Journal of Econometrics 213 (2019) 608–631

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity[☆]

Oliver Linton[a], Zhijie Xiao[b,*,1]

[a] Faculty of Economics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom
[b] Department of Economics, Boston College, Chestnut Hill, MA 02467, USA

ARTICLE INFO

Article history:
Received 29 July 2016
Received in revised form 25 September 2018
Accepted 15 January 2019
Available online 28 August 2019

JEL classification:
C13
C14

Keywords:
Efficiency
Heteroskedasticity
Local polynomial estimation
Nonparametric regression

ABSTRACT

We study the efficient estimation of nonparametric regression in the presence of heteroskedasticity. We focus our analysis on local polynomial estimation of nonparametric regressions with conditional heteroskedasticity in a time series setting. We introduce a weighted local polynomial regression smoother that takes account of the dynamic heteroskedasticity. We show that, although traditionally it is advised that one should not weight for heteroskedasticity in nonparametric regressions, in many popular nonparametric regression models our method has lower asymptotic variance than the usual unweighted procedures. We conduct a Monte Carlo investigation that confirms the efficiency gain over conventional nonparametric regression estimators in finite samples.

1. Introduction

In this paper we consider the nonparametric regression model

$$y_t = m(X_t) + u_t, \;\; t = 1, \ldots, T, \tag{1}$$

where the function $m(\cdot)$ is assumed to be unknown but smooth, while u_t is an error process that is mean zero given the covariate $X_t \in \mathbb{R}^d$ (which may include lagged values of y_t). The parameters of interest include $m(x)$ and partial derivatives of m at x. A popular estimator of $m(x)$ is the local polynomial regression estimator, which minimizes a localized least squares criterion, see, e.g. Fan and Gijbels (1996).

When the error term has some additional structure beyond the conditional moment restriction, it may be possible to improve the estimation of m by taking that structure into account. We consider the regression model (1) where the errors are heteroskedastic, i.e.,

$$u_t = \sigma_t \varepsilon_t, \tag{2}$$

where $\sigma_t^2 = \text{var}\,(u_t|\mathcal{F}_{t-1})$, while ε_t and $\varepsilon_t^2 - 1$ are stationary martingale difference sequences (m.d.s.) i.e., $\text{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, and $\text{E}\big(\varepsilon_t^2 - 1|\mathcal{F}_{t-1}\big) = 0$. Here, $\mathcal{F}_{t-1}$ is the information set that contains X_t and additional information such as lags of

[☆] We thank Yacine Ait-Sahalia (the Co-Editor), an Associate Editor and two referees for their helpful comments on an earlier version of this paper.
[*] Corresponding author.
E-mail addresses: obl20@cam.ac.uk (O. Linton), xiaoz@bc.edu (Z. Xiao).
[1] Financial support from Boston College, United States is gratefully acknowledged.

https://doi.org/10.1016/j.jeconom.2019.01.016
0304-4076/© 2019 Elsevier B.V. All rights reserved.

(X_t, y_t) or possibly other covariates. The specific content of $\mathcal{F}_{t-1}$ may vary over different models, and more details will be given in our later discussion on specific models. We do not assume that the error term is independent of the covariate. As will be more clear later in this paper, it is the information in addition to X_t that brings efficiency improvement. In the special case where $\mathcal{F}_{t-1}$ only contains information about X_t, the conditional variance can be written as

$$\sigma_t = \sigma(X_t), \tag{3}$$

for some measurable function $\sigma(\cdot)$. In this case, it is not possible to improve the asymptotic efficiency (in the sense of Tibshirani (1984)) of the local linear least squares estimator of m, which has variance proportional to $\frac{\sigma^2(x)}{f_X(x)}$, where $f_X(x)$ is the covariate density. This is in contrast to the case of linear regression where the Gauss–Markov theorem assures that GLS improves on OLS except in certain pathological cases, Amemiya (1985, Chapter 6) and Robinson (1987). This is because, locally to $X_t = x$ the process y_t is homoskedastic. For this reason, the traditional advice in the literature is that one should not weight for heteroskedasticity in nonparametric regressions, see, e.g., Jones (1993).

However, in many applications, (3) is not satisfied, and $\text{var}(u_t|X_t) \neq \text{var}\,(u_t|\mathcal{F}_{t-1}) = \sigma_t^2$. The most widely used class of models in economics and finance are the ARCH/GARCH models. In this case, σ_t^2 is characterized by a parametric model that does not satisfy (3). For example, suppose that σ_t^2 follows a GARCH(1,1) process described by unknown parameters $\theta = (\omega, \beta, \gamma)^{\intercal}$:

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma u_{t-1}^2. \tag{4}$$

Since $X_t \in \mathcal{F}_{t-1}$, then $\mathrm{E}(u_t|X_t) = 0$ so that

$$\mathrm{E}(y_t|\mathcal{F}_{t-1}) = \mathrm{E}(y_t|X_t) = m(X_t) \quad \text{and} \quad \sigma_t^2 = \text{var}(y_t|\mathcal{F}_{t-1}). \tag{5}$$

However, $\mathrm{E}(u_t^2|X_t) \neq \sigma_t^2$. We argue in this paper that this type heteroskedasticity will allow efficiency improvements to be made by GLS weighting. In general, if there are variables in $\mathcal{F}_{t-1}$ that affect the volatility but have no influence on the conditional mean, then additional heteroskedasticity can be found even after fixing the value of X_t, and efficiency gain can be achieved by GLS weighting. In essence, we just need some kind of exclusion restriction that the variables driving the variance are not all present in the conditional mean. We show that one can improve efficiency of the conditional mean estimation by taking into account the volatility structure we have described above.

The analysis and proposed approach in this paper actually applies to a wide range of models. A growingly popular approach to volatility modeling is to include additional information either from high frequency data (e.g., Realized Volatility) or from option prices (e.g., the VIX). This case also fits into our framework where we have expanded the definition of $\mathcal{F}_{t-1}$ to include these variables but excluded them from having an influence on the conditional mean. Another case of interest is where the variance is deterministic, perhaps nonparametric, say $\sigma_t^2 = \sigma^2(t/T)$ for some smooth function $\sigma^2(\cdot)$, (Starica, 2003) (which is consistent with the widely used rolling window analysis). In this case, $E(u_t^2|X_t) = E(u_t^2) = \sigma_t^2$, and the covariate has no effect on the evolution of the variance.[2] If $X_t = t/T$, but σ_t^2 is a dynamic heteroskedastic process, we also generally get efficiency improvements. If X_t is a stochastic process independent of u_t, then it is also independent of σ_t^2, we in any case get efficiency gains.

We examine the effect of weighting on nonparametric regressions in this paper. We point to cases where an efficiency gain can be achieved via weighting. and where they cannot. In particular, an efficiency gain can be achieved when the weighting is determined by the correctly specified error volatility structure. In that case the "GLS weighted" least squares smoothing method is shown to have a smaller variance than the variance of the unweighted estimator and yet the bias of the two estimators is the same. In practice, we have to estimate the error variance. We show that this feasible estimator can achieve the same limiting variance and improves the pointwise mean squared error relative to the unweighted estimator. We also propose consistent confidence intervals based on our procedures, which will thereby be shorter than the corresponding ones from the unweighted procedures.

In some applications the efficiency gains may be important. For example, in nonparametrically predicting stock returns one finds that the conditional mean is not very well estimated, but in any case, the memory is relatively short. By contrast, the conditional variance has a very strong nonlinear shape with substantial dynamics or memory (see, e.g. Engle, 1982; Engle and Rangel, 2008). This suggests that conditioning on the variables we include in the mean equation, the variance is still stochastic and may vary substantially such that our GLS procedure will afford substantial efficiency gains in the estimation problem. This may permit shorter confidence intervals and more accurate hypothesis testing. Unfortunately, this efficiency improvement need not translate into improved forecasting, as is well known, (Diebold and Nason, 1990).

Literature Review. There is an extensive literature on efficient estimation of nonparametric models. The simplest case is where the error term is independent of the covariate and is i.i.d. with known density f. This case was considered in Tibshirani's (1984) Phd Thesis where he introduced the local likelihood estimator that replaces the local least squares objective function. The local likelihood estimator has lower large sample variance than the least-squares based local polynomial estimator (and indeed than any other asymptotically normal estimator as follows by the classical Cramér–Rao inequality); under some conditions, the bias of the local likelihood estimator is the same as the bias of the simple local

[2] In this case also, one can say that there are variables (time) that affect volatility but do not affect the covariate (except in the special case where also $X_t = t/T$).

polynomial estimator, so that the local likelihood estimator has lower pointwise mean squared error (MSE). Linton and Xiao (2007) showed that one could achieve the same performance asymptotically, even when f is unknown, by a two step procedure based on estimation of the error density using kernel density techniques. Avramidis (2016) extended this work to cover the estimation of a conditional variance function in the presence of an unknown mean. Linton et al. (2011) consider the case with filtered data, i.e., under repeated left truncation and or right censoring, and established efficient procedures. Wang and Yao (2012) considered the single index model case where $m(x) = g(\beta^{\intercal} x)$. Jin et al. (2015) considered the case $u_t = \sigma(X_t)\varepsilon_t$ where ε_t are i.i.d. and independent of X_t with unknown density f. The efficiency gain here is coming from the shape of the error density that has to be estimated. Chen et al. (2015) have considered adaptive estimation of variable coefficient models where essentially $m(X_t)$ is replaced by $r(z_t)X_t$, where r is an unknown function of the observable quantity z_t. Meanwhile, Yao (2013) has proposed an EM algorithm for implementing the adaptive estimation method. A separate line of work has considered the problem where u_t is serially correlated, i.e., $A(L)u_t = \varepsilon_t$, where ε_t is i.i.d. and independent of X_t with mean zero, while $A(L) = \sum_{j=1}^{\infty} a_j L^j$ is a lag polynomial. Xiao et al. (2003) proposed a more efficient estimator of m based on a prewhitening transformation $Y_t - A(L)(Y_t - m(X_t)) = m(X_t) + \varepsilon_t$, where the right hand side is now a standard nonparametric regression with whitened errors (and replacing the unknown quantities on the left hand side by preliminary estimates of m and the parameters of $A(L)$). The transform implicitly takes account of the autocorrelation structure. They obtained an improvement in terms of variance over the usual kernel smoothers. Linton and Mammen (2005) considered an extension of this model and proposed likelihood based procedures that extended this and showed how one can obtain even higher efficiency; see also Liu et al. (2010), Linton and Wang (2016) and Geller and Neumann (2018). Su and Ullah (2006) constructed efficient estimators in the case where the errors are nonlinearly autodependent. In a panel setting, there are a number of papers that propose more efficient estimators of nonparametric regression curves using weighting schemes, following Wang (2003b). Henderson et al. (2008) extend this work to allow for fixed effects as well, see also Martins-Filho and Yao (2009). To summarize, both parametric and nonparametric structures can be used to improve efficiency of the estimation of $m(x)$.

The rest of this paper is organized as follows: A general discussion on weighted nonparametric regression is given in Section 2. The proposed estimator and leading special cases are studied in Section 3. Section 4 discusses some further issues. Bandwidth selection is considered in Section 5. Some Monte Carlo experiments are reported in Section 6. Section 7 concludes. A supplementary appendix contains some preliminary technical results, details of proofs, an application to the variance ratio test, and some potential extensions.

The basic result of our paper applies to different types of nonparametric estimators. We focus on the local polynomial estimator due to its wide applicability and good properties on the boundary, see, e.g., Fan (1992), and Fan and Gijbels (1996) for discussion on the attractive properties of local polynomials regression. For comparison purpose, we will briefly discuss the Nadaraya–Watson regression in Section 4 and further investigate the impact of weighting on biases. Without loss of generality and for simplicity of derivation, we assume that $d = 1$ in this paper but our result can be easily extended to the general case of multivariate X_t.

2. Weighted nonparametric regressions

In this section we consider a general weighted pth local polynomial regression based on an observed weighting scheme $\{\lambda_t\}$. Suppose that we observe $\{(Y_t, X_t, \lambda_t)\}_{t=1}^{T}$, where λ_t is a (so far unspecified) weighting scheme, and consider the general weighted local polynomial regression based on $\{\lambda_t\}$.

Let $\widehat{\beta}_{\lambda;p}(x) = (\widehat{\beta}_{\lambda 0}(x), \ldots, \widehat{\beta}_{\lambda p}(x))^{\intercal}$ minimize the weighted least squares objective function

$$Q_T\left(\beta; x, K, h, \{\lambda_t\}\right) = \sum_{t=1}^{T} \lambda_t K\left(\frac{x - X_t}{h}\right) \left(Y_t - \sum_{0 \le j \le p} \beta_j ((X_t - x)/h)^j\right)^2 \tag{6}$$

with respect to $\beta = (\beta_0, \ldots, \beta_p)^{\intercal}$. Then, with $w_t = \lambda_t K\left(\frac{x-X_t}{h}\right)$ and $\mathbb{X}_t = (1, (X_t - x)/h, \ldots, ((X_t - x)/h)^p)^{\intercal}$, we have

$$\widehat{\beta}_{\lambda;p}(x) = \left[\sum_{t=1}^{T} w_t \mathbb{X}_t \mathbb{X}_t^{\intercal}\right]^{-1} \sum_{t=1}^{T} w_t \mathbb{X}_t Y_t, \tag{7}$$

provided the matrix $\sum_{t=1}^{T} w_t \mathbb{X}_t \mathbb{X}_t^{\intercal}$ is of full rank.

The special case with $\lambda_t = 1$ corresponds to the standard local polynomial estimator, (Fan and Gijbels, 1996). In particular, the local polynomial estimator of $m(x)$ is given by the component $\widehat{\beta}_{\lambda 0}(x)$ of the estimator $\widehat{\beta}_{\lambda=1;p}(x)$, and we denote this estimator by $\widehat{m}_{LP}(x)$. In the leading case when $p = 1$, this is the local linear regression. Its asymptotic properties are well known.

We next present the asymptotic properties of the weighted estimator in the case where p is an odd integer. We make the following regularity assumptions on the model, the weighting scheme, and the kernel function and bandwidth.

Assumption A1. The data are generated by (1) and (2).

Assumption A2. $\mathrm{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, and $\mathrm{E}\big(\varepsilon_t^2 - 1|\mathcal{F}_{t-1}\big) = 0$, where $\mathcal{F}_{t-1} = \sigma(X_{t-i}, i \geq 0; y_{t-j}, j \geq 1; \lambda_t)$.

Assumption A3. The density $f_X(\cdot)$ of X_t is uniformly bounded and is bounded away from zero on its support $\mathcal{X}$, a compact subset of $\mathbb{R}$. The joint densities of $(X_t, X_{t+\ell})$, $(X_t, X_{t+\ell}, X_{t+j})$, $(X_t, X_{t+\ell}, X_{t+j}, X_{t+s})$ are continuous and bounded. The functions $f_X(\cdot)$ and $m(\cdot)$ are $(p+1)$ times partially differentiable. The derivatives $f_X^{(r)}(x) = d^r f(x)/dx^r$ and $m^{(r)}(x) = d^r m(x)/dx^r$ are bounded and uniformly continuous on $\mathcal{X}$, and there exists $C_1 < \infty$ such that

$$|f_X^{(r)}(u) - f_X^{(r)}(v)| \leq C_2\|u - v\|,$$

$$|m^{(r)}(u) - m^{(r)}(v)| \leq C_1\|u - v\|.$$

Assumption A4. The process $\{W_t\}$ is stationary and absolutely regular, where $W_t = (X_t, \sigma_t, \lambda_t)$. That is,

$$\varrho(\tau) = \sup_s E\left\{\sup_{A\in\mathcal{G}_{s+\tau}^{\infty}} |P(A|\mathcal{G}_{-\infty}^{s}) - P(A)|\right\} \to 0, \text{ as } \tau \to \infty,$$

where $\mathcal{G}_s^t$ is the σ-field generated by $\{W_j : j = s, \ldots, t\}$. In addition, there is a positive δ such that W_t has finite 2+δ moments, and for some $\delta > \delta' > 0$, $\varrho(\tau) = O(\tau^{-(2+\delta')/\delta'})$. The conditional density of $\{\varepsilon_t, \sigma_t, \lambda_t\}$, $f_{\varepsilon_t,\sigma_t,\lambda_t|X_t}(\varepsilon, \sigma, \lambda|x)$ is uniformly bounded and has continuous partial derivatives.

