ADAPTIVE ESTIMATION OF AUTOREGRESSIVE MODELS WITH TIME-VARYING VARIANCES

By

Ke-Li Xu and Peter C.B. Phillips

October 2006
Revised November 27, 2007

COWLES FOUNDATION DISCUSSION PAPER NO. 1585R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Adaptive Estimation of Autoregressive Models with Time-Varying Variances

Ke-Li Xu*and Peter C. B. Phillips†

Yale University

November 27, 2006

Abstract

Stable autoregressive models of known finite order are considered with martingale differences errors scaled by an unknown nonparametric time-varying function generating heterogeneity. An important special case involves structural change in the error variance, but in most practical cases the pattern of variance change over time is unknown and may involve shifts at unknown discrete points in time, continuous evolution or combinations of the two. This paper develops kernel-based estimators of the residual variances and associated adaptive least squares (ALS) estimators of the autoregressive coefficients. These are shown to be asymptotically efficient, having the same limit distribution as the infeasible generalized least squares (GLS). Comparisons of the efficient procedure and ordinary least squares (OLS) reveal that least squares can be extremely inefficient in some cases while nearly optimal in others. Simulations show that, when least squares work well, the adaptive estimators perform comparably well, whereas when least squares work poorly, major efficiency gains are achieved by the new estimators.

Keywords: Adaptive estimation, autoregression, heterogeneity, weighted regression.

JEL classification: C14, C22

*Department of Applied Mathematics, Yale University, 51 Prospect Street, New Haven, Connecticut USA 06520. E-mail address: keli.xu@yale.edu.

†*Corresponding author*. Department of Economics, Cowles Foundation for Research in Economics, Yale University, P. O. Box 208281, New Haven, Connecticut USA 06520-8281. Telephone: +1-203-432-3695. Fax: +1-203-432-6167. E-mail address: peter.phillips@yale.edu.

1 Introduction

The failure of the assumption of homogenous innovations in many time series models has been well documented in the macroeconomics and empirical finance literatures. Ignoring this problem leads to inefficient estimation and unreliable inference on the conditional mean function. To account for conditional heteroskedasticity, it is common practice to assume that the innovations follow some parametric ARCH or GARCH models based on those proposed by Engle (1982) and Bollerslev (1986). Efficient estimation of the mean function in this case is achieved by quasi-maximum likelihood based or other adaptive procedures, and recent developments on this topic have been surveyed by Li, Ling and McAleer (2002).

Although the GARCH-type model is successful in capturing many important features in macroeconomic or financial time series such as volatility clustering and persistent autocorrelation, a crucial weakness is its non-robustness to the stationarity assumption. In typical GARCH-type models, the time-varying volatility is exclusively attributed to the conditional variance or covariance structure, while the unconditional variance is assumed to be constant over time. When this condition fails, ARCH or GARCH-based approaches may lead to serious model mis-specification. For instance, artifical IGARCH effects may be observed due to nonstationary changes in the unconditional volatility (Diebold, 1986, Mikosch and Stărică, 2004). This problem is particularly relevant in view of the strong evidence against constancy of unconditional second moments shown in the empirical literatures, e.g., in time series of exchange rates, interest rates, GDP and other macroeconomic variables (*inter alia*, Loretan and Phillips, 1994, Watson, 1999, McConnell and Perez Quiros, 2000, van Dijk *et al*, 2002). Recently, more complicated GARCH-type models have been proposed to allow for unconditional heteroskedasticity, e.g. varying coefficients GARCH models (Polzehl and Spokoiny, 2006) and spline GARCH models (Engle and Rangel, 2004).

An alternative approach to modeling time-varying volatility is to use a smooth deterministic nonparametric framework, assuming that the unconditional variance is the main time-changing feature to be captured (see, e.g. Hsu, Miller and Wichern, 1974, Officer, 1976, Merton, 1980, and French, Schwert and Stambaugh, 1987). Compared to stochastic heteroskedasticity modeling like GARCH-type models, this deterministic framework is technically easier to handle and allows for nonstationarity. Recently, Drees and Stărică (2002) and Stărică (2003) used a deterministic nonstationary framework to analyze time series of S&P 500 returns, and found that this approach outperforms the GARCH-type models in both fitting the data and forecasting the next-day volatility. However, in the typical setting of this framework, the volatility is specified as a smooth function of time thereby ruling out important practical features like structural breaks in the underlying series. Meanwhile, there are other contributions focusing particularly on modeling structural changes in volatility. For instance, Wichern, Miller and Hsu (1976) investigated the AR(1) model when there are a finite number of step changes at unknown time points in the error variance. These authors used iterative maximum likelihood methods to locate the change points and then estimated the error variances in each block by averaging the squared least squares residuals. The resulting feasible weighted least squares estimator was shown to be efficient for the specific model considered. Alternative methods to detect step changes in the variances of time series models have been studied by Abraham and Wei (1984), Baufays and Rasson (1985), Tsay (1988), Park, Lee and Jeon (2000), Lee and Park (2001), de Pooter and van Dijk (2004) and Galeano and Peña (2004).[1]

However, in practice the pattern of variance changes over time, which may be discrete or continuous, is unknown to the econometrician and it seems desirable to use methods that can adapt for a wide range of possibilities. Accordingly, this paper combines two strands of the literatures mentioned above by providing a general framework to modeling nonparametric deterministic volatility in a stable linear AR(p) model, and develops an efficient estimation procedure that adapts for the presence of different and unknown forms of variance dynamics. Specifically, the model errors are assumed to be martingale differences multiplied by a time-varying scale factor which is a continuous or discontinuous function of time, thereby permitting a spectrum of variance dynamics that include step changes and smooth transitions.

Efficient estimation of linear models under heteroskedasticity with *iid* predictors was earlier investigated by Carroll (1982) and Robinson (1987), and more recently by Kitamura, Tripathi and Ahn (2004) using empirical likelihood methods in a general conditional moment restriction setting. In the time series context, Kuersteiner (2002) developed efficient instrumental variables estimators for autoregressive models under conditional heteroskedasticity but assuming constancy of the unconditional variances over time. Harvey and Robinson (1988) focused on a regression model with deterministically trending regressors only, whose error is an AR(p) process scaled by

[1] Related literature also includes testing and estimation of structural change points of the *mean* function in parametric (Bai, 1994, Bai and Perron, 1998 and references therein) and nonparametric (Yin, 1988, Muller, 1992, Wu and Chu, 1993, Delgado and Hidalgo, 2000) frameworks.

a continuous function of time, thereby allowing for both serial correlation and nonstationarity but ruling out jump behavior in the innovations. In a closely related paper, Hansen (1995) considered the linear regression model, nesting autoregressive models as special cases, when the conditional variance of the model error is a function of a covariate that has the form of a nearly integrated stochastic process with no deterministic drift. Using a kernel-weighted technique similar to ours, he also obtained the adaptive estimation results. There are some important differences between Hansen's paper and ours. The first is model formulation. Instead of focusing on stochastic trends in volatility as in Hansen (1995), we consider deterministic trends in volatility allowing particularly for single or multiple abrupt structural breaks. By doing so, a different scale parameter is employed to obtain sensible limit theory. Second, in constructing the adaptive least squares estimator, we consider two-sided kernel estimates of the residual variances, which are more accurate than Hansen's one-sided kernel estimates when variances are discontinuous over time. For this reason his proof of adaptiveness can not be extended here. Third, we allow for multiple covariates in the mean function by studying pth order autoregressive processes. Fourth, we analyze how specific nonstationary variance patterns, such as shifts and monotone trends in variance, affect the inefficiency of the OLS estimator relative to the GLS estimator. Finally, we also mention that regression models in which the conditional variance of the error is an unscaled function of an integrated time series were recently investigated by Chung and Park (2006) using Brownian local time limit methods developed in Park and Phillips (1999, 2001).

The remainder of the paper proceeds as follows. Section 2 describes the autoregressive model with general nonstationary deterministic volatility. Several assumptions are introduced and discussed. A limit theory is developed in Section 3 for a class of weighted least squares estimators, including efficient (infeasible) generalized least squares (GLS). A range of examples show that OLS can be extremely inefficient asymptotically in some cases while nearly optimal in others. Section 4 proposes a kernel-based estimator of the residual variance and shows the associated adaptive least squares estimator to be asymptotically efficient, in the sense of having the same limit distribution as the infeasible GLS estimator. Simulation experiments are conducted in Section 5 to assess the finite sample performance of the adaptive estimator. Section 6 concludes. Proofs of the main results are collected in two appendices.

2 The Model and Assumptions

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_t\}$ a sequence of increasing $\sigma-$fields of $\mathcal{F}$. Suppose the sample $\{Y_{-p+1}, \cdots, Y_0, Y_1, \cdots, Y_T\}$ from the following data generating process for the time series Y_t is observed

$$A(L)Y_t = u_t \tag{1}$$

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

where L is the lag operator, $A(L) = 1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p$, $\beta_p \neq 0$, is assumed to have all roots outside the unit circle and the lag order p is finite and known. We assume $\{\sigma_t\}$ is a deterministic sequence and $\{\varepsilon_t\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, where $\mathcal{F}_t = \sigma(\varepsilon_s, s \leq t)$ is the $\sigma-$field generated by $\{\varepsilon_s, s \leq t\}$, with unit conditional variance, i.e. $\mathbb{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) = 1$, *a.s.*, for all t. The conditional variance of $\{u_t\}$ is characterized fully by the multiplicative factor σ_t, i.e. $\mathbb{E}(u_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$, *a.s.*. This paper focuses on unconditional heteroskedasticity and σ_t^2 is assumed to be modeled as a general deterministic function, which rules out conditional dependence of σ_t on the past events of Y_t. The autoregressive coefficient vector $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ is the parameter of interest. Ordinary least squares (OLS) estimation gives $\widehat{\beta} = \left(\sum_{t=1}^T X_{t-1}X_{t-1}'\right)^{-1}\left(\sum_{t=1}^T X_{t-1}Y_t\right)$, where $X_{t-1} = (Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p})'$. Throughout the rest of the paper we impose the following conditions.

