\widehat{m}_\theta^{*,F}(y)+\widehat{m}_\theta^{*,G}(y), \tag{69}$$

where $\widehat{m}_\theta^{*,E}$, $\widehat{m}_\theta^{*F}$, and $\widehat{m}_\theta^{*,G}$ satisfy:

$$\begin{aligned}
\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}^{*,E}(y)\right| &= O_p(T^{-2/5}) \text{ with } \widehat{m}^{*,E} \text{ deterministic,}\\
\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^{*,F}(y)\right| &= O_p\left(T^{-2/5}\delta_T^{-1}\right),\\
\sup_{\theta\in\Theta,|y|\leq c}\left|\mathcal{H}_\theta\left(I-\mathcal{H}_\theta\right)^{-1}\widehat{m}_\theta^{*,F}(y)\right| &= o_p(T^{-2/5}),\\
\sup_{\theta\in\Theta,|y|\leq c}\left|\widehat{m}_\theta^{*,G}(y)\right| &= o_p(T^{-2/5}).
\end{aligned}$$

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 A6 follows from similar arguments to those given for 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

$$m_\theta^E = \widehat{m}_\theta^{*,E} + \mathcal{H}_\theta m_\theta^E. \tag{70}$$

As above we get that:

$$\sup_{\theta\in\Theta,|y|\leq c} \left|\widehat{m}_\theta^E(y) - m_\theta^E(y)\right| = o_p(T^{-2/5}), \tag{71}$$

$$\sup_{\theta\in\Theta,|y|\leq c} \left|\widehat{m}_\theta^F(y) - \widehat{m}_\theta^{*,F}(y)\right| = o_p(T^{-2/5}), \tag{72}$$

$$\sup_{\theta\in\Theta,|y|\leq c} \left|\widehat{m}_\theta^G(y)\right| = o_p(T^{-2/5}). \tag{73}$$

We summarize our discussion in the following Proposition.

PROPOSITION 1. *Suppose that conditions (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* $\widehat{m}_\theta$

$$\sup_{\theta\in\Theta,|y|\leq c} \left|\widehat{m}_\theta(y) - m_\theta(y) - m_\theta^B(y) - m_\theta^E(y) - \widehat{m}_\theta^{*,C}(y) - \widehat{m}_\theta^{*,F}(y)\right| = o_p(T^{-2/5}). \tag{74}$$

The terms m_θ^B *and* m_θ^E *have been defined in (63) and (70), respectively.*

Equation (74) 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 A4-A6, so our result applies to any smoothing method that satisfies these conditions. It remains to prove that A4-A6 hold under our primitive conditions B1-B7, and that a central limit theorem (and uniform convergence) applies to $\widehat{m}_\theta^{*,C}(y) + \widehat{m}_\theta^{*,F}(y)$.

A.1.2 Proof of High Level Conditions A1,A3-A6 and CLT

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

We verify A4-A6 with the choice

$$\delta_T = T^{-3/10+\xi} \tag{75}$$

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 A5 and A6 with

$$\begin{aligned}
\widehat{m}_\theta^{*,B}(y) &= \frac{h^2}{2}\mu_2(K)\times\beta_\theta^1(y)\\
\widehat{m}_\theta^{*,C}(y) &= \frac{1}{Tp_0(y)}\sum_{t=1}^{T-\tau_T} K_h\left(y_t-y\right)\eta_{\theta,t}^1\\
\widehat{m}_\theta^{*,E}(y) &= \frac{h^2}{2}\mu_2(K)\times\beta_\theta^2(y)\\
\widehat{m}_\theta^{*,F}(y) &= \frac{1}{Tp_0(y)}\sum_{t=1}^{T-\tau_T} K_h\left(y_t-y\right)\eta_{\theta,t}^2+\frac{1}{T}\sum_{t=1}^{T-\tau_T}\frac{\mu_\theta(y)}{p_0(y)}\left[K_h\left(y_t-y\right)-EK_h\left(y_t-y\right)\right],
\end{aligned}$$

where $\eta_{\theta,t}^1 = \sum_{j=1}^{\infty}\psi_j^\dagger(\theta)\eta_{j,t}$ and $\eta_{\theta,t}^2 = -\sum_{j=\pm1}^{\pm\infty}\psi_j^2(\theta)\zeta_{j,t}(\theta)$, 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 A4. It suffices to show that

$$\sup_{\substack{|x|,|y|\le c\\ 1\le j\le\tau_T}}\left|\widehat{p}_{0,j}(x,y)-p_{0,j}(x,y)\right| = o_p(\delta_T) \tag{76}$$

$$\sup_{|x|\le c}\left|\widehat{p}_0(x)-p_0(x)\right| = o_p(\delta_T). \tag{77}$$

Note that by assumption B4 the density p_0 is bounded from below on $|x|\le c$. For the proof of (76) we make use of an exponential inequality. Using Theorem 1.3 in Bosq (1998) one gets

