Semiparametric Estimator of Time Series Conditional Variance

Santosh Mishra*, Liangjun Su†, and Aman Ullah‡

July 2008

Abstract

We propose a new combined semiparametric estimator, which incorporates the parametric and nonparametric estimators of the conditional variance in a multiplicative way. We derive the asymptotic bias, variance, and normality of the combined estimator under general conditions. We show that under correct parametric specification, our estimator can do as well as the parametric estimator in terms of convergence rates; whereas under parametric mis-specification our estimator can still be consistent. It also improves over the nonparametric estimator of Ziegelman (2002) in terms of bias reduction. The superiority of our estimator is verfied by Monte Carlo simulations and empirical data analysis.

Key Words: Semiparametric Models, Nonparametric Estimator, Conditional Variance

JEL Classifications: C3, C5, G0.

*Department of Economics, Oregon State University, Corvallis, OR, 97330, U.S.A.; e-mail: santosh.mishra@oregonstate.edu.

†School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; e-mail: ljsu@smu.edu.sg.

‡Corresponding Author. Department of Economics, University of California, Riverside, CA 92521-0427, U.S.A., Tel: (951) 827-1591, Fax: (951) 787-5685, e-mail: aman.ullah@ucr.edu.

1 Introduction

There are two main approaches to modeling volatility as the conditional variance σ_t^2 for a stochastic process y_t. The first is based upon the GARCH family of models, which was pioneered by Engle (1982) and soon spawned a plethora of complicated models to capture empirically stylized facts. The second is based upon the stochastic volatility model, which treats σ_t^2 as a latent variable and is expressed as a mixture of predictable and noise components. Both approaches provide parametric models of conditional variance σ_t^2. It is well known that when the parametric model is correctly specified it gives a consistent estimator of σ_t^2, whereas when the parametric model is incorrectly specified the resulting volatility estimator is usually inconsistent with σ_t^2.

Nonparametric and semiparametric estimation of conditional variance can provide consistent estimation of σ_t^2. Pagan and Schwert (1990) and Pagan and Hong (1990) are among the initial works in nonparametric ARCH literature. Härdle and Tsybakov (1997) and Härdle et al. (1998) deal with univariate and multivariate local linear fit for conditional variance, respectively. Fan and Yao (1998) and Ziegelmann (2002) model squared residuals (from the conditional mean equation) nonparametrically to capture the volatility dynamics. Most of the above literature includes one lag in the conditional variance model. Additive and multiplicative models have tried to capture the impact of several period lags in the conditional variance equation. Yang et al. (1999) propose a model where the conditional mean is additive and the conditional variance equation is given as $\sigma_t^2 = \gamma_0 \Pi_{j=1}^d \sigma_j^2(y_{t-j})$, where γ_0 is an unknown parameter and $\sigma_j^2(\cdot)$, $j = 1, ..., d$, are smooth but unknown functions. Linton and Mammen (2005) propose a semiparametric ARCH(∞) model where the conditional variance equation is given as $\sigma_t^2(\boldsymbol{\gamma}, m) = \mu_t + \sum_{j=1}^{\infty} \psi_j(\boldsymbol{\gamma}) m(y_{t-j})$, where $\boldsymbol{\gamma}$ is a finite dimensional parameter, $m(\cdot)$ is a smooth but unknown function and the functional forms of $\psi_j(\cdot)$, $j = 1, 2, ...$, are known. Yang (2006) extends the GJR model (Glosten et al., 1993) to the semiparametric framework.

As explained in the next section, our paper takes a different approach where we combine a parametric estimation with a subsequent nonparametric estimation. We first model the parametric part of the conditional variance and then model the conditional variance of the standardized residual (nonparametric correction factor) nonparametrically capturing some features of σ_t^2 that the parametric model may fail to capture. Thus the combined heteroskedastic model is a multiplicative combination of the parametric model and the nonparametric model for the correction factor. The global parametric estimate of σ_t^2 can be obtained by using any parametric model based on the quasi maximum likelihood estimation (QMLE) principle, say. The estimate of the nonparametric correction factor can be obtained by various nonparametric methods, including the local linear or exponential estimation technique. The idea behind the combined estimation is that if the parametric estimator of σ_t^2 captures some information about the true shape of σ_t^2, the standardized residual will be less volatile

than σ_t^2 itself, which makes it easy to estimate the nonparametric correction factor.

It is worth mentioning that our semiparametric approach has a close analogy with the well known prewhitening method in the time series literature. See Press and Tukey (1956), Andrews and Monahan (1992), among many others. The method has already been applied in the context of kernel density estimation by Hjort and Glad (1995) and in the context of conditional mean regression by Glad (1998). The seminonparametric estimators of Gallant (1981, 1987) and Gallant, and Tauchen (1989) employ a parametric component plus a flexible nonparametric component. Other combined estimators for density function or conditional mean function include Olkin and Spiegelman (1987), and Fan and Ullah (1999).

Built on aforementioned works, our paper provides asymptotic theory for the asymptotic bias, variance, and normality of the semiparametric estimator. It presents several potential improvements over both pure parametric and nonparametric estimators. First, in the case where the parametric model is misspecified so that the parametric estimator for the true volatility is usually inconsistent, our semiparametric estimator can still be consistent with the volatility. Second, in the case where the parametric model is correctly specified, as expected, our semiparametric estimator is generally less efficient than the parametric estimator but it can do as well as the parametric estimator in terms of convergence rates. Third, in comparison with the nonparametric estimator of Ziegelmann (2002), our estimator can result in bias reduction as long as the parametric model can capture some roughness feature of the true volatility function, whereas the two estimators have the same asymptotic variance. In case of correct specification in the first stage parametric modeling, our semiparametric estimator beats the Ziegelmann's estimator. Also, our estimator for the conditional volatility allows the conditioning variable to be estimated from the data and it incorporates the case where the information set is of infinite dimension but can be summarized by a finite dimensional conditioning variable (as in the typical GARCH framework), in sharp contrast with the setup in Ziegelmann (2002).

The structure of the paper is as follows. In Section 2 we introduce the semiparametric model and estimator, and study in detail the asymptotic properties of the semiparametric estimator without and with the assumption of correct parametric specification in the first stage. Section 3 provides simulations and empirical data analysis. Final remarks are contained in Section 4. All technical details are relegated to the Appendix.

2 The Semiparametric Estimation

2.1 The Model and Its Reformulation

In this paper we consider the model

$$y_t = g(U_t, \boldsymbol{\alpha}^0) + \varepsilon_t,\ t = 1, ..., n, \tag{2.1}$$

where the error term ε_t satisfies $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$, $\mathcal{F}_{t-1}$ is the information set at time $t-1$, U_t and $\boldsymbol{\alpha}^0$ are vectors of regressors and pseudo-true parameters, respectively. Since the seminal paper of Engle (1982), a vast literature has developed on the specification of σ_t^2, among which the (G)ARCH family of models play a fundamental role. The majority of the GARCH family of models are specified parametrically, which is subject to the issue of misspecification. In case of misspecification, the conditional variance may not be estimated consistently even though the parameter estimator in the volatility model is still consistent for some pseudo-true parameter. Similar remarks also hold for the class of stochastic volatility models.

Here we propose a semiparametric estimator for the conditional variance based on a multiplicative combined model. The point of divergence from the existing literature is that we represent the conditional variance in two parts: parametric and nonparametric. Analogous to Glad (1998), the idea builds on the simple identity

$$E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma_{p,t}^2 E\left\{\left(\frac{\varepsilon_t}{\sigma_{p,t}}\right)^2 |\mathcal{F}_{t-1}\right\}, \tag{2.2}$$

where $\sigma_{p,t}^2 \in \mathcal{F}_{t-1}$ is determined by a parametric specification of the conditional variance. Let $\sigma_{np,t}^2 = E\{(\varepsilon_t/\sigma_{p,t})^2 |\mathcal{F}_{t-1}\}$. We then have

$$\sigma_t^2 = \sigma_{p,t}^2 \sigma_{np,t}^2. \tag{2.3}$$

The key point is that if the parametric specification $\sigma_{p,t}^2$ captures some roughness features of σ_t^2, the "nonparametric correction factor" $\sigma_{np,t}^2$ will be easier to estimate. In the extreme case when the parametric part $\sigma_{p,t}^2$ is correctly specified, we would hope $\sigma_{np,t}^2$ to be constant over time. To facilitate the presentation, we make the following assumption.

Assumption A0. $\sigma_{p,t}^2 = \sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)$ and $\sigma_{np,t}^2 = E\{(\varepsilon_t/\sigma_{p,t})^2 |\mathcal{F}_{t-1}\} = \sigma_{np}^2(X_{2,t})$, where $\boldsymbol{\gamma}^0$ is the pseudo-true parameter, and $X_{1,t}$ and $X_{2,t}$ are a $d_1 \times 1$ and $d_2 \times 1$ vector, respectively.

Given Assumption A0, $\sigma_t^2 \equiv \sigma^2(X_t) = \sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)\sigma_{np}^2(X_{2,t})$, where X_t, a $d \times 1$ vector, is a disjoint union of $X_{1,t}$ and $X_{2,t}$. We are interested in estimating the volatility function $\sigma^2(\cdot)$ at an interior point, $x \in \mathbb{R}^d$. To see how restrictive Assumption A0 is, we make a definition:

Definition (Minimal reducible dimension d^*) *If $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = E(\varepsilon_t^2|Z_t)$ for some $Z_t \in \mathcal{F}_{t-1}$, a vector of finite dimension d^*, we say that the information set $\mathcal{F}_{t-1}$ is reducible. If further*

$E(\varepsilon_t^2|\mathcal{F}_{t-1}) \neq E(\varepsilon_t^2|\widetilde{Z}_t)$ *for any subvector* $\widetilde{Z}_t$ *of* Z_t*, we say that the dimension* d^* *is minimal for estimating* σ_t^2 *and the set* Z_t *is a minimal reducible set.*

Remark 1. Note that the minimal reducible set Z_t is not necessarily unique. For example, if the true data generating process (DGP) for ε_t is a GARCH(1,1) process:

$$\varepsilon_t = \sqrt{\sigma_t^2} v_t,\ \sigma_t^2 = \gamma_0^0 + \gamma_1^0 \sigma_{t-1}^2 + \gamma_2^0 \varepsilon_{t-1}^2, \tag{2.4}$$

where $\{v_t\}$ is a martingale difference sequence (m.d.s.) with mean 0 and variance 1, then σ_t^2 depends on the entire past information set $\mathcal{F}_{t-1}$ only through σ_{t-1}^2 and ε_{t-1}^2. In this case, one can take $Z_t = \left(\sigma_{t-1}^2, \varepsilon_{t-1}^2\right)^T$ or $Z_t = \left(\sigma_{t-1}^2, \varepsilon_{t-1}\right)^T$ in the above definition. In either case, $d^* = 2$ is minimal.

Remark 2. Clearly, Assumption A0 is a dimension reduction assumption, which is crucial for our asymptotic analysis. On the surface, both parts of the conditional variance, $\sigma_{p,t}^2$ and $\sigma_{np,t}^2$, may depend on a possibly infinite dimension of information in the information set $\mathcal{F}_{t-1}$. Assumption A0 requires that it is possible to summarize the information into a finite dimensional stochastic variable $X_{1,t}$ or $X_{2,t}$. This assumption looks restrictive but is not as restrictive as it appears. Suppose the true DGP for ε_t is a GARCH (1,1) process in (2.4). If we correctly specify the parametric conditional variance model, then $\sigma_{p,t}^2 = \gamma_0^0 + \gamma_1^0 \sigma_{p,t-1}^2 + \gamma_2^0 \varepsilon_{t-1}^2$ and $\sigma_{np,t}^2 \equiv 1$. In this case we can simply set $X_{1,t} = (\sigma_{p,t-1}^2, \varepsilon_{t-1}^2)^T$ and $X_{2,t} = X_{1,t}$ or 1, say; $\sigma_p^2(\cdot,\cdot)$ is affine in both $X_{1,t}$ and $\boldsymbol{\gamma}^0 \equiv (\gamma_0^0, \gamma_1^0, \gamma_2^0)^T$; and $\sigma_{np}^2(\cdot)$ is a constant function. On the other hand, if we specify the parametric conditional variance model as an ARCH(1) model, so that we can write $\sigma_{p,t}^2 = \gamma_0^* + \gamma_2^* \varepsilon_{t-1}^2$, where (γ_0^*, γ_2^*) is the pseudo-true parameter to which the ARCH(1) parametric estimate converges. In this latter case, we can easily identify $X_{1,t} = \varepsilon_{t-1}^2$ and $X_{2,t} = (\sigma_{t-1}^2, \varepsilon_{t-1}^2)^T$; $\sigma_p^2(\cdot,\cdot)$ is affine in both $X_{1,t}$ and $\boldsymbol{\gamma}^0 \equiv (\gamma_0^*, \gamma_2^*)^T$; and $\sigma_{np}^2(X_{2,t}) = \left(\gamma_0^0 + \gamma_1^0 \sigma_{t-1}^2 + \gamma_2^0 \varepsilon_{t-1}^2\right) / (\gamma_0^* + \gamma_2^* \varepsilon_{t-1}^2)$ is a highly nonlinear function.

Remark 3. X_t is not necessarily observable and it summarizes the entire past information set that has influence on the conditional variance. For example, if one specifies $\sigma_{p,t}^2$ as GARCH(1,1) process: $\sigma_{p,t}^2 = \gamma_0^0 + \gamma_1^0 \sigma_{p,t-1}^2 + \gamma_2^0 \varepsilon_{t-1}^2$, then $\boldsymbol{\gamma}^0 = (\gamma_0^0, \gamma_1^0, \gamma_2^0)^T$, $X_{1,t} = \left(\sigma_{p,t-1}^2, \varepsilon_{t-1}^2\right)^T$. So we can re-write the above process as: $\sigma_{p,t}^2 = \sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0) = \gamma_0^0 + \gamma_2^0 \sigma_{p,t-1}^2 + \gamma_2^0 \varepsilon_{t-1}^2$. Given $X_{1,t}$ and $\boldsymbol{\gamma}^0$, we can recover $\sigma_{p,t}^2$ through the map defined by $\left(X_{1,t}, \boldsymbol{\gamma}^0\right) \to \sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)$. This is sufficient for our purpose. If one believes that $X_{1,t}$ summarizes all the past information that affects the conditional variance at time t but has some doubt on the correct specification of the GARCH $(1,1)$ model, one can choose $X_{2,t} = X_{1,t}$. As a matter of fact, either $X_{1,t}$ or $X_{2,t}$ or both may depend on the unknown parameter

$$\boldsymbol{\theta}^0 = \left(\boldsymbol{\alpha}^{0T}, \boldsymbol{\gamma}^{0T}\right)^T.$$

For this reason, we shall write $X_{1,t}$, $X_{2,t}$, and X_t explicitly as $X_{1,t}\left(\boldsymbol{\theta}^0\right)$, $X_{2,t}\left(\boldsymbol{\theta}^0\right)$, and $X_t\left(\boldsymbol{\theta}^0\right)$.

Remark 4. We don't need correct specification for modeling the parametric component. To be concrete, we focus on the case where the set of finite dimensional parameters are estimated by QMLE

technique. Suppose the data $\{y_t, U_t\}$ are generated from the joint density $f(y|u)f(u)$, where $f(u)$ is the marginal density of U_t and $f(y|u)$ is the density of y_t given $U_t = u$. Let $f_{(\boldsymbol{\alpha}^0,\boldsymbol{\gamma}^0)}(y|u)f(u)$ be the density under the chosen parametric assumption. Whether the parametric model is true or not, under some regularity conditions for QMLE, the parameters, $\boldsymbol{\alpha}^0$ and $\boldsymbol{\gamma}^0$, can be estimated consistently at the regular $\sqrt{n}$ rate (White, 1994, Chapter 6). One way is to use QMLE to estimate $\boldsymbol{\alpha}^0$ and $\boldsymbol{\gamma}^0$ jointly. The other is, under some orthogonality conditions, to estimate $\boldsymbol{\alpha}^0$ in (2.1) first, say by nonlinear least squares, and then use the residual series $\{\widehat{\varepsilon}_t\}$ to estimate $\boldsymbol{\gamma}^0$. In either case, we can generally establish $\sqrt{n}$-rate consistency for estimating the pseudo-true parameter $\left(\boldsymbol{\alpha}^{0T}, \boldsymbol{\gamma}^{0T}\right)^T$, which is a minimizer of the Kullback-Leibler distance from the true density $f(y|u)f(u)$ to the suggested density $f_{(\boldsymbol{\alpha}^0,\boldsymbol{\gamma}^0)}(y|u)f(u)$.

2.2 The Semiparametric Estimator

To introduce our semiparametric estimator, we first define some notation.

Let $\widehat{\boldsymbol{\alpha}}$ be a consistent estimator of $\boldsymbol{\alpha}^0$ in the conditional mean model (2.1) and $\widehat{\boldsymbol{\gamma}}$ be a consistent estimator of $\boldsymbol{\gamma}^0$ in the parametric conditional variance model specified through the process $\left\{\sigma_p^2\left(X_{1,t}, \boldsymbol{\gamma}^0\right)\right\}$. Denote $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$. Since the processes $\{X_{1,t}\}$, $\{X_{2,t}\}$, and $\{X_t\}$ may be unobservable and depend on the unknown finite dimensional parameter $\boldsymbol{\theta}^0$, we will write them as functions of $\boldsymbol{\theta}^0$, that is, $X_{1,t} \equiv X_{1,t}\left(\boldsymbol{\theta}^0\right)$, $X_{2,t} \equiv X_{2,t}\left(\boldsymbol{\theta}^0\right)$, and $X_t \equiv X_t\left(\boldsymbol{\theta}^0\right)$. Since $\boldsymbol{\theta}^0$ needs to be estimated, we use $\widehat{X}_{1,t}$, $\widehat{X}_{2,t}$, and $\widehat{X}_t$ to denote $X_{1,t}\left(\boldsymbol{\theta}^0\right)$, $X_{2,t}\left(\boldsymbol{\theta}^0\right)$, and $X_t\left(\boldsymbol{\theta}^0\right)$ with $\boldsymbol{\theta}^0$ being replaced by $\widehat{\boldsymbol{\theta}}$. Let $\varepsilon_t(\boldsymbol{\alpha}) \equiv y_t - g(U_t, \boldsymbol{\alpha})$ and $\widehat{\varepsilon}_t = \varepsilon_t(\widehat{\boldsymbol{\alpha}})$. Then $\varepsilon_t = \varepsilon_t\left(\boldsymbol{\alpha}^0\right)$. Define the "standardized" residual as $\widehat{z}_t \equiv \widehat{\varepsilon}_t/\sigma_p(\widehat{X}_{1,t}, \widehat{\boldsymbol{\gamma}})$. Let $\widehat{r}_t \equiv \sigma_p(x_1, \widehat{\boldsymbol{\gamma}})\widehat{z}_t = \widehat{\varepsilon}_t\sigma_p(x_1, \widehat{\boldsymbol{\gamma}})/\sigma_p(\widehat{X}_{1,t}, \widehat{\boldsymbol{\gamma}})$.