Assumption

(i). The variance term $\sigma_t = g\left(\frac{t}{T}\right)$, *where* $g(\cdot)$ *is a measurable and strictly positive function on the interval* $[0,1]$ *such that* $0 < C_1 < \inf_{r\in[0,1]} g(r) \leq \sup_{r\in[0,1]} g(r) < C_2 < \infty$ *for some positive numbers* C_1 *and* C_2, and $g(r)$ *satisfies a Lipschitz condition except at a finite number of points of discontinuity;*

(ii). $\{\varepsilon_t\}$ *is strong mixing (*α*-mixing) and* $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, $\mathbb{E}(\varepsilon_t^2|\mathcal{F}_{t-1}) = 1$, *a.s., for all* t.

(iii). There exist $\mu > 1$ *and* $C > 0$, *such that* $\sup_t \mathbb{E}|\varepsilon_t|^{4\mu} < C < \infty$.

Under Assumption (i) the function g is integrable on the interval $[0,1]$ to any finite order. For brevity, we write $\int_0^1 g^m(r)dr$ as $\int g^m$ for any finite positive integer m. Formally, of course, the assumption induces a triangular array structure to the processes u_t and Y_t, but we dispense with the additional affix T in the arguments that follow. Assumption (ii) stipulates $\{\varepsilon_t\}$ is a martingale difference (m.d.) sequence and therefore uncorrelated, but may be dependent via higher moments.

In contrast to modeling σ_t in a setting with finitely many parameters, Assumption (i) is nonparametric and σ_t depends only on the relative position of the error in the sample. It allows for a wide range of nonstationary variance dynamics including single or multiple step changes and smooth transitions (e.g. trending or periodic variances. See Examples 1 and 2 below). Assumption (i) excludes the dependence of $E(u_t^2|F_{t-1})$ on past events. A more flexible formulation is to assume σ_t as a function of scaled (T^{-1}) integrated time series with a time trend (see the discussion in the next paragraph).

Our model of nonstationary volatility is related to that of Hansen (1995). In his paper, the volatility process is specified as a function of a first-order nearly integrated process, viz. $E(u_t^2|F_{t-1}) = g^2(c_1 + c_2 S_t/\sqrt{T})$, where $S_t = (1 - c_3/T)S_{t-1} + z_t$ with martingale differences z_t and constants c_i, $i = 1, 2, 3$. Without accounting for structural breaks explicitly, his model focuses on stochastic volatility, which asymptotically reduces to ours in Assumption (i) by a simple extension. To illustrate, suppose a time trend (or drift) $c_4 t$ is added to the nearly unit root process S_t. Since a stochastic trend is dominated by a deterministic trend in the long run at least for a scalar process, Hansen's model in this case is no longer applicable and the normalization factor needs to be adjusted to $1/T$ rather than $1/\sqrt{T}$, as in Hansen's formulation, to achieve a non-degenerate asymptotic theory.

Combining (1) with (2) is particularly useful in accounting for nonstationary volatility that may be present in macroeconomic and financial data. Watson (1999) and McConnell and Perez Quiros (2000) found evidence of monotone trending behavior in variability (corresponding to a monotone version of the function $g(\cdot)$ in Assumption (i)) for US short and long term interest rates and GDP series over specified periods. The volatility structure in (2) was also used by Stărică, Herzel and Nord (2005) in the analysis of the dynamics of stock indexes - see also Stărică and Granger (2005).

We conclude this section by mentioning that much attention has recently been paid to potential structural error variance changes in integrated process models. The effects of step breaks in the innovation variance on unit root tests and stationarity tests were studied by Hamori and Tokihisa (1997), Kim, Leybourne and Newbold (2002), Busetti and Taylor (2003) and Cavaliere (2004a).

A general framework to analyze the effect of time varying variances on unit root tests was given in Cavaliere (2004b) and Cavaliere and Taylor (2004). By contrast, little work of this general nature (as in Assumption (i), which is attributed to Cavaliere, 2004) has been done on autoregressions with coefficients satisfying the stable condition, most of the attention in the literature being concerned with the case of step changes *or* smooth transitions in the error variance, as discussed above. The present paper therefore contributes by focusing on efficient estimation of the AR(p) model with time varying variances of a general form that includes step changes as a special case.

3 Limit Theory

Under the stated assumptions, the process Y_t has the following representation

$$Y_t = \sum_{i=0}^{\infty} \alpha_i u_{t-i}, \tag{3}$$

where the coefficients $\{\alpha_i\}$ satisfy

$$\sum_{i=0}^{\infty} |\alpha_i| < \infty. \tag{4}$$

Under Assumptions (i)-(iii), $\widehat{\beta}$ is asymptotically normal with limit distribution (Phillips and Xu, 2006a): [2]

$$\sqrt{T}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \Lambda\right), \tag{5}$$

where

$$\Lambda = \frac{\int g^4}{(\int g^2)^2} \Gamma^{-1},$$

[2] In a more general framework allowing for both stochastic and deterministic nonstationary volatility, this limit distribution assumes a general form involving stochastic integrals (Xu, 2006, see also Hansen, 1995).

and Γ is the $p \times p$ positive definite matrix with the (i,j)-th element $\gamma_{|i-j|}$, and $\gamma_k = \sum_{i=0}^{\infty} \alpha_i \alpha_{i+k} < \infty$, for $0 \le k \le p-1$. The matrix Γ^{-1} can be consistently estimated by

$$\widehat{\Gamma}^{-1} = \left(\widehat{\gamma}_{|i-j|}\right)_{i,j}^{-1}, \tag{6}$$

where $\widehat{\gamma}_0, \widehat{\gamma}_1, \cdots, \widehat{\gamma}_{p-1}$ are the first p elements in the first column of the $(p^2 \times p^2)$ matrix $[I_{p^2} - F \otimes F]^{-1}$, where $\otimes$ indicates the Kronecker product and

$$F = \begin{pmatrix} \widehat{\beta}_1 & \widehat{\beta}_2 & \cdots & \widehat{\beta}_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix}.$$

Result (5) is a consequence of the following more general theorem.

Theorem 1 *Suppose ω_t^2 is non-stochastic and satisfies (i) $0 < \omega_t^2 < C < \infty$ for all t and some finite positive number $C > 0$; (ii) there exists a function $\omega(\cdot)$ on $[0,1]$, continuous except for a finite number of discontinuities, such that $\omega_{[Tr]}^2 \to \omega^2(r)$ for any $r \in [0,1]$ at which $\omega(\cdot)$ is continuous ; (iii) $\int \omega^2 > 0$. Then, under Assumption (i)-(iii), the weighted least squares (WLS) estimator*

$$\widehat{\beta}_{WLS} = \left(\sum_{t=1}^{T} \omega_t^2 X_{t-1} X_{t-1}'\right)^{-1} \left(\sum_{t=1}^{T} \omega_t^2 X_{t-1} Y_t\right) \tag{7}$$

satisfies

$$\sqrt{T}(\widehat{\beta}_{WLS} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\int \omega^4 g^4}{(\int \omega^2 g^2)^2} \Gamma^{-1}\right), \tag{8}$$

as $T \to \infty$.

Naturally, the estimator with the smallest asymptotic variance matrix in the class (7) is achieved by generalized least squares (GLS)

$$\beta^* = \left(\sum_{t=1}^{T} X_{t-1} X_{t-1}' \sigma_t^{-2}\right)^{-1} \left(\sum_{t=1}^{T} X_{t-1} Y_t \sigma_t^{-2}\right), \tag{9}$$

with weights $\omega_t^2 = \sigma_t^{-2}$(The optimality of β^* can also be justified by the theory of unbiased linear estimating equations, as in Godambe, 1960 and Durbin, 1960) in which case

$$\sqrt{T}(\beta^* - \beta) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}), \tag{10}$$

as $T \to \infty$.

Remarks. Clearly, the asymptotic variance matrix of $\widehat{\beta}$ differs from that of β^* by the factor $\int g^4/(\int g^2)^2$, and since Γ^{-1} is invariant to the function $g(\cdot)$ the inefficiency of the OLS estimator $\widehat{\beta}$ depends crucially on this factor. The following examples[3] show that the factor can be large and OLS can be very inefficient in some cases, whereas in others, the factor is close to unity and OLS is close to optimal.

Example 1 (*A single abrupt shift in the innovation variance*) Let $\tau \in [0,1]$ and $g(r)$ be the step function

$$g(r)^2 = \sigma_0^2 + (\sigma_1^2 - \sigma_0^2)\mathbf{1}_{\{r \geq \tau\}}, r \in [0,1],$$

giving error variance σ_0^2 before the break point $[T\tau]$, and σ_1^2 afterwards. The steepness of the variance shift is measured by the ratio $\delta := \sigma_1/\sigma_0$ of the post-break and pre-break standard deviation. By (5) the asymptotic variance matrix of OLS is

$$\Lambda = \frac{\tau + (1-\tau)\delta^4}{(\tau + (1-\tau)\delta^2)^2}\Gamma^{-1} := f_1^2(\tau, \delta)\Gamma^{-1},$$

where $f_1^2(\tau,\delta) = \left(\tau + (1-\tau)\delta^2\right)^{-2}\left(\tau + (1-\tau)\delta^4\right)$, which is a function of the break date τ and the shift magnitude δ.

Figure 1 plots the value of $f_1(\tau,\delta)$ across $\delta \in [0.01, 100]$ for different values of τ. The variance of the OLS estimator largely depends on where the break in the innovation variance occurs. For the negative ($\delta < 1$) shift, $f_1(\tau,\delta)$ increases steeply as δ decreases when $\tau = 0.1$, and is relatively steady and nearly unity when $\tau = 0.9$. The graph shows that OLS has large variance when the

[3] We follow the formulation of the variance function in Cavaliere (2004) (Section 5, page 271-283), who investigates heteroskedastic unit root testing.

break occurs at the beginning ($\tau = 0.1$) but much smaller variance, and in fact close to that of infeasible GLS, when the break is at the end ($\tau = 0.9$) of the sample. This difference is explained by the fact that when the break in variance occurs early in the sample, the large innovation variance in the early part of the sample affects all later observations via the autoregressive mechanism. By contrast, when the break occurs near the end of the sample, only later observations are directly affected, so the impact of a negative shift is small. This argument applies when there is a negative shift - a shift to a smaller variance at the end of the sample - and a reverse argument applies in the case of a positive shift.