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

where $[x]$ denotes the largest integer smaller or equal to x, and where

$$\begin{aligned}
q &= T^\beta \text{ with } \frac{7}{10}<\beta<1,\ j^2\le T^{1-\beta},\\
b &= CT^{7/10} \text{ for a constant } C,\\
v^2(q) &= 8\frac{q^2}{T^2}\sigma^2(q)+\frac{b}{4}T^{\xi},\\
\sigma^2(q) &= E\left[\sum_{t=1}^{[T/2q]+1}K_h(y_t-x)K_h(y_{t+j}-y)-EK_h(y_t-x)K_h(y_{t+j}-y)\right]^2.
\end{aligned}$$

The variance $\sigma^2(q)$ can be bounded by use of Corollary 1.1. in Bosq (1998). This gives

$$\sigma^2(q) \leq C' T^{2-\beta+(2/5)\gamma} \text{ for } 0 < \gamma < 1$$

with a constant C' depending on γ. This gives with constants $C_1, C_2, \ldots > 0$ for $|x|, |y| \leq c$, $1 \leq j \leq \tau_T$

$$\Pr\left(\left|T^{3/10}\left[\widehat{p}_{0,j}(x,y) - E\widehat{p}_{0,j}(x,y)\right]\right| \geq T^{\xi}\right) \leq C_1 \exp(-C_2 T^{C_3}) + C_4 T^{C_5} \alpha(T^{C_6}).$$

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 $\{|x| \leq c, |y| \leq c\}$, 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

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

for any constant c. By the Bonferroni and Exponential inequalities:

$$\begin{aligned}
\Pr\left[\max_{1\leq q\leq Q, 1\leq j\leq \tau} |V_j(z_q)| > c\delta_T\right] &\leq \sum_{j=1}^{\tau}\sum_{q=1}^{Q} \Pr\left[|V_j(z_q)| > c\delta_T\right] \\
&\leq Q(T)\tau(T)\left[C_1 \exp(-C_2 T^{C_3}) + C_4 T^{C_5}\alpha(T^{C_6})\right] \\
&= o(1),
\end{aligned}$$

provided s_0 in B1 is chosen large enough. By the Lipschitz continuity of K, $|K_h(y_t - x) - K_h(y_t - x_q)| \leq \overline{K}\,|x - x_q|\,/h$, where $\overline{K}$ is finite, and so

$$T^{3/10-\xi}\,|V_j(z_q) - V_j(z)| \leq T^{3/10-\xi}\frac{1}{h^2}\left[c_1\,|x - x_q| + c_2\,|y - y_q|\right] \leq c\epsilon_T T^{7/10-\xi}$$

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 (76) because similar arguments are used in the sequel. Equation (77) follows by the same type of argument.

PROOF OF A5. Claim (58) immediately follows from assumption B4. For the proof of (61) we use the usual variance+bias+remainder term decomposition of the local linear estimates $\widehat{g}_j$ as in Masry (1996). Write $M(y) = p_0(y)\text{diag}(1, \mu_2(K))$ and

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

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

$$B_{Tj}(y) = \begin{bmatrix} B_{Tj,0}(y) \\ B_{Tj,1}(y) \end{bmatrix}, \text{ where } B_{Tj,l}(y) = \frac{1}{Th}\sum_{t=1}^{T}\left(\frac{y - y_{t-j}}{h}\right)^l K\left(\frac{y - y_{t-j}}{h}\right)\Delta_{tj}(y),$$

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

$$U_{Tj}(y) = \begin{bmatrix} U_{Tj,0}(y) \\ U_{Tj,1}(y) \end{bmatrix}, \text{ where } U_{Tj,l}(y) = \frac{1}{Th}\sum_{t=1}^{T}\left(\frac{y - y_{t-j}}{h}\right)^l K\left(\frac{y - y_{t-j}}{h}\right)\eta_{j,t-j}.$$

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,

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

Defining V_{jy} and B_{jy} as $\widehat{V}_{jy}$ and $\widehat{B}_{jy}$ with $M_{Tj}(y)$ replaced by $M_j(y)$, we have

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

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

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

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

$$\sup_{|y| \le c, \theta \in \Theta} |R_{T1}(y,\theta)| = o_p(T^{-2/5}) \tag{78}$$

$$\sup_{|y| \le c, \theta \in \Theta} |R_{T2}(y,\theta)| = o_p(T^{-2/5}) \tag{79}$$

$$\sup_{|y| \le c, \theta \in \Theta} \left| \frac{1}{T p_0(y)} \sum_{t=1}^{T-\tau_T} K_h\left(y_t - y\right) \sum_{j=\tau+1}^{\infty} \psi_j^{\dagger}(\theta) \eta_{j,t} \right| = o_p(T^{-2/5}). \tag{80}$$

It follows that

$$\widehat{m}_{\theta}^{*}(y) - m_{\theta}^{*}(y) = \frac{1}{T p_0(y)} \sum_{t=1}^{T-\tau_T} K_h\left(y_t - y\right) \eta_{\theta,t}^{1} + \sum_{j=1}^{\tau} \psi_j^{\dagger}(\theta) B_{jy} + o_p(T^{-2/5}).$$

First note that $E(A) = 0$, where $A = T^{-1} \sum_{t=1}^{T-\tau_T} K_h\left(y_t - y\right) \sum_{j=\tau+1}^{\infty} \psi_j(\theta)\ \eta_{j,t}/p_0(y)$ and

$$\begin{aligned}
\mathrm{var}(A) &= \frac{1}{T^2 h^2 p_0^2(y)} \sum_{t=1}^{T-\tau_T} \sum_{s=1}^{T-\tau_T} \sum_{j=\tau+1}^{\infty} \sum_{l=\tau+1}^{\infty} \psi_j^{\dagger}(\theta) \psi_l^{\dagger}(\theta) E\left[K\left(\frac{y_t - y}{h}\right) K\left(\frac{y_s - y}{h}\right) \eta_{j,t} \eta_{l,s} \right] \\
&= o(T^{-1} h^{-1})
\end{aligned}$$

by virtue of the decay conditions.