Let $r_t \equiv \varepsilon_t\sigma_p(x_1, \boldsymbol{\gamma}^0)/\sigma_p(X_{1,t}, \boldsymbol{\gamma}^0)$. Then by (2.2)-(2.3), Assumption A0, and the law of iterated expectations,

$$\begin{aligned} E\left(r_t^2 | X_{2,t} = x_2\right) &= \frac{\sigma_p^2(x_1, \boldsymbol{\gamma}^0)}{\sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)} E\left[E\left(\varepsilon_t^2 | \mathcal{F}_{t-1}\right) | X_{2,t} = x_2\right] \\ &= \frac{\sigma_p^2(x_1, \boldsymbol{\gamma}^0)}{\sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)} \sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)\sigma_{np}^2(x_2) = \sigma^2(x). \end{aligned}$$

So in principle we can consider estimating $\sigma^2(x)$ by regressing r_t^2 nonparametrically on $X_{2,t}$, say, by the Nadaraya-Watson (NW) method or the local linear method. The NW method can ensure the nonnegativity of the estimate but it has the boundary bias problem. The local linear method does not have boundary bias problem but it cannot ensure the nonnegativity of the estimate. We now introduce a method that ensures the nonnegativity of the estimate whose bias on the boundary is of the same asymptotic order as that in the interior.

In the general setup, the local linear estimation of $m(x) \equiv E(Y_t | X_t = x)$ is based upon the approximation

$$m(X_t) \approx m(x) + \dot{m}(x)^T (X_t - x) \tag{2.5}$$

when X_t is close to x. Here $\dot{m}(x) \equiv \partial m(x)/\partial x$. When $m(X_t)$ is positive a.s. and X_t is close to x, we can approximate $m(X_t)$ as

$$m(X_t) = \exp(\log(m(X_t))) \approx \exp\left(a(x) + b(x)^T (X_t - x)\right), \tag{2.6}$$

where $a(x) \equiv \log(m(x))$, and $b(x) \equiv \dot{m}(x)/m(x)$ is the first derivative of $\log(m(x))$. In fact, we can replace the exponential function in (2.6) by any well behaved monotone function Ψ and the log function in (2.6) by the inverse function of Ψ. This motives our estimator of $\sigma^2(x)$ below.

To proceed, we first fit the parametric model to obtain $\widehat{\boldsymbol{\theta}}$. Then we estimate $\sigma^2(x)$ by the local smoothing technique:

$$\widehat{\boldsymbol{\beta}} \equiv \arg\min_{\boldsymbol{\beta}} n^{-1} \sum_{t=1}^{n} \left\{ \widehat{r}_t^2 - \Psi\left(\beta_0 + \sum_{j=1}^{d_2} \beta_j (\widehat{X}_{2,tj} - x_{2,j})\right)\right\}^2 K_h(\widehat{X}_{2,t} - x_2), \tag{2.7}$$

where $\boldsymbol{\beta} \equiv \left(\beta_0, \beta_1, ..., \beta_{d_2}\right)^T \in \mathbb{R}^{d_2+1}$, Ψ is a monotone function that has at least two continuous derivatives on its support, $h \equiv (h_1, ..., h_{d_2})^T$ is a vector of bandwidth parameters, $K(\cdot)$ is a non-negative product kernel of $k(\cdot)$, and $K_h(u) = \Pi_{i=1}^{d_2} h_i^{-1} k(u_i/h_i)$. Note that we have suppressed the dependence of $\widehat{r}_t$ and $\widehat{\boldsymbol{\beta}} \equiv \left(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_{d_2}\right)^T$ on x, a disjoint union of x_1 and x_2. We define the volatility estimator as $\widehat{\sigma}^2(x) = \Psi(\widehat{\beta}_0)$.

In Theorem 2.1 below, we show that $\Psi(\widehat{\beta}_0)$ is consistent for $\sigma^2(x)$. We will prove that this estimator, after being appropriately centered and scaled, is asymptotically normally distributed. Note that when $\Psi(u) \equiv u$, we have the local linear estimator for the conditional variance. When $\Psi(u) \equiv \exp(u)$, we have the local exponential estimator of the conditional variance introduced in Ziegelmann (2002) where the conditional variance is modeled fully nonparametrically as a function of a single observable. One obvious advantage of the local exponential approach over the traditional local linear estimation is to ensure the nonnegativity of the estimator of the conditional variance. In contrast, our situation is different than Ziegelmann (2002) mainly in two aspects. First, our volatility function is specified semiparametrically as a product of a parametric component and a nonparametric component. Second, the arguments in our volatility function may depend on all the past information and have to be estimated from the data.

To simplify notation, let $L(x_2, \boldsymbol{\beta}) \equiv \Psi(\beta_0 + \sum_{j=1}^{d_2} \beta_j x_{2,j})$ and $\sigma_p^2(x_1) \equiv \sigma_p^2(x_1, \boldsymbol{\gamma}^0)$. Define

$$\begin{array}{lll} D_{\boldsymbol{\gamma}}\sigma_p^2(x_1,\boldsymbol{\gamma}) \equiv \frac{\partial \sigma_p^2(x_1,\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, & D_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\sigma_p^2(x_1,\boldsymbol{\gamma}) \equiv \frac{\partial^2 \sigma_p^2(x_1,\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T}, & \dot{\sigma}_p^2(x_1,\boldsymbol{\gamma}) \equiv \frac{\partial \sigma_p^2(x_1,\boldsymbol{\gamma})}{\partial x_1}, \\ \dot{\sigma}_{np}^2(x_2) \equiv \frac{\partial \sigma_{np}^2(x_2)}{\partial x_2}, & \ddot{\sigma}_{np}^2(x_2) \equiv \frac{\partial^2 \sigma_{np}^2(x_2)}{\partial x_2 \partial x_2^T}, & \dot{L}(x_2,\boldsymbol{\beta}) \equiv \frac{\partial L(x_2,\boldsymbol{\beta})}{\partial x_2}, \\ \ddot{L}(x_2,\boldsymbol{\beta}) \equiv \frac{\partial^2 L(x_2,\boldsymbol{\beta})}{\partial x_2 \partial x_2^T}. & & \end{array}$$

For i, $j = 0, 1, 2$, let $\kappa_{ij} \equiv \int u^i k(u)^j \, du$. Define a $(d_2+1) \times (d_2+1)$ matrix:

$$S \equiv \begin{pmatrix} 1 & 0 \\ 0 & \kappa_{21} I_{d_2} \end{pmatrix}, \tag{2.8}$$

where I_{d_2} is a $d_2 \times d_2$ identity matrix.

2.3 Asymptotic Theory for the Semiparametric Estimator under General Conditions

To introduce the asymptotic theory, we make the following set of assumptions.

Assumptions

A1. (i) $\boldsymbol{\alpha}^0$ lies in the interior of a compact set $\mathcal{A} \subset \mathbb{R}^{k_1}$. The first derivative $D_{\boldsymbol{\alpha}} g\left(U_t, \boldsymbol{\alpha}\right)$ of $g\left(U_t, \boldsymbol{\alpha}\right)$ with respect to $\boldsymbol{\alpha}$ exists almost surely (a.s.). $D_{\boldsymbol{\alpha}} g\left(U_t, \boldsymbol{\alpha}\right)$ is Lipschitz continuous in $\boldsymbol{\alpha}$ in the neighborhood of $\boldsymbol{\alpha}^0$.

(ii) $\boldsymbol{\gamma}^0$ lies in the interior of a compact set $\Gamma \subset \mathbb{R}^{k_2}$ such that the process $\{\sigma_p^2(X_{1,t}, \boldsymbol{\gamma}^0)\}_{t=1}^{\infty}$ is bounded below by a constant $\underline{\sigma_p^2} > 0$ and is strictly stationary and ergodic. $X_{2,t}$ has a continuous density $f(x_2)$ which is bounded away from zero at x_2, an interior point of $f\left(\cdot\right)$.

(iii) $L\left(x_2, \boldsymbol{\beta}\right)$ and $\dot{L}\left(x_2, \boldsymbol{\beta}\right)$ are bounded uniformly when both x_2 and $\boldsymbol{\beta}$ are restricted in compact sets.

A2. (i) The process $\left\{U_t, X_t\left(\boldsymbol{\theta}^0\right)\right\}$ is stationary and α-mixing with a mixing coefficient $\alpha(j)$ satisfying $\alpha(j) = O\left(j^{-\gamma}\right)$ with $\gamma > (2\nu - 2)/(\nu - 2)$ and $\nu > 2$.

(ii) Let $D_{\boldsymbol{\theta}} X_{1,t}\left(\boldsymbol{\theta}\right) \equiv (\partial/\partial \boldsymbol{\theta}^T) X_{1,t}\left(\boldsymbol{\theta}\right)$ and $D_{\boldsymbol{\theta}} X_{2,t}\left(\boldsymbol{\theta}\right) \equiv (\partial/\partial \boldsymbol{\theta}^T) X_{2,t}\left(\boldsymbol{\theta}\right)$. $D_{\boldsymbol{\theta}} X_{1,t}\left(\boldsymbol{\theta}\right)$ and $D_{\boldsymbol{\theta}} X_{2,t}\left(\boldsymbol{\theta}\right)$ exist and are Lipschitz continuous in the neighborhood of $\boldsymbol{\theta}^0$. $E\left\|D_{\boldsymbol{\theta}} X_{i,t}\left(\boldsymbol{\theta}^0\right)\right\|^{\nu} < \infty$ for some $\nu > 2$, $i = 1, 2$, where $\|\cdot\|$ denotes the Euclidean norm.

(iii) Write $\varepsilon_t = \sqrt{\sigma_t^2} v_t$. The process $\{v_t\}$ is a stationary m.d.s. such that $E\left(v_t | \mathcal{F}_{t-1}\right) = 0$, $E\left(v_t^2 | \mathcal{F}_{t-1}\right) = 1$. $E(|\varepsilon_t|^{2\nu}) < \infty$ and $E\left\|X_t\right\|^{2\nu} < \infty$ for some $\nu > 2$.

A3. (i) $\sigma_p^2(X_{1t}, \boldsymbol{\gamma})$ has two continuous derivatives in $\boldsymbol{\gamma}$ a.s. in the neighborhood of $\boldsymbol{\gamma}^0$. $\dot{\sigma}_p^2(x_1, \boldsymbol{\gamma})$ is Lipschitz continuous in x_1 for $\boldsymbol{\gamma}$ in the neighborhood of $\boldsymbol{\gamma}^0$. $\sigma_{np}^2(x_2)$ has two continuous derivatives in the neighborhood of x_2. $\dot{\sigma}_{np}^2(x_2)$ is Lipschitz continuous in x_2.

(ii) The gradient $(\mu(x_1, \boldsymbol{\gamma}))$ and Hessian matrix $(\upsilon(x_1, \boldsymbol{\gamma}))$ with respect to $\boldsymbol{\gamma}$ of $\log(\sigma_p^2(x_1, \boldsymbol{\gamma}))$ are Lipschitz continuous in the neighborhood of $\boldsymbol{\gamma}^0$, i.e., for some $\epsilon > 0$ such that $\left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right\| \leq \epsilon$, we have: $\left\|\mu(x_1, \boldsymbol{\gamma}) - \mu(\widetilde{x}_1, \boldsymbol{\gamma})\right\| \leq C_1 \left\|x_1 - \widetilde{x}_1\right\|$, and $\left\|\nu(x_1, \boldsymbol{\gamma}) - \upsilon(\widetilde{x}_1, \boldsymbol{\gamma})\right\| \leq C_1 \left\|x_1 - \widetilde{x}_1\right\|$ for some constant C_1 and all x_1, $\widetilde{x}_1 \in \mathbb{R}^{d_1}$, where $\mu(x_1, \boldsymbol{\gamma}) \equiv D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}) / \sigma_p^2(x_1, \boldsymbol{\gamma})$ and $\nu(x_1, \boldsymbol{\gamma}) \equiv \{\sigma_p^2(x_1, \boldsymbol{\gamma}) D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}) - D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}) \left[D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma})\right]^T\} / \sigma_p^2(x_1, \boldsymbol{\gamma})$. Denote $\mu_0(x_1) \equiv \mu(x_1, \boldsymbol{\gamma}^0)$.

A4. $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = n^{-1/2} \sum_{t=1}^n \varphi_t\left(\boldsymbol{\theta}^0\right) + o_p\left(1\right) \xrightarrow{d} N\left(0, V_{\boldsymbol{\theta}^0}\right)$.

A5. (i) The kernel K is a product kernel of k that is a symmetric density with compact support on $\mathbb{R}$.

(ii) $|k(u) - k(v)| \leq C_2 |u - v|$ and $|\dot{k}(u) - \dot{k}(v)| \leq C_2 |u - v|$ for some finite constant C_2 and all u, v on the support of k, where $\dot{k}(u)$ is the first derivative of $k(u)$.

A6. As $n \to \infty$, (i) $h = (h_1, ..., h_{d_2}) \to 0$, (ii) $n \left\|h\right\| \max(\Pi_{i=1}^{d_2} h_i, \left\|h\right\|) \to \infty$, and (iii) $n(\Pi_{i=1}^{d_2} h_i) \left\|h\right\|^4$

$\to C_3 \in [0, \infty)$.

Assumption A1 is standard in the literature. In particular, Assumption A1(ii) can be met for the GARCH family of processes as long as the intercept term is strictly positive, and Assumption A1(iii) is used to show the uniform convergence of some stochastic object. Note that the α-mixing condition in Assumption A2(i) is weaker than β-mixing condition in Hall et al. (1999) and Assumption A2(iii) does not require v_t to be i.i.d. so that its higher order moments may depend on $X_t \in \mathcal{F}_{t-1}$. Assumption A3 imposes the smoothness properties of $\sigma_p^2(x_1, \boldsymbol{\gamma})$ and can easily be satisfied by a variety of GARCH-type models. Assumption A4 follows from asymptotic normality results for QMLE estimators (e.g., Lee and Hansen (1994) and Lumsdaine (1996) for the GARCH(1,1) case and Berkes et al. (2003) and Berkes and Horváth (2003) for the GARCH(p, q) case). Assumptions A5 is similar to Assumption C2 in Hall et al. (1999). As they remark, the requirement that K is compactly supported can be removed at the cost of much lengthier arguments used in the proofs, and in particular, Gaussian kernel is allowed. Assumption A6(i) is standard. Assumption A6(ii) is used in the proof of Theorem 2.1, and given A6(i) it is stronger than the usual requirement $n\Pi_{i=1}^{d_2} h_i \to \infty$ because we use the Taylor expansion in the approximation of $K_h(\widehat{X}_{2,t} - x_2)$ by $K_h(X_{2,t} - x_2)$. Assumption A6(iii) is used toward the end of the proof of Theorem 2.2. Intuitively speaking, it is needed to ensure that the bias order $O(\|h\|^2)$ is of the order no bigger than $(n\Pi_{i=1}^{d_2} h_i)^{-1/2}$.

Theorem 2.1 below shows that the estimator $\widehat{\boldsymbol{\beta}}$ defined in (2.7) converges to $\boldsymbol{\beta}^0$ in probability, where $\boldsymbol{\beta}^0$ is uniquely defined by $\sigma^2(x) = L(0, \boldsymbol{\beta}^0)$ and $\sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2) = \dot{L}(0, \boldsymbol{\beta}^0)$.

Theorem 2.1 *Under Assumptions A0-A6, we have*

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^0,$$

where $\boldsymbol{\beta}^0$ is uniquely defined by $\sigma^2(x) = L(0, \boldsymbol{\beta}^0)$ and $\sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2) = \dot{L}(0, \boldsymbol{\beta}^0)$.

For the proof of the above theorem, see Appendix A. Let $\dot{\Psi}(u)$ denote the first derivative of $\Psi(u)$ with respect to u. Write $\boldsymbol{\beta}^0 = (\beta_0^0, \boldsymbol{\beta}_1^{0T})^T$. Note that $L(0, \boldsymbol{\beta}^0) = \Psi\left(\beta_0^0\right)$ and $\dot{L}(0, \boldsymbol{\beta}^0) = \dot{\Psi}\left(\beta_0^0\right)\boldsymbol{\beta}_1^0$. Then Theorem 2.1 implies that

$$\beta_0^0 \equiv \Psi^{-1}\left(\sigma^2(x)\right) \text{ and } \boldsymbol{\beta}_1^0 \equiv \frac{\sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2)}{\dot{\Psi}\left(\Psi^{-1}\left(\sigma^2(x)\right)\right)},$$

where $\Psi^{-1}(\cdot)$ is the inverse function of $\Psi(\cdot)$.

The following theorem states the asymptotic normality of the estimator $\widehat{\sigma}^2(x)$.

Theorem 2.2 *Under Assumptions A0-A6, we have*

$$\begin{aligned}
&\sqrt{n\Pi_{i=1}^{d_2} h_i}\left\{\widehat{\sigma}^2(x) - \sigma^2(x) - \frac{\kappa_{21}}{2} tr\left\{D_h[\sigma_p^2(x_1)\ddot{\sigma}_{np}^2(x_2) - \ddot{L}\left(0, \boldsymbol{\beta}^0\right)]\right\}\right\} \\
&\xrightarrow{d} N(0, \kappa_{02}^{d_2}(E(v_t^4|X_{2,t} = x_2) - 1)f^{-1}(x_2)\sigma^4(x)),
\end{aligned} \tag{2.9}$$

where recall $\kappa_{ij} \equiv \int u^i k(u)^j du$ and $D_h \equiv diag\left(h_1^2, ..., h_{d_2}^2\right)$.

Remark 5. In the case where $X_{1,t} = X_{2,t} \in \mathbb{R}^1$, we write the bandwidth $h = h_1$. Then the asymptotic variance of our estimator is $\kappa_{02}(E(v_t^4|X_t = x) - 1)f(x)^{-1}\sigma^4(x)/(nh)$ and the bias is $B(x) \equiv (\kappa_{21}h^2/2)[\sigma_p^2(x)\ddot{\sigma}_{np}^2(x) - \ddot{L}\left(0, \boldsymbol{\beta}^0\right)]$. We consider two choices for Ψ and compare our result to that in Theorem 1 of Ziegelmann (2002) for the same bandwidth h and kernel K. For each case, we can see that the two estimators have exactly the same asymptotic variance and this result does not depend on the particular choice of Ψ.