In fact, under a positive ($\delta > 1$) shift, OLS has large variance when the shift occurs late ($\tau = 0.9$) but small variance and more closely approximates infeasible GLS when it is early ($\tau = 0.1$) in the sample. These phenomena are confirmed in the simulation experiment of Gaussian AR(1) case, reported in Section 5.

Example 2 (*Trending variances in the innovations*) Let m be a positive integer and $g(r)$ be

$$g(r)^2 = \sigma_0^2 + (\sigma_1^2 - \sigma_0^2)r^m, r \in [0,1],$$

giving error variance changing from σ_0^2 to σ_1^2 continuously according to an m-th order power function. Then

$$\Lambda = \frac{1 + 2(\delta^2-1)/(m+1) + (\delta^2-1)^2/(2m+1)}{[1 + (\delta^2-1)/(m+1)]^2}\Gamma^{-1} := f_2^2(m,\delta)\Gamma^{-1},$$

where $f_2^2(m,\delta) = \left(1 + \frac{\delta^2-1}{m+1}\right)^{-2}\left(1 + \frac{2(\delta^2-1)}{m+1} + \frac{(\delta^2-1)^2}{2m+1}\right)$ and $\delta = \sigma_1/\sigma_0$.

Figure 2 plots the value of $f_2(m,\delta)$ across $\delta \in [0.01, 100]$ for different values of m, so that both positive ($\delta > 1$) and negative ($\delta < 1$) trending heteroskedasticity is allowed. Compared with the case of a single abrupt shift in the innovation variance (Example 1), the multiplicative factor $f_2(m,\delta)$ changes more steadily for a given value of m, especially when m is small (say, $m = 1$). In the case of large m (say, $m = 6$), much inefficiency in OLS is sustained when there is positive trending heteroskedasticity ($\delta > 1$).

4 Adaptive Estimation

The GLS estimator β^* in (9) is infeasible, since the true values of σ_t are unknown. To produce a feasible procedure, we propose a kernel-based estimator $\widetilde{\beta}$ employing nonparametric estimates of the residual variances and having the same asymptotic distribution as β^*. This entails a preliminary estimate of σ_t^2, denoted by $\widehat{\sigma}_t^2$, which we motivate as follows. Model (1) can be rewritten in the form $(A(L)Y_t)^2 = g^2(t/T) + \epsilon_t$, where $\epsilon_t = \sigma_t^2(\varepsilon_t^2 - 1)$ satisfies $E\epsilon_t = 0$. Let $K(z)$ be a continuous kernel function defined on the real line such that $0 \leq \sup_z K(z) < C$ for some finite real number C and $\int_{-\infty}^{\infty} K(z)dz = 1$. Applying standard kernel-based nonparametric techniques, $g^2(t/T)$ can be estimated by $\sum_{i=1}^{T} w_{ti}(A(L)Y_t)^2$, where

$$w_{ti} = \left(\sum_{i=1}^{T} K_{ti}\right)^{-1} K_{ti},$$

with

$$K_{ti} = \begin{cases} K\left(\frac{t-i}{Tb}\right), \text{ if } t \neq i \\ 0, \text{ if } t = i \end{cases}. \tag{11}$$

Here b is a bandwidth parameter, dependent on T. Since the true value of $A(L)Y_t$ is unknown, the squared OLS residuals $\widehat{u}_t^2 = (Y_t - X_{t-1}'\widehat{\beta})^2$ may be used to define the variance estimator and define $\widehat{\sigma}_t^2$ as the weighted sum

$$\widehat{\sigma}_t^2 = \sum_{i=1}^{T} w_{ti}\widehat{u}_i^2. \tag{12}$$

While (12) is based on the Nadaraya-Watson (or local constant) method, a variety of nonparametric procedures like local polynomial fitting (Fan and Gijbels, 1996, Fan and Yao, 1998) or empirical likelihood re-weighted methods (Phillips and Xu, 2006b) may be used instead. For technical reasons in (12), we use the leave-one-out procedure and omit the observation $\widehat{u}_t^2$. Now we are able to define the adaptive least squares (ALS) estimator of β as

$$\widetilde{\beta} = \left(\sum_{t=1}^{T} X_{t-1}X_{t-1}'\widehat{\sigma}_t^{-2}\right)^{-1} \left(\sum_{t=1}^{T} X_{t-1}Y_t\widehat{\sigma}_t^{-2}\right). \tag{13}$$

The implementation of the estimator $\widehat{\sigma}_t^2$ depends on the choice of kernel function K and the bandwidth b. Commonly used kernels such as the uniform, Epanechnikov, biweight and Gaussian functions can be applied. Bandwidth selection is more crucial. As usual, too small bandwidth produces less bias for the true residual variance but has higher variability. A simple data driven method to choose the parameter b is cross-validation on the average squared error – see Wong (1983). The cross-validatory choice of b is the value b^* which minimizes

$$\widehat{CV}(b) = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{u}_t^2 - \widehat{\sigma}_t^2\right)^2.$$

We use the following assumptions that modify and extend the earlier assumptions to facilitate the development of an asymptotic theory for $\widetilde{\beta}$.

Assumption

(iii'). There exists some finite positive number C such that $\sup_t \mathbb{E}(\varepsilon_t^8) < C < \infty$;

(iv). As $T \to \infty$, $b + \frac{1}{Tb^2} \to 0$.

We replace Assumption (iii) by the stronger assumption (iii'), which requires the existence of eighth moments of ε_t for all t. This moment condition simplifies the proof of the main theorem and is, no doubt, stronger than necessary. Assumption (iv) is a rate condition that requires $b \to 0$ at a slower rate than $T^{-1/2}$.

The main result is as follows.

Theorem 2 *Let $g^2(r-) = \lim_{\bar{r}\uparrow r} g^2(\bar{r})$ and $g^2(r+) = \lim_{\bar{r}\downarrow r} g^2(\bar{r})$. Under Assumptions (i)-(iv) with (iii') instead of (iii), as $T \to \infty$,*

$$\widehat{\sigma}_{[Tr]}^2 \xrightarrow{p} g^2(r-)\int_{-\infty}^{0} K(z)dz + g^2(r+)\int_{0}^{\infty} K(z)dz, \tag{14}$$

and

$$\sqrt{T}(\widetilde{\beta} - \beta) = \sqrt{T}(\beta^* - \beta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1}), \tag{15}$$

where Γ^{-1} *is estimated by (6).*

Result (14) shows that $\widehat{\sigma}^2_{[Tr]}$ converges in probability to $g^2(r)$ at the point $r \in [0,1]$ when the function g is continuous, but in general to a point between $g^2(r-)$ and $g^2(r+)$ depending on the shape of the kernel. The inconsistency of the error variance function estimator at points of discontinuities has a diminishing effect on the behavior of adaptive estimators of the autoregressive coefficients when the sample size is large, as is clear from (15). A one-sided kernel estimator of the residual variance at time t, as proposed by Hansen (1995), can be also constructed by using information up to time $t-1$. But this estimator has larger bias in small samples at discontinuous points since it always converges in probability to $g^2(r-)$, although the difference on adaptive estimation diminishes as the sample size increases.

Another adaptive estimator is suggested by Harvey and Robinson (1988), who dealt with time series regression in the presence of trending regressors. Rather than estimating each σ_t^2 separately, they split the data into K blocks and estimated σ_t^2 in one block by the average of $\widehat{u}_t^2$ in this block. So only K distinct estimators are used. It can be shown under the regularity assumptions, the resulting weighted least squares estimator of β also has the same asymptotic distribution as $\widetilde{\beta}$ if $\frac{1}{T_1} + \frac{T}{T_1^2} + \frac{T_2}{T} \to 0$, as $T \to \infty$, where T_1 and T_2 are the minimum and maximum lengths of the K blocks. Compared to our estimator, this estimator is faster to compute but it does not integrate in an efficient way the information of $\widehat{u}_s^2$ where s is close to t when estimating σ_t^2, especially when t is close to the boundary of the block.

5 Simulations

This section examines the finite sample performance of the ALS efficient procedure proposed in Section 4 using simulations of the heteroskedastic AR(1) model

$$Y_t = \beta Y_{t-1} + u_t, \quad u_t = \sigma_t \varepsilon_t,$$

where $\sigma_t = g\left(\frac{t}{T}\right)$. We use $\beta \in \{0.1, 0.9\}$, and $\varepsilon_t \sim iid\mathcal{N}(0,1)$.

Our simulation design basically follows Cavaliere (2004) and Cavaliere and Taylor (2004). The

g function generating heteroskedasticity is taken as the step function used in Examples 1, viz.,

$$g(r)^2 = \sigma_0^2 + (\sigma_1^2 - \sigma_0^2)\mathbf{1}_{\{r \geq \tau\}}, r \in [0, 1].$$

The break date is chosen from $\{0.1, 0.9\}$ and the ratio of post-break and pre-break standard deviations $\delta = \sigma_1/\sigma_0$ is set to the values $\{0.2, 5\}$. Without loss of generality, we let $\sigma_0 = 1$. The estimates of β are obtained with sample size $T = 50$ and $T = 200$, and the number of replications is set to 10,000. Other models (say the trending variance in Example 2) are also considered in our experiments, although not reported here, and they yield the results similar to those obtained below.

We report estimates for β obtained by OLS, infeasible GLS and ALS. For the ALS estimator (13), we use the Gaussian kernel function, $K(z) = (2\pi)^{-1/2}\exp(-z^2/2)$, for $-\infty < z < \infty$. When a different kernel (such as Epanechnikov kernel) is used, the results do not change much. Five bandwidths are considered, *i.e.*, four fixed bandwidths $h_i = c_i T^{-0.4}$, $i = 1, \cdots, 4$, where $\{c_1, c_2, c_3, c_4\} = \{0.25, 0.4, 0.6, 0.75\}$ as well as a data-driven bandwidth chosen by the cross-validation (CV) procedure described in Section 4.