The uniformity of the bound can be achieved by application of the exponential inequality in Theorem 1.3 of Bosq (1998) used also in the proof of (76).

For the proof of (59) 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

$$\widetilde{\eta}_{\theta,t} = \sum_{j=1}^{\tau_T} \psi_j^\dagger(\theta) \left[\min\{y_{t+j}^2, T^{1/\rho}\} - E(\min\{y_{t+j}^2, T^{1/\rho}\}|y_t)\right].$$

The truncated random variables $\widetilde{\eta}_{\theta,t}$ can be replaced by $\eta_{\theta,t}$ using the fact that

$$\begin{aligned} 1 - \Pr\left(y_t^2 \leq T^{1/\rho} \text{ for } 1 \leq t \leq T\right) &\leq T \Pr\left(y_t^2 > T^{1/\rho}\right) \\ &\leq E\left[y_t^{2\rho} 1(y_t^2 > T^{1/\rho})\right] \\ &\to 0. \end{aligned}$$

It remains to check (60). Define the operator $\mathcal{L}_\theta(x, y)$ by

$$\mathcal{H}_\theta(I - \mathcal{H}_\theta)^{-1} m(x) = \int_{-c}^{c} \mathcal{L}_\theta(x, y) m(y) p_0(y) dy.$$

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

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

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 $\sup_{\theta \in \Theta, j \geq 1} \lambda_{\theta,j} < \gamma$. This shows that

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

Furthermore, it can be checked that $\mathcal{L}_\theta(x, y)$ is continuous in θ, x, y. This follows from 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

$$\nu_\theta(z, x) = \int_{-c}^{c} \mathcal{L}_\theta(x, y) \frac{1}{p_0(y)} K_h(z - y) dy.$$

The function $\nu_\theta(z, x)$ is continuous in θ, z, x. Using this fact, claim (60) can be easily checked, e.g., again by application of the exponential inequality in Theorem 1.3 of Bosq (1998).

PROOF OF A6. Write

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

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 A6 can be checked as in the proof of A5.

PROOF OF CLT FOR $\widehat{m}_\theta^{*,C}(y) + \widehat{m}_\theta^{*,F}(y)$. This follows by an application of Masry and Fan (1997, Theorem 3).

PROOF OF (43) AND (44). The only additionality here is to show that

$$\sup_{\theta \in \Theta, |y| \leq c} |\widehat{m}_\theta^{*,C}(y) + \widehat{m}_\theta^{*,F}(y)| = o_p(T^{-1/4}).$$

This follows from standard arguments for uniform consistency of regression smoothers on mixing processes.

Finally,

$$\begin{aligned}
\sup_{\theta \in \Theta, 1 \leq t} \left| \widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta) \right| &\leq \sup_{\theta \in \Theta} \sum_{j=1}^{\infty} \psi_j(\theta) \sup_{|y| \leq c} |\widehat{m}_\theta(y) - m_\theta(y)| + \sup_{\theta \in \Theta} \sum_{j=\tau_T+1}^{\infty} \psi_j(\theta) \sup_{|y| \leq c} m_\theta(y) \\
&= o_p(T^{-1/4}).
\end{aligned}$$

A.2 Proof of Theorem 2

CONSISTENCY. We apply some general results for semiparametric estimators. Write

$$S_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left\{ y_t^2 - \sigma_t^2(\theta) \right\}^2 \text{ and } S(\theta) = E S_T(\theta).$$

We have

$$\sup_{\theta\in\Theta}|S_T(\theta)-S(\theta)|=o_p(1) \tag{81}$$

by standard arguments. Then

$$\widehat{S}_T(\theta)-S_T(\theta)=-\frac{2}{T}\sum_{t=\tau_T+1}^{T}\eta_t(\theta)\left[\widehat{\sigma}_t^2(\theta)-\sigma_t^2(\theta)\right]+\frac{1}{T}\sum_{t=\tau_T+1}^{T}\left[\widehat{\sigma}_t^2(\theta)-\sigma_t^2(\theta)\right]^2+o_P(1),$$

where $\eta_t(\theta)=y_t^2-\sigma_t^2(\theta)$. Then, because of (44) we have

$$\sup_{\theta\in\Theta}\left|\widehat{S}_T(\theta)-S_T(\theta)\right|\longrightarrow_p 0. \tag{82}$$

Therefore, (81) and (82) we have

$$\sup_{\theta\in\Theta}\left|\widehat{S}_T(\theta)-S(\theta)\right|=o_p(1). \tag{83}$$