Assumption A5. The kernel K has support $[-1, 1]$ and is symmetric about zero. The functions $H_j(u) = u^j K(u)$, for all j with $0 \leq |j| \leq 2p+1$, are Lipschitz continuous, i.e., there exists a positive finite constant C such that $|H_j(u) - H_j(v)| \leq C\|u-v\|$.

Assumption A6. As $T \to \infty$, $h \to 0$ and $Th \to \infty$.

Most of these assumptions are standard in local polynomial nonparametric estimation, (Fan and Gijbels, 1996). These conditions are useful in our technical development and, no doubt some of them could be replaced by a range of similar assumptions. In Assumption A2 we allow for the case that λ_t is not a measurable function of $\{X_{t-i}, i \geq 0; y_{t-j}, j \geq 1\}$; in fact it suffices for consistency here that $E(u_t|X_t, \lambda_t) = 0$. Assumption A3 facilitates the Taylor expansions of the regression function and density function to the required order. Assumption A4 assumes that the data is weakly dependent so that a LLN and CLT apply. Assumption A5 for the kernel function and Assumption A6 for the bandwidth expansion are also quite standard in nonparametric estimation. We introduce the following notations:

$$M(K) = \begin{bmatrix} \mu_0(K) & \cdots & \mu_p(K) \\ \vdots & & \vdots \\ \mu_p(K) & \cdots & \mu_{2p}(K) \end{bmatrix}, \quad \Gamma(K) = \begin{bmatrix} \nu_0(K) & \cdots & \nu_p(K) \\ \vdots & & \vdots \\ \nu_p(K) & \cdots & \nu_{2p}(K) \end{bmatrix},$$

$$B(K) = \left[\mu_{p+1}(K), \cdots, \mu_{2p+1}(K)\right]^{\intercal}, \qquad \gamma(K) = M^{-1}(K)B(K) \;\; ; \;\; \omega(K) = M(K)^{-1}\Gamma(K)M(K)^{-1},$$

$$b_p(x) = \frac{m^{(p+1)}(x)}{(p+1)!} \quad ; \quad \delta_\lambda(x) = \frac{\mathrm{E}\left[\lambda_t^2\sigma_t^2|X_t = x\right]}{\left[\mathrm{E}(\lambda_t|X_t = x)\right]^2},$$

where $\mu_j(K) = \int_{-\infty}^{\infty} u^j K(u)du$ and $\nu_j(K) = \int_{-\infty}^{\infty} u^j K^2(u)du$. Let $\gamma_j(K) = e_j^{\intercal}\gamma(K)$ and $\omega_{jk}(K) = e_j^{\intercal}\omega(K)e_k$, where e_j is the $p+1$ elementary vector with 1 in the jth position and 0 elsewhere. In the univariate local linear case $\omega_{11}(K) = \nu_0(K)$. Let $\beta_0(x) = (\beta_{00}(x), \ldots, \beta_{0p}(x))^{\intercal}$, where $\beta_{0j}(x) = (h^j/j!)m^{(j)}(x)$.

Theorem 1. *Suppose that Assumptions A1–A6 hold. Then, as $T \to \infty$,*

$$\sqrt{Th}\left(\widehat{\beta}_{\lambda;p}(x) - \beta_0(x) - h^{p+1}b_p(x)\gamma(K)\right) \Longrightarrow N\left(0, \frac{\delta_\lambda(x)}{f_X(x)}\omega(K)\right).$$

Furthermore, $\widehat{\beta}_{\lambda;p}(x)$ and $\widehat{\beta}_{\lambda;p}(x')$ are asymptotically independent when $x \neq x'$.

Theorem 1 gives the asymptotic distribution of the local polynomial regression estimator of $m(x)$ and its derivatives for an arbitrary weighting sequence. From the result of Theorem 1 we can see that weighting does not affect the asymptotic bias of the local polynomial regression. The leading bias term of the weighted local polynomial regression is independent of the choice of weights $\{\lambda_t\}$, this is because the influence of λ_t in the numerator and denominator cancels out. The argument is as follows. Notice that the exact conditional bias of the local polynomial estimator is given by

$$\left[\sum_{t=1}^{T} w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T} w_t\mathbb{X}_t\Delta_t(x),$$

where $\Delta_t(x) = m(X_t) - \sum_{0 \le k \le p} \frac{1}{k!} m^{(k)}(x)(X_t - x)^k$. The numerator and the denominator are both affected by the weighting process, in particular:

$$h^{-(p+1)} \frac{1}{Th} \sum_{t=1}^{T} w_t \mathbb{X}_t \Delta_t(x) \xrightarrow{P} \mathrm{E}\{\lambda_t | X_t = x\} f_X(x)\, b_p(x) B(K),$$

$$\frac{1}{Th} \sum_{t=1}^{T} w_t \mathbb{X}_t \mathbb{X}_t^{\top} \xrightarrow{P} f_X(x) \mathrm{E}(\lambda_t | X_t = x)\, M(K).$$

From the above results, we can see how the impact of weighting, which is reflected by the term $\mathrm{E}(\lambda_t|X_t = x)$, is canceled out.

However, the weighting does change the limiting variance except in some special cases; the effect of weighting on the nonparametric regression is captured by the factor $\delta_\lambda(x)$ as indicated by Theorem 1. We next consider some different scenarios with regard to the form of λ_t and σ_t^2 and their effect on $\delta_\lambda(x)$.

If we choose a weight that is a smooth function of the regressor, i.e. $\lambda_t = \lambda(X_t)$, then $\mathrm{E}(\lambda_t|X_t = x) = \lambda(x)$, and $\mathrm{E}\left[\lambda_t^2 \sigma_t^2 | X_t = x\right] = \lambda(x)^2 \mathrm{E}\left[\sigma_t^2 | X_t = x\right]$, so that

$$\delta_\lambda(x) = \frac{\mathrm{E}\left[\lambda_t^2 \sigma_t^2 | X_t = x\right]}{\left[\mathrm{E}(\lambda_t | X_t = x)\right]^2} = \mathrm{E}\left[\sigma_t^2 | X_t = x\right].$$

In this case, the weighted local polynomial estimator has the same limiting variance as the unweighted local polynomial regression estimator. This is because, in the shrinking neighborhood of x, the weights are asymptotically the same, the weighted local polynomial estimator is asymptotically equivalent to the equally weighted local polynomial estimator. In fact, no matter what is the form of σ_t^2, any weights $\lambda(X_t)$ in the form of a smooth function of X_t, would give you the same limiting variance. Combining this result with those on bias, we can see that the weighted local polynomial regression using weights $\lambda(X_t)$ has the same mean-squared error (and limiting distribution) as the ordinary local polynomial estimation.

Suppose that $\sigma_t^2 = \sigma(X_t)^2$. Then, the “optimal” weights $\lambda_t = 1/\sigma(X_t)^2$ deliver the same results as the ordinary nonparametric regression. This is because the assumption $\sigma_t^2 = \sigma^2(X_t)$ implies that the nonparametric regression model is locally-homoskedastic. In this case, unweighted kernel estimators are asymptotically efficient (in the Tibshirani (1984) sense) under normality. In fact, incorrectly weighted regressions are worse than the ordinary nonparametric regressions in this case. To see this, notice that $\mathrm{E}\left[\lambda_t^2 | X_t = x\right] - \left[\mathrm{E}(\lambda_t | X_t = x)\right]^2 = \mathrm{var}(\lambda_t | X_t = x) \ge 0$. Therefore,

$$\delta_\lambda(x) = \frac{\mathrm{E}\left[\lambda_t^2 \sigma_t^2 | X_t = x\right]}{\left[\mathrm{E}(\lambda_t | X_t = x)\right]^2} = \frac{\sigma(x)^2 \mathrm{E}\left[\lambda_t^2 | X_t = x\right]}{\left[\mathrm{E}(\lambda_t | X_t = x)\right]^2} \ge \sigma(x)^2.$$

The equality holds only when $\mathrm{var}(\lambda_t \; | X_t = x) = 0$, which holds when $\lambda_t = \lambda(X_t)$ or $\lambda_t =$ constant. Thus, the ordinary local polynomial estimator is asymptotically the best you can get. For this reason, it is generally advised in the literature that nonparametric regressions should not be weighted, see, e.g. Jones (1993).

Suppose that $\sigma_t^2 \ne \sigma^2(X_t)$. Then, if we choose $\lambda_t = \sigma_t^{-2}$, we have

$$\delta_\lambda(x)|_{\lambda_t = \sigma_t^{-2}} = \frac{\mathrm{E}\left[\sigma_t^{-2} | X_t = x\right]}{\left[\mathrm{E}\left(\sigma_t^{-2} | X_t = x\right)\right]^2} = \frac{1}{\mathrm{E}\left(\sigma_t^{-2} | X_t = x\right)} \le \mathrm{E}\left(\sigma_t^2 | X_t = x\right).$$

This shows the efficiency gain that can be achieved by local GLS regression. In fact, by the Cauchy–Schwarz inequality, for any weights λ_t,

$$\delta_\lambda(x) = \frac{\mathrm{E}\left[\lambda_t^2 \sigma_t^2 | X_t = x\right]}{\left[\mathrm{E}(\lambda_t | X_t = x)\right]^2} \ge \frac{1}{\mathrm{E}\left(\sigma_t^{-2} | X_t = x\right)},$$

and the equality holds only when $\lambda_t = c\sigma_t^{-2}$ for some constant c, indicating that $\lambda_t = \sigma_t^{-2}$ is the optimal weight. We investigate this case further in the next section. In fact, in this case, using the wrong weighting ($\lambda_t \ne \sigma_t^{-2}$) is not necessarily worse than the unweighted estimator: as in linear regression, (Amemiya, 1983), weighting may also improve efficiency. Our standard errors below are consistent whether or not $\lambda_t \ne \sigma_t^{-2}$.

We close with a discussion of standard errors. There are a number of choices for standard errors in nonparametric regression, see Chu et al. (2017), and we just define here the most straightforward and general approach, which is valid provided only $E(u_t|X_t, \lambda_t) = 0$. In fact, it will also be asymptotically valid in some cases we discuss below where this condition is only valid asymptotically. Note that conditional on $\{X_t, \lambda_t\}_{t=1}^T$ the estimator $\widehat{\beta}_{\lambda;p}(x)$ is linear in Y and so its conditional variance is obtainable in closed form, (Fan and Gijbels, 1996, 4.9).

Let $\widehat{u}_t(x) = Y_t - \mathbb{X}_t^{\top} \widehat{\beta}_{\lambda;p}(x)$ and

$$\widehat{V}(x) = \left[\sum_{t=1}^{T} w_t \mathbb{X}_t \mathbb{X}_t^{\top}\right]^{-1} \sum_{t=1}^{T} w_t^2 \mathbb{X}_t \mathbb{X}_t^{\top} \widehat{u}_t(x)^2 \left[\sum_{t=1}^{T} w_t \mathbb{X}_t \mathbb{X}_t^{\top}\right]^{-1}. \tag{8}$$

Then, similarly to Fan and Gijbels (1996, 4.11), we can show that

$$\left(e_j^{\top}\widehat{V}(x)e_j\right)^{-1}\left(\widehat{\beta}_{\lambda j}(x)-\beta_{0j}(x)-h^{p+1}b_p(x)\gamma_j(K)\right)\Longrightarrow N\left(0,1\right) \tag{9}$$

under the conditions A1–A6, i.e., whether or not $\lambda_t = \sigma_t^{-2}$. From this we can obtain confidence intervals for $\beta_{0j}(x)$ (assuming undersmoothing). More sophisticated pointwise and uniform confidence intervals can be constructed by using bias correction/bootstrap, see for example, Hall (1992a,a) and Calonico et al. (2014), and we expect similar improvements to carry over to these cases due to the more efficient estimation.

3. The local GLS estimator

The previous section provides a general discussion on weighted nonparametric regressions. We now specialize the discussion to the case where the weighting $\lambda_t = \sigma_t^{-2}$, where $\sigma_t^2 = E(u_t^2|\mathcal{F}_{t-1})$ is the conditional variance of the error process. We first give a general result for this estimator. Then in two subsections we consider particular models for the error variance, one parametric, and one nonparametric, which allow us to estimate consistently the optimal weighting and thereby to achieve asymptotically the same efficiency.

Define $\widehat{m}(x) = \widehat{\beta}_{\lambda 0}(x)$ from (7) with $\lambda_t = \sigma_t^{-2}$. We call this the local GLS estimator. In this case, the objective function (6) can be given the interpretation of a local likelihood, under Gaussianity, see Tibshirani (1984), and so the estimation method can be given an optimality justification along the lines he gave.

We slightly modify Assumption A4 to accommodate the special case where $\lambda_t = \sigma_t^{-2}$.

Assumption A4′. Let $W_{1t} = \{X_t, \sigma_t\}$, $\{W_{1t}\}$ be a stationary absolutely regular process. That is,

$$\varrho(\tau) = \sup_s E\left\{\sup_{A\in\mathcal{G}_{s+\tau}^{\infty}} |P(A|\mathcal{G}_{-\infty}^{s}) - P(A)|\right\} \to 0, \text{ as } \tau\to\infty,$$

where $\mathcal{G}_s^t$ is the σ-field generated by $\{W_{1j} : j = s, \dots, t\}$. In addition, there is a positive δ such that $E(|W_{1t}|^{2+\delta}) < \infty$, and for some δ' with $\delta > \delta' > 0$, $\varrho(\tau) = O(\tau^{-(2+\delta')/\delta'})$. The conditional density of $\{\varepsilon_t, \sigma_t\}$, $f_{\varepsilon_t,\sigma_t|X_t}(\varepsilon, \sigma|x)$ is uniformly bounded and has continuous partial derivatives.

Corollary 1. *Suppose that Assumptions A1–A3, A4′, A5 and A6 hold. Then, as $T\to\infty$,*

$$\sqrt{Th}\left[\widehat{m}(x) - m(x) - h^{(p+1)}b(x)\right] \Longrightarrow N\left(0, \frac{\omega_{11}(K)}{f_X(x)E\left[\sigma_t^{-2}|X_t = x\right]}\right).$$

Corollary 1 indicates that the asymptotic variance of the infeasible weighted local estimator $\widehat{m}(x)$ is proportional to $1/E\left[\sigma_t^{-2}|X_t = x\right]$, which is less than $\mathrm{E}\left[\sigma_t^2|X_t = x\right]$, unless precisely (3) holds. We next discuss some concrete special cases.