Table 1 reports the ratios of the root mean squared errors (RMSE) of estimators considered relative to the RMSE of GLS. The levels (rather than the ratios) of RMSE are reported for GLS in brackets. Clearly, OLS is inefficient and the ALS estimator works reasonably well in all cases considered. The largest inefficiency in OLS is observed when an early shift in the innovation variance is negative, for instance, $(\tau, \delta) = (0.1, 0.2)$, and when a late shift is positive, for instance, $(\tau, \delta) = (0.9, 5)$. The former is explained by the fact that the large variance early in the sample affects all later observations and the latter is explained by the fact that the large variance in the last part of the sample means that the OLS estimator is more closely approximated by the terms involving the last few observations, thereby effectively reducing the sample size. In both these cases, substantial efficiency gains are achieved by the ALS estimator. In contrast, when there is a positive early shift or a negative late shift in the innovation variance, for instance, $(\tau, \delta) = (0.1, 5)$ or $(0.9, 0.2)$, OLS works nearly as well as GLS, especially when the sample size is large. The ALS estimator performs comparably well with OLS in those cases. When the sample size is increased from $T = 50$ to $T = 200$, the ALS estimators have the smaller ratio of RSME, while no improvement (or even larger inefficiency) is observed for OLS.

We also note that the cross-validation procedure to choose the bandwidth of the ALS estimator works satisfactorily. Sometime the ALS estimator with the cross-validated bandwidth is outperformed by certain specified fixed bandwidth in certain cases (in most case by h_2), but is not uniformly dominated by a single fixed bandwidth from the four we considered. In practice we recommend using the cross-validated bandwidth or the fixed bandwidth h_2.

Simulations results, along with those not reported here, also show that, in both models the improvement of the ALS procedure relative to OLS is insensitive to the location of the true value of the autoregressive parameter β, as long as $|\beta| < 1$.

We also check the homoskedastic case when $\delta = 1$ and show results in Table 1. OLS is equivalent to GLS when the errors are homoskedastic, so the ratio of RMSE of OLS relative to GLS is unity. We observe that in this case the the ALS estimator is also close to one, so that ALS may be used satisfactorily even when the errors are homoskedastic.

Furthermore, to check the robustness of our ALS procedure to skewed or heavy-tailed error distributions, we let ε_t be subject to a $\chi^2(5)$ or a $t(5)$ distribution each with degree of freedom five, normalized so that it has zero mean and unit variance. Apparently when $\varepsilon_t \sim t(5)$, the technical assumption *(iii')* is violated. This model is incorporated to illustrate that the conclusion of Theorem 2 extends to more general error distributions. The corresponding results are reported only for the case of a positive late shift (*i.e.* $\tau = 0.9$, $\delta = 5$) in Table 2. Again, we can see that major efficiency gains are achieved by the ALS estimator compared to the OLS procedure. Just as the cases with Gaussian errors we consider above, ALS is almost as efficient as the infeasible GLS estimator when T is increased from 50 to 200.

In summary, our kernel-based ALS estimator and cross-validation procedure both appear to perform reasonablly well, at least within the simulation design considered. The advantages are clear - they are convenient for practical use and have uniformly good performance over the parameter space.

6 Further Remarks

This paper considers efficient estimation of finite order autoregressive models under unconditional heteroskedasticity of unknown form. Several extensions of the approach taken in the paper are possible. One of these is to consider efficient estimation of unconditionally heteroskedastic stable autoregressions of possible infinite order. The issue here is whether the nonparametric feasible GLS estimator considered here is still asymptotically efficient when the order of autoregression, p, increases with the sample size, T. We leave this and other extensions for future research.

7 Appendix A: Proofs of the Theorems.

This section gives the proofs of Theorem 1 and Theorem 2. In what follows, C is a generic positive constant. We use $|\cdot|$ to denote the Euclidean norm $|X| = (X_1^2 + \cdots + X_n^2)^{1/2}$ for $X = (X_1, \cdots, X_n)'$, and $||\cdot||_K$ to denote the L^K-norm, so that $||\xi||_K = (E|\xi|^K)^{1/K}$ for a random vector ξ.

The Proof of the Theorem 1. The WLS estimator $\widehat{\beta}_{WLS}$ satisfies

$$\sqrt{T}(\widehat{\beta}_{WLS} - \beta) = \left(\frac{1}{T}\sum_{t=1}^{T}\omega_t^2 X_{t-1}X_{t-1}'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\omega_t^2 X_{t-1}u_t\right). \tag{16}$$

It is easy to show that under Assumption (i)-(iii), $\{\omega_t^2 Y_{t-h}Y_{t-h-k} - \omega_t^2\mathbb{E}(Y_{t-h}Y_{t-h-k})\}$ is mean-zero L^1-NED (near-epoch dependent) on $\{\varepsilon_t\}$ (see e.g. Theorem 17.9 in Davidson, 1994) for $1 \le h \le p$, $0 \le k \le p-h$, and therefore a L^1-mixingale with respect to $\mathcal{F}_t$. It is uniform integrable by applying Lemma A (a) with $\mu = 2$. By the law of large numbers for L^1-mixingales (Andrews, 1988) we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(\omega_t^2 Y_{t-h}Y_{t-h-k} - \omega_t^2\mathbb{E}(Y_{t-h}Y_{t-h-k})\right) \xrightarrow{p} 0. \tag{17}$$

Lemma A(ii) of Phillips and Xu (2006a) shows that for every continuous point r of $g(\cdot)$, $\lim_{T\to\infty}\mathbb{E}Y_{[Tr]-h}\cdot Y_{[Tr]-h-k} = g^2(r)\gamma_k$, where $[\cdot]$ refers to the integer part. Let $r_1 < r_2 <, \cdots, < r_Q$ be the discontinuous points of $g(\cdot)$ and $w(\cdot)$, where Q is a finite number (independent of T). So by (17), for sufficiently large T, $T^{-1}\sum_{t=1}^{T}\omega_t^2 Y_{t-h}Y_{t-h-k} = T^{-1}\sum_{t=1}^{T}\omega_t^2\mathbb{E}(Y_{t-h}Y_{t-h-k}) + o_p(1) = \sum_{t=1}^{T}\int_{\frac{t}{T}}^{\frac{t+1}{T}}\omega_{[Tr]}^2\mathbb{E}Y_{[Tr]-h}Y_{[Tr]-h-k}dr + o_p(1) = \int_{\frac{1}{T}}^{r_1}\omega_{[Tr]}^2\mathbb{E}Y_{[Tr]-h}Y_{[Tr]-h-k}dr + \sum_{j=1}^{Q-1}\int_{r_j}^{r_{j+1}}\omega_{[Tr]}^2\cdot \mathbb{E}Y_{[Tr]-h}Y_{[Tr]-h-k}dr + \int_{r_Q}^{\frac{T+1}{T}}\omega_{[Tr]}^2\mathbb{E}Y_{[Tr]-h}Y_{[Tr]-h-k}dr + o_p(1) \xrightarrow{p} (\int\omega^2 g^2)\gamma_k$. So we have $T^{-1}\sum_{t=1}^{T}X_{t-1}\cdot X_{t-1}'\sigma_t^{-2} \xrightarrow{p} (\int\omega^2 g^2)\Gamma$. Next we show that $T^{-1}\sum_{t=1}^{T}\omega_t^4\cdot X_{t-1}X_{t-1}'u_t^2 \xrightarrow{p} (\int\omega^4 g^4)\Gamma$, which holds if $T^{-1}\sum_{t=1}^{T}\omega_t^4 Y_{t-h}Y_{t-h-k}u_t^2 \xrightarrow{p} \gamma_k$ for $1 \le h \le p$, $0 \le k \le p-h$. Indeed, since $\{\omega_t^4 Y_{t-h}Y_{t-h-k}u_t^2 -$

$\omega_t^4\sigma_t^2\mathbb{E}Y_{t-h}Y_{t-h-k},\mathcal{F}_t\}$ are martingale differences, so $T^{-1}\sum_{t=1}^T\omega_t^4Y_{t-h}Y_{t-h-k}u_t^2 = T^{-1}\sum_{t=1}^T\omega_t^4\cdot$ $\sigma_t^2\mathbb{E}Y_{t-h}Y_{t-h-k}+o_p(1)\xrightarrow{p}(\int\omega^4g^4)\gamma_k$ by similar arguments used above. Furthermore, $\mathbb{E}|\omega_t^2X_{t-1}u_t|^4$ $<\infty$ by Lemma A (b) with $\mu=2$. By the central limit theorem for vector martingale differences, $T^{-1/2}\sum_{t=1}^T\omega_t^2X_{t-1}u_t\xrightarrow{d}\mathcal{N}\left(0,(\int\omega^4g^4)\Gamma\right)$. Then Theorem 1 follows from (16).