(1) If we take $\Psi(u) \equiv \exp(u)$, then Theorem 2.1 implies that $\boldsymbol{\beta}^0 = (\beta_0^0, \beta_1^0)^T = (\log(\sigma^2(x)), \dot{\sigma}_{np}^2(x)/\sigma_{np}^2(x))^T$. Hence

$$\ddot{L}\left(0, \boldsymbol{\beta}^0\right) = \left.\frac{\partial^2 \exp\left(\beta_0^0 + \beta_1^0 x\right)}{\partial x^2}\right|_{x=0} = \exp(\beta_0^0)\left(\beta_1^0\right)^2 = \frac{\sigma_p^2(x)\left(\dot{\sigma}_{np}^2(x)\right)^2}{\sigma_{np}^2(x)},$$

so that $B(x) = \frac{h^2\kappa_{21}}{2}\frac{\partial^2 \log(\sigma_{np}^2(x))}{\partial x^2}\sigma^2(x)$ for our estimator. To compare with the bias for Ziegelmann's (2002) estimator, one can write the bias of his estimator as $\widetilde{B}(x) = \frac{h^2\kappa_{21}}{2}\frac{\partial^2 \log(\sigma^2(x))}{\partial x^2}\sigma^2(x)$. This means that our estimator will achieve bias reduction if one can choose $\sigma_p^2(x)$ in such a way that

$$\left|\frac{\partial^2 \log(\sigma_{np}^2(x))}{\partial x^2}\right| < \left|\frac{\partial^2 \log(\sigma^2(x))}{\partial x^2}\right|. \tag{2.10}$$

In other words, if $\sigma_p^2(x)$ can capture some of the shape features of $\sigma^2(x)$ in the neighborhood of x, $\log\big(\sigma_{np}^2(x)\big)$ will be less rough than $\log\big(\sigma^2(x)\big)$ itself so that (2.10) can be easily satisfied and we achieve bias reduction. Also, in terms of global approximated weighted mean squared error, our estimator is better than that of Ziegelmann's (2002) if

$$\int\left\{\frac{\partial^2 \log(\sigma_{np}^2(x))}{\partial x^2}\right\}^2 w(x)dx < \int\left\{\frac{\partial^2 \log(\sigma^2(x))}{\partial x^2}\right\}^2 w(x)dx, \tag{2.11}$$

where $w(x)$ is a nonnegative weighting function.

(2) If we take $\Psi(u) \equiv u$, then $\ddot{L}\left(0, \boldsymbol{\beta}^0\right) = 0$ so that the bias for our estimator is $B(x) = \frac{h^2\kappa_{21}}{2}\sigma_p^2(x)\ddot{\sigma}_{np}^2(x)$ and that for Ziegelmann's (2002) estimator is $\widetilde{B}(x) = \frac{h^2\kappa_{21}}{2}\ddot{\sigma}^2(x)$, where $\ddot{\sigma}^2(x)$ is the second derivative of $\sigma^2(x)$. We will achieve bias reduction provided that

$$\left|\sigma_p^2(x)\ddot{\sigma}_{np}^2(x)\right| < \left|\ddot{\sigma}^2(x)\right|. \tag{2.12}$$

Like in Glad (1998), if the initial parametric choice $\sigma_p^2(x)$ happens to be proportional to the true volatility $\sigma^2(x)$, the nonparametric correction factor $\sigma_{np}^2(x)$ is constant and hence the bias reduces to a negligible order for all potential values of C_3 in Assumption A6. If $\sigma_p^2(x)$ captures some of the shape features of $\sigma^2(x)$, $\sigma_{np}^2(x)$ will be less rough than $\sigma^2(x)$ itself so that (2.12) can be maintained.

For general function Ψ, bias reduction can be achieved with similar weak requirement. That is, the parametric component $\sigma_p^2(\cdot)$ bears some information on the shape of $\sigma^2(\cdot)$ in the neighborhood

of x. In the case when the parametric component is correctly specified, i.e., $\sigma^2(x) = \sigma_p^2(x_1, \boldsymbol{\gamma}^0)$ so that $\sigma_{np}^2(x_2) \equiv 1$, our estimator is bias-free asymptotically ($\sqrt{nh^{d_2}}B(x) = 0$) while the asymptotic bias ($\sqrt{nh^{d_2}}\widetilde{B}(x)$) for the Ziegelmann's (2002) estimator does not vanish for the case where $C_3 > 0$ in Assumption A6. In the special case when $C_3 = 0$, the bias for both estimators are asymptotically negligible but different in higher order bias. In short, in case of correct specification for the parametric component, our semiparametric estimator always beats the Ziegelmann's estimator in terms of (integrated) mean squared error.

Remark 6. Using the notation and theories developed in Masry (1996a, 1996b), one can easily generalize our theory to allow for the use of higher order polynomials in (2.7). Generally speaking, higher order local polynomial will help reduce the bias but demand more data due to "sparsity".

Remark 7. Based on Theorem 2.2, we can develop a nonparametric test for the adequacy of the parametric conditional variance model. The null hypothesis is

$$H_0 : \sigma_{np}^2 (X_{2,t}) = 1 \text{ a.s.},$$

and the alternative hypothesis H_1 is the negation of H_0. Under H_0, the parametric conditional variance model is correctly specified so that $\sigma_{np,t}^2 \equiv \sigma_{np}^2 (X_{2,t}) = 1$ a.s. $\sigma_{np}^2 (\cdot)$ is a non-unity function under the alternative. Let $u_t \equiv \varepsilon_t^2/\sigma_{p,t}^2 - 1$. Then the null hypothesis can be written as $H_0 : E(u_t|X_{2,t}) = 0$ a.s. We can construct consistent tests of H_0 versus H_1 using various distance measures. One convenient measure is

$$J \equiv E\left[u_t E(u_t|X_{2,t}) f(X_{2,t})\right],$$

because $J = E\left\{\left[E(u_t|X_{2,t})\right]^2 f(X_{2,t})\right\} \geq 0$ and $J = 0$ if and only if H_0 holds. The sample analog of J is

$$J_n \equiv \frac{1}{n^2 \Pi_{i=1}^{d_2} b_i} \sum_{t=1}^{n} \sum_{s \neq t}^{n} \widehat{u}_t \widehat{u}_s K_b \left(\widehat{X}_{2,t} - \widehat{X}_{2,s}\right)$$

where $\widehat{u}_t$ and $\widehat{X}_{2,t}$ consistently estimate u_t and $X_{2,t}$ under H_0 and $b = (b_1, ..., b_{d_2})$ is the bandwidth sequence. A statistic of this type has been recently used by Hsiao and Li (2001) to test for conditional heteroskedasticity where u_t can be estimated at $\sqrt{n}$-rate under the null. We conjecture that one can extend their analysis and show that after being suitably scaled, J_n converges to the standard normal distribution under the null and it diverges to infinity under the alternative. The detailed analysis is beyond the scope of this paper.

Remark 8. In the above analysis, we did not restrict the density function $f(\cdot)$ of $X_{2,t}$ to be compactly supported. In the case where $f(\cdot)$ is compactly supported, it is worthwhile to study the behavior of the estimator at the boundary of the support. Without loss of generality, assume that $d_2 = 1$ and the support of $f(\cdot)$ is $[0, 1]$. In this case, we denote the bandwidth simply as $h \equiv h(n)$ and consider the left boundary point $x_{n2} = ch$, where c is a positive constant. Also, following the

literature, we assume that $f(0) \equiv \lim_{x_2 \downarrow 0} f(x_2)$ exists and is strictly positive. In this case, we can show that the asymptotic bias (Abias) and variance (Avar) of $\widehat{\sigma}^2(x)$ are given by

$$\text{Abias}\left(\widehat{\sigma}^2(x)\right) = \frac{h^2}{2}\frac{\mu_{c2}^2 - \mu_{c1}\mu_{c3}}{\mu_{c0}\mu_{c2} - \mu_{c1}^2}[\sigma_p^2(x_1)\ddot{\sigma}_{np}^2(0) - \ddot{L}\left(0, \boldsymbol{\beta}^0\right)] \tag{2.13}$$

and

$$\text{Avar}\left(\widehat{\sigma}^2(x)\right) = \frac{\int_{-c}^{\infty}\left(\mu_{c2} - \mu_{c1}z\right)^2 k^2(z)\,dz}{nh\left(\mu_{c0}\mu_{c2} - \mu_{c1}^2\right)^2}\left[E(v_t^4|X_{2,t} = x_2) - 1\right] f^{-1}(0)\sigma^4(x_1, 0), \tag{2.14}$$

where $\mu_{cj} = \int_{-c}^{\infty} z^j k(z)\,dz$ for $j = 0, 1, 2, 3$. See Appendix C for an outline of the proof.

2.4 Asymptotic Theory for the Semiparametric Estimator under Correct Parametric Specification

In this subsection, we will derive the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2(x)$ under the additional assumption that the parametric conditional variance model is correctly specified, i.e., $P(\sigma^2(X_t) = \sigma_p^2\left(X_{1t}, \boldsymbol{\gamma}^0\right)) = 1$ for some $\boldsymbol{\gamma}^0 \in \Gamma$. We will hold the bandwidth sequence $h \equiv (h_1, ..., h_d)$ fixed and demonstrate that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2(x)$ converges to their population true values at the parametric $\sqrt{n}$-rate.

First, we state the consistency of $\widehat{\boldsymbol{\beta}}$.

Theorem 2.3 *Suppose* $P\left(\sigma^2(X_t) = \sigma_p^2\left(X_{1t}, \boldsymbol{\gamma}^0\right)\right) = 1$. *Let* $h \equiv (h_1, ..., h_{d_2})$ *be fixed. Under Assumptions A0-A5, we have*

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^0,$$

where $\boldsymbol{\beta}^0 \equiv (\Psi^{-1}\left(\sigma_p^2(x_1)\right), 0_{1\times d_2})^T$, *and* $\Psi^{-1}(\cdot)$ *is the inverse function of* $\Psi(\cdot)$.

Next, we study the asymptotically normality of the estimator $\widehat{\boldsymbol{\beta}}$.

Theorem 2.4 *Suppose* $P\left(\sigma^2(X_t) = \sigma_p^2\left(X_{1t}, \boldsymbol{\gamma}^0\right)\right) = 1$. *Let* $h \equiv (h_1, ..., h_{d_2})$ *be fixed. Under Assumptions A0-A5, we have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \xrightarrow{d} N\left(0, \left[\dot{\Psi}(\beta_0^0)\right]^{-2}\Sigma_{\boldsymbol{\beta}^0}^{-1}\Omega_{\boldsymbol{\beta}^0}\Sigma_{\boldsymbol{\beta}^0}^{-1}\right),$$

where

$$\Sigma_{\boldsymbol{\beta}^0} \equiv E\left[\begin{pmatrix} 1 & (X_{2,t} - x_2)^T \\ X_{2,t} - x_2 & (X_{2,t} - x_2)(X_{2,t} - x_2)^T \end{pmatrix} K_{ht}\right],$$

$$\Omega_{\boldsymbol{\beta}^0} \equiv \sigma_p^4(x_1)E\left\{\left[E(v_t^4|X_{2,t}{=}x_2) - 1\right] K_{ht}^2\begin{pmatrix} 1 & (X_{2,t} - x_2)^T \\ X_{2,t} - x_2 & (X_{2,t} - x_2)(X_{2,t} - x_2)^T \end{pmatrix}\right\} + \Upsilon V_{\boldsymbol{\theta}^0}\Upsilon^T,$$

$$\Upsilon = E\left[\sigma_p^2(x_1)\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix} K_{ht}\left\{\left[0_{1\times k_1}, (\mu_0(x_1) \text{-} \mu_0(X_{1,t}))^T\right] - \frac{\dot{\sigma}_p^2(X_{1,t}, \boldsymbol{\gamma}^0)^T}{\sigma_p^2(X_{1,t})} D_{\boldsymbol{\theta}} X_{1,t}\left(\boldsymbol{\theta}^0\right)\right\}\right],$$

$K_{ht} \equiv K_h(X_{2,t} - x_2)$, $V_{\boldsymbol{\theta}^0}$ *is defined in Assumption A4, and recall* $\dot{\Psi}(u)$ *denotes the first derivative of* $\Psi(u)$ *with respect to* u.

Note that when h is held fixed, $\Sigma_{\boldsymbol{\beta}^0}$ and $\Omega_{\boldsymbol{\beta}^0}$ are generally non-diagonal matrices. By the delta method, we can prove the following corollary.

Corollary 2.5 *Under the conditions of Theorem 2.4,*

$$\sqrt{n}\left(\widehat{\sigma}^2(x) - \sigma^2(x)\right) \xrightarrow{d} N\left(0, e_1^T \Sigma_{\boldsymbol{\beta}^0}^{-1} \Omega_{\boldsymbol{\beta}^0} \Sigma_{\boldsymbol{\beta}^0}^{-1} e_1\right),$$

where e_1 *is a* $(d_2+1) \times 1$ *vector with 1 in the first entry and 0 elsewhere.*

Remark 9. Corollary 2.5 states that in the case where the parametric conditional variance model is correctly specified, the combined estimator $\widehat{\sigma}^2(x)$ converges to $\sigma^2(x)$ at the parametric rate and is asymptotically normally distributed. A nice feature about $\sqrt{n}\widehat{\sigma}^2(x)$ in Corollary 2.5 is that it is asymptotically unbiased and its asymptotic variance does not depend on the particular tilting function Ψ in use. A second feature about $\sqrt{n}\widehat{\sigma}^2(x)$ is that its asymptotic variance is affected by the first stage estimation of the conditional mean and variance models. The impact of the first stage estimation is accounted for through the term $\Upsilon V_{\boldsymbol{\theta}^0} \Upsilon^T$ in the definition of $\Omega_{\boldsymbol{\beta}^0}$.

Remark 10. To compare our estimator with the parametric estimator of conditional variance, first note that when the parametric component is correctly specified, as expected, our estimator is usually less efficient than the parametric one since our estimator has a slower convergence rate than the parametric estimator when $h \to 0$ as in Theorem 2.2, and generally has a larger asymptotic variance when h is kept fixed as in Corollary 2.5. To see the last point more clearly, we now explicitly calculate the asymptotic variance of the parametric conditional variance estimator $\widehat{\sigma}^2(x_1, \widehat{\boldsymbol{\gamma}})$. Under the correct specification of the parametric conditional variance model,

$$\sqrt{n}\left(\widehat{\sigma}^2(x_1, \widehat{\boldsymbol{\gamma}}) - \sigma^2(x)\right) \xrightarrow{d} N\left(0, \left[D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}^0)\right]^T V_{\boldsymbol{\gamma}_0} \left[D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}^0)\right]\right),$$

where $V_{\boldsymbol{\gamma}_0}$ is the lower-right $k_2 \times k_2$ submatrix of $V_{\boldsymbol{\theta}_0}$. The difference between the two asymptotic variances is given by

$$\begin{aligned} & \operatorname{Avar}\left(\sqrt{n}\widehat{\sigma}^2(x)\right) - \operatorname{Avar}\left(\sqrt{n}\widehat{\sigma}^2(x_1, \widehat{\boldsymbol{\gamma}})\right) \\ = \; & e_1^T \Sigma_{\boldsymbol{\beta}^0}^{-1} \Omega_{\boldsymbol{\beta}^0} \Sigma_{\boldsymbol{\beta}^0}^{-1} e_1 - \left[D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}^0)\right]^T V_{\boldsymbol{\gamma}_0} \left[D_{\boldsymbol{\gamma}} \sigma_p^2(x_1, \boldsymbol{\gamma}^0)\right], \end{aligned}$$

which is difficult to simplify unless both $\Omega_{\boldsymbol{\beta}^0}$ and Υ are diagonal. In this sense, we say that our estimator is as good as the parametric estimator in terms of convergence rates when h is kept fixed, which is consistent with Glad (1998) even though she did not explicitly point this fact out. In contrast, Fan and Ullah (1999) consider a combined estimator of the regression mean in the i.i.d. framework. Their combined estimator is a linear combination of a parametric estimator and a

nonparametric estimator with the weights automatically determined by the data. The parametric rate of convergence of their estimator in case of correct parametric specification can be achieved by letting the bandwidth approach zero.

Remark 11. Like Fan and Ullah (1999), in case of misspecification the parametric conditional variance estimator is usually inconsistent (even though $\widehat{\boldsymbol{\theta}}$ is consistent for some pseudo-true parameter $\boldsymbol{\theta}^0$) while our semiparametric estimator is still consistent. Moreover, our semiparametric estimator can capture some shape structure of the conditional variance function that the parametric estimator fails to capture. Fan and Ullah (1999) also considered the case where the parametric model is approximately correct. In principle, we can extend their work to our framework and study the behavior of our combined estimator in the case where the parametric conditional variance model is approximately correct, that is,

$$\sigma^2(X_t) = \sigma_p^2\left(X_{1,t}, \boldsymbol{\gamma}^0\right) + \delta_n \Delta(X_t) \text{ a.s.},$$

where $\delta_n \to 0$ as $n \to \infty$, and $\Delta(x)$ is continuously differentiable with $|\Delta(x)| < \infty$. We conjecture that the rate of convergence of $\widehat{\sigma}^2(x)$ depends crucially on the magnitude of δ_n in relation to $n^{-1/2}$ and $(n\Pi_{i=1}^{d_2} h_i)^{-1/2}$. For example, if $\delta_n = o\left(n^{-1/2}\right)$, we can show that the result in Corollary 2.5 continues to hold by using fixed bandwidth; if $\delta_n \propto n^{-1/2}$, then $\widehat{\sigma}^2(x)$ continues to converge to $\sigma^2(x)$ at the parametric rate by using fixed bandwidth and it has non-negligible asymptotic bias determined by $\Delta(x)$; if $\delta_n n^{1/2} \to \infty$ and $\delta_n = o((n\Pi_{i=1}^{d_2} h_i)^{-1/2})$, the result of Theorem 2.2 continues to hold for diminishing bandwidth; if $\delta_n(n\Pi_{i=1}^{d_2} h_i) \to c \in (0, \infty]$, then $\delta_n \Delta(x)$ also contributes to the bias of $\widehat{\sigma}^2(x)$ in Theorem 2.2. For brevity, we omit the details.

2.5 Bandwidth Selection

As is the case for all nonparametric curve estimation, the bandwidth parameter plays an essential role in practice. It is desirable to have a reliable data-driven and yet easily-implementable bandwidth selection procedure.