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

$$\begin{aligned}\frac{\partial\widehat{S}_T(\theta)}{\partial\theta} &= -\frac{2}{T}\sum_{t=\tau_T+1}^{T}\widehat{\eta}_t(\theta)\frac{\partial\widehat{\sigma}_t^2(\theta)}{\partial\theta}\\ \frac{\partial^2\widehat{S}_T(\theta)}{\partial\theta^2} &= \frac{2}{T}\sum_{t=\tau_T+1}^{T}\left[\frac{\partial\widehat{\sigma}_t^2(\theta)}{\partial\theta}\right]^2-\widehat{\eta}_t(\theta)\frac{\partial^2\widehat{\sigma}_t^2(\theta)}{\partial\theta^2},\end{aligned}$$

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

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

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

$$\inf_{|\theta-\theta_0|\le\epsilon_T}\left|\frac{\partial^2\widehat{S}_T(\theta)}{\partial\theta^2}\right|>C+o_p(1) \tag{84}$$

$$\sqrt{T}\frac{\partial\widehat{S}_T(\theta_0)}{\partial\theta}=O_p(1). \tag{85}$$

This implies that (46) holds.

To establish the results (84) and (85) we use some expansions given in Lemma 1 below.

PROOF OF (84). By straightforward but tedious calculation we show that

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

Specifically, it suffices to show that

$$\sup_{|\theta-\theta_0|\leq\epsilon_T,1\leq t\leq T}\left|\frac{\partial^j\widehat{\sigma}_t^2(\theta)}{\partial\theta^j}-\frac{\partial^j\sigma_t^2(\theta)}{\partial\theta^j}\right|=o_p(1),\ j=0,1,2. \tag{86}$$

For $j=0,1$ this follows from (44)-(45). 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,

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

by standard arguments. Therefore, by the triangle inequality

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

PROOF OF (85). Write

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

and let with $\eta_t=\eta_t(\theta_0)$

$$\begin{aligned}
\sqrt{T}E_T(\theta_0) &= E_{T1}+E_{T2},\\
E_{T1} &= -\frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\eta_t\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta},\\
E_{T2} &= \frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left[\widehat{\sigma}_t^2(\theta_0)-\sigma_t^2(\theta_0)\right]\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\\
&\quad -\frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\eta_t\left[\frac{\partial\widehat{\sigma}_t^2(\theta_0)}{\partial\theta}-\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right].
\end{aligned}$$

Then

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

by (44)-(45).

The term E_{T1} is asymptotically normal with mean zero and finite variance by standard central limit theorem for mixing processes. Note that

$$E\left[\eta_t\frac{\partial\sigma_t^2(\theta_0)}{\partial\theta}\right]=0$$

by definition of θ_0.

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

$$\begin{aligned}E_{T2} &= \frac{h^2}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left\{\sum_{j=1}^{\tau_T}\psi_j(\theta_0)b^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)+\eta_t\sum_{j=1}^{\tau_T}\psi_j'(\theta_0)b^0(y_{t-j})\right\} \\ &+\frac{h^2}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j(\theta_0)b^1(y_{t-j})\right\} \\ &+\frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left\{\sum_{j=1}^{\tau_T}\psi_j(\theta_0)s^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)\right\} \\ &+\frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j'(\theta_0)s^0(y_{t-j})\right\} \\ &+\frac{1}{\sqrt{T}}\sum_{t=\tau_T+1}^{T}\left\{\eta_t\sum_{j=1}^{\tau_T}\psi_j(\theta_0)s^1(y_{t-j})\right\}+o_P(1),\end{aligned} \tag{87}$$

where

$$\begin{aligned}b_\theta(y) &= h^{-2}\left[m_\theta^B(y)+m_\theta^E(y)\right], \\ s_\theta(y) &= (I-\mathcal{H}_\theta)^{-1}(m_\theta^{*,C}+m_\theta^{*,F})(y), \\ b^j(y) &= \frac{\partial^j}{(\partial\theta)^j}b_{\theta_0}(y), \\ s^j(y) &= \frac{\partial^j}{(\partial\theta)^j}s_{\theta_0}(y).\end{aligned}$$

By tedious calculations it can be shown that the last three terms on the right hand side of (87) 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 (87) note that b^0 and b^1 are deterministic functions. Furthermore, we will show that

$$E\left[\sum_{j=1}^{\infty}\psi_j(\theta_0)b^0(y_{t-j})\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)+\eta_t\psi_j'(\theta_0)b^0(y_{t-j})\right]=0, \tag{88}$$

$$E\left[\eta_t\sum_{j=1}^{\infty}\psi_j(\theta_0)b^1(y_{t-j})\right]=0. \tag{89}$$

Note that in (88)-(89) we have replaced the upper index of the sum by ∞. Thus, with (88)-(89) we see that the first two terms on the right hand side of (87) are sums of variables with mean geometrically tending to zero. The sums are multiplied by factors $h^2T^{-1/2}$. 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 (88)-(89). By definition for each function g

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

is minimized for $\delta=0$. By taking derivatives with respect to δ we get that

$$E\left\{\left[y_t^2-\sigma_t^2(\theta)\right]\sum_{j=1}^{\infty}\psi_j(\theta)g(y_{t-j})\right\}=0. \tag{90}$$