Example. Suppose that $\{X_t\}$ and $\{u_t\}$ are independent processes (included in this case is the situation where $X_t = t/T$ and u_t is a stochastic process; also included is the case where σ_t^2 is the stochastic volatility class of processes without leverage effects, e.g., (Shephard, 1996) that is independent of the process X). In this case, $\mathrm{E}\left[\sigma_t^{-2}|X_t = x\right] = \mathrm{E}\left[\sigma_t^{-2}\right]$ and $\mathrm{E}\left[\sigma_t^2|X_t = x\right] = E\left[\sigma_t^2\right]$, and for any nontrivial stochastic process

$$E\left[\sigma_t^2\right] > \frac{1}{E\left[\sigma_t^{-2}\right]}$$

by the Cauchy–Schwarz inequality.

Example. Suppose that $X_t = y_{t-j}$ so that the processes $\{X_t\}$ and $\{u_t\}$ are not independent . In that case, $\mathrm{E}\left[\sigma_t^2|X_t = x\right]$ and $\mathrm{E}\left[\sigma_t^{-2}|X_t = x\right]$ are not constant, but we may also have an efficiency gain because these quantities are not exact reciprocals of each other unless σ_t^2 only depends on y_{t-j}.

In practice, σ_t^2 may be unknown in which case $\widehat{m}(x)$ is infeasible. However, the infeasible procedure defines an efficiency standard against which we should measure our feasible estimator. We next consider the case where estimated weights are allowed for.

Let $\widehat{\sigma}_t^2$ be a consistent estimator of σ_t^2; we will consider several examples below depending on model structure. Then define the feasible weighted local polynomial estimator $\widetilde{m}(x)$ as $\widehat{\beta}_{\lambda 0}(x)$ from (7) with $\lambda_t = \widehat{\sigma}_t^{-2}$. Letting $\widehat{w}_t = K\left((x - X_t)/h\right)/\widehat{\sigma}_t^2$, then the proposed estimator has the representation (provided the denominator matrix has full rank)

$$\widetilde{\beta}(x) = \left[\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\top}\right]^{-1}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tY_t, \tag{10}$$

and $\widetilde{m}(x) = \widetilde{\beta}_0(x) = e_1^{\top}\widetilde{\beta}(x)$. We call this the local FGLS estimator.

We add the following high level Assumption A7 to take into account the preliminary estimation of weights.

Assumption A7. Let $w_t = K\left((x - X_t)/h\right)/\sigma_t^2$. Then :

(a) $\left\| (Th)^{-1} \sum_{t=1}^{T} (\widehat{w}_t - w_t)\, \mathbb{X}_t \mathbb{X}_t^{\intercal} \right\| = o_p(1)$;

(b) $\left\| (Th)^{-1/2} \sum_{t=1}^{T} (\widehat{w}_t - w_t)\, \mathbb{X}_t \mathbf{\Delta_t(x)} \right\| = o_p(1)$;

(c) $\left\| (Th)^{-1/2} \sum_{t=1}^{T} (\widehat{w}_t - w_t)\, \mathbb{X}_t u_t \right\| = o_p(1)$.

The result for the proposed estimator is summarized in Theorem 2.

Theorem 2. *Suppose that Assumptions A1–A3, A4′, A5–A7 hold. Then, as $T \to \infty$,*

$$\sqrt{Th}\left[\widetilde{m}(x) - m(x) - h^{(p+1)} b_p(x)\right] \Longrightarrow N\left(0, \frac{\omega_{11}(K)}{f_X(x) E\left[\sigma_t^{-2} | X_t = x\right]}\right).$$

Theorem 2 shows that the proposed estimator is asymptotically equivalent to the infeasible weighted local estimator $\widehat{m}(x)$ and thus is more efficient than the conventional local polynomial estimator. The relative efficiency of $\widetilde{m}(x)$ is given by

$$\mathrm{E}\left[\sigma_t^2 | X_t = x\right] \times \mathrm{E}\left[\sigma_t^{-2} | X_t = x\right], \tag{11}$$

which varies with x. If the process σ_t^2 were independent of the covariate, then the relative efficiency is $\mathrm{E}\left[\sigma_t^2\right] \times \mathrm{E}\left[\sigma_t^{-2}\right]$.

The efficiency gains above can deliver smaller nonparametric confidence intervals for the regression function. One can construct confidence intervals using (8) with $\widehat{w}_t$ replacing w_t and under the conditions of Theorem 2 these will have correct asymptotic coverage. One may also use the GLS structure to define alternative confidence intervals based on explicitly estimating $f_X(x)$ and $\mathrm{E}\left[\sigma_t^{-2} | X_t = x\right]$, see Chu et al. (2017), although this will not improve the confidence interval to first order.

In the next two subsections we consider two different models for the heteroskedasticity and show how one can construct the local FGLS estimator in each case and how one can establish the equivalence of the FGLS estimator with the GLS estimator.

3.1. The case with GARCH model

We consider in more detail the special case where errors terms satisfy a GARCH(1,1) process. In particular, without loss of generality, we assume that the model is given by (1), (2), and (4). Given model (4), and under Assumption A1′,

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \gamma \sum_{j=1}^{\infty} \beta^{j-1} u_{t-j}^2.$$

The proposed estimation procedure for the GARCH case is as follows:

1. First, we construct a preliminary local polynomial estimator $\breve{m}(x)$ using bandwidth h_1 by minimizing $Q_T(\beta; x, K, h_1, \{1\})$ from (6) with respect to β
2. Then estimate σ_t^2 using $\widehat{u}_t = y_t - \breve{m}(X_t)$, denote the estimated variance by

$$\widehat{\sigma}_t^2 = \frac{\widehat{\omega}}{1-\widehat{\beta}} + \widehat{\gamma} \sum_{j=1}^{\min\{t-1,\tau\}} \widehat{\beta}^{j-1} \widehat{u}_{t-j}^2,$$

   where $\widehat{\theta} = (\widehat{\omega}, \widehat{\beta}, \widehat{\gamma})^{\intercal}$ are preliminary root-T consistent estimators of $\theta = (\omega, \beta, \gamma)^{\intercal}$, and $\tau = \tau(T) = \ln(T)$ is a truncation parameter. For example, $\widehat{\theta}$ could be the Gaussian QMLE constructed from the residuals, (Bollerslev, 1986).
3. The feasible weighted local polynomial estimator $\widetilde{m}(x)$ is constructed by minimizing $Q_T\left(\beta; x, K, h, \{\widehat{\sigma}_t^2\}\right)$ from (6) with respect to β, where h is the bandwidth in the final estimation.

For simplicity, we use the same kernel function in both the preliminary estimation and the final estimation. In the presence of a general GARCH(p,q) model, see, e.g. Francq and Zakoian (2004, 2010) for more details on QMLE estimation.

Notice that although σ_t^2 is characterized by a parametric model, estimation of σ_t^2 uses $\widehat{u}_t = y_t - \breve{m}(X_t)$, which is based on a preliminary nonparametric regression estimator $\breve{m}(X_t)$ of the conditional mean function. Consequently, the estimation of σ_t^2 depends on the bandwidth h_1.

We modify Assumptions A1, A6 and A7 to accommodate the GARCH case. We assume that the GARCH process is stationary, and we undersmooth in the preliminary estimation.

Assumption A1′. The data is given by (1), (2) and (4), where $\omega > 0$, $\beta \geq 0$, $\gamma \geq 0$, $\beta + \gamma < 1$.

Assumption A6′. As $T \to \infty$, $h \to 0$, $h_1 \to 0$ and $h_1/h \to 0$. $Th_1h^{1+p} \to \infty$, $Th_1^2h^{-1} \to \infty$, $\sqrt{Th}h_1^{p+1} \to 0$. $\tau = c \log T$ for some constant $c > 0$.

Assumption A7′. $\widehat{\theta}$ is a root-T estimator of $\theta = (\omega, \beta, \gamma)^{\intercal}$.

Under Assumption A1′, the GARCH parameters θ can be estimated at rate root-T. Assumption A6′ for the bandwidth expansion is standard in nonparametric estimation. Assumption A7′ implies Assumption A7 under this GARCH setting.

The result for the GARCH case is summarized in Theorem 3.

Theorem 3. *Suppose that Assumptions A1′, A2, A3, A4′, A5, A6′, A7′ hold. Then, as $T \to \infty$,*

$$\sqrt{Th}\left[\widetilde{m}(x) - m(x) - h^{(p+1)}b(x)\gamma_1(K)\right] \Longrightarrow N\left(0, \frac{\omega_{11}(K)}{f_X(x)E\left[\sigma_t^{-2}|X_t = x\right]}\right).$$

Theorem 3 shows that, in the presence of the GARCH effect, the proposed estimator is asymptotically equivalent to the infeasible weighted local estimator $\widehat{m}(x)$ and thus is more efficient than the conventional local polynomial estimator. We note that in this case the condition $E(u_t|X_t, \lambda_t) = 0$ fails, but in large samples $\lambda_t = \widehat{\sigma}_t^2 \simeq \sigma_t^2 \in \mathcal{F}_{t-1}$ and since we have assumed that $E(u_t|\mathcal{F}_{t-1}) = 0$ the consistency and asymptotic normality follow. Indeed the standard errors constructed from (8) are consistent in this case.

Remark. For the preliminary estimator $\widehat{\theta}$, several methods exist for estimating parameters in GARCH models with unknown innovation distributions. The QMLE is arguably the most frequently used estimator in practice. The asymptotic properties of the QMLE have been studied in the literature under regularity conditions similar to ours. When the innovation distribution is heavy tailed, Peng and Yao (2003) propose a least absolute deviations estimator (LADE) as an alternative which is robust with respect to the heavy tails of the innovation distribution. In fact, the LADE is asymptotically normal with the standard convergence rate under weaker assumptions.

Remark. The above analysis and results can be easily extended to the case of general parametric volatility when $\sigma_t^2 = \mathrm{var}(y_t|\mathcal{F}_{t-1}) = \sigma_t^2(\theta)$, where θ is the vector of unknown parameters. For example, the well-known location-scale type model where σ_t^2 is equal to a parametric function of covariate Z_t, say $\sigma_t^2 = \rho_0 + \rho_1 Z_t^2$.

3.2. Nonparametric deterministic volatility

Although our analysis in this paper focuses on nonparametric regressions with stationary stochastic conditional heteroskedasticity, the approach can also be applied to the nonstationary case. In this subsection, we illustrate such extensions for nonparametric regressions with locally varying unconditional volatilities, or long run components. Suppose that $\sigma_t^2 = \sigma^2(t/T)$ with $\sigma^2(\cdot)$ a smooth unknown function, that is,

$$u_t = \sigma_t\varepsilon_t, \sigma_t = \sigma(t/T), \tag{12}$$

where ε_t and $\varepsilon_t^2 - 1$ are stationary martingale difference sequences. In this case, the process u_t is not stationary, although it is locally stationary, (Dahlhaus, 1997). We assume that Assumption A4 holds with $W_t = X_t$ being a stationary absolutely regular process.

For this model, under regularity conditions, the asymptotic distribution of the conventional local-polynomial regression estimator is given by

$$\sqrt{Th}\left[\widehat{m}_{LP}(x) - m(x) - h^{p+1}b_p(x)\gamma_1(K)\right] \Longrightarrow N\left(0, \frac{\int_0^1 \sigma(r)^2 dr}{f_X(x)}\omega_{11}(K)\right). \tag{13}$$

In this section, we proposed a weighted local-polynomial regression estimator along the lines of the previous sections and showed that the proposed weighted local-polynomial regression estimator has the same bias but a smaller variance.

A feasible weighted local-polynomial regression estimator $\widetilde{m}(x)$ requires estimates of σ_t^2, which can be estimated nonparametrically. We consider the following estimation procedure:

1. First, we construct a preliminary local polynomial estimator $\check{m}(\cdot)$ using bandwidth h_1 by minimizing $Q_T(\beta; x, K, h_1, \{1\})$ from (6) with respect to β
2. Then estimate $\sigma(t/T)^2$ by nonparametric smoothing on $\widehat{u}_t = y_t - \check{m}(X_t)$,

   $$\widehat{\sigma}(r)^2 = \frac{\sum_{t \neq Tr, t=1}^{T} G\left((r - t/T)/h_\sigma\right)\widehat{u}_t^2}{\sum_{t \neq Tr, t=1}^{T} G\left((r - t/T)/h_\sigma\right)},$$

   where G is a kernel function and h_σ is a bandwidth for the estimation of volatility.
3. The feasible weighted local polynomial estimator $\widetilde{m}(x)$ is constructed by minimizing $Q_T\left(\beta; x, K, h, \{\widehat{\sigma}_t^2\}\right)$ from (6) with respect to β, where h is the bandwidth in the final estimation.

In step 2, we use the leave-one-out estimator here to obtain a martingale difference sequence structure that simplifies the proof, see, e.g. Xu and Phillips (2008). We also suppose the following:

Assumption A1″. The data is given by (1), (2) and $\sigma_t = \sigma(t/T)$, where the function $\sigma(\cdot)$ is continuous and $0 < c_L \leq \inf_{u\in[0,1]} \sigma(u) \leq \sup_{u\in[0,1]} \sigma(u) \leq c_U < \infty$, such that $\int_0^1 \sigma(r)^2 dr$ and $\int_0^1 \sigma(r)^{-2} dr$ exist.

Assumption A4″. $\{X_t\}$ is a stationary absolutely regular process. That is,

$$\varrho(\tau) = \sup_s E\left\{\sup_{A\in\mathcal{G}_{s+\tau}^{\infty}} |P(A|\mathcal{G}_{-\infty}^{s}) - P(A)|\right\} \to 0, \text{ as } \tau \to \infty,$$

where $\mathcal{G}_s^t$ is the σ-field generated by $\{X_j : j = s, \ldots, t\}$. In addition, there is a positive δ such that $E(|X_t|^{2+\delta}) < \infty$, and for some δ' with $\delta > \delta' > 0$, $\varrho(\tau) = O(\tau^{-(2+\delta')/\delta'})$. The conditional density of ε_t, $f_{\varepsilon_t|X_t}(\varepsilon|x)$ is uniformly bounded and has continuous partial derivatives.

Assumption A5′. The kernels $K(\cdot)$ and $\mathrm{G}(\cdot)$ have support $[-1, 1]$ and are symmetric about zero.

Assumption A6″. As $T \to \infty$, $h \to 0$, $h_1 \to 0$, $h_\sigma \to 0$ and $h_1/h \to 0$, $h_1^{2p} h_\sigma^{-1} \to 0$, $T^{-1} h_1^{-1} h_\sigma^{-1} \log(T) \to 0$, $Th_1 h^{1+p} \to \infty$, $Th_1^2 h^{-1} \to \infty$, $\sqrt{Th} h_1^{p+1} \to 0$, $Th_\sigma h^{1/2} \to \infty$, $Th_\sigma^2 \to \infty$.

We obtain the following result.