One approach is to apply a "plug-in" method to obtain an estimate of h as described for example in Fan and Gijbels (1996, Ch 4.2 for the single regressor case). Without assuming the correct specification of the parametric conditional variance model, the asymptotic mean integrated squared error (AMISE) of $\widehat{\sigma}^2(x)$ is

$$\text{AMISE}(h) \equiv \int \left[B_h^2(x) + V_h(x)\right] w(x)\, dx \tag{2.15}$$

where $B_h(x) = \frac{\kappa_{21}}{2} \text{tr}\left\{D_h[\sigma_p^2(x_1)\ddot{\sigma}_{np}^2(x_2) - \ddot{L}\left(0, \boldsymbol{\beta}^0\right)]\right\}$, $V_h(x) = \kappa_{02}^{d_2}(E(v_t^4|X_{2,t}=x_2)-1)\, f^{-1}(x_2)\sigma^4(x)$ $/(n\Pi_{i=1}^{d_2} h_i)$, and $w(x)$ is a weighting function. In principle, one can choose the vector h to minimize AMISE(h) but this may not result in an analytic solution. If we restrict that $h_1 = \cdots = h_{d_2} = h_n$,

then we only need to choose a scalar h_n to minimize AMISE(h_n), the solution of which is given by

$$h_n^* = \left\{ \frac{\kappa_{02}^{d_2} \int (E(v_t^4 | X_{2,t}{=}x_2) - 1) f^{-1}(x_2) \sigma^4(x) w(x)\, dx}{\kappa_{21}^2 \int \operatorname{tr}\left\{ [\sigma_p^2(x_1) \ddot{\sigma}_{np}^2(x_2) - \ddot{L}\left(0, \boldsymbol{\beta}^0\right)] \right\} w(x)\, dx} \right\}^{1/(d_2+4)} \times n^{-1/(d_2+4)}. \tag{2.16}$$

Hence h_n^* converges to zero at the rate $n^{-1/(d_2+4)}$. Since h_n^* depends on several unknown quantities, to obtain an estimate for h_n^*, we need to estimate these unknown quantities first. This will generally require the choice of some pilot bandwidth. The performance of our estimate $\widehat{\sigma}^2(x)$ will be contingent upon the choice of such a pilot bandwidth and the estimates of these unknown quantities, which is a disadvantage of this approach. Another drawback of this approach is that it can never be the optimal bandwidth in the case of correct parametric specification. In this latter case, Corollary 2.5 suggests that we should hold the bandwidth fixed in order to achieve the parametric convergence rate of $\widehat{\sigma}^2(x)$. This parallels the case of a general local linear regression where underlying model may be linear or not. If the underlying model is indeed linear, it is well known that we should not let the bandwidth tend to zero. Instead, the bandwidth should be kept fixed.

So we propose to apply the leave-one-out least squares cross validation (LSCV) to obtain the data-driven bandwidth. Let $\widehat{\sigma}_{(-t)}^2(\widehat{X}_t)$ be the leave-one-out analog of $\widehat{\sigma}^2(\widehat{X}_t)$ without using the tth observation in the estimation. We choose $h = (h_1, ..., h_{d_2})$ to minimize the following LSCV criterion function

$$CV(h) = \frac{1}{n} \sum_{t=1}^{n} \left(\widehat{r}_t^2 - \widehat{\sigma}_{(-t)}^2(\widehat{X}_t) \right)^2 w(\widehat{X}_t), \tag{2.17}$$

where $w(\cdot)$ is a nonnegative weight function, e.g., $w(\widehat{X}_t) = \Pi_{i=1}^{d} 1(|\widehat{X}_{t,i} - \overline{x}_i| \leq 2s_i)$ with $\overline{x}_i$ and s_i being the sample mean and standard deviation of $\widehat{X}_{t,i}$, respectively. Let $\widehat{h}$ denote the minimizer of $CV(h)$. We conjecture $\widehat{h}$ converges to the minimizer of AMISE(h) in (2.15) in case of misspecification of the parametric model and is not convergent to zero in case of correct parametric specification. A formal study of the theoretical property of $\widehat{h}$ is beyond the scope of this paper.

3 Simulation and Empirical Analysis

3.1 Simulation

To consider the data generating processes (DGPs), we focus on the case where $y_t = \varepsilon_t$, $\varepsilon_t = \sigma_t v_t$, v_t ~ $i.i.d.N(0, 1)$, $E(\varepsilon_t | \mathcal{F}_{t-1}) = 0$, and $E(\varepsilon_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2$. The dynamics considered for σ_t^2 are specified as follows.

DGP 1: The ARCH(1) model of Engle (1982): $\sigma_t^2 = \gamma_0 + \gamma_1 \varepsilon_{t-1}^2$, where $\gamma_0 = 0.6$ and $\gamma_1 = 0.4$.

DGP 2: The GARCH(1, 1) model of Bollerslev (1986): $\sigma_t^2 = \gamma_0 + \gamma_1 \sigma_{t-1}^2 + \gamma_2 \varepsilon_{t-1}^2$, where $\gamma_0 = 0.03$, $\gamma_1 = 0.94$, and $\gamma_2 = 0.03$.

DGP 3: The threshold GARCH (GJR) model of Glosten et al. (1993):

$$\sigma_t^2 = \gamma_0 + \gamma_1\sigma_{t-1}^2 + \gamma_2\varepsilon_{t-1}^2 + \gamma_3\varepsilon_{t-1}^2\mathbf{1}(\varepsilon_{t-1} \leq 0),$$

where $\gamma_0 = 0.03$, $\gamma_1 = 0.93$, $\gamma_2 = 0.02$, and $\gamma_3 = 0.03$.

DGP 4: Combined GARCH (CGARCH): $\sigma_t^2 = (\gamma_0+\gamma_1\sigma_{t-1}^2+\gamma_2\varepsilon_{t-1}^2)\exp(\gamma_3+\gamma_4\sigma_{t-1}^2+\gamma_5|\varepsilon_{t-1}|)$, where $\gamma_0 = 0.03$, $\gamma_1 = 0.94$, and $\gamma_2 = 0.03$, $\gamma_3 = 0.56$, $\gamma_4 = -0.55$, and $\gamma_5 = 0.02$.

DGP 5: Combined ARCH (CARCH): $\sigma_t^2 = (\gamma_0 + \gamma_1\varepsilon_{t-1}^2)\exp(\gamma_2 + \gamma_3|\varepsilon_{t-1}|)$, where $\gamma_0 = 0.6$, $\gamma_1 = 0.4$, $\gamma_2 = 0.6$, and $\gamma_3 = -0.4$.

DGP 6: Stochastic Volatility (SV): $\log\sigma_t^2 = \gamma_0+\gamma_1\log\sigma_{t-1}^2+\gamma_2\varsigma_t$, where $\gamma_0 = -0.04$, $\gamma_1 = 0.96$, and $\gamma_2 = 0.345$. We also assume that (v_t, ς_t) ~ $i.i.d.N(0, I_2)$, where I_2 is a 2×2 identity matrix.

For each DGP, we estimate the conditional variance by ARCH(1), GARCH(1,1), the Nonparametric Local Exponential (NPLE) estimator of Ziegelmann (2002) and two versions of semiparametric estimators, SPGARCH and SPARCH, which correspond to the first stage GARCH(1,1) and ARCH(1) parametric models, respectively. For both the SPARCH estimation and the NPLE estimation, we choose the conditioning variable $X_{2,t} = y_{t-1}$, while for the SPGARCH estimation, we choose $X_{2,t} = (y_{t-1}, \widehat{\sigma}_{p,t-1}^2)^T$ where $\widehat{\sigma}_{p,t-1}^2$ is the fitted parametric conditional variance in the first stage. In all cases we choose $\Psi(u) = \exp(u)$ for our semiparametric estimator to ensure the nonnegativity of the conditional variance estimator.

To obtain the NPLE estimator and our semiparametric estimator, we need to choose both the kernel and bandwidth. It is well known that the choice of kernel does not play any significant role in nonparametrics so that in all cases we use the normalized Epanechnikov kernel:

$$k(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}u^2\right)1\left(|u| \leq \sqrt{5}\right).$$

In contrast, the choice of bandwidth is very important in nonparametric or semiparametric estimation. To avoid any ambiguity, we use the least squares cross-validation (LSCV) method for both estimators. The LSCV function for our second-stage nonparametric estimator is given in (2.17) and that for the NPLE estimator is similarly defined.

We use i, j, and t to denote the index of replications, models and time, respectively. We draw replications of ν_{it} independently (across both i and t) from the standard normal distribution, and use them to generate ε_{it} through the above specified conditional variance DGPs for $t = -n_0 + 1$, $-n_0 + 2, ..., 1, ..., n$. We throw away the first $n_0 = 500$ observations to avoid the starting-out effect and use sample sizes $n = 100$, 200, and 300. The number of replications is $M = 200$ for each case. Let

$$B_t^j = \frac{1}{M}\sum_{i=1}^{M}[(\sigma_t^j)^2 - (\hat{\sigma}_{it}^j)^2],\ t = 1, ..., n \text{ and } j = 1, ..., 6,$$

$$S_t^j = \frac{1}{M}\sum_{i=1}^{M}[(\hat{\sigma}_{it}^j)^2 - \frac{1}{M}\sum_{i=1}^{M}(\hat{\sigma}_{it}^j)^2]^2,\ t = 1, ..., n \text{ and } j = 1, ..., 6,$$

Table 1: Comparison of mean square errors (MSEs) for various estimators of conditional variance

DGPs\Models	n	ARCH(1)	GARCH(1,1)	NPLE	SPARCH	SPGARCH
ARCH(1)	100	**0.084**	0.096	0.254	0.098	0.228
	200	**0.026**	0.036	0.112	0.040	0.101
	300	**0.010**	0.024	0.076	0.026	0.069
GARCH(1,1)	100	0.077	**0.027**	0.566	0.231	0.086
	200	0.029	**0.013**	0.302	0.086	0.044
	300	0.021	**0.007**	0.189	0.054	0.028
GJR	100	0.075	0.067	0.435	0.092	**0.044**
	200	0.064	0.042	0.420	0.087	**0.040**
	300	0.060	0.039	0.211	0.054	**0.021**
CGARCH	100	0.082	0.029	0.583	0.037	**0.012**
	200	0.076	0.009	0.324	0.031	**0.006**
	300	0.062	0.008	0.215	0.029	**0.004**
CARCH	100	0.131	0.121	0.973	**0.112**	0.119
	200	0.059	0.053	0.491	**0.051**	0.053
	300	0.029	0.042	0.296	**0.026**	0.040
SV	100	0.987	0.606	0.978	1.040	**0.786**
	200	0.415	0.300	0.638	0.606	**0.102**
	300	0.193	0.141	0.225	0.221	**0.037**

NOTE: The DGPs correspond to DGPs 1-6 in the text. We use five models to fit each DGP. The sample sizes are $n = 100$, 200, and 300, and the number of simulations is $M = 200$. The main entries are the mean square errors (MSEs)$\times 10$ associated with different DGPs. The comparison for a particular DGP in a given row is done across the columns, and given the DGP, the best model is the model with the lowest MSE value (boldfaced).

$B^j = [B_1^j, ..., B_n^j]'$, and $S^j = [S_1^j, ..., S_n^j]'$, where $(\sigma_t^j)^2$ is the true conditional variance for model j at time t, and $(\hat{\sigma}_{it}^j)^2$ is the estimated conditional variance for the ith replication and the jth model at time t. Here, B and S stand for bias and variance, respectively. Let $MSE_t^j = (B_t^j)^2 + S_t^j$ be the mean square error of estimates of $(\sigma_t^j)^2$. Thus we calculate the average MSE for model j by $MSE^j = n^{-1}\sum_{t=1}^n MSE_t^j$. We compare MSE^j for all the models and DGPs under study. For a given DGP, the lowest MSE value suggests the best fit of the model.

Table 1 provides the comparison of the MSEs. The MSEs for each DGP are presented across the columns (for example the first row corresponds to DGP ARCH(1)). We have multiplies each MSE value by 10 for the convenience of presentation. The model corresponding to the minimum MSE for a given DGP is the best model. We find the following interesting points. (i) One of the main theoretical findings of the paper is that as long as the first stage parametric conditional variance can capture some shape features of the true conditional variance function, the semiparametric estimators, namely SPGARCH and SPARCH, always dominate the NPLE estimator in terms of the MSEs. This is observed throughout our simulations. (ii) When the true DGP coincides with the fitted parametric model, then the parametric model is the dominant model. For example, ARCH(1) and

Table 2: Mean and mean square error (MSE) for the estimator of the nonparametric component

DGPs\Models	n	SPARCH		SPGARCH	
		mean	mse$\times 10^2$	mean	mse$\times 10^2$
ARCH(1)	100	1.058	0.184	1.130	0.457
	200	1.044	0.083	1.111	0.185
	300	1.040	0.050	1.103	0.104
GARCH(1,1)	100	1.363	0.978	1.086	0.147
	200	1.360	0.263	1.021	0.060
	300	1.395	0.185	1.019	0.038
GJR	100	0.848	0.524	0.921	0.158
	200	0.858	0.208	0.925	0.140
	300	0.862	0.065	0.932	0.036
CGARCH	100	0.860	1.221	1.094	0.057
	200	0.870	0.542	1.089	0.019
	300	0.880	0.316	1.090	0.012
CARCH	100	1.210	0.354	1.312	0.745
	200	1.201	0.188	1.310	0.400
	300	1.194	0.116	1.311	0.234
SV	100	1.782	6.500	1.630	1.840
	200	1.734	5.608	1.564	0.806
	300	1.742	4.308	1.597	0.563

NOTE: This table reports the mean and mse for the estimator of the nonparametric component in the SPGARCH and SPARCH models. A mean value that is close to 1 and a small MSE value indicates a good first stage fit of the parametric conditional variance model to the true model.

GARCH(1,1) are the best models for DGPs ARCH (1) and GARCH (1,1), respectively. (iii) The parametric GARCH model outperforms SPARCH model for all DGPs except for the CARCH DGP. This can be explained by the fact that as the conditioning set of SPARCH model includes only y_{t-1} it cannot suitably capture the conditional variance of DGPs that have conditioning set $\{y_{-\infty}, ..., y_{t-1}\}$. However, we find that SPGARCH model improves over GARCH model in all DGPS except for the GARCH(1,1) and ARCH(1) DGPs. For example, when the sample size is 300 and the true DGP is GJR, the MSE for SPGARCH is 0.0021, and it is 0.0039 for GARCH(1,1). (iv) When the true DGP is the stochastic volatility model, a model fundamentally different in structure compared to the GARCH class of models, we find that SPGARCH is the best model. (v) When the DGPs mimic the class of semiparametric model considered here (i.e. CGARCH and CARCH), we find that either the SPARCH or SPGARCH model dominate all other models. (vi) We note that whenever the first stage parametric model is misspecified SPGARCH model is the best model except for the CARCH case.

Another issue of relevance is the presence of residual nonlinearity. It is useful to analyze how close the second stage nonparametric estimation is to the true values. Since our second stage estimator $\widehat{\sigma}^2(X_t)$ estimates the conditional variance $\sigma_t^2 \equiv \sigma^2(X_t)$ directly, we can recover the estimator of

$\sigma^2_{np}(X_{2,t})$ by $\widehat{\sigma}^2_{np}(X_{2,t}) \equiv \widehat{\sigma}^2(X_t)/\sigma^2_p(X_{1,t},\widehat{\gamma})$. If the first stage parametric conditional variance model is correctly specified, one expects that $\widehat{\sigma}^2_{np}(X_{2,t})$ vary in the neighborhood of unity. This is indeed observed in our simulations. In Table 2 we report the means and MSEs of the $\widehat{\sigma}^2_{np}(X_{2,t})$ for the SPGARCH and SPARCH models under investigation. The MSE values are multiplied by 10^2 for the convenience of presentation. We observe that: (i) When the first stage parametric conditional variance model is correctly specified, the mean of the estimator for the nonparametric component is close to 1 and the MSE of the estimator for the nonparametric component is lower than the case where the first stage model is misspecified. For example, when the DGP is GARCH(1,1) and sample size $n = 300$, the MSE of $\widehat{\sigma}^2_{np}(X_{2,t})$ for the SPGARCH model is 0.00038, which is much lower than 0.00185 for the SPARCH model. (ii) When the first stage parametric model is misspecified, the mean of the estimator for the nonparametric component can be quite different from 1 and the performance of the SPGARCH and SPARCH estimators is mixed. For the CARCH process, we find that the MSE is lower for the SPARCH estimators, but for the GJR, CGARCH and SV processes, we find that the SPGARCH estimator outperforms SPARCH.

3.2 Empirical Data Analysis

In this subsection we fit the semiparametric model to the real data (collected from Datastream) and do some diagnostic checking to see whether the model is able to capture the time-varying conditional variance and leverage effect.

We consider the S&P500 daily returns from January 3rd, 2002 through January 3rd, 2007, a total of 1258 observations. It is sometimes found that the conditional mean of stock return has AR or MA components. We therefore first carry out a diagnostic check for the return series to look for possible AR or MA components. The ACF and PACF of the series indicate no such memory structure. We also fit AR and MA models of lags of order 1 through 4 to check for the statistical significance of the parameter estimates. We also find that all the parameter estimates are statistically insignificant at the 5% level. It is possible to have nonlinear dynamics in the conditional mean equation, but we have not explored this avenue here. Consequently, we model the return series $\{y_t\}_{t=1}^n$ of a financial asset as: $y_t = \alpha + \varepsilon_t$, where $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ and $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$. In addition to the two semiparametric models considered in the simulation, we consider a third semiparametric model where we fit the GJR model in the first stage parametric estimation. It is useful to check whether the second stage estimation can capture some additional nonlinearity in the data.

In the first stage parametric modeling, we estimate α by the sample average of y_t, and estimate the three parametric conditional variance models: ARCH(1), GARCH(1,1), and GJR by using the Gaussian QMLE method. In the second-stage nonparametric estimation, we use y_{t-1} as the conditioning variable for the ARCH(1) case and $(y_{t-1}, \widehat{\sigma}^2_{p,t-1})^T$ for the other two cases; and we denote

Table 3: Estimated parameters in the first stage estimation and decomposition of conditional variance

	$\widehat{\gamma}_0$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	% variation explained by the parametric component
ARCH(1)	0.720 (0.037)	0.27 (0.043)			88.2
GARCH(1,1)	0.004 (0.00025)	0.92 (0.0116)	0.062 (0.0105)		84.4
GJR	0.004 (0.00018)	0.94 (0.0136)	0.023 (0.0085)	0.040 (0.0061)	81.1

NOTE: The parameter interpretations are the same as given in the simulation. Numbers in parentheses are standard errors. The last column indicates the percentage of total variation of the volatility estimator explained by the first stage parametric fit.

the resulting semiparametric model as SPARCH, SPGARCH and SPGJR, respectively. We find that estimate of α is statistically insignificant. Thus we don't provide any result for this estimate. As in the simulations, we use the Epanechnikov kernel for all nonparametric estimation and use the least squares cross validation to choose the bandwidth for our second stage nonparametric estimation.