With $g=b^0$ and θ_0 this gives (89). For the proof of (88) we now take the difference of (90) for θ and θ_0. This gives

$$E\left[y_t^2-\sigma_t^2(\theta_0)\right]\sum_{j=1}^{\infty}\left[\psi_j(\theta)-\psi_j(\theta_0)\right]g(y_{t-j})-E\left[\sigma_t^2(\theta)-\sigma_t^2(\theta_0)\right]\sum_{j=1}^{\infty}\psi_j(\theta)g(y_{t-j})=0.$$

Taking derivatives with respect to θ gives

$$E\left\{u_t\sum_{j=1}^{\infty}\psi_j'(\theta_0)g(y_{t-j})-\frac{\partial\sigma_t^2}{\partial\theta}(\theta_0)\sum_{j=1}^{\infty}\psi_j(\theta_0)g(y_{t-j})\right\}=0.$$

With $g=b^0$ this gives (88). ■

A.3 Proof of Theorems 3 and 4

We only give a proof of Theorem 3. Theorem 4 follows along the same lines. For a proof of (47) one shows that for $C > 0$

$$\sup_{|\theta-\theta_0|\leq CT^{-1/2}} |\widehat{m}_\theta(y) - \widehat{m}_{\theta_0}(y)| = o_P[(Th)^{-1/2}].$$

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 (47) for a proof of (48) it suffices to show

$$\sqrt{Th}\left[\widehat{m}_{\theta_0}(y) - m_{\theta_0}(y) - h^2 b(y)\right] \Longrightarrow N\left(0, \omega(y)\right). \tag{91}$$

So it remains to show (91). Put

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

$$\widehat{p}^2_0(y) = \frac{1}{T}\sum_{t=1}^{T}(y_t - y)^2 K_h(y_t - y).$$

Then, by using similar arguments as in the proof of Theorem 1, we have for $\gamma > 0$

$$\sup_{|y|\leq c}\left|\widehat{p}^1_0(y) - h^2\,\mu_2(K)\;p'_0(y)\right| = O_p(h^{1/2}\,T^{-1/2+\gamma} + h^3),$$

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

Furthermore,

$$\sup_{|y|\leq c}|\widehat{p}_0(y) - p_0(y)| = O_p(h^2 + h^{-1/2}T^{-1/2+\gamma}).$$

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

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

By plugging this into

$$\widehat{m}_{\theta_0}^*(y) - (I-\widehat{\mathcal{H}}_{\theta_0})m_0(y) = \sum_{j=1}^{\tau_T}\psi_j^\dagger(\theta_0)\left[\widehat{g}_j(y)-\mu\right] - m_0(y) - \sum_{0<|j|<\tau_T}\psi_j^*(\theta_0)\int\frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)}\,m_0(x)dx,$$

we get

$$\begin{aligned}
\widehat{m}_{\theta_0}^*(y) - (I-\widehat{\mathcal{H}}_{\theta_0})m_0(y) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\psi_j^\dagger(\theta_0)\frac{K_h(y_{t-j}-y)}{p_0(y)}\,\sigma_t^2 u_t \\
&+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty}\psi_j^\dagger(\theta_0)\psi_\ell(\theta_0)\frac{K_h(y_{t-j}-y)}{\widehat{p}_0(y)}\,m_0(y_{t-\ell}) \\
&+h^2\mu_2(K)\frac{p_0^1(y)}{p_0(y)}[\frac{\partial}{\partial y}(\mathcal{H}_{\theta_0}m_0(y)-m_0(y))] \\
&-m_0(y) - \sum_{j\neq 0}\psi_j^*(\theta_0)\int\frac{\widehat{p}_{0,j}(y,x)}{\widehat{p}_0(y)}\,m(x)dx + o_p(T^{-1/2}) \\
&= S_1 + S_2 + S_3 - m(y) + S_4 + o_p(T^{-1/2}).
\end{aligned}$$

We have

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

Therefore we get uniformly in $\mid y \mid \leq c$

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

with

$$
w_t(y)=\sum_{j=1}^{\tau_T}\psi_j^{\dagger}(\theta_0)\frac{K_h(y_{t-j}-y)}{p_0(y)}.
$$

From this stochastic expansion we immediately get an expansion for the asymptotic bias. For the

calculation of the asymptotic variance note that

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

■

A.4 Proofs of Theorems 5 and 6

The proof make 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 $\sup_{|y|\leq c}\left|\widetilde{m}_{\theta_0}-\widetilde{m}_{\widetilde{\theta}}\right|=o_P(T^{-2/5})$. 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 following integral equation.

$$\widetilde{\overline{m}}_\theta=\widetilde{\overline{m}}_\theta^*+\widetilde{\overline{\mathcal{H}}}_\theta\widetilde{\overline{m}}_\theta,$$

where

$$\widetilde{\overline{m}}_\theta^*=\frac{\sum_{j=1}^{\tau_T}\psi_j(\theta)\widetilde{g}_j^a(y)}{\sum_{j=1}^{\tau_T}\psi_j^2(\theta)\widetilde{g}_j^b(y)} \quad ; \quad \widetilde{\overline{\mathcal{H}}}_\theta(x,y)=-\sum_{\substack{j=1\\ l\neq j}}^{\tau_T}\sum_{l=1}^{\tau_T}\psi_j(\theta)\psi_l(\theta)\widetilde{w}_{j,l}^\theta(x,y)\frac{\widehat{p}_{0,l-j}(x,y)}{\widehat{p}_0(y)\widehat{p}_0(y)}$$

$$\widetilde{w}_{j,l}^\theta(x,y)=\frac{\widetilde{g}_{l,j}^c(x,y)}{\sum_{j=1}^{\tau_T}\psi_j^2(\theta)\widetilde{g}_j^b(y)}.$$