Theorem 4. *Suppose that Assumptions A1″, A2, A3, A4″, A5′, A6″ hold. Then, as $T \to \infty$*

$$\sqrt{Th}\left[\widetilde{m}(x) - m(x) - h^{p+1} b_p(x) \gamma_1(K)\right] \Longrightarrow N\left(0, \frac{\omega_{11}(K)}{f_X(x) \int_0^1 \sigma(r)^{-2} dr}\right).$$

Theorem 4 shows that, in nonparametric regressions with locally varying volatilities, the weighted local estimator $\widetilde{m}(x)$ is more efficient than the conventional local polynomial estimator. The relative efficiency (ratio of variances) of $\widetilde{m}(x)$ to $\breve{m}(x)$ is

$$v_{eff} = \int_0^1 \sigma(r)^2 dr \times \int_0^1 \sigma(r)^{-2} dr \geq 1, \tag{14}$$

where the inequality follows by the Cauchy Schwarz inequality $(1 = \sigma \times \sigma^{-1})$ — this is just the ratio of the arithmetic mean to the harmonic mean of $\sigma(r)^2$. The magnitude of the efficiency gain increases with the variability of $\sigma(r)^2$ and is unbounded. The feasible weighted local-polynomial regression estimator $\widetilde{m}(x)$ is asymptotically equivalent (same asymptotic variance) to the infeasible weighted local-polynomial regression estimator $\widehat{m}(x)$. Muller and Stadtmuller (1987) consider the case where $X_t = t/T$ and confirm the equivalence of the unweighted and weighted kernel regression smoothers. We note that in this case the condition $E(u_t|X_t, \lambda_t) = 0$ fails, but in large samples $\lambda_t = \widehat{\sigma}_t^{-2} \simeq \sigma_t^{-2}$ is deterministic and the consistency and asymptotic normality follow. Indeed the standard errors constructed from (8) are valid in this case.

Remark. Following Vogt (2013) one may also allow the covariate to be locally stationary, i.e., to have a time varying density, which changes the variance formula a little.

4. A discussion on the bias of weighted kernel regressions

The idea of weighted regression and the previous analysis may be extended to many other nonparametric methods and models. The local polynomial estimator is widely used due to its attractive properties. For this reason, we focus our analysis on the local polynomial regression. Similar analysis on weighted regression can be applied to other nonparametric methods, say, the well-known Nadaraya–Watson regression. In general, under the assumption $\mathrm{E}(u_t^2|\mathcal{F}_{t-1}) \neq \mathrm{E}(u_t^2|X_t)$, GLS regression reduces the variances of nonparametric regressions. However, the weighting effect on biases is different among different types of nonparametric regressions. For comparison purposes and to further illustrate the effect of weighting, we briefly discuss weighted Nadaraya–Watson regression in this section. We show that although weighting has similar effects on variance, it has a different impact on biases for different nonparametric regression estimators. In particular, the weighted local polynomial regression with odd order does not change the bias, but the weighted Nadaraya–Watson kernel regression estimator (even order polynomial) does change the bias.

Consider the weighted Nadaraya–Watson regression that minimizes the following criterion:

$$\sum_{t=1}^{T} \lambda_t K\left(\frac{x - X_t}{h}\right) (Y_t - \beta)^2, \tag{15}$$

where λ_t are weights associated to the tth observation in the local polynomial regression. Let $w_t = \lambda_t K\left((x - X_t)/h\right)$. To compare the kernel estimator with pth order local polynomial regression, we consider $(p+1)$th order kernel in the Nadaraya–Watson regression, thus $\int K(u)u^r du = 0$ for $r = 1, \dots, p$, and $\int K(u)u^{p+1}du = 1$. The weighted Nadaraya–Watson estimator is given by

$$\hat{m}(x) = \frac{\sum_{t=1}^{T} \lambda_t K_h(X_t - x)y_t}{\sum_{t=1}^{T} \lambda_t K_h(X_t - x)}. \tag{16}$$

Again, suppose that the model is given by (1) and (2), it can be verified that the variance of the limiting distribution is given by $(\delta_\lambda(x)/f_X(x))\nu_0(K)$. The impact of weighting on the limiting variance of the Nadaraya–Watson regression is the same as that of the local polynomial regression. In particular, any weights λ in the form of a smooth function of X_t would give the same limiting variance. If $\sigma_t^2 \neq \sigma(X_t)^2$, GLS regressions will reduce the limiting variance. In particular, the limiting variance of GLS regression is determined by $\nu_0(K)/f_X(x)\mathrm{E}\left(\sigma_t^{-2}|X_t = x\right)$, which is smaller than the limiting variance of the unweighted NW kernel estimator $\nu_0(K)\mathrm{E}\left(\sigma_t^2|X_t = x\right)/f_X(x)$, as long as $\mathrm{E}\left(u_t^2|\mathcal{F}_{t-1}\right) \neq \mathrm{E}\left(u_t^2|X_t\right)$.

To analyze the bias term, let the joint density of (λ_t, X_t) be $g(v, x)$, notice that K is $(p+1)$th order kernel, it can be verified that

$$m^{(r)}(X_t)\frac{1}{Th}\sum_{t=1}^{T} \lambda_t K_h(X_t - X_t)u^r \approx h^{p+1-r}m^{(r)}(X_t)\frac{1}{(p+1-r)!}\int v g_x^{(p+1-r)}(v, x)dv\mu_{p+1}(K)$$

where $g_x^{(p+1-r)}(v, x) = \frac{\partial^{p+1-r}g(v,x)}{\partial x^{p+1-r}}$. The leading bias of the weighted Nadaraya–Watson estimator is given by

$$h^{p+1}\frac{\mu_{p+1}(K)}{f_X(x)\mathrm{E}\left(\lambda_t|\, X_t = x\right)}\left[\sum_{r=1}^{p+1}\frac{1}{r!(p+1-r)!}\left(m^{(r)}(x)\int v g_x^{(p+1-r)}(v, x)dv\right)\right].$$

Although weighting does not change the bias in the local polynomial regression, it does change the bias term in the Nadaraya–Watson regression. Bias reduction is possible by appropriately chosen weights. In the special case where $\lambda_t = \lambda(X_t)$ and the kernel is second order, i.e., $p + 1 = 2$, the leading bias is

$$\frac{1}{2}h^2\mu_2(K)\left[2m^{(1)}(X_t)\frac{(\lambda f)'(x)}{(\lambda f)(x)} + m^{(2)}(x)\right], \tag{17}$$

where we denote $\lambda(x)f(x)$ by $(\lambda f)(x)$, which is the result given by Jones (1993).

5. Bandwidth selection

The proposed weighted nonparametric estimator involves the use of bandwidth parameter h, and the preliminary estimation of weights also involves a bandwidth h_1 in the unweighted local regressions. In practice, a data-driven smoothing parameter selection is highly appreciated. Although in principle the bandwidth could be selected by minimizing the second order effects in MSE of the nonparametric estimator, the second order term is quite complicated and messy, and it is practically difficult to select an optimal bandwidth along this direction. Cross-validation has been widely used in selecting tuning parameters in econometrics and statistics, see, e.g. Hall and Racine (2015). In this section, we propose the following cross-validation type procedure for selecting smoothing parameters.

1. First, we construct a preliminary local polynomial estimator $\check{m}(\cdot)$ using bandwidth h_1 by minimizing $Q_T\left(\beta; x, K, h_1, \{1\}\right)$ from (6) with respect to β
2. Then estimate σ_s^2 using $\widehat{u}_s = y_s - \check{m}(X_s)$, denote the estimated variance by $\widehat{\sigma}_s^2$.
3. For each t, we estimate $m(X_t)$ using observations $\{(Y_s, X_s), |s - t| > \kappa\}$, for some large κ. More specifically, we construct the leave-k -out $(k = 2\kappa + 1)$ weighted local polynomial estimator $\widetilde{m}_{-t}(X_t)$ by minimizing:

   $$T^{-1}\sum_{s:|s-t|>\kappa}\frac{K\left((X_s - X_t)/h\right)}{\widehat{\sigma}_s^2}\left(Y_s - \sum_{0\leq j\leq p}\beta_j\left(\frac{X_t - X_s}{h}\right)^j\right)^2,$$

   where h is the bandwidth in final estimation.
4. Calculate

   $$CV\left(h, h_1\right) = \sum_{t=1}^{T}\left(Y_t - \widetilde{m}_{-t}(X_t)\right)^2$$

We may choose (h, h_1) to minimize $CV\left(h, h_1\right)$.

Remark. (1) This is a cross-validation type estimator. Since the data is weakly dependent over time, we construct the final estimator of $m(X_t)$ based on observations separated away from time t. Under weak dependence, the tth observation is almost independent with the dataset based on which we estimate it. However, $\widehat{\sigma}_t^2$ is constructed based on the whole sample for two reasons: since σ_t^2 is captured by a parametric model and the parameters are estimated based on the whole sample, we expect that the impact of the tth observation on the parameter estimation is relatively small due to weak dependence; on the other hand, the dependence structure is maintained when estimating the volatility parameters. (2) The proposed cross-validation type estimator can be easily extended to the case when the volatility is nonparametrically estimated. For the case of nonparametric deterministic volatility discussed in Section 3.2, the second step of estimating σ_s^2 in the above procedure will then be replaced by the nonparametric volatility estimator, which is dependent on h_σ, as a result, the criterion in step 4 will now become $CV\ (h, h_1, h_\sigma)$.

For convenience in practice, we also propose a simple rule of thumb method following Fan and Gijbels (1996, p 111). Specifically, to estimate bias terms we use a global polynomial curve

$$\overline{m}(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{p+1} x^{p+1}, \tag{18}$$

which is estimated by least squares, yielding estimates $\widehat{\alpha}_j$, $j = 0, \ldots, p+1$. We propose the following rule of thumb bandwidth estimator

$$h_{ROT} = C_{0,p}(K)\left[\frac{\left(\max_{1\le t\le T} X_t - \min_{1\le t\le T} X_t\right)}{\left(\frac{\widehat{\alpha}_{p+1}}{p+1!}\right)^2 \times \frac{1}{T}\sum_{t=1}^{T}\widehat{\sigma}_t^{-2}}\right]^{1/(2p+3)} T^{-1/(2p+3)}, \tag{19}$$

where $C_{0,p}(K)$ is taken from Fan and Gijbels (1996, Table 3.2). This bandwidth approximates the minimizer of the asymptotic integrated mean square error of the odd order local polynomial regression function estimator under specific conditions, which includes the specification (18) as well as the mean independence of σ_t^{-2} from X_t. When these conditions are violated h_{ROT} still converges to zero at the right rate but may not be optimal.

6. Simulation study

We conducted a Monte Carlo simulation to evaluate the finite sample performance of the proposed estimation procedure. In particular, we compare the finite sample performance between the proposed estimator $\widetilde{m}(x)$ and the conventional unweighted nonparametric estimator $\check{m}(\cdot)$. We also report the performance of the infeasible weighted local polynomial estimator $\widehat{m}(\cdot)$ based on known σ_t^2 to illustrate the potential of efficiency gain. Thus, the three estimators we consider are:

1. The conventional unweighted local polynomial estimator $\check{m}(\cdot)$ based on minimizing $Q_T\ (\beta; x, K, h_1, \{1\})$ from (6) with respect to β.
2. The proposed weighted local polynomial estimator $\widetilde{m}(\cdot)$ based on minimizing $Q_T\left(\beta; x, K, h_1, \{\widehat{\sigma}_t^2\}\right)$ from (6) with respect to β, where $\widehat{\sigma}_t^2$ is calculated based on estimated ARCH/GARCH parameters.
3. The infeasible weighted local polynomial estimator $\widehat{m}(\cdot)$ based on minimizing $Q_T\left(\beta; x, K, h_1, \{\sigma_t^2\}\right)$ from (6) with respect to β.

The data were generated from the model $Y_t = m(X_t) + \sigma_t \varepsilon_t$, where ε_t are i.i.d. standard normal distributions. Several specifications of $m(x)$ were investigated in generating the data and qualitatively similar results were obtained. Thus we report the results for the case $m(x) = x^2$ at $x = 0$.

6.1. ARCH

Our first model is the ARCH(1) model

$$\sigma_t^2 = \omega + \gamma u_{t-1}^2,$$

with $\omega = 1$. We consider a range of ARCH parameter values: $\gamma = 0.5, 0.7, 0.9$. The ARCH parameters are estimated based on OLS regression: $\widehat{u}_t^2 = \omega + \gamma \widehat{u}_{t-1}^2 + \check{\eta}_t$, where $\widehat{u}_t$ is the conventional local polynomial regression residual $\widehat{u}_t = y_t - \check{m}(X_t)$, and thus σ_t^2 can be estimated by $\widehat{\sigma}_t^2 = \widehat{\omega} + \widehat{\gamma}\widehat{u}_{t-1}^2$.

For the regressor X_t, we consider three cases: Case (I) X_t are i.i.d. standard normal; $\{X_t\}_{t=1}^T$ and $\{\varepsilon_t\}_{t=1}^T$ are independent. Case (II) X_t are i.i.d. U[0,1]; $\{X_t\}_{t=1}^T$ and $\{\varepsilon_t\}_{t=1}^T$ are independent. Case (III) $X_t = Y_{t-1}$. We report the results of the case $T = 100$. The results with $T = 500$ are qualitatively similar. The number of replications is 2000 in each case. We investigated both local linear estimation and the third order ($p = 3$) local polynomial estimation with kernel $\mathcal{K}(u) = 0.75(1-u^2)1(|u| \le 1)$, again, similar results were obtained and thus we only report the results of the case $p = 3$. Different bandwidth values were considered for the case $p = 3$. In particular, we consider bandwidth choices $h = d_0 \times s_X T^{-1/9}$ and $h_1 = d_1 \times s_X T^{-1/6}$, where s_X is the sample standard deviation of X, for 5 different sets of values

Table 1
(Case I: X_t = i.i.d. N(0,1)).

h		$\gamma = 0.5$			$\gamma = 0.7$			$\gamma = 0.9$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1	$\check{m}$	0.0001	0.0421	0.0421	−0.0015	0.0708	0.0708	0.0103	0.1802	0.1803
	$\widehat{m}$	0.0018	0.0327	0.0327	0.0003	0.0394	0.0394	0.0002	0.0447	0.0447
	$\widetilde{m}$	0.0009	0.0344	0.0344	−0.0005	0.0450	0.0450	0.0107	0.0766	0.0767
2	$\check{m}$	−0.0025	0.0335	0.0335	−0.0030	0.0550	0.0550	−0.0018	0.1221	0.1221
	$\widehat{m}$	−0.0035	0.0259	0.0259	−0.0010	0.0308	0.0308	−0.0046	0.0329	0.0329
	$\widetilde{m}$	−0.0041	0.0272	0.0272	−0.0034	0.0353	0.0353	−0.0028	0.0909	0.0909
3	$\check{m}$	−0.0043	0.0355	0.0355	−0.0062	0.0679	0.0679	−0.0007	0.1213	0.1213
	$\widehat{m}$	−0.0023	0.0261	0.0261	−0.0055	0.0292	0.0293	−0.0003	0.0328	0.0328
	$\widetilde{m}$	−0.0023	0.0286	0.0286	−0.0043	0.0440	0.0440	0.0192	0.0890	0.0894
4	$\check{m}$	0.0013	0.0311	0.0311	0.0050	0.0511	0.0511	0.0146	0.1079	0.1081
	$\widehat{m}$	−0.0019	0.0239	0.0239	0.0007	0.0274	0.0274	0.0110	0.0308	0.0310
	$\widetilde{m}$	−0.0027	0.0283	0.0283	0.0009	0.0321	0.0321	0.0136	0.0516	0.0518
5	$\check{m}$	−0.0047	0.0329	0.0329	0.0071	0.0557	0.0557	0.0140	0.1437	0.1439
	$\widehat{m}$	−0.0026	0.0250	0.0250	−0.0003	0.0267	0.0267	0.0033	0.0315	0.0315
	$\widetilde{m}$	−0.0025	0.0270	0.0270	0.0055	0.0458	0.0459	0.0060	0.0462	0.0462
6	$\widetilde{m}_{ROT}$	−0.0108	0.0221	0.0223	0.0109	0.0317	0.0318	−0.0116	0.0422	0.0423
7	$\widetilde{m}_{cv}$	−0.0153	0.0205	0.0207	0.0117	0.0289	0.0290	−0.0101	0.0401	0.0402