Table 3 reports the estimates for the three parametric models with corresponding standard errors in parentheses. All the parameter estimates appear significant at the conventional 5% significance level, which means the corresponding pseudo-true values are statistically different from 0. The last column in Table 3 indicates the percentage of variation of our combined volatility estimator that is captured by the first stage parametric estimator, that is, it is the sample variance of $\widehat{\sigma}_p^2(\widehat{X}_{1,t}, \widehat{\gamma})$ over that of $\widehat{\sigma}^2(\widehat{X}_t)$ multiplied by 100. The numbers indicate that the second stage nonparametric estimation can capture 12-19% of the variation of the conditional variance which the parametric model fails to capture.

Another issue of interest is the impact of y_{t-1} (innovation) on the current conditional variance when all other elements of the information set are kept constant. This is often referred to as news impact curve, see Engle and Ng (1993). Figure 1 plots the news impact curve for the ARCH, GARCH, and GJR models, where we have normalized the conditional variance term to be zero when y_{t-1} is zero. For example, for the ARCH model this implies that it is a plot of $\widehat{\sigma}_{p,t}^2 - \widehat{\gamma}_0$ versus y_{t-1}, where $\widehat{\sigma}_{p,t}^2 = \widehat{\gamma}_0 + \widehat{\gamma}_1 y_{t-1}^2$. All the plots have the expected shapes. To understand the role of second stage estimation we also add the plot of $\widehat{\sigma}_{np,t}^2$ versus y_{t-1} to the news impact curves for each parametric model. We find that for all three cases, i.e., SPARCH, SPGARCH, and SPGJR, the second stage nonparametric fit is essentially downward sloping. Thus the multiplicative factor is higher for the negative shocks than that of the positive shocks. The second stage asymmetry captures the leverage effect associated with negative news. Thus if any asymmetry is not explained in the first stage then it is captured in the second stage estimation (even in the case of SPGJR).

Figure 2 provides the pointwise 95% confidence intervals for the conditional variance estimates for all three semiparametric models as a function of y_{t-1} (innovation). For the SPARCH case, there

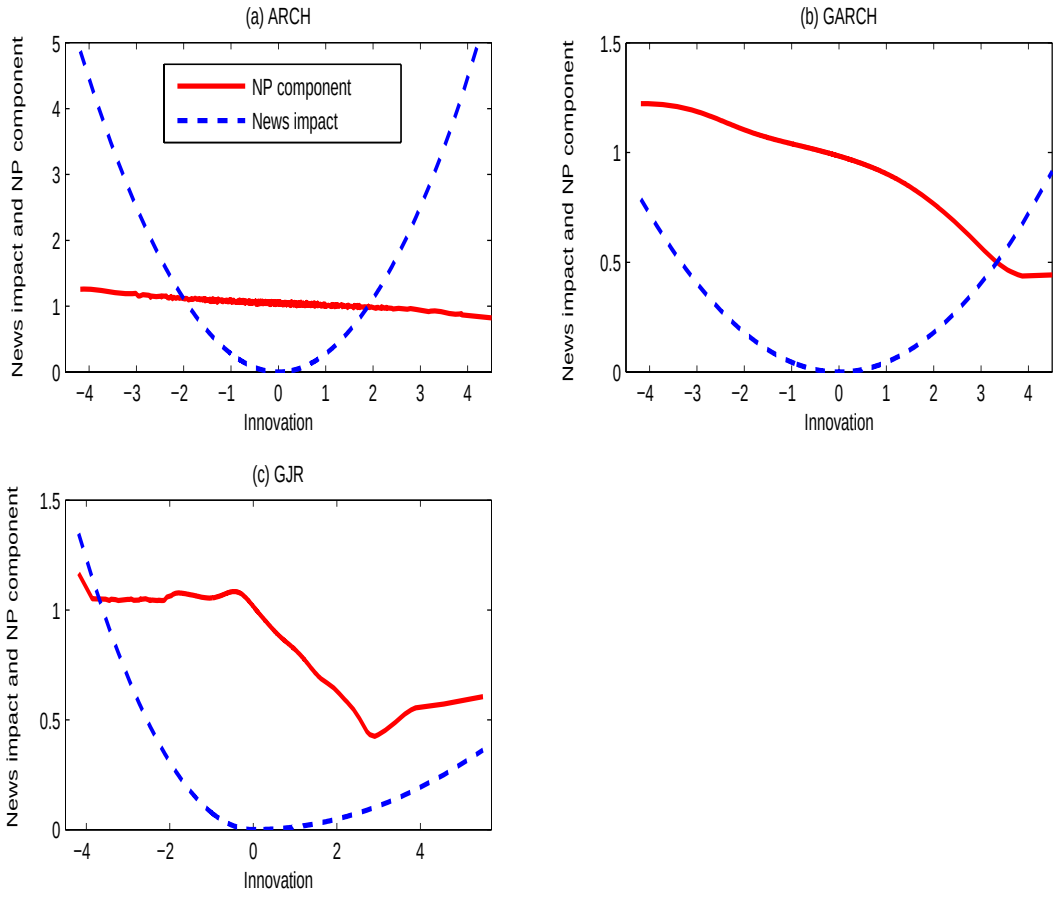

Figure 1: News impact curve and second-stage nonlinearity for SP 500 daily returns

is only one conditioning variable, i.e., y_{t-1}, in the construction of the semiparametric estimator. For the SPGARCH and SPGJR cases, we have two conditioning variables (y_{t-1} and $\widehat{\sigma}^2_{p,t-1}$) to construct the semiparametric estimator and have held $\widehat{\sigma}^2_{p,t-1}$ fixed at their sample mean level in order to obtain plots (b)-(c) in Figure 2. As we can tell from Figure 2, the three semiparametric estimates share similar shapes despite the fact that different parametric models are fitted in the first stage. All of them can capture the leverage effects that are well documented in the finance literature.

Diagnostic check for the standardized residuals is an important aspect of modeling exercise. The basic diagnostic checks include visual inspection of the ACF and PACF plots of standardized residuals. It is also useful to check correlations of polynomials involving standardized residuals to detect nonlinear dependence. It is well known that the standard Box-Pierce statistic is not applicable in our framework due to the presence of conditional heteroskedasticity and nonparametric estimates. Li and Mak (1994) propose an alternative chi-square-based test to account for conditional heteroskedasticity. Nevertheless, their test statistic depends on the estimates of the Hessian matrix from a log-likelihood function (see Lemma 3.3 of Ling and Li (1997) for details), which is not applicable in the presence of nonparametric estimation. Wong and Ling (2005) propose a modification to Li and Mak (1994) which allows for potential parametric misspecification. But the test statistic is posited in a one-step likelihood based framework and not applicable to our case either. Hence we eschew the Box-Pierce type of chi-square test and limit ourselves to the ACF and PACF tests indicated above. We find that SPARCH model doesn't pass the diagnostic test, but the ACF and PACF tests on standardized residuals for the SPGARCH and SPGJR models indicate that the standardized residuals are uncorrelated in levels and for polynomials up to order 3.

4 Conclusion

This paper proposes a new semiparametric estimator for time varying conditional variance. This is accomplished by combining the parametric and nonparametric estimators in a multiplicative way. We provide asymptotic theory for our semiparametric estimator and show that it can improve upon the pure parametric and nonparametric estimators in different scenarios. We also include simulations and empirical applications. We find from the simulations that our semiparametric estimators are superior to their parametric and nonparametric counterparts in terms of MSE criteria. From the empirical analysis we see that semiparametric models can capture the asymmetric effect in the data even beyond that captured by the parametric component in the model.

ACKNOWLEDGEMENTS

The authors gratefully thank Arthur Lewbel, the associate editor, and two anonymous referees

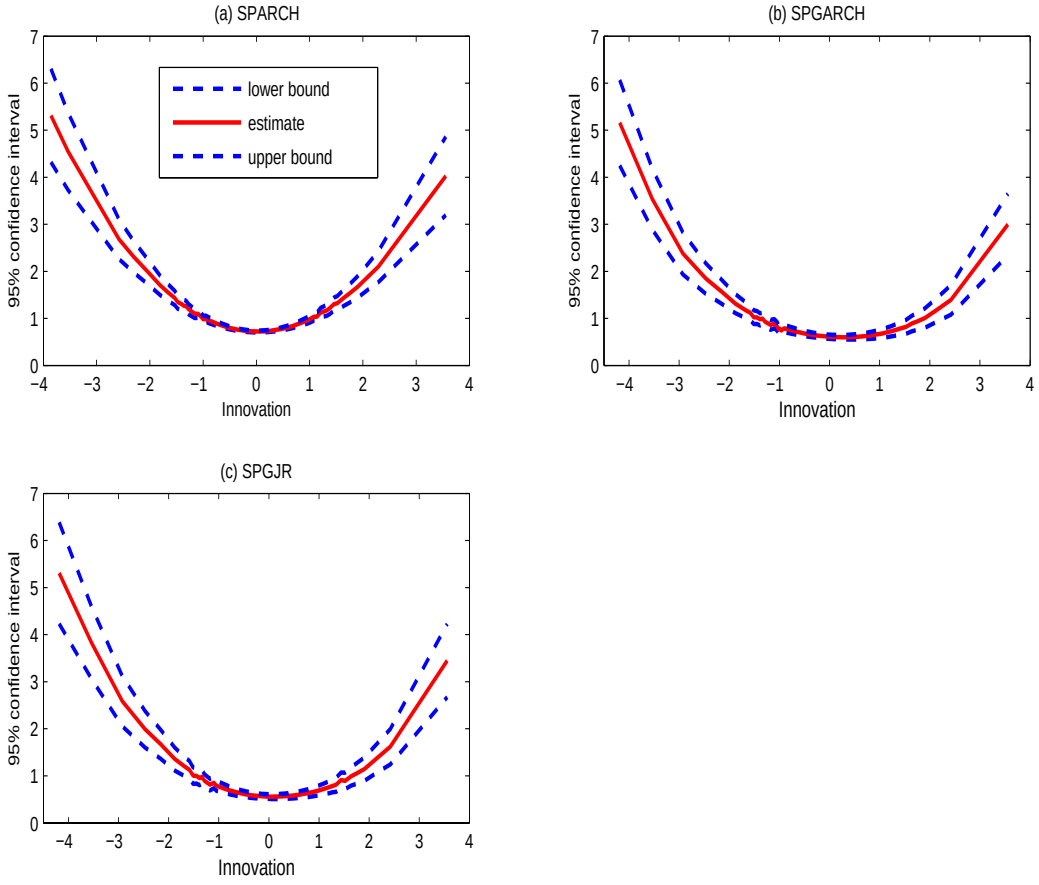

Figure 2: 95% confidence interval for the SPARCH, SPGARCH and SPGJR estimates of conditional variance for SP 500 daily returns

for their many helpful comments and advice that have led to considerable improvements of the presentation. They are grateful to Torben G. Andersen, John Galbraith, Gloria González-Rivera, Qi Li, Essie Maasoumi, Nour Meddahi, and Victoria Zinde-Walsh for their discussions and comments on the subject matter of this paper. They are also thankful for the comments by the participants of the seminars at McGill University, Southern Methodist University, Syracuse University, Texas A&M University, Vanderbilt University, Université de Montréal, and University of Guelph. The second author gratefully acknowledges financial support from the NSFC under the grant numbers 70501001 and 70601001. The third author acknowledges the support from the academic senate, UCR.

Appendix

We use $\|\cdot\|$ to denote the Euclidean norm of $\cdot$, C to signify a generic constant whose exact value may vary from case to case, and a^T to denote the transpose of a. For ease of presentation, we assume that $h_1 = \cdots = h_{d_2}$ and with a little abuse of notation, we further denote $h_1 = \cdots = h_{d_2} = h$. Recall $K_{ht} \equiv K_h(X_{2,t} - x_2)$ and $r_t \equiv \varepsilon_t \sigma_p(x_1)/\sigma_p(X_{1,t})$. Let $\xi_t \equiv r_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(X_{2,t})$.

A Proof of Theorem 2.1

Recall that $\widehat{\boldsymbol{\beta}}$ minimizes

$$\frac{1}{n}\sum_{t=1}^{n}\left\{\widehat{r}_t^2 - L\left(\widehat{X}_{2,t} - x_2, \boldsymbol{\beta}\right)\right\}^2 K_h(\widehat{X}_{2,t} - x_2).$$

It also minimizes the criterion function

$$\widehat{G}_n(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{t=1}^{n}\left\{\left[\widehat{r}_t^2 - L\left(\widehat{X}_{2,t} - x_2, \boldsymbol{\beta}\right)\right]^2 - \left[\left(\widehat{r}_t^2\right)^2 - \left(r_t^2\right)^2\right]\right\} K_h(\widehat{X}_{2,t} - x_2). \tag{A.1}$$

By Theorem 3.4 of White (1994), it suffices to show that

$$\widehat{G}_n(\boldsymbol{\beta}) - G_n(\boldsymbol{\beta}) \xrightarrow{p} 0 \text{ uniformly in } \boldsymbol{\beta} \text{ on a compact set } \mathcal{B}, \tag{A.2}$$

and

$$\limsup_{n\to\infty} \max_{\boldsymbol{\beta}\in N_\epsilon^c(\boldsymbol{\beta}^0)} \left[G_n(\boldsymbol{\beta}) - G_n\left(\boldsymbol{\beta}^0\right)\right] > 0 \text{ for any } \epsilon > 0, \tag{A.3}$$

where

$$G_n(\boldsymbol{\beta}) \equiv E\left\{\left[\sigma_p^2(x_1)\sigma_{np}^2(X_{2,t}) - \varsigma_{1t}(\boldsymbol{\beta})\right]^2 K_{ht}\right\} + E\left\{\xi_t^2 K_{ht}\right\} \tag{A.4}$$

$\varsigma_{1t}(\boldsymbol{\beta}) \equiv L(0,\boldsymbol{\beta}) - \dot{L}(0,\boldsymbol{\beta})^T (X_{2,t} - x_2)$, $N_\epsilon^c\left(\boldsymbol{\beta}^0\right)$ is the complement of an open neighborhood of $\boldsymbol{\beta}^0$ on $\mathcal{B}$ of diameter ϵ, and $\boldsymbol{\beta}^0$ is uniquely defined by $\sigma^2(x) = L(0,\boldsymbol{\beta}^0)$ and $\sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2) = \dot{L}(0,\boldsymbol{\beta}^0)$. Write

$$\begin{aligned}
\widehat{G}_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{t=1}^{n}\left[r_t^2 - L(X_{2,t} - x_2, \boldsymbol{\beta})\right]^2 K_h(\widehat{X}_{2,t} - x_2) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left\{L^2(\widehat{X}_{2,t} - x_2, \boldsymbol{\beta}) - L^2(X_{2,t} - x_2, \boldsymbol{\beta})\right\} K_h(\widehat{X}_{2,t} - x_2) \\
&\quad - \frac{2}{n}\sum_{t=1}^{n}\left\{\widehat{r}_t^2 L(\widehat{X}_{2,t} - x_2, \boldsymbol{\beta}) - r_t^2 L(X_{2,t} - x_2, \boldsymbol{\beta})\right\} K_h(\widehat{X}_{2,t} - x_2) \\
&\equiv \widehat{G}_{n1}(\boldsymbol{\beta}) + \widehat{G}_{n2}(\boldsymbol{\beta}) - 2\widehat{G}_{n3}(\boldsymbol{\beta}),
\end{aligned} \tag{A.5}$$

We will show that

$$\widehat{G}_{n1}(\boldsymbol{\beta}) = G_n(\boldsymbol{\beta}) + o_p(1) \text{ uniformly in } \boldsymbol{\beta}, \tag{A.6}$$

and

$$\widehat{G}_{nj}(\boldsymbol{\beta}) = o_p(1) \text{ uniformly in } \boldsymbol{\beta} \text{ for } j = 2, 3. \tag{A.7}$$

By Assumptions A2 and A4-A6, and the Taylor expansion, it is easy to show that $\widehat{G}_{n1}(\boldsymbol{\beta}) = \widetilde{G}_{n1}(\boldsymbol{\beta}) + O_p\left(n^{-1/2}h^{-1} + n^{-1}h^{-(d_2+1)}\right) = \widetilde{G}_{n1}(\boldsymbol{\beta}) + o_p(1)$ uniformly in $\boldsymbol{\beta}$, where $\widetilde{G}_{n1}(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{t=1}^{n} K_{ht}[r_t^2 - L(X_{2,t} - x_2, \boldsymbol{\beta})]^2$. By the ergodic theorem, $\widetilde{G}_{n1}(\boldsymbol{\beta}) \xrightarrow{p} E[q_{nt}(\boldsymbol{\beta})]$, where $q_{nt}(\boldsymbol{\beta}) \equiv \left[r_t^2 - L(X_{2,t} - x_2, \boldsymbol{\beta})\right]^2 \times K_{ht}$. To show that this convergence is also uniform in $\boldsymbol{\beta}$, it suffices to show the stochastic equicontinuity of $\widetilde{G}_{n1}(\boldsymbol{\beta})$. We do this by verifying the two conditions of Lemma 3 in Andrews (1992). By the compactness of $\mathcal{B}$, the fact that $K(\cdot)$ is compactly supported (Assumption A5), and Assumption A1(iii). $q_{nt}(\boldsymbol{\beta}) \leq 2r_t^2 K_{ht} + 2CK_{ht} \equiv \overline{q}_{nt}$. Clearly, $E[\overline{q}_{nt}] < \infty$, which implies Assumption DM of Andrews (1992). Similarly,

$$\begin{aligned}
\left|q_{nt}(\boldsymbol{\beta}) - q_{nt}(\boldsymbol{\beta}')\right| &\leq 2r_t^2 \left|L(X_{2,t} - x_2, \boldsymbol{\beta}) - L(X_{2,t} - x_2, \boldsymbol{\beta}')\right| K_{ht} \\
&\quad + \left|L^2(X_{2,t} - x_2, \boldsymbol{\beta}) - L^2(X_{2,t} - x_2, \boldsymbol{\beta}')\right| K_{ht} \\
&\leq \left(r_t^2 \overline{C}_1 + \overline{C}_2\right) \left\|\boldsymbol{\beta} - \boldsymbol{\beta}'\right\| \text{ for some constants } \overline{C}_1 \text{ and } \overline{C}_2,
\end{aligned}$$

which implies Assumption TSE of Andrews (1992) by the Markov inequality. Consequently, $\widetilde{G}_{n1}(\boldsymbol{\beta})$ is stochastic equicontinuous on $\mathcal{B}$ and it converges to $E[q_{nt}(\boldsymbol{\beta})]$ uniformly in $\boldsymbol{\beta}$.