The Proof of the Theorem 2. First we prove (14). Recall that $\widehat{u}_i$'s are the OLS residuals. Let $\overline{\sigma}_t^2=\sum\limits_{i=1}^T w_{ti}\sigma_i^2$, and it is easy to see that

$$\left|\left(\tfrac{1}{Tb}\sum_{i=1}^T K_{ti}\right)\left(\widehat{\sigma}_t^2-\overline{\sigma}_t^2\right)\right|\le\left|\tfrac{1}{Tb}\sum_{i=1}^T K_{ti}\left(u_i^2-\sigma_i^2\right)\right|+o_p(1)=o_p(1). \tag{18}$$

Actually, if we let $a_i=u_i^2-\sigma_i^2$, then $\{a_i\}$ is an m.d. sequence and $\mathbb{E}(\frac{1}{Tb}\sum_{i=1}^T K_{ti}a_i)^2=\frac{1}{(Tb)^2}\sum_{i=1}^T K_{ti}^2Ea_i^2\le\frac{1}{Tb}\left(\sup_i K_{ti}\right)\left(\sup_i Ea_i^2\right)\left(\frac{1}{Tb}\sum_{i=1}^T K_{ti}\right)=O(\frac{1}{Tb})\to 0$, in view of Lemma A (c). On the other hand, we have

$$\begin{aligned}\frac{1}{Tb}\sum_{i=1}^T K_{[Tr]i}\sigma_i^2 &= \frac{1}{b}\int_{1/T}^{(T+1)/T}K(\frac{[Ts]-[Tr]}{Tb})g^2(\frac{[Ts]}{T})ds+o(1)\\ &\overset{z=(s-r)/b}{=}\int_{(1-T)/Tb}^{(T+1-Tr)/Tb}K(\frac{[T(r+bz)]-[Tr]}{Tb})g^2(\frac{[T(r+bz)]}{T})dz+o(1)\\ &\to g^2(r-)\int_{-\infty}^0K(z)dz+g^2(r+)\int_0^\infty K(z)dz. \end{aligned}\tag{19}$$

Combining (18) and (19) gives $\widehat{\sigma}_{[Tr]}^2=\left(\frac{1}{Tb}\sum_{i=1}^T K_{[Tr]i}\right)\overline{\sigma}_{[Tr]}^2+o_p(1)=\frac{1}{Tb}\sum_{i=1}^T K_{[Tr]i}\sigma_i^2+o_p(1)\xrightarrow{p}g^2(r-)\int_{-\infty}^0K(z)dz+g^2(r+)\int_0^\infty K(z)dz$ as claimed.

Now we prove (15). We follow closely the proof of the theorem in Robinson (1987) using some of his notation. First, note that $\widetilde{\beta}$ satisfies

$$\sqrt{T}(\widetilde{\beta}-\beta)=\left(\tfrac{1}{T}\sum_{t=1}^T X_{t-1}X_{t-1}'\widehat{\sigma}_t^{-2}\right)^{-1}\left(\tfrac{1}{\sqrt{T}}\sum_{t=1}^T X_{t-1}u_t\widehat{\sigma}_t^{-2}\right).$$

Define $a(f)=\frac{1}{\sqrt{T}}\sum_{t=1}^T X_{t-1}u_tf_t^{-2}$ and $A(f)=\frac{1}{T}\sum_{t=1}^T X_{t-1}X_{t-1}'f_t^{-2}$, then we have $\sqrt{T}(\beta^*-\beta)=A(\sigma)^{-1}a(\sigma)$ and $\sqrt{T}(\widetilde{\beta}-\beta)=A(\widehat{\sigma})^{-1}a(\widehat{\sigma})=A(\sigma)^{-1}a(\sigma)+A(\widehat{\sigma})^{-1}(a(\widehat{\sigma})-a(\sigma))-A(\sigma)^{-1}(A(\widehat{\sigma})-A(\sigma))A(\widehat{\sigma})^{-1}a(\sigma)$. We have $A(\sigma)\xrightarrow{p}\Gamma$ which is positive definite, and $a(\sigma)=O_p(1)$, which follows from Markov's inequality and $\mathbb{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_{t-h}u_t\sigma_t^{-2}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^{-4}\mathbb{E}Y_{t-h}^2u_t^2 \leq C\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}Y_{t-h}^2u_t^2 < \infty$, by Lemma A (b). Hence (15) follows if we prove

$$A(\widehat{\sigma}) - A(\sigma) \xrightarrow{p} 0,\ a(\widehat{\sigma}) - a(\sigma) \xrightarrow{p} 0. \tag{20}$$

Define $\widetilde{\sigma}_t^2 = \sum_{i=1}^{T} w_{ti}u_i^2$ and $\overline{\sigma}_t^2 = \sum_{i=1}^{T} w_{ti}\sigma_i^2$, and (20) follows from the following six results as in Robinson (1987): (a) $a(\widehat{\sigma}) - a(\widetilde{\sigma}) \xrightarrow{p} 0$; (b) $a(\widetilde{\sigma}) - a(\overline{\sigma}) \xrightarrow{p} 0$; (c) $a(\overline{\sigma}) - a(\sigma) \to_p 0$; (d) $A(\widehat{\sigma}) - A(\widetilde{\sigma}) \xrightarrow{p} 0$; (e) $A(\widetilde{\sigma}) - A(\overline{\sigma}) \xrightarrow{p} 0$; (f) $A(\overline{\sigma}) - A(\sigma) \xrightarrow{p} 0$. These will be shown as follows:

(a) Since $a(\widehat{\sigma}) - a(\widetilde{\sigma}) = \frac{1}{\sqrt{T}}\sum_t X_{t-1}u_t\frac{\widetilde{\sigma}_t^2-\widehat{\sigma}_t^2}{\widehat{\sigma}_t^2\widetilde{\sigma}_t^2}$, we have $|a(\widehat{\sigma}) - a(\widetilde{\sigma})| \leq (\min_t \widetilde{\sigma}_t^2)^{-1}(\min_t \widehat{\sigma}_t^2)^{-1} \cdot \sum_{t=1}^{T}\frac{|X_{t-1}u_t|}{\sqrt{T}}|\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2| \leq (\min_t \widetilde{\sigma}_t^2)^{-1}(\min_t \widehat{\sigma}_t^2)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}|X_{t-1}u_t|^2\right)^{1/2}\left(\sum_{t=1}^{T}|\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2|^2\right)^{1/2} = O_p(\frac{1}{Tb}) \xrightarrow{p} 0$, by Lemma A (b, h, j, k).

(b) We write

$$\begin{aligned} a(\widetilde{\sigma}) - a(\overline{\sigma}) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{t-1}u_t(\widetilde{\sigma}_t^{-2} - \overline{\sigma}_t^{-2}) \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2\widetilde{\sigma}_t^{-2}\overline{\sigma}_t^{-4}, \end{aligned} \tag{21}$$

which holds since for two any nonzero real numbers p and q we have the following equality $p^{-1} - q^{-1} = (q-p)q^{-2} + (q-p)^2p^{-1}q^{-2}$. We will show the two terms of (21) vanishes in probability. For the first term, we note that $\{X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}, \mathcal{F}_t\}$ is an m. d. sequence. Indeed, we have

$$\mathbb{E}(X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}|\mathcal{F}_{t-1})$$

$$= \overline{\sigma}_t^{-2}\mathbb{E}(X_{t-1}u_t|\mathcal{F}_{t-1}) - \overline{\sigma}_t^{-4}\mathbb{E}(X_{t-1}u_t\sum_{i<t}w_{ti}u_i^2|\mathcal{F}_{t-1}) - \overline{\sigma}_t^{-4}\mathbb{E}(X_{t-1}u_t\sum_{i>t}w_{ti}u_i^2|\mathcal{F}_{t-1}). \tag{22}$$

Both the last two terms are zero, since for the term $i > t$, $\mathbb{E}(X_{t-1}u_t u_i^2|\mathcal{F}_{t-1}) = X_{t-1}\mathbb{E}(u_t u_i^2|\mathcal{F}_{t-1}) = X_{t-1}\mathbb{E}(u_t\mathbb{E}(u_i^2|\mathcal{F}_{i-1})|\mathcal{F}_{t-1}) = X_{t-1}\mathbb{E}(u_t|\mathcal{F}_{t-1}) = 0$, and for the term $i < t$, $\mathbb{E}(X_{t-1}u_t u_i^2|\mathcal{F}_{t-1}) = X_{t-1}u_i^2 \cdot \mathbb{E}(u_t|\mathcal{F}_{t-1}) = 0$. Thus, by (22) $\mathbb{E}(X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}|\mathcal{F}_{t-1}) = 0$. So the first term of (21) converges to zero in probability by the Markov inequality and $\mathbb{E}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}\right|^2 \leq \frac{C}{T}\sum_{t=1}^T \mathbb{E}|X_{t-1}u_t|^2(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2 \leq \frac{C}{T}\sum_{t=1}^T(\mathbb{E}|X_{t-1}u_t|^4)^{1/2}\cdot(\mathbb{E}(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^4)^{1/2} \leq (\max_t \mathbb{E}(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^4)^{1/2}\cdot \frac{C}{T}\sum_{t=1}^T(\mathbb{E}|X_{t-1}u_t|^4)^{1/2} = O_p(\frac{1}{Tb}) \xrightarrow{p} 0$, by Lemma A (a, f). For the second term of (21), $\left|\sum_{t=1}^T \frac{X_{t-1}u_t}{\sqrt{T}}(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2\widetilde{\sigma}_t^{-2}\overline{\sigma}_t^{-4}\right| \leq C\left(\frac{1}{T}\sum_{t=1}^T |X_{t-1}u_t|^2\right)^{1/2}\left(\sum_{t=1}^T(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^4\right)^{1/2} = O_p(\frac{1}{T^{1/2}b})$ $\xrightarrow{p} 0$, by Lemma A (a, f). This completes the proof of (b).

(c) First we note

$$\sigma_t^2\left(\overline{\sigma}_t^{-2} - \sigma_t^{-2}\right)^2 \leq \overline{\sigma}_t^{-4}\sigma_t^{-2}\left|\overline{\sigma}_t^2 + \sigma_t^2\right| \cdot \left|\overline{\sigma}_t^2 - \sigma_t^2\right| \leq C\left|\overline{\sigma}_t^2 - \sigma_t^2\right|. \tag{23}$$

Since $\{X_{t-1}u_t\}$ is an m.d. sequence, we get $\mathbb{E}|a(\overline{\sigma}) - a(\sigma)|^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}(|X_{t-1}|^2 u_t^2)(\overline{\sigma}_t^{-2} - \sigma_t^{-2})^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}(|X_{t-1}|^2\mathbb{E}(u_t^2|\mathcal{F}_{t-1}))(\overline{\sigma}_t^{-2} - \sigma_t^{-2})^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}|X_{t-1}|^2\sigma_t^2|\overline{\sigma}_t^{-2} - \sigma_t^{-2}|^2 \leq \frac{C}{T}\sum_{t=1}^T \mathbb{E}|X_{t-1}|^2\cdot |\overline{\sigma}_t^2 - \sigma_t^2| \leq C\max_t \mathbb{E}|X_{t-1}|^2 \cdot \frac{1}{T}\sum_{t=1}^T |\overline{\sigma}_t^2 - \sigma_t^2| = o_p(1)$, by Lemma A (a, l).