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 exist a bounded function b such that uniformly for $|y|\leq c,|\theta-\theta_0|\leq CT^{-1/2}$

$$\widetilde{\overline{m}}_\theta(y)-\widetilde{m}_\theta(y)-h^2b(y)=o_P(T^{-1/2}).$$

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 stochastic term of $\widetilde{\overline{m}}_{\theta_0}(y)$ we get

$$\frac{1}{T}\sum_{t=1}^{T} \overline{w}_t(y)\sigma_t^{-4}(y_t^2 - \sigma_t^2) = \frac{1}{T}\sum_{t=1}^{T} \overline{w}_t(y)\sigma_t^{-2}u_t,$$

where

$$\overline{w}_t(y) = \frac{\sum_{j=1}^{\tau_T} \psi_j(\theta_0) K_h(y_{t-j} - y)}{p_0(y)\sum_{j=1}^{\tau_T} \psi_j^2(\theta_0)\mathrm{E}[\sigma_t^{-4}|y_{t-j} = y]}.$$

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

$$\begin{aligned} & h\mathrm{E}\overline{w}_t(y)^2\sigma_t^{-4}u_t^2 \\ & \quad = \frac{1}{p_0(y)}\left[\sum_{j=1}^{\infty}\psi_j^2(\theta_0)\mathrm{E}(\sigma_j^{-4}|y_0 = y)\right]^{-2} \nu_0(K)\sum_{j=1}^{\infty}\psi_j^2(\theta_0)\mathrm{E}(\sigma_j^{-4}u_j^2|y_0 = y) + o(1) \\ = & \frac{1}{p_0(y)}\left[\sum_{j=1}^{\infty}\psi_j^2(\theta_0)\mathrm{E}(\sigma_j^{-4}|y_0 = y)\right]^{-1} \nu_0(K)(2 + \kappa_4 + o(1). \end{aligned}$$

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

$$\frac{\partial\widetilde{l}}{\partial\theta}(\theta_0) = -\frac{1}{T}\sum_{t=1}^{T}\sigma_t^{-2}u_t\frac{\partial\overline{\sigma}_t^2}{\partial\theta}(\theta_0) + o_P(T^{-1/2}), \tag{92}$$

$$\frac{\partial^2\widetilde{l}}{\partial\theta^2}(\theta) = -\mathrm{E}\sigma_t^{-4}\left[\frac{\partial\overline{\sigma}_t^2}{\partial\theta}(\theta_0)\right]^2 + o_P(1), \tag{93}$$

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

$$\widetilde{\theta} = \theta_0 - \left\{\mathrm{E}\sigma_t^{-4}\left[\frac{\partial\overline{\sigma}_t^2}{\partial\theta}(\theta_0)\right]^2\right\}^{-1}\frac{1}{T}\sum_{t=1}^{T}\sigma_t^{-2}u_t\frac{\partial\overline{\sigma}_t^2}{\partial\theta}(\theta_0) + o_P(T^{-1/2}). \tag{94}$$

This expansion can be used to show the desired asymptotic normal limit for $\widetilde{\theta}$. It remains to show (92)-(93). This can be done by using similar arguments as for the proof of (84) and (85). ■

A.5 Lemmas

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

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

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 now $\left(\widehat{m}_\theta^{*,C} + \widehat{m}_\theta^{*,F} \right)(y)$ is replaced by $(I - \mathcal{H}_\theta)^{-1} \left(\widehat{m}_\theta^{*,C} + \widehat{m}_\theta^{*,F} \right)(y)$. The difference of these terms is of order $O_P(T^{-1/2})$. For the proof for $j = 1$ we make use of the following integral equation for $\widehat{m}_\theta^1 = \frac{\partial}{\partial\theta}\widehat{m}_\theta$

$$\widehat{m}_\theta^1 = \frac{\partial}{\partial\theta}\widehat{m}_\theta^* + \left[\frac{\partial}{\partial\theta}\widehat{\mathcal{H}}_\theta \right] \widehat{m}_\theta + \widehat{\mathcal{H}}_\theta \widehat{m}_\theta^1.$$

Thus with

$$\widehat{m}_\theta^{*,1} = \frac{\partial}{\partial\theta}\widehat{m}_\theta^* + \left[\frac{\partial}{\partial\theta}\widehat{\mathcal{H}}_\theta \right] \widehat{m}_\theta$$

the derivative $\widehat{m}_\theta^1$ fulfills

$$\widehat{m}_\theta^1 = \widehat{m}_\theta^{*,1} + \widehat{\mathcal{H}}_\theta \widehat{m}_\theta^1.$$

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. ■

REFERENCES

Atkinson, K. (1976). An automatic program for linear Fredholm integral equations of the second kind. ACM Transactions on Mathematical Software 2,2 154-171.