Table 2
(Case II: X_t = i.i.d. U[0,1]).

h		$\gamma = 0.5$			$\gamma = 0.7$			$\gamma = 0.9$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1	$\check{m}$	−0.0381	0.8241	0.8256	0.0027	1.3371	1.3371	0.0204	2.7215	2.7219
	$\widehat{m}$	−0.0327	0.6700	0.6711	0.0031	0.7752	0.7752	−0.0098	0.8281	0.8282
	$\widetilde{m}$	−0.0338	0.6932	0.6943	0.0005	0.8998	0.8998	0.0028	1.2184	1.2184
2	$\check{m}$	−0.0106	0.4061	0.4062	0.0078	0.6196	0.6197	0.0121	1.4287	1.4289
	$\widehat{m}$	−0.0177	0.3142	0.3145	0.0159	0.3531	0.3533	0.0192	0.4029	0.4033
	$\widetilde{m}$	−0.0141	0.3322	0.3324	0.0160	0.3895	0.3898	0.0122	0.5589	0.5590
3	$\check{m}$	−0.0032	0.3802	0.3802	−0.0062	0.5617	0.5617	0.0002	1.0959	1.0959
	$\widehat{m}$	−0.0111	0.2622	0.2624	−0.0063	0.2888	0.2888	0.0064	0.3556	0.3556
	$\widetilde{m}$	−0.0130	0.2938	0.2940	−0.0107	0.3282	0.3284	0.0118	0.5264	0.5266
4	$\check{m}$	−0.0145	0.3429	0.3431	0.0012	0.6861	0.6861	−0.0078	1.0942	1.0942
	$\widehat{m}$	−0.0085	0.2495	0.2496	0.0015	0.3201	0.3201	−0.0091	0.3693	0.3694
	$\widetilde{m}$	−0.0105	0.2638	0.2639	−0.0019	0.3748	0.3748	−0.0059	0.4924	0.4924
5	$\check{m}$	−0.0085	0.3458	0.3459	−0.0194	0.5715	0.5718	−0.0278	1.2512	1.2520
	$\widehat{m}$	−0.0078	0.2723	0.2724	−0.0045	0.3083	0.3083	0.0053	0.3445	0.3446
	$\widetilde{m}$	−0.0046	0.3120	0.3120	−0.0028	0.3874	0.3874	−0.0070	0.5302	0.5303
6	$\widetilde{m}_{ROT}$	−0.0753	0.2792	0.2848	0.0433	0.3441	0.3460	−0.0326	0.5179	0.5189
7	$\widetilde{m}_{cv}$	−0.0128	0.2426	0.2427	0.0285	0.3167	0.3175	−0.0060	0.4014	0.4014

of (d_0, d_1): $(3, 2)$, $(5.5, 3.5)$, $(8, 5)$, $(15, 10)$, $(25, 16)$. We also examine the performance of the estimator based on the ROT bandwidth (denoted by $\widetilde{m}_{ROT}$ in the tables) and the cross-validation based estimator (denoted by $\widetilde{m}_{cv}$ in the tables) proposed in Section 5. For the estimator based on the ROT bandwidth, we simply used $h_1 = 10 \times s_X T^{-1/6}$ in the first stage preliminary estimation.

We compared the biases, variances, and mean squared errors of these estimators given different choices of innovation processes and bandwidth values. Tables 1, 2, 3 report results for cases (I), (II), (III). The efficiency gain from weighted regression is quite significant. In addition, it is apparent that as the conditional heteroskedasticity increases (as γ increases from 0.5 to 0.9), the efficiency gain from weighted nonparametric regression also increases. Third, the efficiency gain in the case with independent regressors is generally larger than that of the autoregressions.

6.2. GARCH

We next consider the GARCH model. We consider the same regression function, i.e. the data were generated from the model $Y_t = m(X_t) + \sigma_t \varepsilon_t$, with $m(x) = x^2$. Now σ_t follows a GARCH(1,1) process

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma u_{t-1}^2$$

with $\omega = 1$. We consider a range of GARCH parameter values given as follows

Table 3
(Case III: $X_t = Y_{t-1}$).

h		$\gamma = 0.5$			$\gamma = 0.7$			$\gamma = 0.9$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1	$\breve{m}$	0.0076	2.2742	2.2743	−0.0316	1.3927	1.3937	−0.0222	2.2900	2.2905
	$\widehat{m}$	−0.0020	2.2605	2.2605	−0.0227	1.0982	1.0987	−0.0180	2.2896	2.2900
	$\widetilde{m}$	0.0030	2.1859	2.1859	−0.0392	1.2358	1.2374	−0.0443	2.2112	2.2132
2	$\breve{m}$	0.0029	0.8111	0.8111	0.0236	1.0902	1.0907	−0.0147	1.3587	1.3590
	$\widehat{m}$	−0.0042	0.6403	0.6404	0.0044	0.6797	0.6798	−0.0132	0.5275	0.5276
	$\widetilde{m}$	−0.0001	0.7448	0.7448	0.0144	0.9944	0.9946	−0.0186	0.9854	0.9858
3	$\breve{m}$	−0.0156	0.6050	0.6053	−0.0107	0.7629	0.7630	0.0159	1.3672	1.3675
	$\widehat{m}$	−0.0044	0.4796	0.4797	0.0038	0.5601	0.5601	−0.0089	0.4358	0.4359
	$\widetilde{m}$	−0.0054	0.5729	0.5729	−0.0009	0.7510	0.7510	−0.0028	1.0082	1.0082
4	$\breve{m}$	0.0113	0.6568	0.6569	−0.0091	0.8096	0.8097	−0.0057	1.3241	1.3242
	$\widehat{m}$	0.0055	0.5193	0.5193	−0.0061	0.4770	0.4771	0.0109	0.5879	0.5880
	$\widetilde{m}$	0.0103	0.6284	0.6285	−0.0059	0.6604	0.6605	0.0038	1.0216	1.0216
5	$\breve{m}$	−0.0073	0.6495	0.6495	0.0002	0.6982	0.6982	0.0253	1.5548	1.5555
	$\widehat{m}$	−0.0112	0.5258	0.5259	0.0014	0.4649	0.4649	0.0065	0.6399	0.6399
	$\widetilde{m}$	−0.0090	0.6260	0.6260	0.0032	0.5950	0.5950	0.0394	1.3590	1.3606
6	$\widetilde{m}_{ROT}$	0.0284	0.5140	0.5148	0.0230	0.9529	0.9534	0.2184	0.9304	0.9781
7	$\widetilde{m}_{cv}$	−0.0019	0.3838	0.3839	0.0291	0.3927	0.3936	−0.0083	0.5996	0.5997

β	0.1	0.3	0.5	0.7	0.3	0.5	0.1	0.3	0.5	0.9	0.05
γ	0.8	0.6	0.4	0.2	0.5	0.3	0.6	0.4	0.2	0.05	0.9

We have investigated the sampling properties for similar designs on X. Qualitatively very similar results to the ARCH model are obtained. For this reason, we report the results for the case where $\{\varepsilon_t\}_{t=1}^T$ are i.i.d. N(0,1), and $\{X_t\}_{t=1}^T$ are i.i.d. U[0,1] random variables that are independent with $\{\varepsilon_s\}_{s=1}^T$. Again, $T = 100$, and the number of replications is 2000 in each case.

Since there are more parameters in the GARCH case, and results are similar to the ARCH case, for simplicity, we only report the biases and mean squared errors of the local polynomial estimators with $p = 3$ and $x = 0$. We consider the same bandwidth choices 1–5, as well as the ROT and cross-validation bandwidth as in the previous case. The results are contained in Table 4. In particular, we find that: Given each γ, as β increases, the relative efficiency gain increases. Similarly, given each β, as γ increases, the relative efficiency gain increases.

6.3. Locally varying volatility

We finally look at the locally varying volatility model. We consider the same regression function, i.e. the data were generated from the model $Y_t = m(X_t) + \sigma_t\varepsilon_t$, with $m(x) = x^2$. Now σ_t follows a locally varying volatility process:

$$\sigma_t^2 = \omega + \gamma\ \sin(t\pi/T)^2$$

with $\omega = 0.1$, and we consider different values for $\gamma = 1, 2, 5$. The choices of the regressor X_t are similar to the previous cases, i.e., we again consider the same three cases where (i) X_t are i.i.d. standard normal and independent with ε_t; (ii) X_t are i.i.d. U[0,1] and independent with ε_t; and (iii) $X_t = Y_{t-1}$. We investigated the third order ($p = 3$) local polynomial estimation with kernel $\mathcal{K}(u) = 0.75(1 - u^2)1(|u| \leq 1)$, again, similar results were obtained in local linear estimation and we only report the results of the case $p = 3$. We use the same kernel function in estimating the volatility $\sigma(\cdot)$ as the one used in estimating the mean function. In addition to $h = d_0 \times s_X T^{-1/9}$ and $h_1 = d_1 \times s_X T^{-1/6}$, that we used before for the second stage and first stage nonparametric estimation of the mean, we simply use $h_\sigma = s_X T^{-1/6}$. The same 5 different sets of values of (d_0, d_1) were considered. We also examine the performance of the ROT and cross-validation based estimator (again, denoted by $\widetilde{m}_{cv}$ in the tables) proposed in Section 5.

The number of replications is the same as before. We report the results of the case $T = 100$.

The Monte Carlo results that we obtained are very similar to the previous cases. For this reason, we only report the result for the case when X_t are generated by i.i.d. U[0,1]. In particular, Table 5 reports results for the biases, variances, and mean squared errors of these estimators given different choices of bandwidth values. Results of Table 5 show the potential of efficiency gain from weighted nonparametric regression in the locally varying volatility models.

6.4. Additional discussion: The effect of weighting near IGARCH

The Monte Carlo simulation above indicates that the weighted nonparametric regression generally brings efficiency gain for models with a GARCH error process. In particular, the efficiency gain from weighted regression is quite significant when γ is large.

Table 4
GARCH.

h		$(\beta,\gamma)=(0.1,0.8)$		$(\beta,\gamma)=(0.3,0.6)$		$(\beta,\gamma)=(0.5,0.4)$		$(\beta,\gamma)=(0.7,0.2)$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
1	$\breve{m}$	0.0385	1.4338	0.0003	1.6126	0.0329	1.8480	0.0326	1.7066
	$\widehat{m}$	0.0135	0.4775	0.0101	0.7220	0.0160	1.0654	0.0237	1.4932
	$\widetilde{m}$	0.0195	0.5816	0.0064	0.8363	0.0176	1.2015	0.0313	1.5684
2	$\breve{m}$	−0.0057	1.5221	0.0507	1.4295	0.0150	1.5971	0.0072	1.7455
	$\widehat{m}$	0.0109	0.4340	0.0311	0.6825	0.0268	0.9586	0.0110	1.4731
	$\widetilde{m}$	−0.0033	0.5508	0.0361	0.8192	0.0073	1.1020	0.0056	1.5527
3	$\breve{m}$	−0.0125	1.5600	0.0245	1.7434	0.0033	1.7712	−0.0168	1.7405
	$\widehat{m}$	−0.0023	0.4424	0.0304	0.6869	−0.0118	1.0316	−0.0091	1.4144
	$\widetilde{m}$	−0.0013	0.5324	0.0276	0.7789	−0.0156	1.1829	−0.0249	1.5416
4	$\breve{m}$	0.0279	1.5065	0.0103	1.6511	−0.0106	1.7384	−0.0180	1.6859
	$\widehat{m}$	0.0110	0.4743	0.0261	0.6901	−0.0414	1.0157	−0.0203	1.5276
	$\widetilde{m}$	0.0193	0.5548	0.0315	0.7658	−0.0343	1.1487	−0.0163	1.5599
5	$\breve{m}$	0.0115	1.7135	−0.0022	1.7232	−0.0251	1.6672	0.0489	1.8362
	$\widehat{m}$	−0.0047	0.4540	−0.0129	0.6624	−0.0001	1.0662	0.0601	1.5518
	$\widetilde{m}$	−0.0096	0.5750	−0.0185	0.7865	−0.0143	1.1997	0.0559	1.6254
6	$\widetilde{m}_{ROT}$	0.0145	0.4772	−0.0164	0.7216	0.0169	1.2048	0.1402	1.2641
7	$\widetilde{m}_{cv}$	−0.0152	0.4161	0.0030	0.6798	−0.1266	1.0034	0.1337	1.2252
h		$(\beta,\gamma)=(0.5,0.3)$		$(\beta,\gamma)=(0.1,0.6)$		$(\beta,\gamma)=(0.3,0.4)$		$(\beta,\gamma)=(0.5,0.2)$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
1	$\breve{m}$	−0.0269	0.9331	−0.0030	0.6135	0.0109	0.6110	−0.0154	0.6353
	$\widehat{m}$	−0.0221	0.7452	−0.0063	0.3685	−0.0047	0.4632	−0.0152	0.5919
	$\widetilde{m}$	−0.0223	0.7924	−0.0037	0.4014	−0.0009	0.4925	−0.0148	0.6057
2	$\breve{m}$	−0.0037	0.9313	−0.0186	0.6028	−0.0164	0.5611	−0.0151	0.5959
	$\widehat{m}$	0.0023	0.7549	−0.0202	0.3608	−0.0305	0.4278	−0.0153	0.5566
	$\widetilde{m}$	0.0053	0.8082	−0.0214	0.3925	−0.0269	0.4473	−0.0204	0.5701
3	$\breve{m}$	−0.0336	0.8765	0.0146	0.6235	−0.0112	0.6082	0.0340	0.5991
	$\widehat{m}$	−0.0233	0.7011	0.0103	0.3456	−0.0214	0.4696	0.0275	0.5568
	$\widetilde{m}$	−0.0227	0.7448	0.0029	0.3769	−0.0132	0.4893	0.0348	0.5714
4	$\breve{m}$	−0.0145	0.8483	0.0098	0.5578	−0.0161	0.5707	0.0454	0.5986
	$\widehat{m}$	−0.0042	0.7075	0.0195	0.3415	−0.0105	0.4405	0.0458	0.5390
	$\widetilde{m}$	−0.0092	0.7374	0.0128	0.3669	−0.0110	0.4623	0.0469	0.5619
5	$\breve{m}$	−0.0171	0.9010	−0.0039	0.5655	0.0043	0.5607	−0.0090	0.5937
	$\widehat{m}$	−0.0208	0.7364	−0.0005	0.3501	0.0196	0.4324	−0.0013	0.5469
	$\widetilde{m}$	−0.0275	0.7777	0.0003	0.3769	0.0176	0.4549	−0.0016	0.5568
6	$\widetilde{m}_{ROT}$	−0.0113	0.6964	0.0382	0.3674	0.0371	0.4795	0.0013	0.5332
7	$\widetilde{m}_{cv}$	−0.0027	0.6957	−0.0532	0.3589	−0.0276	0.4394	−0.0001	0.5057

h		$(\beta,\gamma)=(0.3,0.5)$		$(\beta,\gamma)=(0.05,0.9)$		$(\beta,\gamma)=(0.9,0.05)$	
		Bias	MSE	Bias	MSE	Bias	MSE
1	$\breve{m}$	−0.0289	1.2714	0.0075	2.2866	−0.0815	3.2651
	$\widehat{m}$	−0.0178	0.5508	−0.0300	0.3856	−0.0644	3.2037
	$\widetilde{m}$	−0.0194	0.6025	−0.0263	0.5018	−0.0795	3.2220
2	$\breve{m}$	−0.0512	0.8763	−0.0396	1.4264	−0.0495	3.6965
	$\widehat{m}$	−0.0217	0.5442	0.0463	0.4416	−0.0996	3.6524
	$\widetilde{m}$	−0.0207	0.6002	0.0722	0.7835	−0.0819	3.6634
3	$\breve{m}$	0.0031	0.8514	−0.1526	1.6438	−0.0637	4.1541
	$\widehat{m}$	0.0076	0.5296	−0.0178	0.4140	−0.0401	4.0120
	$\widetilde{m}$	0.0041	0.5655	0.0066	0.6197	−0.0376	4.0504
4	$\breve{m}$	0.0519	0.8754	−0.0016	1.3620	−0.1599	3.3553
	$\widehat{m}$	0.0384	0.5477	0.0328	0.4094	−0.1581	3.2812
	$\widetilde{m}$	0.0328	0.6046	0.0318	0.4833	−0.1357	3.3467
5	$\breve{m}$	−0.0376	0.8586	−0.0359	1.9688	0.0301	3.6197
	$\widehat{m}$	−0.0169	0.5437	−0.0310	0.4007	0.0104	3.5215
	$\widetilde{m}$	−0.0197	0.5870	−0.0013	0.4575	0.0078	3.6016
6	$\widetilde{m}_{ROT}$	0.0265	0.5785	−0.0084	0.4687	0.1985	3.3681
7	$\widetilde{m}_{cv}$	−0.0039	0.5586	−0.0121	0.3998	−0.1595	3.2911