By the second order Taylor expansion, we have

$$\begin{aligned}
E[q_{nt}(\boldsymbol{\beta})] &= E\left\{\left[r_t^2 - \varsigma_{1t}(\boldsymbol{\beta}) - \varsigma_{2t}(\boldsymbol{\beta})\right]^2 K_{ht}\right\} \\
&= E\left\{\left[\xi_t + \left(\sigma_p^2(x_1)\sigma_{np}^2(X_{2,t}) - \varsigma_{1t}(\boldsymbol{\beta})\right) - \varsigma_{2t}(\boldsymbol{\beta})\right]^2 K_{ht}\right\} \\
&= G_n(\boldsymbol{\beta}) + E\left\{\varsigma_{2t}^2(\boldsymbol{\beta}) K_{ht}\right\} + 2E\left\{\xi_t\left[\sigma_p^2(x_1)\sigma_{np}^2(X_{2,t}) - \varsigma_{1t}(\boldsymbol{\beta})\right] K_{ht}\right\} \\
&\quad -2E\left\{\xi_t \varsigma_{2t}(\boldsymbol{\beta}) K_{ht}\right\} - 2E\left\{\left[\sigma_p^2(x_1)\sigma_{np}^2(X_{2,t}) - \varsigma_{1t}(\boldsymbol{\beta})\right]\varsigma_{2t}(\boldsymbol{\beta}) K_{ht}\right\} \\
&\equiv G_n(\boldsymbol{\beta}) + \overline{G}_{n1a}(\boldsymbol{\beta}) + \overline{G}_{n1b}(\boldsymbol{\beta}) - 2\overline{G}_{n1c}(\boldsymbol{\beta}) - 2\overline{G}_{n1d}(\boldsymbol{\beta}),
\end{aligned}$$

where $\varsigma_{2t}(\boldsymbol{\beta}) \equiv \frac{1}{2}(X_{2,t} - x_2)^T \ddot{L}(C_t(X_{2,t} - x_2), \boldsymbol{\beta})(X_{2,t} - x_2)$, and C_t is a $d_2 \times d_2$ matrix with elements that lie on the interval $[0, 1]$. We can easily verify that uniformly in $\boldsymbol{\beta}$, $\overline{G}_{n1a}(\boldsymbol{\beta}) = O(h^4)$, $\overline{G}_{n1b}(\boldsymbol{\beta}) = \overline{G}_{n1c}(\boldsymbol{\beta}) = 0$ because $E(\xi_t | \mathcal{F}_{t-1}) = 0$ and $X_t \in \mathcal{F}_{t-1}$, and $\overline{G}_{n1d}(\boldsymbol{\beta}) = O\left(h^2\right)$. Consequently,

$$E[q_{nt}(\boldsymbol{\beta})] = G_n(\boldsymbol{\beta}) + o(1) \text{ uniformly in } \boldsymbol{\beta}$$

and (A.6) follows. Now, by the Taylor expansion and Assumptions A1(iii)-A2 and A4-A6,

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|\widehat{G}_{n2}\left(\boldsymbol{\beta}\right)\right| &\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\frac{1}{n}\sum_{t=1}^{n}\left|L^{2}(\widehat{X}_{2,t}-x_{2},\boldsymbol{\beta})-L^{2}\left(X_{2,t}-x_{2},\boldsymbol{\beta}\right)\right|K_{h}(\widehat{X}_{2,t}-x_{2}) \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{B}}\frac{2}{n}\sum_{t=1}^{n}\left|L\left(X_{2,t}^{*}-x_{2},\boldsymbol{\beta}\right)\dot{L}\left(X_{2,t}^{*}-x_{2},\boldsymbol{\beta}\right)\left(\widehat{X}_{2,t}-X_{2,t}\right)\right|K_{h}(\widehat{X}_{2,t}-x_{2}) \\
&\leq \frac{2C}{n}\sum_{t=1}^{n}\left\|\widehat{X}_{2,t}-X_{2,t}\right\|K_{h}(\widehat{X}_{2,t}-x_{2}) \\
&= \frac{2C}{n}\sum_{t=1}^{n}\left\|D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}_{t}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{0}\right)\right\|\left|K_{ht}+\left[\dot{K}_{h}\left(X_{2,t}^{**}-x_{2}\right)\right]^{T}\left(\widehat{X}_{2,t}-X_{2,t}\right)\right| \\
&\leq \frac{2C}{n}\sum_{t=1}^{n}\left(\left\|D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^{0}\right)\right\|\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{0}\right\|+O_{p}\left(n^{-1}\right)\right) \\
&\quad\times\left|K_{ht}+\left[\dot{K}_{h}\left(X_{2,t}-x_{2}\right)\right]^{T}D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^{0}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{0}\right)+O_{p}\left(n^{-1}h^{-d_{2}-1}\right)\right| \\
&\leq \frac{2C}{n}\sum_{t=1}^{n}\left\|D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^{0}\right)\right\|K_{ht}\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{0}\right\| \\
&\quad+\frac{2C}{n}\sum_{t=1}^{n}\left\|D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^{0}\right)\right\|^{2}\left\|\dot{K}_{h}\left(X_{2,t}-x_{2}\right)\right\|\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{0}\right\|^{2}+O_{p}\left(n^{-3/2}h^{-d_{2}-1}\right) \\
&= O_{p}(n^{-1/2})+O_{p}\left(n^{-1}h^{-1}\right)+O_{p}\left(n^{-3/2}h^{-d_{2}-1}\right)=o_{p}\left(1\right),
\end{aligned}$$

where $X_{2,t}^{*}$ and $X_{2,t}^{**}$ lie between $\widehat{X}_{2,t}$ and $X_{2,t}$, $\boldsymbol{\theta}_{t}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{0}$ and $\dot{K}_{h}\left(u\right)\equiv\left(\partial/\partial u\right)K_{h}\left(u\right)$. Similarly, we can show that $\sup_{\boldsymbol{\beta}\in\mathcal{B}}|\widehat{G}_{n3}\left(\boldsymbol{\beta}\right)|=o_{p}\left(1\right)$. Consequently,

$$\widehat{G}_{n}\left(\boldsymbol{\beta}\right)=G_{n}\left(\boldsymbol{\beta}\right)+o_{p}\left(1\right)\ \text{uniformly in }\boldsymbol{\beta}.$$

Note that choosing $\boldsymbol{\beta}$ to minimize $G_{n}\left(\boldsymbol{\beta}\right)$ is equivalent to choosing $L\left(0,\boldsymbol{\beta}\right)$ and $\dot{L}\left(0,\boldsymbol{\beta}\right)$ to minimize

$$G_{n}^{*}\left(\boldsymbol{\beta}\right)\equiv E\left\{\left[\sigma_{p}^{2}\left(x_{1}\right)\sigma_{np}^{2}\left(X_{2,t}\right)-L\left(0,\boldsymbol{\beta}\right)-\dot{L}\left(0,\boldsymbol{\beta}\right)^{T}\left(X_{2,t}-x_{2}\right)\right]^{2}K_{ht}\right\},$$

It is easy to verify that when $L\left(0,\boldsymbol{\beta}\right)=\sigma_{p}^{2}\left(x_{1}\right)\sigma_{np}^{2}\left(x_{2}\right)$ and $\dot{L}\left(0,\boldsymbol{\beta}\right)=\sigma_{p}^{2}\left(x_{1}\right)\dot{\sigma}_{np}^{2}\left(x_{2}\right)$, $G_{n}^{*}\left(\boldsymbol{\beta}\right)=O\left(h^{4}\right)$ is minimized. By the monotonicity of $\Psi\left(\cdot\right)$, $\boldsymbol{\beta}^{0}$ is uniquely determined by $L\left(0,\boldsymbol{\beta}^{0}\right)=\sigma_{p}^{2}\left(x_{1}\right)\sigma_{np}^{2}\left(x_{2}\right)$ and $\dot{L}\left(0,\boldsymbol{\beta}^{0}\right)=\sigma_{p}^{2}\left(x_{1}\right)\dot{\sigma}_{np}^{2}\left(x_{2}\right)$. This implies (A.3).■

B Proof of Theorem 2.2

Let

$$R_{n}(x,\boldsymbol{\beta})\equiv\sum_{t=1}^{n}\left\{\widehat{r}_{t}^{2}-L\left(\widehat{X}_{2,t}-x_{2},\boldsymbol{\beta}\right)\right\}^{2}K_{h}(\widehat{X}_{2,t}-x_{2}). \tag{B.1}$$

A second order Taylor expansion of $L(\widehat{X}_{2,t}-x_2,\boldsymbol{\beta})$ around $L\left(0,\boldsymbol{\beta}\right)$ yields:

$$\begin{aligned} R_n(x,\boldsymbol{\beta}) &= \sum_{t=1}^{n}\Big\{\widehat{r}_t^2 - L\left(0,\boldsymbol{\beta}\right) - \dot{L}\left(0,\boldsymbol{\beta}\right)^T(\widehat{X}_{2,t}-x_2) \\ &\quad - \frac{1}{2}(\widehat{X}_{2,t}-x_2)^T\ddot{L}\left(C_t(\widehat{X}_{2,t}-x_2),\boldsymbol{\beta}\right)(\widehat{X}_{2,t}-x_2)\Big\}^2 K_h(\widehat{X}_{2,t}-x_2), \end{aligned}$$

where C_t is a $d_2\times d_2$ matrix with elements that lie on the interval $[0,1]$. Define $R_n^*(x,\boldsymbol{\beta})$ as $R_n(x,\boldsymbol{\beta})$ with $\boldsymbol{\beta}$ in $\ddot{L}(C_t(\widehat{X}_{2,t}-x_2),\boldsymbol{\beta})$ replaced by $\widehat{\boldsymbol{\beta}}$. Let $\widehat{\boldsymbol{\beta}}^*$ denote the minimizer of $R_n^*(x,\boldsymbol{\beta})$ and put $\widehat{\sigma}_*^2(x)=L(0,\widehat{\boldsymbol{\beta}}^*)$. It suffices to show that

$$\sqrt{nh^{d_2}}\left\{\widehat{\sigma}_*^2\left(x\right)-\sigma^2\left(x\right)-B(x)\right\}\xrightarrow{d}N(0,\kappa_{02}^{d_2}(E(v_t^4|X_{2,t}=x_2)-1)f(x_2)^{-1}\sigma^4(x)). \tag{B.2}$$

and

$$\widehat{\sigma}^2(x)=\widehat{\sigma}_*^2(x)+o_p\left(n^{-1/2}h^{-d_2/2}\right), \tag{B.3}$$

where $B(x)\equiv\frac{\kappa_{21}h^2}{2}\mathrm{tr}\Big\{\sigma_p^2(x_1)\ddot{\sigma}_{np}^2(x_2)-\ddot{L}\left(0,\boldsymbol{\beta}^0\right)\Big\}$.

By Ruppert and Wand (1994, p.1348) or Fan and Gijbels (1996, p.298),

$$\widehat{\sigma}_*^2(x)=e_1^T\left(\widehat{X}_x^T\widehat{W}_x\widehat{X}_x\right)^{-1}\widehat{X}_x^T\widehat{W}_x\widehat{U}_x,$$

where e_1 is the (d_2+1) vector having 1 in the first entry and all other entries 0, $\widehat{X}_x$ is an $n\times(d_2+1)$ matrix whose tth row is given by $\widehat{X}_{x,t}\equiv(1,(\widehat{X}_{2,t}-x_2)^T)$, $\widehat{W}_x$ =diag$\{K_h(\widehat{X}_{2,1}-x_2),...,K_h(\widehat{X}_{2,n}-x_2)\}$, and $\widehat{U}_x$ is an $n\times1$ vector with typical element

$$\widehat{U}_{x,t}\equiv\widehat{r}_t^2-\frac{1}{2}(\widehat{X}_{2,t}-x_2)^T\ddot{L}\left(C_t(\widehat{X}_{2,t}-x_2),\widehat{\boldsymbol{\beta}}\right)(\widehat{X}_{2,t}-x_2).$$

Define the $(d_2+1)\times(d_2+1)$ matrix

$$S_n(x_2)=\begin{pmatrix} n^{-1}\sum_{t=1}^n K_h(\widehat{X}_{2,t}-x_2) & n^{-1}\sum_{t=1}^n K_h(\widehat{X}_{2,t}-x_2)\frac{(\widehat{X}_{2,t}-x_2)^T}{h} \\ n^{-1}\sum_{t=1}^n K_h(\widehat{X}_{2,t}-x_2)\frac{\widehat{X}_{2,t}-x_2}{h} & n^{-1}\sum_{t=1}^n K_h(\widehat{X}_{2,t}-x_2)\frac{\widehat{X}_{2,t}-x_2}{h}\frac{(\widehat{X}_{2,t}-x_2)^T}{h} \end{pmatrix} \tag{B.4}$$

and the equivalent kernel

$$K_n^*(u,x_2)=e_1^TS_n(x_2)^{-1}(1,(u-x_2)^T/h)^TK_h(u-x_2). \tag{B.5}$$

Then by Ruppert and Wand (1994, p.1351), $\widehat{\sigma}_*^2(x)=\sum_{t=1}^n K_n^*(\widehat{X}_{2,t},x_2)\widehat{U}_{x,t}$, $n^{-1}\sum_{t=1}^n K_n^*(\widehat{X}_{2,t},x_2)$

$= 1$, and $n^{-1}\sum_{t=1}^{n} K_n^*(\widehat{X}_{2,t}, x_2)(\widehat{X}_{2,t} - x_2) = 0$. Hence

$$\begin{aligned}
& \widehat{\sigma}_*^2(x) - \sigma^2(x) \\
= \; & n^{-1}\sum_{t=1}^{n} K_n^*(\widehat{X}_{2,t}, x_2)\left\{\widehat{r}_t^2 - \sigma^2(x) - \sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2)^T(\widehat{X}_{2,t} - x_2)\right. \\
& \left. -\frac{1}{2}(\widehat{X}_{2,t} - x_2)^T\ddot{L}\left(C_t(\widehat{X}_{2,t} - x_2), \widehat{\boldsymbol{\beta}}\right)(\widehat{X}_{2,t} - x_2)\right\} \\
= \; & n^{-1}\sum_{t=1}^{n} K_n^*(\widehat{X}_{2,t}, x_2)\left\{\left[\widehat{r}_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(\widehat{X}_{2,t})\right] + \frac{1}{2}\widehat{A}_{nt}\right\} \\
= \; & n^{-1}\sum_{t=1}^{n} K_n^*(X_{2,t}, x_2)\left[\widehat{r}_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(\widehat{X}_{2,t})\right] + \frac{1}{2}n^{-1}\sum_{t=1}^{n} K_n^*(X_{2,t}, x_2)\widehat{A}_{nt} \\
& + n^{-1}\sum_{t=1}^{n}\left[K_n^*(\widehat{X}_{2,t}, x_2) - K_n^*(X_{2,t}, x_2)\right]\left[\widehat{r}_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(\widehat{X}_{2,t})\right] \\
& + \frac{1}{2}n^{-1}\sum_{t=1}^{n}\left[K_n^*(\widehat{X}_{2,t}, x_2) - K_n^*(X_{2,t}, x_2)\right]\widehat{A}_{nt} \\
\equiv \; & T_{n1} + T_{n2} + T_{n3} + T_{n4}, \qquad \text{(B.6)}
\end{aligned}$$

where

$$\widehat{A}_{nt} = (\widehat{X}_{2,t} - x_2)^T\left[\sigma_p^2(x_1)\ddot{\sigma}_{np}^2(\widetilde{C}_t(\widehat{X}_{2,t} - x_2) + x_2) - \ddot{L}\left(C_t(\widehat{X}_{2,t} - x_2), \widehat{\boldsymbol{\beta}}\right)\right](\widehat{X}_{2,t} - x_2), \qquad \text{(B.7)}$$

and $\widetilde{C}_t$ is a $d_2 \times d_2$ matrix with elements that lie on the interval $[0, 1]$. $T_{nj}, j = 1, 2, 3, 4$, are analyzed in Lemmas B.1-B.4 below. Collecting the results in these lemmas, we obtain (B.2).

Next we show (B.3). By Theorem 2.1, under Assumptions A0-A6, $\boldsymbol{\beta}^0 \equiv \boldsymbol{\beta}^0(x)$ is uniquely defined by $\sigma^2(x) = L(0, \boldsymbol{\beta}^0)$ and $\sigma_p^2(x_1)\dot{\sigma}_{np}^2(x_2) = \dot{L}(0, \boldsymbol{\beta}^0)$, and $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|| \xrightarrow{p} 0$. Further, following the arguments of Hall et al. (1999, p.163), one can show that $\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}} = o_p(h^2)$. Consequently, $\widehat{\sigma}^2(x) - \widehat{\sigma}_*^2(x) = L(0, \widehat{\boldsymbol{\beta}}) - L(0, \widehat{\boldsymbol{\beta}}^*) = o_p(h^2) = o_p\left(n^{-1/2}h^{-d_2/2}\right)$.■

Let $T_{nj}, j = 1, 2, 3, 4$, be defined as in (B.6), we prove the following lemmas under the conditions of Theorem 2.2.

Lemma B.1 *Recall* $T_{n1} \equiv n^{-1}\sum_{t=1}^{n} K_n^*(X_{2,t}, x_2)\left[\widehat{r}_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(\widehat{X}_{2,t})\right]$. *Then* $\sqrt{nh^{d_2}}T_{n1} \xrightarrow{d} N(0,$ $\kappa_{02}^{d_2}(E(v_t^4|X_{2,t}{=}x_2) - 1)f(x_2)^{-1}\sigma^4(x))$.