(d) It follows from $|A(\widehat{\sigma}) - A(\widetilde{\sigma})| \leq (\min_t \widetilde{\sigma}_t^2)^{-1}(\min_t \widehat{\sigma}_t^2)^{-1}\frac{1}{T}\sum_{t=1}^T |X_{t-1}|^2|\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2| \leq C \cdot \max_t |\widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2| \cdot \frac{1}{T}\sum_{t=1}^T |X_{t-1}|^2 = O_p(\frac{1}{\sqrt{Tb}})$, by Lemma A (a, h, i, j).

(e) This can be proved in the same way as (d) by employing Lemma A (g).

(f) It follows from $|A(\overline{\sigma}) - A(\sigma)| \leq (\min_t \overline{\sigma}_t^2)^{-1}(\min_t \sigma_t^2)^{-1}\frac{1}{T}\sum_{t=1}^T |X_{t-1}|^2|\overline{\sigma}_t^2 - \sigma_t^2| \leq (\min_t \overline{\sigma}_t^2)^{-1}\cdot (\min_t \sigma_t^2)^{-1} \cdot \max_t |X_{t-1}|^2 \cdot \frac{1}{T}\sum_{t=1}^T |\overline{\sigma}_t^2 - \sigma_t^2| = o_p(1)$,

by Lemma A (a, e, l).

8 Appendix B: Supplementary Results and Proofs.

This section states and proves some results (Lemma A) used in the proofs of the theorems.

Lemma A (a) If $\sup_{1\leq t\leq T} \mathbb{E}|\varepsilon_t|^{2\mu} < \infty$, $1 \leq \mu < \infty$, then $\sup_{1\leq t\leq T} \mathbb{E}|Y_{t-h}|^{2\mu} < \infty$ holds for

$1 \le h \le p$;

(b) If $\sup\limits_{1\le t\le T} \mathbb{E}|\varepsilon_t|^{4\mu} < \infty$, $1 \le \mu < \infty$, then $\sup\limits_{1\le t\le T} \mathbb{E}|Y_{t-h}u_t|^{2\mu} < \infty$ holds for $1 \le h \le p$;

(c) Let $t = [Tr]$ for any fixed $r \in (0,1]$, then $\frac{1}{Tb}\sum_{i=1}^T K_{ti} \to \int_{-\infty}^{\infty} K(z)dz = 1$, where K_{ti} is defined in (11);

(d) $\max\limits_{t,i} w_{ti} = O(\frac{1}{Tb})$;

(e) $\min\limits_{1\le t\le T} \overline{\sigma}_t^2 \ge C > 0$;

(f) $\max\limits_{1\le t\le T} \mathbb{E}|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^4 = O\left(\frac{1}{(Tb)^2}\right)$;

(g) $\max\limits_{t} |\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^\delta = O_p(T^{-\delta/4}b^{-\delta/2})$, for $\delta = 1, 2$;

(h) $(\min\limits_{1\le t\le T} \widetilde{\sigma}_t^2)^{-1} = O_p(1)$, as $T \to \infty$;

(i) $\max\limits_{1\le t\le T} \left|\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2\right| = O_p(\frac{1}{\sqrt{Tb}})$;

(j) $(\min\limits_{1\le t\le T} \widehat{\sigma}_t^2)^{-1} = O_p(1)$, as $T \to \infty$;

(k) $\sum_{t=1}^T (\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2 = O_p(\frac{1}{(Tb)^2})$;

(l) $\frac{1}{T}\sum\limits_{t=1}^T |\overline{\sigma}_t^2 - \sigma_t^2| = o(1)$.

The Proof of Lemma A. (a) Note that $Y_{t-h}^2 = \sum_{k=0}^\infty \sum_{l=0}^\infty \alpha_k\alpha_l u_{t-h-k}u_{t-h-l}$ and $\mathbb{E}|u_{t-h-k}\cdot u_{t-h-l}|^\mu \le (\mathbb{E}|u_{t-h-k}|^{2\mu}\mathbb{E}|u_{t-h-l}|^{2\mu})^{1/2} < \infty$. So we have $\mathbb{E}|Y_{t-h}|^{2\mu} = ||Y_{t-h}^2||_\mu^\mu \le (\sum_{k=0}^\infty \sum_{l=0}^\infty |\alpha_k\alpha_l|\cdot ||u_{t-h-k}u_{t-h-l}||_\mu)^\mu \le C(\sum_{k=0}^\infty \sum_{l=0}^\infty |\alpha_k\alpha_l|)^\mu = C(\sum_{k=0}^\infty |\alpha_k|)^{2\mu} < \infty$.

(b) Since $Y_{t-h}^2 u_t^2 = \sum_{k=0}^\infty \sum_{l=0}^\infty \alpha_k\alpha_l u_{t-h-k}u_{t-h-l}u_t^2$ and $\mathbb{E}|u_{t-h-k}u_{t-h-l}u_t^2|^\mu \le (\mathbb{E}|u_{t-h-k}|^{4\mu}\cdot \mathbb{E}|u_{t-h-l}|^{4\mu})^{1/4}\cdot(\mathbb{E}|u_t|^{4\mu})^{1/2} < \infty$, so $\mathbb{E}|Y_{t-h}u_t|^{2\mu} = ||Y_{t-h}^2u_t^2||_\mu^\mu \le \left(\sum_{k=0}^\infty \sum_{l=0}^\infty |\alpha_k\alpha_l|\cdot ||u_{t-h-k}u_{t-h-l}u_t^2||_\mu\right)^\mu \le C\cdot\left(\sum_{k=0}^\infty \sum_{l=0}^\infty |\alpha_k\alpha_l|\right)^\mu < \infty$.

(c) Let $t - i = [Tx]$, where x is a real number, $|x| < 1$. Then $\frac{1}{Tb}\sum_{i=1}^T K_{ti} = \frac{1}{Tb}\sum_{i=1}^T K(\frac{t-i}{Tb}) + o(1) = \sum_{i=1}^T \int_{(t-i)/T}^{(t-i+1)/T} K(\frac{[Tx]}{Tb})d\left(\frac{x}{b}\right) + o(1) \overset{z=x/b}{=} \sum_{i=1}^T \int_{(t-i)/Tb}^{(t-i+1)/Tb} K(\frac{[Tbz]}{Tb})dz + o(1)$
$= \int_{(t-T)/Tb}^{t/Tb} K(\frac{[Tbz]}{Tb})dz + o(1) \to \int_{-\infty}^\infty K(z)dz = 1.$

(d) It follows from $w_{ti} = \left(\frac{1}{Tb}\sum_{i=1}^T K_{ti}\right)^{-1}\frac{K_{ti}}{Tb}$ and (c).

(e) It follows from $\min\limits_{1\le t\le T} \overline{\sigma}_t^2 \ge \min\limits_{1\le i\le T} \sigma_i^2\cdot(\sum_{i=1}^T w_{ti}) \ge \inf\limits_{s\in[0,1]} g^2(s) \ge C > 0$.

(f) We make use of the Burkholder's inequality (BI) (c.f. Shiryaev (1995), p499): for the m.d. sequence $\xi_1, \cdots, \xi_T$ and $p > 1$, there exists constant A_p and B_p, such that

$$A_p \left\|(\textstyle\sum_{t=1}^T \xi_t^2)^{1/2}\right\|_p \le \left\|\textstyle\sum_{t=1}^T \xi_t\right\|_p \le B_p \left\|(\textstyle\sum_{t=1}^T \xi_t^2)^{1/2}\right\|_p.$$

Let $a_i = u_i^2 - \sigma_i^2$, then a_i is a m.d. sequence and $\mathbb{E}a_i^4 < \infty$. Then $\mathbb{E}(\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2)^4 = \mathbb{E}\left(\sum_{i=1}^T w_{ti}a_i\right)^4$ $\overset{BI(p=4)}{\leq} \mathbb{E}\left(\sum_{i=1}^T w_{ti}^2 a_i^2\right)^2 \overset{(d)}{\leq} \frac{1}{(Tb)^2}\mathbb{E}\left(\sum_{i=1}^T w_{ti} a_i^2\right)^2 \overset{Jensen}{\leq} \frac{1}{(Tb)^2}\sum_{i=1}^T w_{ti}\mathbb{E}a_i^4 = O\left(\frac{1}{(Tb)^2}\right)$, where the last inequality is by Jensen's $f(\sum_{i=1}^T w_{ti}a_i^2) \leq \sum_{i=1}^T w_{ti} f(a_i^2)$ with convex function $f(x) = x^2$.

(g) It holds since for arbitrary $C > 0$, $\mathbb{P}(\max_t |\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^\delta > CT^{-\delta/4}b^{-\delta/2}) \leq \sum_{t=1}^T \mathbb{P}(|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^\delta > CT^{-\delta/4}b^{-\delta/2})$

$$\overset{Markov}{\leq} \begin{cases} C^{-4}Tb^2\sum\limits_{t=1}^T \mathbb{E}|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^4 \overset{(f)}{=} O(C^{-4}), \ \delta = 1; \\ C^{-2}Tb^2\sum\limits_{t=1}^T \mathbb{E}|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2|^4 \overset{(f)}{=} O(C^{-2}), \ \delta = 2. \end{cases}$$

(h) It follows from $0 < C \overset{(e)}{\leq} \min\limits_{1\leq t\leq T} \overline{\sigma}_t^2 \leq \min\limits_{1\leq t\leq T} \widetilde{\sigma}_t^2 + \max\limits_t |\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2| = \min\limits_{1\leq t\leq T} \widetilde{\sigma}_t^2 + o_p(1)$.