The focus of this paper is on stationary time series. Although not the focus of this paper, an interesting case is the IGARCH model, i.e., σ_t^2 obeys (4) with $\beta + \gamma = 1$. Then provided $\omega > 0$ and $E[\ln\left(\beta + \gamma\varepsilon_t^2\right)] < 0$, the process σ_t^2 is

Table 5
Locally varying volatility model.

h		$\gamma = 1$			$\gamma = 2$			$\gamma = 5$		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1	$\breve{m}$	0.0149	0.2483	0.2486	−0.0768	0.6294	0.6353	−0.0693	1.1792	1.1840
	$\widehat{m}$	0.0154	0.1473	0.1475	−0.0545	0.3412	0.3442	−0.0415	0.5314	0.5331
	$\widetilde{m}$	0.0156	0.1787	0.1789	−0.0577	0.4608	0.4642	−0.0839	0.8203	0.8274
2	$\breve{m}$	−0.0198	0.1211	0.1215	0.0020	0.2411	0.2411	−0.0083	0.5346	0.5346
	$\widehat{m}$	−0.0201	0.0712	0.0716	0.0195	0.1333	0.1337	−0.0363	0.2283	0.2296
	$\widetilde{m}$	−0.0165	0.0904	0.0907	0.0156	0.1669	0.1671	−0.0173	0.3322	0.3325
3	$\breve{m}$	−0.0124	0.1217	0.1219	−0.0005	0.2123	0.2123	−0.0111	0.5075	0.5076
	$\widehat{m}$	−0.0140	0.0806	0.0808	0.0027	0.0958	0.0958	−0.0076	0.1587	0.1587
	$\widetilde{m}$	−0.0142	0.0913	0.0915	0.0014	0.1470	0.1470	−0.0051	0.3128	0.3128
4	$\breve{m}$	−0.0052	0.1093	0.1093	−0.0324	0.1768	0.1778	0.0416	0.4227	0.4245
	$\widehat{m}$	−0.0021	0.0619	0.0620	0.0069	0.0930	0.0930	−0.0183	0.1581	0.1584
	$\widetilde{m}$	−0.0021	0.0802	0.0802	−0.0012	0.1176	0.1176	0.0148	0.2700	0.2703
5	$\breve{m}$	0.0092	0.1026	0.1027	0.0314	0.1780	0.1790	−0.0098	0.4810	0.4810
	$\widehat{m}$	−0.0162	0.0617	0.0619	0.0049	0.0923	0.0923	−0.0265	0.1781	0.1788
	$\widetilde{m}$	−0.0050	0.0782	0.0782	0.0261	0.1212	0.1219	−0.0039	0.3051	0.3051
6	$\widetilde{m}_{ROT}$	−0.0333	0.0876	0.0887	−0.0391	0.1178	0.1194	−0.0526	0.2783	0.2811
7	$\widetilde{m}_{cv}$	−0.0214	0.0699	0.0703	−0.0009	0.1177	0.1177	0.0085	0.2700	0.2701

strictly stationary and ergodic, Nelson (1990), Theorem 2), while Nelson (1990, Theorem 3), implies that $E(\sigma_t^2) = \infty$, and $E(u_t^2) = \infty$ (but $E(|u_t|^{1+\alpha}) < \infty$ for some $\alpha \in (0, 1)$). In this case, the Nadaraya–Watson smoother may be consistent but its asymptotic variance is infinite, i.e., the rate of convergence is slower than $\sqrt{Th}$. However, the weighted smoother can be asymptotically normal at the usual rates, since under strong stationarity we may have for example $E(\sigma_t^{-2}) < \infty$.

In the case of IGARCH, under appropriate regularity assumptions, in particular, if

$$\mathrm{E}[\ln\left(\beta + \gamma\varepsilon_{t-1}^2\right)] < 0, \text{ and } \mathrm{E}\left[\left(\beta + \gamma\varepsilon_t^2\right)\ln\left(\beta + \gamma\varepsilon_t^2\right)\right] < \infty$$

and we assume that $\omega > 0$, then there exists a stationary solution to the GARCH model, and the stationary solution is regularly varying and strong mixing with geometric rate. In this case, $E\left(\sigma_t^2\right) = \infty$. the unweighted local least-squares estimator will converge at a slower rate. However, the weighted smoother can be asymptotically normal at the usual rates, since under strong stationarity we may have $E(\sigma_t^{-2}) < \infty$. In addition, a root-n consistent estimator of the IGARCH parameter can be obtained via the QMLE (see, e.g. Lumsdaine 1996), and can be used in constructing the weighted nonparametric estimator. Thus, efficiency gain of weighted nonparametric regression may be extended to the IGARCH case.

We provide a preliminary Monte Carlo investigation below on the relative efficiency between the unweighted nonparametric regression and the weighted nonparametric regression for the GARCH model when the summation of parameters is close to unity. We consider the same GARCH model as in the previous section, i.e. the data were generated from the model $Y_t = m(X_t) + \sigma_t\varepsilon_t$, with $m(x) = x^2$, and

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma u_{t-1}^2$$

with $\omega = 1$. We consider some GARCH parameter values that $(\beta + \gamma)$ are near unity. Again, ε_t are i.i.d. N(0,1), and X_t are i.i.d. U[0,1] random variables that are independent with $\{\varepsilon_t\}_{t=1}^T$. $T = 100$.

Table 6 reports the biases and mean squared errors of the local polynomial estimators with $p = 3$ and $x = 0$. We consider the same bandwidth choices 1–5, as well as the ROT and cross-validation bandwidth as in the previous section.

These Monte Carlo results indicate that, efficiency gain of the weighted nonparametric regression over unweighted nonparametric regression can also be obtained in the near IGARCH and IGARCH cases.

7. Conclusions

We have shown that the efficiency of local linear regression estimators can be improved by weighting factors that take account of the heteroskedasticity where that heteroskedasticity is partly driven by factors different from those that influence the mean. In some applications this may deliver substantial efficiency gains in estimation. In this paper, we focus our analysis on stationary models. We expect that the method can be extended to nonstationary volatility models such as IGARCH. Monte Carlo evidence indicates efficiency gains from the weighted nonparametric regressions in this case. However, the asymptotic analysis requires different techniques. The analysis in our paper can also be extended to nonparametric quantile regression with heteroskedastic errors. We wish to explore these extensions in future research.

Table 6
Near IGARCH.

h	(β, γ) :	$(0.05, 0.95)$		$(0.05, 0.94)$		$(0.5, 0.5)$		$(0.94, 0.05)$		$(0.95, 0.05)$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
1	$\breve{m}$	−0.0582	2.7579	−0.1313	1.9291	−0.1765	6.2080	−0.0260	7.6961	−0.2222	9.7180
	$\widehat{m}$	0.0006	0.4925	−0.0829	0.4897	0.0255	1.6471	0.0136	6.0870	−0.1141	6.8226
	$\widetilde{m}$	−0.0157	0.6585	−0.0985	0.6095	0.0231	2.0672	0.0099	6.5904	−0.1635	7.6730
2	$\breve{m}$	0.0233	1.7763	−0.0170	2.2681	−0.0459	4.8465	−0.2219	7.9403	−0.0503	9.5961
	$\widehat{m}$	−0.0101	0.4720	−0.0055	0.4172	0.0568	1.4943	−0.1590	6.2429	0.0032	6.0484
	$\widetilde{m}$	−0.0031	0.5727	0.0209	0.5001	0.0586	2.0535	−0.1983	6.8638	−0.0393	6.6404
3	$\breve{m}$	0.1142	6.8679	0.0257	2.8847	−0.0363	4.8474	−0.1087	7.3660	−0.1352	10.1685
	$\widehat{m}$	−0.0650	0.4678	0.0184	0.4853	0.0153	1.6351	0.0078	5.4626	0.0255	6.5706
	$\widetilde{m}$	−0.0347	0.6831	0.0224	0.5838	0.0075	1.9647	0.0157	5.9843	0.0359	7.1606
4	$\breve{m}$	0.0006	1.6362	−0.0370	1.7776	−0.4752	14.3330	−0.0487	6.9447	−0.2477	10.7011
	$\widehat{m}$	−0.0188	0.4217	0.0305	0.3899	−0.1528	1.6073	−0.0066	5.0597	−0.1492	7.3720
	$\widetilde{m}$	−0.0176	0.5554	0.0072	0.5165	−0.2036	2.3974	−0.0392	5.7262	−0.1860	8.3218
5	$\breve{m}$	−0.2755	8.3777	0.1285	2.4898	−0.1210	5.0230	−0.1928	6.5399	0.0006	10.1353
	$\widehat{m}$	−0.0288	0.3953	0.0274	0.4726	0.0311	1.5307	−0.1194	5.1529	0.0344	6.9045
	$\widetilde{m}$	−0.0551	0.5066	0.0147	0.5834	−0.0181	1.8940	−0.1366	5.8575	0.0723	7.7006
6	$\widetilde{m}_{ROT}$	−0.0541	0.5193	0.0045	0.5169	−0.0143	1.9056	−0.0388	5.7157	0.0509	6.9218
7	$\widetilde{m}_{cv}$	−0.0421	0.4575	0.0131	0.4675	−0.0077	1.6259	0.0001	5.1482	−0.0449	6.6402

8. A sketch of proofs

We provide a sketch of proofs for our theorems in the paper. A more detailed proof can be found in the supplementary technical appendix.

8.1. Some preliminary results

Let $M_{T,h}(x)$ be a $(p+1) \times (p+1)$ matrix with the (j, k) element defined as:

$$M_{T,h,j,k} = \frac{1}{Th} \sum_{i=1}^{T} \left(\frac{x - X_i}{h} \right)^{j+k} K \left(\frac{x - X_i}{h} \right), j, k = 0, 1, \ldots, p,$$

and $\Psi_T(x)$ be a $(p+1) \times 1$ vector with the jth element:

$$\Psi_{T,h,j} = \frac{1}{Th} \sum_{i=1}^{T} \left(\frac{x - X_i}{h} \right)^{j} K \left(\frac{x - X_i}{h} \right) Y_i, j = 0, 1, \ldots, p,$$

then, the local polynomial estimator $\breve{m}(x)$ can be written as $\breve{m}(x) = \breve{\beta}_0(x) = e_1^\top M_{T,h_1}^{-1} \Psi_{T,h_1}$.

To analyze the bias and variance effects of $\breve{m}(x)$, we define the stochastic term $U_{T,h_1}(x)$ and the bias term $B_{T,h_1}(x)$ as $(p+1) \times 1$ vectors with the jth elements:

$$U_{T,h,j} = \frac{1}{Th} \sum_{i=1}^{T} \left(\frac{x - X_i}{h} \right)^{j} K \left(\frac{x - X_i}{h} \right) u_i, j = 0, 1, \ldots, p,$$

$$B_{T,h,j} = \frac{1}{Th} \sum_{i=1}^{T} \left(\frac{x - X_i}{h} \right)^{j} K \left(\frac{x - X_i}{h} \right) \Delta_i(x), j = 0, 1, \ldots, p,$$

where $\Delta_i(x) = m(X_i) - \sum_{0 \le k \le p} \frac{1}{k!} m^{(k)}(x)(X_i - x)^k$. Then,

$$\sqrt{Th_1} \left[\breve{m}(x) - m(x) - e_1^\top M_{T,h_1}^{-1}(x) B_{T,h_1}(x) \right] = e_1^\top M_{T,h_1}^{-1}(x) \sqrt{Th_1} U_{T,h_1}(x).$$

8.2. Proof of Theorem 1

The weighted local linear regression minimizes the following criterion:

$$Q_n(x; \beta) = T^{-1} \sum_{t=1}^{T} w_t \left(Y_t - \beta^\top \mathbb{X}_t \right)^2$$

and

$$\sqrt{Th}\left(\widehat{\beta}_{\lambda}-\beta-\left[\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\right)=\left[\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\left[\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t\right].$$

Notice that, under Assumption A3, and by a Taylor expansion of $m(X_t)$ around x, it can be verified that the leading bias term is given by

$$\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\approx h^{p+1}\frac{m^{(p+1)}(x)f_X(x)\mathrm{E}\left(\lambda_t|X_t=x\right)}{(p+1)!}B(K).$$

Second, under Assumptions A4 and A5,

$$\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\to f_X(x)\mathrm{E}\left(\lambda_t|X_t=x\right)M(K),$$

thus

$$\frac{1}{h^{p+1}}\left[\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\to\frac{m^{(p+1)}(x)}{(p+1)!}M(K)^{-1}B(K).$$

Finally, we look at the effect of weighting on variance. Notice that $\{\varepsilon_t\}$ is a m.d.s., and $\mathrm{E}\varepsilon_t^2=1$, under Assumptions A2–A4, by central limiting theorem for m.d.s. and application of the Cramer–Wold device, we have

$$\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t\Longrightarrow N\left(0,f_X\left(x\right)\mathrm{E}\left[\lambda_t^2\sigma_t^2|X_t=x\right]\Gamma(K)\right).$$

Thus,

$$\sqrt{Th}\left(\widehat{\beta}_{\lambda}-\beta-h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}M(K)^{-1}B(K)\right)\Longrightarrow N\left(0,\frac{\mathrm{E}\left[\lambda_t^2u_t^2|X_t=x\right]}{\left[\mathrm{E}\left(\lambda_t|X_t=x\right)\right]^2}\frac{1}{f_X\left(x\right)}M(K)^{-1}\Gamma(K)M(K)^{-1}\right).$$

8.3. *Proof of Corollary 1*

The results can be obtained from Theorem 1 by taking $\lambda_t=\sigma_t^{-2}$ and calculating the corresponding expectations.