Proof. Noting that $\widehat{\varepsilon}_t = \varepsilon_t + (g\left(U_t, \boldsymbol{\alpha}^0\right) - g\left(U_t, \widehat{\boldsymbol{\alpha}}\right))$, we have

$$\begin{aligned}
\widehat{r}_t^2 &= \frac{\varepsilon_t^2\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}})} - \frac{\varepsilon_t^2\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}})}\left(1 - \frac{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(\widehat{X}_{1,t},\widehat{\boldsymbol{\gamma}})}\right) \\
&\quad + \frac{2\varepsilon_t\left[g\left(U_t,\boldsymbol{\alpha}^0\right) - g\left(U_t,\widehat{\boldsymbol{\alpha}}\right)\right]\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(\widehat{X}_{1,t},\widehat{\boldsymbol{\gamma}})} + \frac{\left[g\left(U_t,\boldsymbol{\alpha}^0\right) - g\left(U_t,\widehat{\boldsymbol{\alpha}}\right)\right]^2\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(\widehat{X}_{1,t},\widehat{\boldsymbol{\gamma}})} \\
&\equiv \xi_{n1t} - \xi_{n2t} + \xi_{n3t} + \xi_{n4t}. \qquad \text{(B.8)}
\end{aligned}$$

Then

$$\begin{aligned}
T_{n1} &= n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\left[\xi_{n1t} - \sigma_p^2(x_1)\sigma_{np}^2(X_{2,t})\right] \\
&\quad - n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\sigma_p^2(x_1)\left[\sigma_{np}^2(\widehat{X}_{2,t}) - \sigma_{np}^2(X_{2,t})\right] \\
&\quad - n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\xi_{n2t} + n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\xi_{n3t} + n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\xi_{n4t} \\
&\equiv T_{n11} - T_{n12} - T_{n13} + T_{n14} + T_{n15}.
\end{aligned}$$

It suffices to show that

$$\sqrt{nh^{d_2}}T_{n11} \xrightarrow{d} N(0, \kappa_{02}^{d_2}(E(v_t^4|X_{2,t}=x_2) - 1)f(x_2)^{-1}\sigma^4(x))) \qquad \text{(B.9)}$$

and

$$T_{n1j} = o_p(n^{-1/2}h^{-d_2/2}) \text{ for } j = 2,3,4,5. \qquad \text{(B.10)}$$

First, by the second order Taylor expansion,

$$\xi_{n1t} = \frac{\varepsilon_t^2\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}})} = r_t^2\left\{1 + (\mu_0(x_1) - \mu_0(X_{1,t}))^T\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right) + \frac{1}{2}\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right)^T G_t\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right)\right\},$$

where recall $\mu_0(x_1) \equiv \mu(x_1,\boldsymbol{\gamma}^0)$, $G_t \equiv (\nu_*(x_1) - \nu_*(X_{1,t})) + (\mu_*(x_1) - \mu_*(X_{1,t}))^T(\mu_*(x_1) - \mu_*(X_{1,t}))$, and $\mu_*(x_1)$ and $\nu_*(x_1)$ are the gradient and the Hessian matrix with respect to $\boldsymbol{\gamma}$ of $\log\sigma_p^2(x_1,\boldsymbol{\gamma})$ evaluated at $\boldsymbol{\gamma}_t^*$, respectively. Here $\boldsymbol{\gamma}_t^*$ is the intermediate value that lies between $\boldsymbol{\gamma}^0$ and $\widehat{\boldsymbol{\gamma}}$. By Assumption A4, $\xi_{n1t} = r_t^2\{1 + (\mu_0(x_1) - \mu_0(X_{1,t}))^T\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right) + O_p(n^{-1})\|G_t\|\}$, and

$$\begin{aligned}
T_{n11} &= n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)\xi_t + n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)r_t^2(\mu_0(x_1) - \mu_0(X_{1,t}))^T\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right) \\
&\quad + \frac{O_p(n^{-1})}{2}n^{-1}\sum_{t=1}^n K_n^*(X_{2,t},x_2)r_t^2\|G_t\| \\
&\equiv T_{n11a} + T_{n11b} + T_{n11c}, \qquad \text{(B.11)}
\end{aligned}$$

where recall $\xi_t \equiv r_t^2 - \sigma_p^2(x_1)\sigma_{np}^2(X_{2,t})$.

Following Masry (1996b), we can show that $S_n(x_2) = f(x_2)S\left(1 + O_p\left(h\right)\right)$, where S is defined in (2.8). Then $\sqrt{nh^{d_2}}T_{n11a} = n^{-1/2}h^{d_2/2}f(x_2)^{-1}\sum_{t=1}^n K_{ht}\xi_t + o_p(1)$. Noting that $\{K_{ht}\xi_t, \mathcal{F}_t\}$ is an m.d.s., $E[K_{ht}\xi_t] = 0$, and

$$\begin{aligned}\text{Var}\left(n^{-1/2}h^{d_2/2}f(x_2)^{-1}\sum_{t=1}^n K_{ht}\xi_t\right) &= h^{d_2}f(x_2)^{-2}E\left\{K_{ht}\xi_t\right\}^2 \\ &\to \kappa_{02}^{d_2}(E(v_t^4|X_{2,t} = x_2) - 1)f(x_2)^{-1}\sigma^4(x),\end{aligned}$$

by the dominated convergence theorem. By Assumptions A2 and A5-A6 and the arguments of Masry (1996a, Theorem 3),

$$\sqrt{nh^{d_2}}T_{n11a} \xrightarrow{d} N(0, \kappa_{02}^{d_2}(E(v_t^4|X_{2,t} = x_2) - 1)f(x_2)^{-1}\sigma^4(x)). \tag{B.12}$$

By the ergodic theorem and Assumptions A2-A6,

$$\begin{aligned}T_{n11b} &= n^{-1}f(x_2)^{-1}\sum_{t=1}^n K_h(X_{2,t} - x_2)r_t^2(\mu_0(x_1) - \mu_0(X_{1,t}))^T\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right) \\ &= O_p\left(1\right)O_p(n^{-1/2}) = o_p(n^{-1/2}h^{-d_2/2}).\end{aligned} \tag{B.13}$$

Now by Assumptions A2-A6 and the weak law of large numbers, $\|G_t\| \leq C\sum_{i=1}^2 \|X_{1,t} - x_1\|^i$, and

$$\begin{aligned}T_{n11c} &= \frac{O_p(n^{-1})}{2}n^{-1}\sum_{t=1}^n K_n^*(X_{2,t}, x_2)r_t^2\|G_t\| \\ &\leq \frac{O_p(n^{-1})}{2}Cn^{-1}\sum_{t=1}^n K_n^*(X_{2,t}, x_2)r_t^2\sum_{i=1}^2\|X_{1,t} - x_1\|^i \\ &= \frac{O_p(n^{-1})}{2}Cf(x_2)^{-1}n^{-1}\sum_{t=1}^n K_{ht}r_t^2\sum_{i=1}^2\|X_{1,t} - x_1\|^i\left\{1 + O_p\left(h\right)\right\} \\ &= O_p(n^{-1})O_p(1) = O_p(n^{-1}).\end{aligned} \tag{B.14}$$

Hence (B.9) follows from (B.11)-(B.14).

By the Taylor expansion and Assumptions A2 and A4,

$$\begin{aligned}|T_{n12}| &= \left|n^{-1}\sum_{t=1}^n K_n^*(X_{2,t}, x_2)\sigma_p^2(x_1)\dot{\sigma}_{np}^2(X_{2,t}^*)\left(\widehat{X}_{2,t} - X_{2,t}\right)\right| \\ &= \sigma_p^2(x_1)n^{-1}f^{-1}\left(x_2\right)\sum_{t=1}^n\left|K_{ht}\dot{\sigma}_{np}^2(X_{2,t})D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)\right| + O_p\left(n^{-1}h^{-d_2}\right) \\ &= O_p(n^{-1/2}) + O_p\left(n^{-1}h^{-d_2}\right) = o_p(n^{-1/2}h^{-d_2/2}),\end{aligned}$$

where $X_{2,t}^*$ lies between $\widehat{X}_{2,t}$ and $X_{2,t}$. Next,

$$\begin{aligned}\xi_{n2t} &= \frac{\varepsilon_t^2\sigma_p^2(x_1,\widehat{\boldsymbol{\gamma}})}{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}})}\left(1-\frac{\sigma_p^2(X_{1,t},\widehat{\boldsymbol{\gamma}}}{\sigma_p^2(\widehat{X}_{1,t},\widehat{\boldsymbol{\gamma}})}\right)\\ &= r_t^2\{1+(\mu_0(x_1)-\mu_0(X_{1,t}))^T\left(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right)+O_p(n^{-1})\left\|G_t\right\|\}\\ &\quad\times\frac{\dot{\sigma}_p^2(X_{1,t},\boldsymbol{\gamma}_t^*)^T}{\sigma_p^2(X_{1,t},\boldsymbol{\gamma}_t^*)}D_{\boldsymbol{\theta}}X_{1,t}\left(\boldsymbol{\theta}_t^*\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right),\end{aligned}$$

where $\boldsymbol{\theta}_t^*$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$, and $\boldsymbol{\gamma}_t^*$ lies between $\widehat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}^0$.

$$\begin{aligned}T_{n13} &= n^{-1}\sum_{t=1}^{n}K_n^*(X_{2,t},x_2)r_t^2\frac{\dot{\sigma}_p^2(X_{1,t},\boldsymbol{\gamma}_t^*)^T}{\sigma_p^2(X_{1,t},\boldsymbol{\gamma}_t^*)}D_{\boldsymbol{\theta}}X_{1,t}\left(\boldsymbol{\theta}_t^*\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right)+O_p\left(n^{-1}\right)\\ &= n^{-1}f\left(x_2\right)^{-1}\sum_{t=1}^{n}K_{ht}r_t^2\frac{\dot{\sigma}_p^2(X_{1,t},\boldsymbol{\gamma}^0)^T}{\sigma_p^2(X_{1,t})}D_{\boldsymbol{\theta}}X_{1,t}\left(\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right)+O_p\left(n^{-1}\right)\\ &= O_p\left(1\right)O_p(n^{-1/2})+O_p\left(n^{-1}\right)=o_p(n^{-1/2}h^{-d_2/2}).\end{aligned}$$

Similarly, we can show that $T_{n14}=o_p(n^{-1/2}h^{-d_2/2})$ and $T_{n15}=o_p(n^{-1/2}h^{-d_2/2})$. This completes the proof of the lemma. ■

Lemma B.2 $T_{n2}\equiv\frac{1}{2}n^{-1}\sum_{t=1}^{n}K_n^*(X_{2,t},x_2)\widehat{A}_{nt}=B\left(x\right)+o_p(n^{-1/2}h^{-d_2/2})$.

Proof. Let A_{nt} be defined as $\widehat{A}_{nt}$ in (B.7) but with $X_{2,t}$ replacing $\widehat{X}_{2,t}$. Decompose

$$\begin{aligned}T_{n2} &\equiv \frac{1}{2}n^{-1}\sum_{t=1}^{n}K_n^*(X_{2,t},x_2)A_{nt}+\frac{1}{2}n^{-1}\sum_{t=1}^{n}K_n^*(X_{2,t},x_2)\left(\widehat{A}_{nt}-A_{nt}\right)\\ &\equiv T_{n21}+T_{n22}.\end{aligned}$$

By the ergodic theorem, the dominated convergence theorem, and Assumptions A2-A6,

$$\begin{aligned}T_{n21} &= \frac{1}{2}n^{-1}f(x_2)^{-1}\sum_{t=1}^{n}K_{ht}A_{nt}\left\{1+O_p\left(h\right)\right\}\\ &= B(x)+o_p(h^2)=B(x)+o_p(n^{-1/2}h^{-d_2/2}).\end{aligned}\tag{B.15}$$

Now, by Taylor expansions,

$$\begin{aligned}T_{n22} &= \frac{1}{2}n^{-1}f(x_2)^{-1}\sum_{t=1}^{n}K_{ht}\left(\widehat{A}_{nt}-A_{nt}\right)\left\{1+O_p\left(h\right)\right\}\\ &= O_p(n^{-1/2})=o_p(n^{-1/2}h^{-d_2/2}).\end{aligned}$$

Hence $T_{n2}=B(x)+o_p(n^{-1/2}h^{-d_2/2})$. ■

Lemma B.3 $T_{n3}\equiv n^{-1}\sum_{t=1}^{n}[K_n^*(\widehat{X}_{2,t},x_2)-K_n^*(X_{2,t},x_2)][\widehat{r}_t^2-\sigma_p^2(x_1)\sigma_{np}^2(\widehat{X}_{2,t})]=o_p(n^{-1/2}\ h^{-d_2/2})$.

Proof. Let $\eta_{nt} = f(x_2)^{-1}\{K_h(\widehat{X}_{2,t} - x_2) - K_h(X_{2,t} - x_2)\}$. We can show $K_n^*(\widehat{X}_{2,t}, x_2) - K_n^*(X_{2,t}, x_2) = \eta_{nt}\{1 + O_p(h)\}$ where the order $O_p(h)$ does not depend on t. Then by the notation used in the proof of Lemma B.1 (eq. (B.8) in particular),

$$\begin{aligned} T_{n3} &= n^{-1}\sum_{t=1}^{n}\eta_{nt}\left[\xi_t + \sigma_p^2(x_1)\left(\sigma_{np}^2(X_{2,t}) - \sigma_{np}^2(\widehat{X}_{2,t})\right) - \xi_{n2t} + \xi_{n3t} + \xi_{n4t}\right]\{1 + O_p(h)\} \\ &= \left\{n^{-1}\sum_{t=1}^{n}\eta_{nt}\xi_t + n^{-1}\sum_{t=1}^{n}\eta_{nt}\sigma_p^2(x_1)\left(\sigma_{np}^2(X_{2,t}) - \sigma_{np}^2(\widehat{X}_{2,t})\right)\right. \\ &\quad \left. -n^{-1}\sum_{t=1}^{n}\eta_{nt}\xi_{n2t} + n^{-1}\sum_{t=1}^{n}\eta_{nt}\xi_{n3t} + n^{-1}\sum_{t=1}^{n}\eta_{nt}\xi_{n4t}\right\}\{1 + O_p(h)\} \\ &\equiv \{T_{n31} + T_{n32} + T_{n33} + T_{n34} + T_{n35}\}\{1 + O_p(h)\}. \end{aligned}$$

It suffices to show that $T_{n3j} = o_p(n^{-1/2}h^{-d_2/2})$, $j = 1, 2, ..., 5$. First,

$$\begin{aligned} T_{n31} &= n^{-1}f(x_2)^{-1}\sum_{t=1}^{n}\xi_t\dot{K}_h(X_{2,t} - x_2)^T D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right) + O_p(n^{-1}h^{-d_2+2}) \\ &\equiv \widetilde{T}_{n31}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + O_p\left(n^{-1}h^{-d_2+2}\right), \end{aligned}$$

where $\dot{K}_h(u) \equiv (\partial/\partial u)K_h(u)$. It is easy to show that $E(\widetilde{T}_{n31}) = 0$ and $||\mathrm{Var}(\widetilde{T}_{n31})|| = O\left(n^{-1}h^{-d_2-2}\right)$. It follows from the Chebyshev inequality $\widetilde{T}_{n31} = O_p\left(n^{-1/2}h^{-d_2/2-1}\right)$ and hence $T_{n31} = O_p(n^{-1}h^{-d_2/2-1} + n^{-1}h^{-d_2+2}) = o_p(n^{-1/2}h^{-d_2/2})$. Similarly, we can show that $T_{n3j} = O_p(n^{-1}h^{-1} + n^{-1}h^{-d_2+1}) = o_p(n^{-1/2}h^{-d_2/2})$, for $j = 2, 3, 4, 5$. This completes the proof. ■

Lemma B.4 $T_{n4} \equiv \frac{1}{2}n^{-1}\sum_{t=1}^{n}\left[K_n^*(\widehat{X}_{2,t}, x_2) - K_n^*(X_{2,t}, x_2)\right]\widehat{A}_{nt} = o_p(n^{-1/2}h^{-d_2/2})$.

Proof. Using the arguments and notation in the proof of Lemmas B.1-B.3, we have

$$\begin{aligned} T_{n4} &= \left\{\frac{1}{2}n^{-1}\sum_{t=1}^{n}\eta_{nt}A_{nt} + \frac{1}{2}n^{-1}\sum_{t=1}^{n}\eta_{nt}\left(\widehat{A}_{nt} - A_{nt}\right)\right\}\{1 + O_p(h)\} \\ &\equiv \{T_{n41} + T_{n42}\}\{1 + O_p(h)\}. \end{aligned}$$

First,

$$\begin{aligned} T_{n41} &= \frac{1}{2}n^{-1}f(x_2)^{-1}\sum_{t=1}^{n}\dot{K}_h(X_{2,t} - x_2)^T D_{\boldsymbol{\theta}}X_{2,t}\left(\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)A_{nt} + O_p\left(n^{-1}h^{-d_2+2}\right) \\ &= O_p(n^{-1/2}h) + O_p(n^{-1}h^{-d_2+2}) = o_p(n^{-1/2}h^{-d_2/2}). \end{aligned}$$

Similarly, we can show that $T_{n42} = O_p\left(n^{-1/2}h\right) + O_p\left(n^{-1}h^{-d_2+2}\right) = o_p\left(n^{-1/2}h^{-d_2/2}\right)$. This completes the proof. ■

C Proof of (2.13) and (2.14)

Since the proof parallels that of Theorem 2.2, we only sketch the difference. Recall $S_n(x)$ is defined in B.4. When $x_{n2} = ch$, we can show that $S_n(x_{n2}) \xrightarrow{p} f(0) S_c$, where

$$S_c \equiv \begin{pmatrix} \mu_{c0} & \mu_{c1} \\ \mu_{c1} & \mu_{c2} \end{pmatrix}.$$

Following the proof of Lemma B.2, the dominant term in the asymptotic bias is determined by

$$\begin{aligned} \widetilde{T}_{n21} &= \frac{1}{2} n^{-1} \sum_{t=1}^{n} e_1^T S_n^{-1}(x_{n2}) \begin{pmatrix} 1 \\ \frac{X_{2,t}-x_{n2}}{h} \end{pmatrix} K_h(X_{2,t}-x_{n2}) A_{nt} \{1 + o_p(1)\} \\ &= \frac{1}{2} n^{-1} \frac{1}{f(0)(\mu_{c0}\mu_{c2} - \mu_{c1}^2)} \sum_{t=1}^{n} \left(\mu_{c2} - \mu_{c1} \frac{X_{2,t}-x_{n2}}{h} \right) K_h(X_{2,t}-x_{n2}) A_{nt} \{1 + o_p(1)\} \\ &\xrightarrow{p} \frac{h^2}{2} \frac{\mu_{c2}^2 - \mu_{c1}\mu_{c3}}{\mu_{c0}\mu_{c2} - \mu_{c1}^2} [\sigma_p^2(x_1) \ddot{\sigma}_{np}^2(0) - \ddot{L}(0, \boldsymbol{\beta}^0)]. \end{aligned}$$

Following the proof of Lemma B.1, the dominant term in the asymptotic variance is determined by

$$\begin{aligned} & \sqrt{nh^{d_2}} \widetilde{T}_{n11a} \\ = \; & n^{-1/2} h^{d_2/2} \sum_{t=1}^{n} e_1^T S_n^{-1}(x_{n2}) \begin{pmatrix} 1 \\ \frac{X_{2,t}-x_{n2}}{h} \end{pmatrix} K_h(X_{2,t}-x_{n2}) \xi_t + o_p(1) \\ = \; & n^{-1/2} h^{d_2/2} \frac{1}{f(0)(\mu_{c0}\mu_{c2} - \mu_{c1}^2)} \sum_{t=1}^{n} \left(\mu_{c2} - \mu_{c1} \frac{X_{2,t}-x_{n2}}{h} \right) K_h(X_{2,t}-x_{n2}) \xi_t + o_p(1) \\ \equiv \; & V_n + o_p(1). \end{aligned}$$

Noting that $\{[\mu_{c2} - \mu_{c1}(X_{2,t}-x_{n2})/h] K_h(X_{2,t}-x_{n2}) \xi_t, \mathcal{F}_t\}$ is an m.d.s., simple calculations show that $E[V_n] = 0$, and

$$\mathrm{Var}(V_n) = \frac{\int_{-c}^{\infty} (\mu_{c2} - \mu_{c1} z)^2 k^2(z) dz}{(\mu_{c0}\mu_{c2} - \mu_{c1}^2)^2} \left[E(v_t^4 | X_{2,t} = x_2) - 1 \right] f^{-1}(0) \sigma^4(x_1, 0)) + o(1).$$

This completes the proof. ■

D Proof of Theorem 2.3

The proof parallels that of Theorem 2.1. The major difference lies in three aspects: (a) now we hold h as fixed; (b) we use the fact that $\sigma_{np}^2(\cdot) \equiv 1$ under the correct specification of the first stage parametric conditional variance model; (c) the probability limit $G_n(\boldsymbol{\beta})$ now becomes

$$G_n(\boldsymbol{\beta}) \equiv E\left\{ [\sigma_p^2(x_1) - L((X_{2,t}-x_2), \boldsymbol{\beta})]^2 K_{ht} \right\} + E\left\{ \xi_t^2 K_{ht} \right\}.$$

We proceed as in the proof of Theorem 2.1 until (A.7).