(i) Note that $\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 = \sum_{i=1}^T w_{ti}(\widehat{u}_i^2 - u_i^2) = \sum_{i=1}^T w_{ti}\left((\widehat{\beta}-\beta)'X_{i-1}X_{i-1}'(\widehat{\beta}-\beta) - 2u_iX_{i-1}'(\widehat{\beta}-\beta)\right)$, and $\max\limits_{t,i}\sum_{i=1}^T w_{ti}^2 \leq \max\limits_{t,i} w_{ti} \cdot \sum_{i=1}^T w_{ti} = O(\frac{1}{Tb})$. We also have $\widehat{\beta} - \beta = O(T^{-1/2})$ by (5). Thus $\max\limits_{1\leq t\leq T}|\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2| \leq \max\limits_{1\leq t\leq T}\sum_{i=1}^T w_{ti}|(\widehat{\beta}-\beta)'X_{i-1}X_{i-1}'(\widehat{\beta}-\beta) - 2u_iX_{i-1}'(\widehat{\beta}-\beta)|$
$\leq \max\limits_{1\leq t\leq T}\sum_{i=1}^T w_{ti}|\widehat{\beta}-\beta|^2|X_{i-1}|^2 + 2\max\limits_{1\leq t\leq T}\sum_{i=1}^T w_{ti}|u_iX_{i-1}'|\cdot|\widehat{\beta}-\beta|$
$\leq \max\limits_{t,i} w_{ti}\cdot|\widehat{\beta}-\beta|^2\sum_{i=1}^T|X_{i-1}|^2 + 2|\widehat{\beta}-\beta|\cdot(\max\limits_{t,i}\sum_{i=1}^T w_{ti}^2)^{1/2}\cdot\left(\sum_{i=1}^T|u_iX_{i-1}'|\right)^{1/2}$
$= O_p(\frac{1}{Tb}) + O_p(\frac{1}{\sqrt{Tb}}) = O_p(\frac{1}{\sqrt{Tb}})$.

(j) It follows from $0 < C \overset{(h)}{\leq} \min\limits_{1\leq t\leq T}\widetilde{\sigma}_t^2 \leq \min\limits_{1\leq t\leq T}\widehat{\sigma}_t^2 + \max\limits_t|\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2|| = \min\limits_{1\leq t\leq T}\widehat{\sigma}_t^2 + o_p(1)$.

(k) Since $\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 = \sum_{i=1}^T w_{ti}(\widehat{u}_i^2 - u_i^2) = (\widehat{\beta}-\beta)'(\sum_{i=1}^T w_{ti}^2X_{i-1}X_{i-1}')(\widehat{\beta}-\beta) - 2(\sum_{i=1}^T w_{ti}^2u_iX_{i-1}')\cdot(\widehat{\beta}-\beta)$, then $\sum_{t=1}^T(\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2$ is bounded by

$$\begin{aligned} &\sum_{t=1}^T C\left(|\widehat{\beta}-\beta|^4\left|\sum_{i=1}^T w_{ti}^2X_{i-1}X_{i-1}'\right|^2 + \left|\sum_{i=1}^T w_{ti}^2u_iX_{i-1}'\right|^2|\widehat{\beta}-\beta|^2\right) \\ \leq \quad & |\widehat{\beta}-\beta|^4\sum_{t=1}^T C\left(\sum_{i=1}^T w_{ti}^2|X_{i-1}|^2\right)^2 + |\widehat{\beta}-\beta|^2\sum_{t=1}^T C\left(\sum_{i=1}^T w_{ti}^2|u_iX_{i-1}'|\right)^2 \end{aligned} \tag{24}$$

The first term of (24) is bounded by

$$|\widehat{\beta}-\beta|^4\sum_{t=1}^{T}C\left(\sup_i|X_{i-1}|^2\cdot\max_{t,i}w_{ti}\cdot\sum_{i=1}^{T}w_{ti}\right)^2=O_p(\frac{1}{T^3b^2}),$$

by (a) and (d), and similarly the second term of (24) is $O_p(\frac{1}{T^2b^2})$. So (k) follows.

(l) Let $r_1 < r_2 <, \cdots, < r_D$ be the discontinuous points of $g(\cdot)$, where D is finite. Then for sufficiently large T,

$\frac{1}{T}\sum_{t=1}^{T}|\overline{\sigma}_t^2-\sigma_t^2| = \sum_{t=1}^{T}\int_{t/T}^{(t+1)/T}|\overline{\sigma}_{[nr]}^2-\sigma_{[nr]}^2|dr = \int_{1/T}^{r_1}|\overline{\sigma}_{[nr]}^2-\sigma_{[nr]}^2|dr+\sum_{j=1}^{D-1}\int_{r_j}^{r_{j+1}}|\overline{\sigma}_{[nr]}^2-\sigma_{[nr]}^2|dr+\int_{r_D}^{(T+1)/T}|\overline{\sigma}_{[nr]}^2-\sigma_{[nr]}^2|dr \to 0$, provided that

$$\overline{\sigma}_{[nr]}^2 \to g^2(r) \tag{25}$$

when g is continuous at r. Indeed, following the proof of (c) we can similarly have $\frac{1}{Tb}\sum_{i=1}^{T}K_{ti}\sigma_i^2 \to g^2(r)$ when g is continuous at r. Thus (25) holds by (c).

9 Acknowledgement

Phillips gratefully acknowledges partial support from the Kelly Foundation and the NSF under Grant No. SES 04-142254. The authors would like to thank the Editor, Peter M. Robinson, the Associate Editor and the two anonymous referees for their detailed comments on the original draft. The authors would also like to thank Donald W. K. Andrews and other participants at the Econometrics Research Seminar at Yale University, participants at the Singapore Econometric Study Group (SESG) at Singapore Management University and the International Symposium on Econometric Theory and Application (SETA) at Xiamen University for their useful remarks and suggestions.

References

[1] Abaraham, B., Wei, W., 1984. Inference about the parameters of a time series model with changing variance. Metrika 31, 183-194.

[2] Andrews, D. W. K., 1988. Laws of large numbers for dependent non-identically distributed random variables. Econometric Theory 4, 458-467.

[3] Bai, J., 1994. Least squares estimation of a shift in linear processes. Journal of Time Series Analysis 15, 453– 472.

[4] Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66, 47–78.

[5] Baufays, P., Rasson, J. P., 1985. Variance changes in autoregressive models. Time Series Analysis: Theory and Practice. 2nd Ed. Springer, New York.

[6] Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

[7] Busetti, F., Taylor, A. M. R., 2003. Variance shifts, structural breaks, and stationarity tests. Journal of Business and Economic Statistics 21(4), 510-31.

[8] Carroll, R. J., 1982. Adapting for heteroskedasticity in linear models. Annals of Statistics 10, 1224-1233.

[9] Cavaliere, G., 2004a. Testing stationarity under a permanent variance shift. Economics Letters 82, 403-408.

[10] Cavaliere, G., 2004b. Unit root tests under time-varying variance shifts. Econometric Reviews 23, 259-292.

[11] Cavaliere, G., Taylor, A. M. R., 2004. Testing for unit roots in time series models with non-stationary volatility. Working paper, University of Birmingham.

[12] Chu, J.S., Wu, C.K., 1993. Kernel-type estimators of jump points and values of a regression function. Annals of Statistics 21, 1545-1566.

[13] Chung, H., Park, J. Y., 2006. Nonstationary nonlinear heteroskedasticity in regression. forthcoming in Journal of Econometrics.

[14] Davidson, J., 1994. Stochastic limit theory: an introduction for econometricians. Oxford University Press.

[15] Delgado, M. A., Hidalgo, J., 2000. Nonparametric inference on structural breaks. Journal of Econometrics 96, 113–144.

[16] de Pooter, M., van Dijk, D., 2004. Testing for changes in volatility in heteroskedastic time series - a further examination. Erasmus University Rotterdam, Econometric Institute Report EI 2004-38.

[17] Diebold, F. X., 1986. Modeling the persistence of the conditional variances: a comment. Econometric Reviews 5, 51-56.

[18] Drees, H., Stărică, C., 2002. A simple non-stationary model for stock returns. working paper, Chalmers University of Technology.

[19] Durbin, J., 1960. Esitmation of parameters in time series regression models. Journal of the Royal Statistical Society, Series A 22, 139-153.

[20] Engle, R. F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. Econometrica 50, 987–1008.

[21] Engle, R. F., Rangel, J. G., 2004. The spline GARCH model for unconditional volatility and its global macroeconomic causes. Working paper, New York University and University of California, San Diego.

[22] Fan, J., Gijbels, I., 1996. Local Polynomial Modeling and its Applications. London: Chapman and Hall.

[23] Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645-660.

[24] French, K., Schwert, W., Stambaugh, R., 1987. Expected stock returns and volatility. Journal of Financial Economics 19, 3–29.

[25] Galeano, P., Peña, D., 2004. Variance changes detection in multivariate time series. Working paper 04-13, Statistics and Econometrics Series 05, Universidad Carlos III de Madrid, Spain.

[26] Godambe, V. P., 1960. An optimum property of regular maximum likelihood equation. Annals of Mathematical Statistics 31, 1208-1211.

[27] Hansen, B. E., 1995. Regression with nonstationary volatility. Econometrica 63, 1113-1132.

[28] Hamori, S., Tokihisa, A., 1997. Testing for a unit root in the presence of a variance shift. Economics Letters 57, 245-253.

[29] Harvey, A. C., Robinson, P. M., 1988. Efficient estimation of nonstationary time series regression. Journal of Time Series Analysis 9, 201-214.

[30] Hsu, D. A., Miller, R., Wichern, D., 1974. On the stable Paretian behavior of stock-market prices. Journal of American Statistical Association 69, 108–113.

[31] Kim, T. H., Leybourne, S., Newbold, P., 2002. Unit root tests with a break in innovation variance. Journal of Econometrics 109, 365-387.

[32] Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical likelihood-based inference in conditional moment restriction models. Econometrica 72, 1667-1714.

[33] Kuersteiner, G. M., 2002. Efficient IV estimation for autoregressive models with conditional heteroskedasticity. Econometric Theory 18 (3), 547-583.

[34] Lee, S., Park, S., 2001. The cusum of squares test for scale changes in infinite order moving average models. Scandinavian Journal of Statistics 28(4), 625-644.

[35] Li, W. K., Ling, S., McAleer, M, 2002. Recent theoretical results for time series models with GARCH errors. Journal of Economic Surveys 16. 245-269.