8.4. *Proof of Theorem 2*

Notice that $\sqrt{Th}\left[\widetilde{m}(x)-m(x)\right]=\sqrt{Th}\left[\widehat{m}(x)-m(x)\right]+\sqrt{Th}\left[\widetilde{m}(x)-\widehat{m}(x)\right]$, by result of Theorem 1, we only need to show:

$$\sqrt{Th}\left[\widetilde{m}(x)-\widehat{m}(x)\right]=o_p(1).$$

By definition, $\widetilde{m}(x)$ is obtained by minimizing

$$\widetilde{Q}_T\left(x;\beta\right)=T^{-1}\sum_{t=1}^{T}\widehat{w}_t\left(Y_t-\beta^{\intercal}\mathbb{X}_t\right)^2,$$

and

$$\widetilde{\beta}=\beta+\left[\sum_{\mathbf{t=1}}^{\mathbf{T}}\widehat{\mathbf{w}}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}^{\intercal}\right]^{\mathbf{-1}}\sum_{\mathbf{t=1}}^{\mathbf{T}}\widehat{\mathbf{w}}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}\boldsymbol{\Delta}_{\mathbf{t}}(\mathbf{x})+\left[\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tu_t.$$

In addition, notice that

$$\widehat{\beta}=\beta+\left[\sum_{\mathbf{t=1}}^{\mathbf{T}}\mathbf{w}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}^{\intercal}\right]^{\mathbf{-1}}\sum_{\mathbf{t=1}}^{\mathbf{T}}\mathbf{w}_{\mathbf{t}}\mathbb{X}_{\mathbf{t}}\boldsymbol{\Delta}_{\mathbf{t}}(\mathbf{x})+\left[\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t,$$

Thus,

$$\begin{aligned}\sqrt{Th}\left(\widetilde{\beta}-\widehat{\beta}\right) &= \left[\frac{1}{Th}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tY_t-\left[\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}\frac{1}{\sqrt{Th}}w_t\mathbb{X}_tY_t\\ &= \left[\mathbf{\frac{1}{Th}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal}}\right]^{\mathbf{-1}}\mathbf{\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\Delta_t(x)}-\left[\mathbf{\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}}\right]^{\mathbf{-1}}\mathbf{\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)}\\ &+\left[\frac{1}{Th}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tu_t-\left[\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t\end{aligned}$$

We need to analyze the following terms:

$$\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal},\ \sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\Delta_t(x),\ \frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tu_t.$$

Denote

$$\widehat{A}=\frac{1}{Th}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbb{X}_t^{\intercal},\text{ and }A=\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal},$$

$$\widehat{B}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_t\mathbf{\Delta_t(x)},\text{ and }B=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbf{\Delta_t(x)},$$

$$\widehat{C}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\widehat{w}_t\mathbb{X}_tu_t,\text{ and }C=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t,$$

then, notice that

$$\widehat{A}^{-1}=A^{-1}-A^{-1}\left(\widehat{A}-A\right)A^{-1}+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1},$$

we have

$$\begin{aligned}&\sqrt{Th}\left(\widetilde{\beta}-\widehat{\beta}\right)\\ &=\widehat{A}^{\mathbf{-1}}\widehat{B}-A^{-1}B+\widehat{A}^{-1}\widehat{C}-A^{-1}C\\ &=\left[A^{-1}-A^{-1}\left(\widehat{A}-A\right)A^{-1}+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}\right]\left(B+\widehat{B}-B\right)-A^{-1}B\\ &+\ \left[A^{-1}-A^{-1}\left(\widehat{A}-A\right)A^{-1}+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}\right]\left(C+\widehat{C}-C\right)-A^{-1}C\\ &=A^{-1}\left(\widehat{B}-B\right)-A^{-1}\left(\widehat{A}-A\right)A^{-1}B+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}B-A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{B}-B\right)\\ &+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}\left(\widehat{B}-B\right)\\ &+A^{-1}\left(\widehat{C}-C\right)-A^{-1}\left(\widehat{A}-A\right)A^{-1}C+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}C-A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{C}-C\right)\\ &+A^{-1}\left(\widehat{A}-A\right)A^{-1}\left(\widehat{A}-A\right)\widehat{A}^{-1}\left(\widehat{C}-C\right)\end{aligned}$$

which is $o_p(1)$ by (1) Assumption A7 and the fact that $A=\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\to f_X(x)\mathrm{E}(\lambda_t|X_t=x)\,M(K)$ which is positive definite.

Thus $\sqrt{Th}\left(\widehat{\beta}-\widetilde{\beta}\right)=o_p(1)$. Consequently,

$$\sqrt{Th}\left[\widetilde{m}(x)-m(x)-h^{(p+1)}b(x)\right]\Longrightarrow N\left(0,\frac{\omega_{11}(K)}{f_X(x)\mathrm{E}\left[\sigma_t^{-2}|X_t=x\right]}\right).$$

8.5. *Proof of* Theorem 3

Notice that the conditional variance follows a GARCH(1,1) process, then under Assumption A1′,

$$\sigma_t^2=\omega+\beta\sigma_{t-1}^2+\gamma u_{t-1}^2=\frac{\omega}{1-\beta}+\gamma\sum_{j=1}^{\infty}\beta^{j-1}u_{t-j}^2.$$

Let $\widehat{\theta}=(\widehat{\omega},\widehat{\beta},\widehat{\gamma})^{\intercal}$ be a preliminary root-T consistent estimator of $\theta=(\omega,\beta,\gamma)$, and $\widehat{u}_t=y_t-\check{m}(X_t)$, we estimate σ_t^2 by

$$\widehat{\sigma}_t^2=\frac{\widehat{\omega}}{1-\widehat{\beta}}+\widehat{\gamma}\sum_{j=1}^{\min\{t-1,\tau\}}\widehat{\beta}^{j-1}\widehat{u}_{t-j}^2.$$

Thus, for $t > \tau$,

$$\begin{aligned}
&\widehat{\sigma}_t^2 - \sigma_t^2 \\
&= \frac{\widehat{\omega}}{1-\widehat{\beta}} - \frac{\omega}{1-\beta} + (\widehat{\gamma} - \gamma)\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\widehat{u}_{t-j}^2 \\
&+ \gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}^2 - u_{t-j}^2\right) + \gamma\sum_{j=1}^{\tau}\left(\widehat{\beta}^{j-1} - \beta^{j-1}\right)u_{t-j}^2 - \gamma\sum_{j=\tau+1}^{\infty}\beta^{j-1}u_{t-j}^2.
\end{aligned}$$

We use the proof of Theorem 2 to the GARCH case. In particular, we verify that under the assumptions of Theorem 3, Assumption A7(a), (b) and (c) (that were used in the proof of Theorem 2) still hold in the GARCH case. Let

$$R_{T1} = \frac{1}{Th}\sum_{t=1}^{T}(\widehat{w}_t - w_t)\,\mathbb{X}_t\mathbb{X}_t^{\intercal},\ R_{T2} = \frac{1}{\sqrt{Th}}\sum_{t=1}^{T}(\widehat{w}_t - w_t)\,\mathbb{X}_t u_t,\ R_{T3} = \frac{1}{\sqrt{Th}}\sum_{t=1}^{T}(\widehat{w}_t - w_t)\,\mathbb{X}_t\boldsymbol{\Delta}_{\mathbf{t}}(\mathbf{x})$$

we show each of these terms is $o_p(1)$.

For

$$\begin{aligned}
R_{T1} &= \frac{1}{Th}\sum_{t=1}^{T}(\widehat{w}_t - w_t)\,\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&= -\frac{1}{Th}\sum_{t=1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}(\widehat{\sigma}_t^2 - \sigma_t^2)\mathbb{X}_t\mathbb{X}_t^{\intercal} + \frac{1}{Th}\sum_{t=1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4\widehat{\sigma}_t^2}(\widehat{\sigma}_t^2 - \sigma_t^2)^2\mathbb{X}_t\mathbb{X}_t^{\intercal},
\end{aligned}$$

and under Assumption A1′ we have

$$\begin{aligned}
&-\frac{1}{Th}\sum_{t=1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}(\widehat{\sigma}_t^2 - \sigma_t^2)\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&= -\left(\frac{\widehat{\omega}}{1-\widehat{\beta}} - \frac{\omega}{1-\beta}\right)\frac{1}{Th}\sum_{t=1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&- \frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\widehat{\gamma}\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\widehat{u}_{t-j}^2 - \gamma\sum_{j=1}^{\infty}\beta^{j-1}u_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&+ o_p(1)
\end{aligned}$$

It is easy to verify that, under Assumption A7′, the first term in the above expression is $O_p(T^{-1/2}) = o_p(1)$, and the second term can be decomposed into

$$\begin{aligned}
&(\widehat{\gamma} - \gamma)\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\widehat{u}_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&+ \frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}^2 - u_{t-j}^2\right)\right)\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&+ \frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\left(\widehat{\beta}^{j-1} - \beta^{j-1}\right)u_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal} \\
&- \frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=\tau+1}^{\infty}\beta^{j-1}u_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal}.
\end{aligned}$$

Again, under Assumptions A1′, A6′, and A7′, the first and the third term above are $o_p(1)$, and the second term can be written as

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}^2-u_{t-j}^2\right)\right)\mathbb{X}_t\mathbb{X}_t^\top$$

$$=\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}\left(2\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}u_{t-j}\left(\widehat{u}_{t-j}-u_{t-j}\right)\right)\mathbb{X}_t\mathbb{X}_t^\top$$

$$+\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}-u_{t-j}\right)^2\right)\mathbb{X}_t\mathbb{X}_t^\top.$$

We first consider

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}u_{t-j}\left(\widehat{u}_{t-j}-u_{t-j}\right)\mathbb{X}_t\mathbb{X}_t^\top$$

$$=\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}u_{t-j}e_1^\top M_{T,h_1}^{-1}(X_{t-j})B_{T,h_1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top$$

$$+\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}u_{t-j}e_1^\top M_{T,h_1}^{-1}(X_{t-j})U_{T,h_1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top$$

By a direct calculation of the first and second moments, we can verify that the first (bias) term is of order h_1^{p+1}, which is $o_p(1)$. The second term is asymptotically equivalent to

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4 f_X(X_{t-j})}e_1^\top M(K)^{-1}U_{T,h_1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top\sigma_{t-j}\varepsilon_{t-j}.$$

For convenience, denote the (i,j)th element of M^{-1} by $\mu^{i,j}(K)$, then the above term can be written as

$$\frac{1}{Th}\sum_{l=1}^{p+1}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4 f_X(X_{t-j})}\mu^{1,l}(K)U_{T,h_1,l-1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top\sigma_{t-j}\varepsilon_{t-j}.$$

For $l=1,\ldots,p+1$,

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4 f_X(X_{t-j})}\mu^{1,l}(K)U_{T,h_1,l-1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top\sigma_{t-j}\varepsilon_{t-j}$$

$$=\frac{1}{T^2h_1h}\sum_{t=\tau+1}^{T}\sum_{s=1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4 f_X(X_{t-j})}\mu^{1,l}(K)\left(\frac{X_{t-j}-X_s}{h_1}\right)^{l-1}K\left(\frac{X_{t-j}-X_s}{h_1}\right)\mathbb{X}_t\mathbb{X}_t^\top\sigma_s\varepsilon_s\sigma_{t-j}\varepsilon_{t-j},$$

and we consider three cases (1) $t-j=s$; (2) $t-j>s$; (3) $t-j<s$. In particular, when $t-j=s$, only when $l=1$ this term is non-zero, by a calculation of moments, it can be verified that its first moment is of order $O\left(T^{-1}h_1^{-1}\right)$, and the second moment is $O\left(T^{-2}h_1^{-2}\right)$. Thus this term is $o_p(1)$ under the bandwidth assumption. For the other cases, using the inequality of Yoshihara (1976), we can verify that the term is of order $O(T^{-3/2}h_1^{-1}h^{-1/2}+T^{-3/2}h^{1/2})=O(T^{-3/2}h_1^{-1}h^{-1/2})$. Thus we can verify that

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4 f_X(X_{t-j})}e_1^\top M(K)^{-1}U_{T,h_1}(X_{t-j})\mathbb{X}_t\mathbb{X}_t^\top\sigma_{t-j}\varepsilon_{t-j}=O_p\left(T^{-1}h_1^{-1}+T^{-1}h_1^{-1/2}h^{-1/2}\right)=o_p\left(1\right).$$

Notice that $\tau=O(\log T)$, by similar methods, one can verify

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}-u_{t-j}\right)^2\right)\mathbb{X}_t\mathbb{X}_t^\top=o_p(1),$$

Thus,

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left(\left(x-X_t\right)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=1}^{\tau}\widehat{\beta}^{j-1}\left(\widehat{u}_{t-j}^2-u_{t-j}^2\right)\right)\mathbb{X}_t\mathbb{X}_t^\top=o_p\left(1\right).$$

Next, consider

$$\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=\tau+1}^{\infty}\beta^{j-1}u_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal}.$$

Notice that $|\beta|<1$, direct calculations show that

$$\mathrm{E}\left\|\frac{1}{Th}\sum_{t=\tau+1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4}\left(\gamma\sum_{j=\tau+1}^{\infty}\beta^{j-1}u_{t-j}^2\right)\mathbb{X}_t\mathbb{X}_t^{\intercal}\right\|=O\left(|\beta|^{\tau}\right),$$

which is $o(1)$ since $\tau=c\ln T\to\infty$.

By similar analysis we can show that

$$\frac{1}{Th}\sum_{t=1}^{T}\frac{K\left((x-X_t)/h\right)}{\sigma_t^4\widehat{\sigma}_t^2}(\widehat{\sigma}_t^2-\sigma_t^2)^2\mathbb{X}_t\mathbb{X}_t^{\intercal}=o_p(1).$$

Combining the above results, we have $R_{T1}=o_p(1)$.

The analysis of R_{T2} and R_{T3} is parallel to the analysis of R_{T1}.