Audrino, F., and Bühlmann, P. (2001), "Tree-structured GARCH models," Journal of The Royal Statistical Society, 63, 727-744.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and J. A. Wellner (1993). Efficient and adaptive estimation for semiparametric models. The John Hopkins University Press, Baltimore and London.

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

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994), "ARCH Models," in Handbook of Econometrics, volume IV, eds. R. F. Engle and D. L. McFadden, Elsevier Science, 2959-3038.

Bollerslev, T. and Wooldridge, J. M. (1992). "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models With Time Varying Covariances," Econometric Reviews, 11, 143-172.

Bosq, D (1998). Nonparametric Statistics for Stochastic Processes. Estimation and Prediction. Springer, Berlin.

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.

Buja, A., T. Hastie, and R. Tibshirani, (1989). Linear smoothers and additive models (with discussion). Ann. Statist. 17, 453-555.

Carrasco, M. and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," Econometric Theory, 18, 17-39.

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.

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 XLVIII, 1749-1778.

Fan, J. (1992). Design-adaptive nonparametric regression. J. Am. Statist Soc. 82, 998-1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. Annals of Statistics 21, 196-216.

Fan, J., and Q. Yao (1998). Efficient estimation of conditional variance functions in Stochastic Regression. Biometrika forthcoming.

Friedman, J.H., and W. Stuetzle (1981). Projection pursuit regression. Journal of the American Statistical Association 76, 817-823.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (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.

Hafner, C. (1998). *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility.* Heidelberg, Physica.

Hall, P. and J.L. Horowitz (2003). Nonparametric methods for inference in the presence of instrumental variables. Manuscript, Northwestern University.

Hannan, E.J. (1973). The asymptotic theory of linear time-series models. Journal of Applied Probability 10, 130-145.

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, (1998) Nonparametric vector autoregression . Discussion Paper, J. Stat. Planning. Inference, 68, 221-245.

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.

Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall, London.

Hong, Y., and Y.J Lee (2003). "Generalized Spectral Test for Conditional Mean Models with Conditional Heteroskedasticity of Unknown Form". Manuscript, Cornell University.

Horowitz, J.L., and E. Mammen (2002). Nonparametric Estimation of an additive model with a link function. Manuscript, Northwestern University.

Kim, W., and O. Linton (2002). A Local Instrumental Variable Estimation method for Generalized Additive Volatility Models.

Lee, S., and Hansen, B. (1994), "Asymptotic Theory for the GARCH(1,1) Quasi-Maximum Likelihood Estimator," Econometric Theory, 10, 29-52.

Linton, O. (1993) Adaptive estimation in ARCH models. Econometric Theory 9, 539-569.

Linton, O.B. (1996). Efficient estimation of additive nonparametric regression models. Biometrika 84, 469-474.

Linton, O.B. (2000). Efficient estimation of generalized additive nonparametric regression models. Econometric Theory 16, 502-523.

Linton, O, E. Mammen, J. Nielsen, and C. Tanggaard (2001). Estimating the Yield Curve by Kernel Smoothing Methods. Journal of Econometrics 105/1 185-223.

Linton, O.B. and J.P. Nielsen, (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika 82, 93-100.

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 Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Annals of Statistics.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. J. Time Ser. Anal. 17, 571-599.

Masry, E., and J. Fan (1997). Local Polynomial Estimation of Regression Functions for Mixing Processes. Scandinavian Journal of Statistics 24, 165-179.

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.

Masry, E., and D. Tjøstheim (1997). Additive nonlinear ARX time series and projection estimates. Econometric Theory.

Nelsen, D. (1990). Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347-370.

Nielsen, J.P., and O.B. Linton (1997). An optimization interpretation of integration and backfitting estimators for separable nonparametric models. Journal of the Royal Statistical Society, Series B

Newey, W. K. and Powell, J. L. (1989,2003). Instrumental variables estimation for nonparametric regression models. Forthcoming in Econometrica.

Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. Annals of Statistics 25, 186 - 211.

O'Sullivan, F. (1986). Ill posed inverse problems (with discussion). Statistical Science 4, 503-527.

Pagan, A.R., and G.W. Schwert (1990): "Alternative models for conditional stock volatility," Journal of Econometrics 45, 267-290.

Pagan, A.R., and Y.S. Hong (1991): "Nonparametric Estimation and the Risk Premium," in W. Barnett, J. Powell, and G.E. Tauchen (eds.) Nonparametric and Semiparametric Methods in Econometrics and Statistics, Cambridge University Press.

Perron, B. (1998), "A Monte Carlo Comparison of Non-parametric Estimators of the Conditional Variance," Unpublished manuscript, Université de Montréal.

Powell, J. (1994), "Estimation of Semiparametric Models," in Handbook of Econometrics, volume IV, eds. R. F. Engle and D. L. McFadden, Elsevier Science, 2443-2521.

Riesz, F. and Sz.-Nagy, B. (1990). Functional Analysis. Dover, New York.

Stone, C.J. (1985). Additive regression and other nonparametric models. Ann. Statist. 13, 685-705.

Rust, J. (1997). Using randomization to break the curse of dimensionality. Econometrica 65, 487-516.