By Assumptions A1-A2 and A5, the Taylor expansion and the weak uniform law of large numbers, it is easy to show that $\widehat{G}_{n1}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n}\left[r_t^2 - L(X_{2,t}-x_2,\boldsymbol{\beta})\right]^2 K_{ht} + O_p\left(n^{-1/2}\right) = \overline{G}_{n1}(\boldsymbol{\beta}) + o_p(1)$ uniformly in $\boldsymbol{\beta}$, where $\overline{G}_{n1}(\boldsymbol{\beta}) = E\{\left[r_t^2 - L(X_{2,t}-x_2,\boldsymbol{\beta})\right]^2 K_{ht}\}$. Now $\xi_t \equiv r_t^2 - \sigma_p^2(x_1)$, and

$$\begin{aligned}
\overline{G}_{n1}(\boldsymbol{\beta}) &= E\left\{\left[r_t^2 - L((X_{2,t}-x_2),\boldsymbol{\beta})\right]^2 K_{ht}\right\} \\
&= E\left\{\left[\xi_t + \left(\sigma_p^2(x_1) - L((X_{2,t}-x_2),\boldsymbol{\beta})\right)\right]^2 K_{ht}\right\} \\
&= G_n(\boldsymbol{\beta}),
\end{aligned}$$

where the last equality follows from the fact that $\{\xi_t\left[\sigma_p^2(x_1) - L((X_{2,t}-x_2),\boldsymbol{\beta})\right]K_{ht}, \mathcal{F}_t\}$ is an m.d.s. As in the proof of Theorem 2.1, $\widehat{G}_{n2}(\boldsymbol{\beta}) = o_p(1)$ and $\widehat{G}_{n3}(\boldsymbol{\beta}) = o_p(1)$ uniformly in $\boldsymbol{\beta}$. Hence

$$\widehat{G}_n(\boldsymbol{\beta}) = G_n(\boldsymbol{\beta}) + o_p(1) \text{ uniformly in } \boldsymbol{\beta}.$$

As in the proof of Theorem 2.1, choosing $\boldsymbol{\beta}$ to minimize $G_n(\boldsymbol{\beta})$ is equivalent to choosing $L(0,\boldsymbol{\beta})$ and $\dot{L}(\cdot,\boldsymbol{\beta})$ to minimize

$$G_n^*(\boldsymbol{\beta}) \equiv E\left\{\left[\sigma_p^2(x_1) - L(0,\boldsymbol{\beta}) - \dot{L}\left(\overline{C}_t(X_{2,t}-x_2),\boldsymbol{\beta}\right)^T(X_{2,t}-x_2)\right]^2 K_{ht}\right\},$$

where $\overline{C}_t$ is a $d_2 \times d_2$ matrix with elements that lie on the interval $[0,1]$. Clearly, when $L(0,\boldsymbol{\beta}) = \sigma_p^2(x_1)$ and $\dot{L}(\cdot,\boldsymbol{\beta}) \equiv 0$, $G_n^*(\boldsymbol{\beta}) = 0$ is minimized. Consequently, $\boldsymbol{\beta}^0$ is uniquely determined by $L\left(0,\boldsymbol{\beta}^0\right) = \sigma_p^2(x_1)$ and $\dot{L}\left(\cdot,\boldsymbol{\beta}^0\right) \equiv 0$. Note that $L\left(0,\boldsymbol{\beta}^0\right) = \Psi\left(\beta_0^0\right)$, $\dot{L}\left(x_2,\boldsymbol{\beta}^0\right) = \dot{\Psi}\left(\beta_0^0 + \sum_{i=1}^{d_2}\beta_i^0 x_{2,i}\right)\boldsymbol{\beta}_1^0$ and $\dot{\Psi}(\cdot) > 0$, thus we conclude that $\beta_0^0 = \Psi^{-1}\left(\sigma_p^2(x_1)\right)$ and $\boldsymbol{\beta}_1^0 = 0$. ■

E Proof of Theorem 2.4

The proof parallels that of Theorem 2.2. The major difference lies in three aspects: (a) now we hold h as fixed; (b) we use the fact that $\sigma_{np}^2(\cdot) \equiv 1$ under the correct specification of the first stage parametric conditional variance model; (c) we use the fact $\dot{L}\left(\cdot,\boldsymbol{\beta}^0\right) \equiv 0$.

Let $R_n(x,\boldsymbol{\beta})$ be defined as in (B.1). By a Taylor expansion of the first order conditions from minimizing $R_n(x,\boldsymbol{\beta})$, we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) = -\left(\frac{1}{n}\frac{\partial^2 R_n(x,\widetilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}}, \tag{E.1}$$

where $\widetilde{\boldsymbol{\beta}}$ lies between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^0$ and thus converges to $\boldsymbol{\beta}^0$ in probability by Theorem 2.3,

$$\begin{aligned}
\frac{\partial R_n(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\sum_{t=1}^{n}\left\{\widehat{r}_t^2 - L\left(\widehat{X}_{2,t}-x_2,\boldsymbol{\beta}\right)\right\}\dot{\Psi}_{\boldsymbol{\beta},t}\begin{pmatrix}1\\ \widehat{X}_t - x\end{pmatrix}K_h(\widehat{X}_{2,t}-x_2),\\
\frac{\partial^2 R_n(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -2\sum_{t=1}^{n} s_t(\boldsymbol{\beta})\begin{pmatrix}1 & \widehat{X}_{2,t}-x_2\\ \widehat{X}_{2,t}-x_2 & (\widehat{X}_{2,t}-x_2)(\widehat{X}_{2,t}-x_2)^T\end{pmatrix}K_h(\widehat{X}_{2,t}-x_2)
\end{aligned}$$

where $s_t(\boldsymbol{\beta}) \equiv [\widehat{r}_t^2 - L(\widehat{X}_{2,t}-x_2,\boldsymbol{\beta})]\ddot{\Psi}_{\boldsymbol{\beta},t} - [\dot{\Psi}_{\boldsymbol{\beta},t}]^2$, $\dot{\Psi}_{\boldsymbol{\beta},t} \equiv \dot{\Psi}((1,(\widehat{X}_{2,t}-x_2)^T)\boldsymbol{\beta}))$, $\ddot{\Psi}_{\boldsymbol{\beta},t} \equiv \ddot{\Psi}((1,(\widehat{X}_{2,t}-x_2)^T)\boldsymbol{\beta}))$, and $\dot{\Psi}(u)$ and $\ddot{\Psi}(u)$ denote the first and second derivatives of $\Psi(u)$ with respect to u, respectively. Note that Theorem 2.3 implies

$$L(\widehat{X}_{2,t}-x_2,\boldsymbol{\beta}^0) = L\left(0,\boldsymbol{\beta}^0\right) = \sigma_p^2(x_1),\ \dot{\Psi}_{\boldsymbol{\beta}^0,t} \equiv \dot{\Psi}(\beta_0^0),\ \text{and}\ \ddot{\Psi}_{\boldsymbol{\beta}^0,t} \equiv \ddot{\Psi}(\beta_0^0). \tag{E.2}$$

Let $N_\epsilon(\boldsymbol{\beta}^0)$ denote the ϵ-neighborhood of $\boldsymbol{\beta}^0$ where $\epsilon \equiv \epsilon(n) \to 0$ as $n \to \infty$. The proof is complete if we can show

$$\frac{1}{n}\frac{\partial^2 R_n(x,\widetilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - \frac{1}{n}\frac{\partial^2 R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = o_p(1)\ \text{ uniformly in } \widetilde{\boldsymbol{\beta}} \in N_\epsilon(\boldsymbol{\beta}^0), \tag{E.3}$$

$$\frac{1}{n}\frac{\partial^2 R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - 2\left[\dot{\Psi}(\beta_0^0)\right]^2\Sigma_{\boldsymbol{\beta}^0} = o_P(1), \tag{E.4}$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}} \xrightarrow{d} N\left(0, 4\left[\dot{\Psi}(\beta_0^0)\right]^2\Omega_{\boldsymbol{\beta}^0}\right). \tag{E.5}$$

We first show (E.3). By Theorem 2.3, (E.2) and Assumptions A1-A5, we can show that $\sup_{\widetilde{\boldsymbol{\beta}}\in N_\epsilon(\boldsymbol{\beta}^0)}|s_t(\widetilde{\boldsymbol{\beta}}) - s_t\left(\boldsymbol{\beta}^0\right)| = o_p(1)$ uniformly in t. (E.3) then follows from this fact by applications of Taylor expansion and weak law of large numbers. Next, by Assumptions A2-A4, Taylor expansions and weak law of large numbers,

$$\begin{aligned}
\frac{\partial^2 R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -\frac{2}{n}\sum_{t=1}^{n}\left\{\left[r_t^2 - \sigma_p^2(x_1)\right]\ddot{\Psi}(\beta_0^0) - \left[\dot{\Psi}(\beta_0^0)\right]^2\right\}\\
&\quad\times\begin{pmatrix}1 & X_{2,t}-x_2\\ X_t - x & (X_{2,t}-x_2)(X_{2,t}-x_2)^T\end{pmatrix}K_{ht} + o_p(1)\\
&= 2\left[\dot{\Psi}(\beta_0^0)\right]^2\Sigma_{\boldsymbol{\beta}^0} + o_p(1).
\end{aligned} \tag{E.6}$$

Now by (B.8), we make the following decomposition:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}} &= \frac{-2\dot{\Psi}(\beta_0^0)}{\sqrt{n}}\sum_{t=1}^{n}\left\{\widehat{r}_t^2 - \sigma_p^2(x_1)\right\}\widetilde{X}_{2,t}K_h(\widehat{X}_{2,t}-x_2)\\
&= -2\dot{\Psi}(\beta_0^0)\sum_{j=1}^{5}V_{n,j},
\end{aligned}$$

where $\widetilde{X}_{2,t} \equiv (1,\ (\widehat{X}_{2,t} - x_2)^T)^T$, $V_{n,1} \equiv n^{-1/2}\sum_{t=1}^{n}\left\{r_t^2 - \sigma_p^2(x_1)\right\}\widetilde{X}_{2,t}K_h(\widehat{X}_{2,t} - x_2)$, $V_{n,2} \equiv n^{-1/2}\sum_{t=1}^{n}\left\{\xi_{n1t} - r_t^2\right\}\widetilde{X}_{2,t}K_h(\widehat{X}_{2,t} - x_2)$, and $V_{n,j+1} \equiv n^{-1/2}\sum_{t=1}^{n}\xi_{n2j}\widetilde{X}_{2,t}K_h(\widehat{X}_{2,t} - x_2)$, $j = 2,3,4$. We can show that

$$V_{n,1} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{r_t^2 - \sigma_p^2(x_1)\right\}\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + o_p(1),$$

$$V_{n,2} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sigma_p^2(x_1)\left(\mu_0(x_1) - \mu_0(X_{1,t})\right)^T\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right)\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + o_p(1),$$

$$V_{n,3} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\sigma_p^2(x_1)\,\dot{\sigma}_p^2(X_{1,t},\boldsymbol{\gamma}^0)^T}{\sigma_p^2(X_{1,t})}D_{\boldsymbol{\theta}}X_{1,t}\left(\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + o_p(1),$$

$$V_{n,4} = -\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{2\varepsilon_t\sigma_p^2(x)\left[D_{\boldsymbol{\alpha}}g\left(U_t,\boldsymbol{\alpha}^0\right)\right]^T\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\right)}{\sigma_p^2(X_{1,t})}\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + o_p(1) = o_p(1),$$

and

$$V_{n,5} = -\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\sigma_p^2(x)\left\{\left[D_{\boldsymbol{\alpha}}g\left(U_t,\boldsymbol{\alpha}^0\right)\right]^T\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\right)\right\}^2}{\sigma_p^2(X_{1,t})}\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + o_p(1) = o_p(1).$$

Hence,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\frac{\partial R_n(x,\boldsymbol{\beta}^0)}{\partial\boldsymbol{\beta}} \\
= \ & \frac{-2\dot{\Psi}(\beta_0^0)}{\sqrt{n}}\sum_{t=1}^{n}\left\{r_t^2 - \sigma_p^2(x_1)\right\}\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} \\
& + \frac{-2\dot{\Psi}(\beta_0^0)}{n}\sum_{t=1}^{n}\sigma_p^2(x_1)\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} \\
& \times\left\{\left[0_{1\times k_1},\left(\mu_0(x_1)\text{-}\mu_0(X_{1,t})\right)^T\right] - \frac{\dot{\sigma}_p^2(X_{1,t},\boldsymbol{\gamma}^0)^T}{\sigma_p^2(X_{1,t})}D_{\boldsymbol{\theta}}X_{1,t}\left(\boldsymbol{\theta}^0\right)\right\}n^{-1/2}\sum_{s=1}^{n}\varphi_s\left(\boldsymbol{\theta}^0\right) + o_p(1) \\
= \ & \frac{-2\dot{\Psi}(\beta_0^0)}{\sqrt{n}}\sum_{t=1}^{n}\left\{\left[r_t^2 - \sigma_p^2(x_1)\right]\begin{pmatrix} 1 \\ X_{2,t} - x_2 \end{pmatrix}K_{ht} + \Upsilon\varphi_t\left(\boldsymbol{\theta}^0\right)\right\} + o_p(1) \\
& \xrightarrow{d} N\left(0, 4\left[\dot{\Psi}(\beta_0^0)\right]^2\Omega_{\boldsymbol{\beta}^0}\right).
\end{aligned}$$

Consequently, $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \xrightarrow{d} N\left(0, \left[\dot{\Psi}(\beta_0^0)\right]^{-2}\Sigma_{\boldsymbol{\beta}^0}^{-1}\Omega_{\boldsymbol{\beta}^0}\Sigma_{\boldsymbol{\beta}^0}^{-1}\right)$.■

References

Andrews, D. W. K. (1992), Generic uniform convergence, *Econometric Theory* 8, 241-257.

Andrews, D. W. K. and J. C. Monahan (1992), An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica* 60, 953-966.

Berkes, I. and L. Horváth (2003), The rate of consistency of the quasi-maximum likelihood estimator. *Statistics & Probability Letters* 61, 133-143.

Berkes, I., L. Horváth, and P. Koloszka (2003), GARCH processes: structure and estimation. *Bernoulli* 9, 201-227.

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

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

Engle, R. F. and V. K. Ng. (1993), Measuring and testing the impact of news on volatility, *Journal of Finance* 48, 1749-1778.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*. Champman and Hall.

Fan, J. and Q. Yao. (1998), Efficient estimation of conditional variance functions in stochastic regression, *Biometrika* 85, 645-660.

Fan, Y. and A. Ullah. (1999), Asymptotic normality of a combined regression estimator, *Journal of Multivariate Analysis* 85, 191-240.

Gallant, A. R. (1981), On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form. *Journal of Econometrics* 15, 211-245.

Gallant, A. R. (1987), Identification and consistency in seminonparametric regression, in T. F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress* Vol 1, 145-170, Cambridge University Press.

Gallant, A. R. and G. Tauchen (1989), Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications. *Econometrica* 57, 1091-1120.

Glad, I. K. (1998), Parametrically guided non-parametric regression, *Scandinavian Journal of Statistics* 25, 649-668.

Glosten, L. R., R. Jagannathan, and D. Runkle (1993), On the relationship between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* 48, 1779-1801.

Hall, P., R. C. L. Wolf, and Q. Yao (1999), Methods of estimating a conditional distribution function, *Journal of the American Statistical Association 94*, 154-163.

Härdle, W., and A. B. Tsybakov (1997), Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics 81*, 223-242.

Härdle, W., A. B. Tsybakov, and L. Yang (1998), Nonparametric vector autoregression, *Journal of Statistical Planning and Inference* 68, 221-245.

Hjort, N. and I. K. Glad (1995), Nonparametric density estimation with parametric start, *Annals of Statistics* 23, 882-904.

Hsiao, C. and Li, Q. (2001), A consistent test for conditional heteroskedasticity in time-series regression models, *Econometric Theory* 17, 188-221.

Lee, S. W. and B. E. Hansen, (1994), Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator, *Econometric Theory* 10, 29-52.

Li, W and T. K. Mak (1994), On the squared residual autocorrelations in nonlinear time series with conditional heteroskedasticity, *Journal of Time series Analysis* 15, 627-636.

Ling S and W. K. Li (1997), Diagnostic checking of nonlinear multivariate time series with multivariate ARCH errors, *Journal of Time series Analysis* 18, 447-464.

Linton, O. and E. Mammen, (2005), Estimating semiparametric ARCH(∞) models by kernel smoothing methods, *Econometrica 73,* 771-836.

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.

Masry, E. (1996a), Multivariate regression estimation: local polynomial fitting for time series, *Stochastic Processes and Their Applications* 65, 81-101.

Masry, E. (1996b), Multivariate local polynomial regression for time series: uniform strong consistency rates, *Journal of Time Series Analysis* 17, 571-599.

Olkin, I. and C. Spiegelman (1987), A Semiparametric approach to density estimation, *Journal of American Statistical Association*, 88, 858-865.

Pagan, A. and Y. Hong (1990), 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.

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

Press, H. and J. W. Tukey (1956), Power spectral methods of analysis and their application to problems in airline dynamics, *Flight Test Manual*, NATO, Advisory Group for Aeronautical Research and Development, Vol., IV-C, 1-41.

Ruppert, D., and M. P. Wand (1994), Multivariate weighted least-squares regression, *Annals of Statistics* 22, 1346-1370.

White, H. (1994), *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Wong, H. and S. Ling (2005), Mixed Portmanteau tests for time series models, *Journal of Time series Analysis* 26, 569-579.

Yang, L. (2006), A semiparametric GARCH model for foreign exchange volatility, *Journal of Econometrics* 130, 365-384.

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.

Ziegelmann, F. (2002), Nonparametric estimation of volatility functions: the local exponential estimator, *Econometric Theory 18, 985-991.*