8.6. Proof of Theorem 4

Again, notice that

$$\sqrt{Th}\left[\widetilde{m}(x)-m(x)-h^{p+1}b(x)\right]=\sqrt{Th}\left[\widehat{m}(x)-m(x)-h^{p+1}b(x)\right]+\sqrt{Th}\left[\widetilde{m}(x)-\widehat{m}(x)\right],$$

we show that, under our assumptions,

$$\sqrt{Th}\left[\widehat{m}(x)-m(x)-h^{p+1}b(x)\right]\Longrightarrow N\left(0,\frac{1}{f_X(x)\int_0^1\sigma(r)^{-2}dr}\omega^2\right). \tag{20}$$

and

$$\sqrt{Th}\left[\widetilde{m}(x)-\widehat{m}(x)\right]=o_p(1). \tag{21}$$

For result (20) for $\widehat{\beta}$, notice that

$$\sqrt{Th}\left(\widehat{\beta}-\beta-\left[\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\right)$$
$$=\left[\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\left[\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t\right].$$

It can be verified that

$$\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\to\int_0^1\sigma(r)^{-2}drf_X(x)M(K),$$

and, by Taylor expansion,

$$\frac{1}{Th}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\approx h^{p+1}\frac{m^{(p+1)}(x)f_X(x)\int_0^1\sigma(r)^{-2}dr}{(p+1)!}B(K),$$

thus,

$$\frac{1}{h^{p+1}}\left[\sum_{t=1}^{T}w_t\mathbb{X}_t\mathbb{X}_t^{\intercal}\right]^{-1}\sum_{t=1}^{T}w_t\mathbb{X}_t\Delta_t(x)\to\frac{m^{(p+1)}(x)}{(p+1)!}M(K)^{-1}B(K),$$

For the stochastic component, notice that:

$$\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\mathbb{X}_tK\left(\frac{x-X_t}{h}\right)\sigma_t^{-1}\varepsilon_t,$$

First,

$$\frac{1}{Th}\sum_{i=1}^{T}\left[\frac{K\left((x-X_i)/h\right)}{\sigma_i^2}\right]^2\left(\frac{x-X_i}{h}\right)^{j+l}u_i^2 \to f_X(x)\left[\int_0^1\sigma(r)^{-2}dr\right]\int K(u)^2u^{j+l}du.$$

in addition, for every fixed $(p+1)$-vector λ

$$\frac{1}{Th}\sum_{i=1}^{T}\lambda^{\intercal}\mathbb{X}_t\mathbb{X}_t^{\intercal}\lambda\left[\frac{K\left((x-X_i)/h\right)}{\sigma_i^2}\right]^2u_i^2 \to f_X(x)\int_0^1\sigma(r)^{-2}dr\lambda^{\intercal}\Gamma(K)\lambda$$

notice that $\left\{\lambda^{\intercal}\mathbb{X}_tK\left(\frac{x-X_t}{h}\right)\sigma_t\varepsilon_t,\mathcal{F}_t\right\}$ is a martingale difference sequence,

$$\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\lambda^{\intercal}\mathbb{X}_t\frac{K\left((x-X_i)/h\right)}{\sigma_i^2}u_t \Longrightarrow N\left(0,f_X(x)\int_0^1\sigma(r)^{-2}dr\lambda^{\intercal}\Gamma(K)\lambda\right).$$

Thus, by the Cramér–Wold device, we have

$$\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}w_t\mathbb{X}_tu_t \to N\left(0,f_X(x)\int_0^1\sigma(r)^{-2}dr\Gamma(K)\right).$$

Thus, together with the analysis with the bias effect, we obtain

$$\sqrt{Th}\left(\widehat{\beta}-\beta-h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}M(K)^{-1}B(K)\right)\Longrightarrow N\left(0,\frac{1}{f_X(x)\cdot\int_0^1\sigma(r)^{-2}dr}M(K)^{-1}\Gamma(K)M(K)^{-1}\right),$$

and

$$\sqrt{Th}\left[\widehat{m}(x)-m(x)-h^{p+1}b(x)\right]\Longrightarrow N\left(0,\frac{1}{f_X(x)\int_0^1\sigma(r)^{-2}dr}\omega^2\right).$$

Next we prove (21). Notice that

$$\widehat{w}_t=\frac{K\left((x-X_t)/h\right)}{\widehat{\sigma}_t^2},$$

following a similar argument as the previous theorems, we only need to show the following results hold for the locally varying volatility model:

$$R_{T1}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\left(\widehat{w}_t-w_t\right)\mathbb{X}_tY_t=o_p(1),$$

$$R_{T2}=\left(\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\left(\widehat{w}_t-w_t\right)\mathbb{X}_t\mathbb{X}_t^{\intercal}\right)=o_p(1).$$

Notice that $R_{T1}=R_{T11}+R_{T12}$, where

$$R_{T11}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\left(\widehat{w}_t-w_t\right)\mathbb{X}_tm(X_t),\ R_{T12}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}\left(\widehat{w}_t-w_t\right)\mathbb{X}_tu_t.$$

We first consider R_{T12}. Let

$$W_{ts}=\frac{G\left(((s-t)/T)/h_\sigma\right)}{\sum_{i=1}^{T}G\left(((i-t)/T)/h_\sigma\right)}$$

then $R_{T12}=R_{T12A}+R_{T12B}+R_{T12C}$, where

$$R_{T12A}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}K((X_t-x_0)/h)\left[\frac{1}{\widehat{\sigma}_t^2}-\frac{1}{\widetilde{\sigma}_t^2}\right]\mathbb{X}_tu_t,$$

$$R_{T12B}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}K((X_t-x_0)/h)\left[\frac{1}{\widetilde{\sigma}_t^2}-\frac{1}{\overline{\sigma}_t^2}\right]\mathbb{X}_tu_t,$$

$$R_{T12C}=\frac{1}{\sqrt{Th}}\sum_{t=1}^{T}K((X_t-x_0)/h)\left[\frac{1}{\overline{\sigma}_t^2}-\frac{1}{\sigma_t^2}\right]\mathbb{X}_tu_t.$$

and

$$\widehat{\sigma}_t^2 = \sum_{s=1}^T W_{ts}\widehat{u}_s^2,\ \widetilde{\sigma}_t^2 = \sum_{s=1}^T W_{ts}u_s^2,\ \overline{\sigma}_t^2 = \sum_{s=1}^T W_{ts}\sigma_s^2.$$

We can show each of these terms is $o_p(1)$. For R_{T12A}, notice that, under Assumption A1″, we have

$$0 < c_L \le \min_t \overline{\sigma}_t^2 \le \min_t \widetilde{\sigma}_t^2 + \max_t \left|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2\right| = \min_t \widetilde{\sigma}_t^2 + o_p(1)$$

and

$$0 < c_L \le \min_t \widetilde{\sigma}_t^2 \le \min_t \widehat{\sigma}_t^2 + \max_t \left|\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2\right| = \min_t \widehat{\sigma}_t^2 + o_p(1)$$

In addition, $\sum_{t=1}^T \left(\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2\right)^2$ is bounded by

$$C_1 \sum_{t=1}^T \left(\sum_{s=1}^T W_{ts}\left(\widehat{u}_s - u_s\right)u_s\right)^2 + C_2 \sum_{t=1}^T \left(\sum_{s=1}^T W_{ts}\left(\widehat{u}_s - u_s\right)^2\right)^2,$$

where C_1 and C_2 are constants. It can be verified that

$$\sum_{s=1}^T W_{ts}^2 u_s^2 \le \max |W_{ts}| \sum_{s=1}^T W_{ts}u_s^2 = O\left(\frac{1}{Th_\sigma}\right).$$

Denote C to be a generic constant term, then

$$\begin{aligned}&\sum_{t=1}^T \left(\sum_{s=1}^T W_{ts}\left(\widehat{u}_s - u_s\right)u_s\right)^2 \\ &\le C\sum_{t=1}^T \left(\left(\max_t |\widehat{u}_s - u_s|\right)^2 \cdot \sum_{s=1}^T W_{ts}^2 u_s^2\right) \\ &= O_p\left(h_1^{2q}h_\sigma^{-1} + T^{-1}h_1^{-1}h_\sigma^{-1}\log(T)\right) = o_p(1).\end{aligned}$$

The other term can be analyzed similarly. Thus, $\sum_{t=1}^T \left(\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2\right)^2 = o_p(1)$.

For any $j = 0, 1, \ldots, p$,

$$\begin{aligned}&\left|\frac{1}{\sqrt{Th}}\sum_{t=1}^T \left[\frac{\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2}{\widehat{\sigma}_t^2\widetilde{\sigma}_t^2}\right] K((X_t - x_0)/h)\left(\frac{X_t - x}{h}\right)^j u_t\right| \\ &\le \left[\frac{1}{\left(\min_t \widehat{\sigma}_t^2\right)\left(\min_t \widetilde{\sigma}_t^2\right)}\right]\left[\sum_{t=1}^T \left(\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2\right)^2\right]^{1/2}\left[\frac{1}{Th}\sum_{t=1}^T K((X_t - x_0)/h)^2\left(\frac{X_t - x}{h}\right)^{2j} u_t^2\right]^{1/2} \\ &\to 0\end{aligned}$$

thus, $R_{T12A} \to 0$.

The second term R_{T12B},

$$\begin{aligned}R_{T12B} &= \frac{1}{\sqrt{Th}}\sum_{t=1}^T K((X_t - x_0)/h)\left[\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2\right]\overline{\sigma}_t^{-4}\mathbb{X}_t u_t \\ &+ \frac{1}{\sqrt{Th}}\sum_{t=1}^T K((X_t - x_0)/h)\left[\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2\right]^2 \widetilde{\sigma}_t^{-2}\overline{\sigma}_t^{-4}\mathbb{X}_t u_t\end{aligned}$$

It can be verified that both of these two terms are $o_p(1)$.

For R_{T12C},

$$R_{T12C} = -\frac{1}{\sqrt{Th}}\sum_{t=1}^T K((X_t - x_0)/h)\left[\frac{\overline{\sigma}_t^2 - \sigma_t^2}{\overline{\sigma}_t^2\sigma_t^2}\right]\mathbb{X}_t u_t,$$

notice that $\overline{\sigma}_t^2$ and σ_t^2 are deterministic functions of t, for $j = 0, \ldots, p$, $K((X_t - x_0)/h)\left[\frac{\overline{\sigma}_t^2 - \sigma_t^2}{\overline{\sigma}_t^2 \sigma_t^2}\right]\left(\frac{X_t - x}{h}\right)^j u_t$ are martingales, and

$$\mathrm{E}\left|\frac{C}{\sqrt{Th}}\sum_{t=1}^{T} K((X_t - x_0)/h)\left[\overline{\sigma}_t^2 - \sigma_t^2\right]\left(\frac{X_t - x}{h}\right)^j u_t\right|^2 = O\left(h_\sigma^2\right) \to 0,$$

thus $R_{T12C} \to 0$. Consequently, $R_{T12} = R_{T12A} + R_{T12B} + R_{T12C} \to 0$.

The proofs for R_{T11} and R_{T2} are similar.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.01.016.

References

Amemiya, Takeshi, 1983. Partially generalized least squares and two-stage least squares estimators. J. Econometrics 23.

Amemiya, Takeshi, 1985. Advanced Econometrics. Harvard university press.

Avramidis, Panagiotis, 2016. Adaptive likelihood estimator of conditional variance function. J. Nonparametr. Stat. 28 (1), 132–151. http://dx.doi.org/10.1080/10485252.2015.1122189.

Bollerslev, Tim, 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31 (3), 307–327.

Calonico, S., Cattaneo, M.D., Farrell, M.H., 2014. On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Estimation, Working Paper, University of Michigan.

Chen, Y., Wang, Q., Yao, W., 2015. Adaptive estimation for varying coefficient models. J. Multivariate Anal..

Chu, B., Jacho-Chavez, D., Linton, O., 2017. Standard errors for nonparametric regression. Econometric Rev. Forthcoming.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Statist. 25, 1–3.

Diebold, F., Nason, J.A., 1990. Nonparametric exchange rate prediction?. J. Int. Econ. (ISSN: 0022-1996) 28 (3), 315–332.

Engle, Robert F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50 (4), 987–1007.

Engle, Robert F., Rangel, Jose Gonzalo, 2008. The spline-GARCH model for low-frequency volatility and its global macroeconomic Causes. Rev. Financ. Stud. 21 (3), 1187–1222. http://dx.doi.org/10.1093/rfs/hhn004.

Fan, J., 1992. Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. 87, 998–1004.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall.

Francq, C., Zakoian, J.M., 2004. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli 10, 605–637.

Francq, C., Zakoian, J.-M., 2010. GARCH Models: Structure, Statistical Inference and Financial Applications. John Wiley & Sons, Chichester, UK.

Geller, Juliane, Neumann, Michael H., 2018. Improved local polynomial estimation in time series regression. J. Nonparametr. Stat. 30 (1), 1–27. http://dx.doi.org/10.1080/10485252.2017.1402118.

Hall, P., 1992a. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann. Statist. 20 (2), 675–694.

Hall, Peter G., Racine, Jeffrey S., 2015. Infinite order cross-validated local polynomial regression. J. Econometrics 185 (2), 510–525.

Henderson, Daniel J., Carroll, Raymond J., Li, Qi, 2008. Nonparametric estimation and testing of fixed effects panel data models. J. Econometrics 144 (1), 257–275.

Jin, S., Su, L., Xiao, Z., 2015. Adaptive nonparametric regression with conditional heteroskedasticity. Econometric Theory 31 (06), 1153–1191.

Jones, M.C., 1993. Do not weight for heteroscedasticity in nonparametric regression. Aust. N. Z. J. Statist. 35 (1), 89–92.

Linton, Oliver, Mammen, Enno, 2005. Estimating semiparametric ARCH (∞) models by kernel smoothing methods1. Econometrica 73 (3), 771–836.

Linton, Oliver, Mammen, Enno, Nielsen, Perch Jens, Keilegom, Van, Ingrid, 2011. Nonparametric regression with filtered data. Bernoulli 17 (1), 60–87. http://dx.doi.org/10.3150/10-BEJ260, https://projecteuclid.org/euclid.bj/1297173833.

Linton, O., Wang, Q., 2016. Nonparametric transformation regression with nonstationary data. Econometric Theory 32 (1), 1–29. http://dx.doi.org/10.1017/S026646661400070X.

Linton, O., Xiao, Z., 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. Econometric Theory 23, 371–413.

Liu, Jun M., Chen, Rong, Yao, Qiwei, 2010. Nonparametric transfer function models. J. Econometrics 157 (1), 151–164.

Martins-Filho, Carlos, Yao, Feng, 2009. Nonparametric regression estimation with general parametric error covariance. J. Multivariate Anal. 100 (3), 309–333.

Muller, Hans-Georg, Stadtmuller, Ulrich, 1987. Estimation of heteroscedasticity in regression analysis. Ann. Statist. 15 (2), 610–625. http://dx.doi.org/10.1214/aos/1176350364, http://projecteuclid.org/euclid.aos/1176350364.

Nelson, D., 1990. Stationarity and persistence in the GARCH(1,1) Model. Econometric Theory 6 (3), 318–334, Retrieved from http://www.jstor.org/stable/3532198.

Peng, L., Yao, Q., 2003. Least absolute deviations estimation for arch and garch models. Biometrika 90 (4), 967–975.

Robinson, P.M., 1987. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. Econometrica 55, 875–891.

Shephard, Neil, 1996. Statistical aspects of ARCH and stochastic volatility. Monogr. Stat. Appl. Probab. 65, 1–68.

Starica, C., 2003. Is GARCH (1, 1) as good a model as the Nobel prize accolades would imply? Working paper.

Su, Liangjun, Ullah, A., 2006. More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. Econometric Theory 22 (1), 98–126, Retrieved from http://www.jstor.org/stable/4093190.

Tibshirani, R., 1984. Local Likelihood Estimation Ph.D. thesis. Stanford University.

Wang, Naisyin, 2003b. Marginal nonparametric kernel regression accounting for within-subject correlation. Biometrika 90 (1), 43–52. http://dx.doi.org/10.1093/biomet/90.1.43.

Wang, Q., Yao, W., 2012. An adaptive estimation of MAVE. J. Multivariate Anal. 104 (1), 88–100.

Xiao, Z., Linton, O.B., Carroll, R.J., Mammen, E., 2003. More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. J. Amer. Statist. Assoc. 98 (464), 980–992.

Xu, Ke-Li, Phillips, Peter C.B., 2008. Adaptive estimation of autoregressive models with time-varying variances. J. Econometrics 142 (1), 265–280.

Yao, W., 2013. A note on EM algorithm for mixture models. Statist. Probab. Lett. 83 (2), 519–526.

Yoshihara, Ken-ichi, 1976. Limiting behavior of u-statistics for stationary, absolutely regular processes. Probab. Theory Related Fields 35 (3), 237–252.