Rust (2000). Nested Fixed Point Algorithm Documentation Manual. Version 6, Yale University.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 592-606.

Tjøstheim, D., and Auestad, B. (1994a). Nonparametric identification of nonlinear time series: projections. J. Am. Stat. Assoc. 89, 1398-1409.

Tjøstheim, D., and Auestad, B. (1994b). Nonparametric identification of nonlinear time series: selecting significant lags. J. Am. Stat. Assoc. 89, 1410-1419.

Tong, H. (1990). Nonlinear Time Series Analysis: A dynamic Approach, Oxford University Press, Oxford.

Wu, G., and Z. Xiao (2002). A generalized partially linear model for asymmetric volatility. Journal of Empirical Finance 9, 287-319.

Xia, Y., H. Tong, W.K. Li, and L.X Zhu (2002). An adaptive estimation of dimension reduction space (with discussion). Journal of the Royal Statistical Society, Series B 64, 1-28.

B Tables and Figures

n	tau/c_h	$E\widehat{\theta}$	stdc$(\widehat{\theta})$	med$(\widehat{\theta})$	iqr$(\widehat{\theta})$	$\int(\mathrm{E}\widehat{\sigma}-\sigma)^2$	$\int\mathrm{var}(\widehat{\sigma})$	$\int\lvert\widehat{\sigma}-\sigma\rvert$
	3/0.5	0.6420	0.1566	0.5700	0.2900	0.0087	**0.0858**	**0.2395**
	3/1	0.6871	0.1757	0.6550	0.3900	0.0110	0.1046	0.2546
	3/2	0.6715	0.1649	0.6400	0.3600	0.0090	0.1350	0.2883
	8/0.5	0.6225	0.1436	0.5000	0.2500	**0.0050**	0.2051	0.2894
50	8/1	0.6731	0.1584	0.6700	0.3300	0.0113	0.1105	0.2668
	8/2	0.6679	0.1611	0.6800	0.3300	0.0120	0.1099	0.2811
	12/0.5	0.6334	**0.1390**	0.6100	**0.2400**	0.0087	0.1352	0.2636
	12/1	0.6566	0.1582	0.6200	0.3200	0.0109	0.1481	0.2822
	12/2	0.6590	0.1654	0.5950	0.3400	0.0082	0.1319	0.2836
	6/0.5	0.6355	0.1202	0.6300	0.2200	0.0013	0.1679	0.2479
	6/1	0.7010	0.1373	0.7300	0.2600	0.0022	0.1171	0.2554
	6/2	0.7341	0.1585	0.7900	0.3500	0.0052	0.1527	0.2861
	10/0.5	0.6155	0.1129	0.6100	**0.1900**	0.0098	**0.0630**	**0.2100**
100	10/1	0.7365	0.1380	0.7900	0.2200	0.0052	0.1337	0.2549
	10/2	0.7341	0.1615	0.7900	0.3900	0.0073	0.1114	0.2642
	15/0.5	0.6308	**0.1102**	0.6300	0.2000	**0.0011**	0.1760	0.2459
	15/1	0.7109	0.1411	0.7400	0.2500	0.0060	0.1353	0.2728
	15/2	0.7512	0.1468	0.8000	0.2200	0.0070	0.1622	0.2766
	10/0.5	0.6177	0.0945	0.6100	0.1500	**0.0030**	0.0648	0.1891
	10/1	0.7248	0.0957	0.7400	0.1100	0.0059	0.1111	0.2288
	10/2	0.7904	0.1178	0.8300	0.1800	0.0052	0.1534	0.2764
	15/0.5	0.5989	**0.0873**	0.5950	0.1600	0.0036	**0.0542**	**0.1835**
200	15/1	0.7336	0.0955	0.7500	**0.1000**	0.0069	0.0766	0.2267
	15/2	0.7777	0.1281	0.8350	0.1900	0.0040	0.1582	0.2777
	25/0.5	0.6138	0.0980	0.6150	0.1900	0.0038	0.0595	0.1875
	25/1	0.7374	0.1008	0.7650	0.1200	0.0073	0.0756	0.2191
	25/2	0.7994	0.1206	0.8500	0.1100	0.0083	0.1143	0.2666

Table 1: $\theta = 0.75$

Table 2. Cumulants by Frequency

	Daily	Weekly	Monthly
Mean (×100)	0.0293	0.1406	0.6064
St. Deviation (×100)	0.0381	0.1999	0.9034
Skewness	-1.5458	-0.3746	-0.5886
Excess Kurtosis	43.3342	6.5215	5.5876

Note: Descriptive statistics for the returns on the S&P500 index for the period 1955-2002 for three different data frequencies.

Table 3. Parametric Estimation

	Daily	Weekly	Monthly
ω	0.009183	0.032703	0.463794
	(0.000798)	(0.006052)	(0.121070)
θ	0.921486	0.848581	0.466191
	(0.002349)	(0.015381)	(0.156040)
γ	0.035695	0.054402	−0.076207
	(0.002892)	(0.013415)	(0.039192)
δ	0.071410	0.130121	0.266446
	(0.003100)	(0.018849)	(0.092070)

Note: Standard errors in parentheses. These estimates are for the standardized data series and refer to the AGARCH model

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