[36] Loretan, M., Phillips, P. C. B., 1994. Testing covariance stationarity under moment condition failure with an application to common stock returns. Journal of Empirical Finance 1, 211–248.

[37] McConnell, M. M., Perez Quiros, G., 2000. Output fluctuations in the United States: what has changed since the early 1980s? American Economic Review 90, 1464–1476.

[38] Merton, R., 1980. On estimatingt he expected return on the market: an exploratory investigation. Journal of Financial Economics 8, 323-361.

[39] Mikosch, T., Stărică, C., 2004. Non-stationarities in financial time series, the long range dependence and the IGARCH effects. Review of Economics and Statistics 86, 378-390.

[40] Müller, H.-G., 1992. Change-points in nonparametric regression analysis. Annals of Statistics 20, 737-761.

[41] Officer, R., 1976. The variability of the market factor of the New York Stock Exchange. Journal of Business 46, 434–453.

[42] Park, J., Phillips, P. C. B., 1999. Asymptotics for nonlinear transformations of integrated time series. Econometric Theory 15, 269-298.

[43] Park, J., Phillips, P. C. B., 2001. Nonlinear regression with integrated time series. Econometrica 69, 117-161.

[44] Park, S., Lee, S., Jeon J., 2000. The cusum of squares test for variance changes in infinite order autoregressive models. Journal of the Korean Statistical Society 29, 351-361.

[45] Phillips, P. C. B., Xu, K.-L., 2006a. Inference in autoregression under heteroskedasticity. Journal of Time Series Analysis, 27, 289-308.

[46] Phillips, P. C. B., Xu, K.-L., 2006b. Tilted nonparametric estimation of volatility functions. working paper, Yale University, 2006.

[47] Polzehl, J., Spokoiny, V. 2006. Varying coefficient GARCH versus local constant volatility modeling: comparison of predictive power. working paper, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany.

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

[49] Robinson, P. M., 1989. Nonparametric estimation of time-varying parameters. In Statistical Analysis and Forecasting of Economic Structural Change (P. Hackl ed.). Amsterdam: North-Holland, 253-264.

[50] Robinson, P. M., 1991. Time-varying nonlinear regression. In Economic Structural Change (P. Hackl and A. H. Westlund eds.). Berlin: Springer-Verlag, 179-190.

[51] Shiryaev, A. N., 1995. Probability. New York: Springer-Verlag.

[52] Stărică, C. 2003. Is GARCH (1,1) as good a model as the Nobel prize accolades would imply? working paper, Chalmers University of Technology.

[53] Stărică, C. Herzel, S., Nord, T. 2005. Why does the GARCH(1,1) model fail to provide sensible longer-horizon volatility forecasts? working paper, Chalmers University of Technology.

[54] Stărică, C., Granger, C., 2005. Non-stationarities in stock returns. Review of Economics and Statistics 87, 503-522.

[55] Tsay, R. S., 1988. Outliers, level shifts and variance changes in time series. Journal of Forecasting 7, 1-20.

[56] van Dijk, D., Osborn, D. R., Sensier, M., 2002. Changes in variability of the business cycle in the G7 countries. Erasmus University Rotterdam, Econometric Institute Report EI 2002-28.

[57] Watson, M. W. (1999). Explaining the increased variability in long-term interest rates. Federal Reserve Bank Richmond Economic Quarterly 85:71–96.

[58] Wichern, D., Miller, R., Hsu, D., 1976. Changes of variance in first order autoregression time series models - with an application. Applied Statistics 25, 248-256.

[59] Wong, W. H., 1983. On the consistency of cross validation in kernel nonparametric regression. Annals of Statistics 11, 1136–1141.

[60] Xu, K. -L., 2006. Bootstrapping autoregression under nonstationary volatility. working paper, Yale University.

[61] Yin, Y.Q., 1988. Detection of the number, locations and magnitudes of jumps. Communications in Statistics: Stochastic Models 4, 445-455.

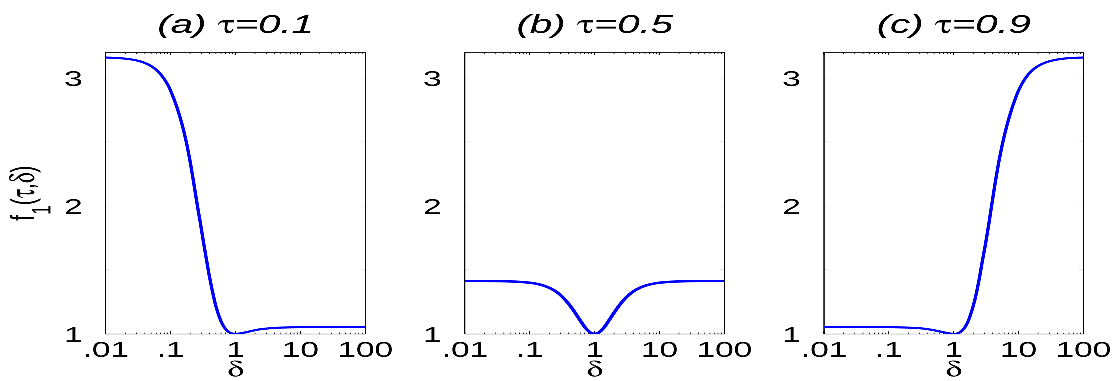

Figure 1: *The values of* $f_1(\tau, \delta)$ *(y-axis) in Example 1 across* δ *(x-axis) for different values of* τ*: (a)* $\tau = 0.1$; *(b)* $\tau = 0.5$; *(c)* $\tau = 0.9$.

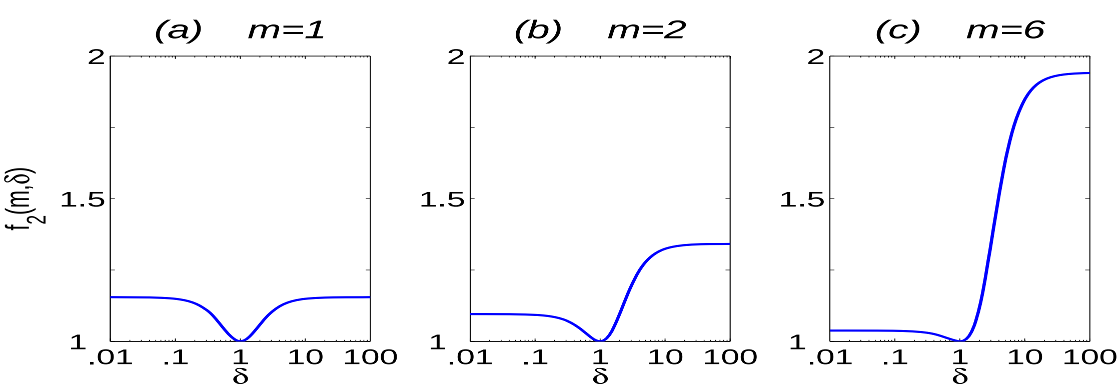

Figure 2: *The values of* $f_2(m, \delta)$ *(y-axis) in Example 2 across* δ *(x-axis) for different values of* m*: (a)* $m = 1$; *(b)* $m = 2$; *(c)* $m = 6$.

Table 1: The ratios of the RMSEs of OLS estimator and ALS estimators using four fixed bandwidths and cross-validated bandwidth, relative to that of GLS (The levels of RMSE are reported for GLS in brackets). Error distribution: normal. Parameter values: $\beta \in \{0.1, 0.9\}$, $\tau \in \{0.1, 0.9\}$, $\delta \in \{0.2, 5\}$ and the sample size $T = \{50, 200\}$.

				ALS					
τ	δ	T	OLS	h_1	h_2	h_3	h_4	CV	GLS
					$\beta = 0.1$				
0.1	0.2	50	1.9749	1.5029	1.5278	1.6169	1.6865	1.5612	[.1236]
		200	2.4751	1.1501	1.1501	1.1830	1.2182	1.1538	[.0636]
	1	50	1.0000	1.1586	1.0745	1.0375	1.0241	1.0329	[.0885]
		200	1.0000	1.0466	1.0280	1.0187	1.0151	1.0155	[.0374]
	5	50	1.0333	1.1220	1.0754	1.0561	1.0498	1.0612	[.1471]
		200	1.0351	1.0780	1.0676	1.0631	1.0600	1.0594	[.0667]
0.9	0.2	50	1.1801	1.3196	1.2625	1.2339	1.2199	1.2359	[.1170]
		200	1.1100	1.1253	1.1172	1.1164	1.1151	1.1198	[.0691]
	5	50	1.9576	1.1925	1.1958	1.2583	1.3177	1.2555	[.1433]
		200	2.2333	1.0859	1.0784	1.0952	1.1208	1.0795	[.0701]
					$\beta = 0.9$				
0.1	0.2	50	2.0748	1.4599	1.4968	1.5742	1.6417	1.5380	[.0633]
		200	2.3822	1.1994	1.2020	1.2270	1.2450	1.1995	[.0283]
	1	50	1.0000	1.0931	1.0374	1.0172	1.0110	1.0191	[.0851]
		200	1.0000	1.0398	1.0213	1.0115	1.0080	1.0103	[.0346]
	5	50	1.0427	1.1260	1.0749	1.0628	1.0592	1.0754	[.0885]
		200	1.0225	1.0571	1.0425	1.0380	1.0354	1.0362	[.0374]
0.9	0.2	50	1.2853	1.2581	1.2763	1.2875	1.2904	1.2838	[.0664]
		200	1.1856	1.1315	1.1540	1.1781	1.1866	1.1844	[.0291]
	5	50	2.0607	1.2049	1.1773	1.2188	1.2769	1.2068	[.0887]
		200	2.2663	1.0903	1.0748	1.0825	1.0983	1.0823	[.0346]

Table 2: The ratios of the RMSEs of OLS estimator and ALS estimators using four fixed bandwidths and cross-validated bandwidth, relative to that of GLS (The levels of RMSE are reported for GLS in brackets). Error distribution: $\chi^2(5)$ or t_5. Parameter values: $\beta \in \{0.1, 0.9\}$, $\tau = 0.9$, $\delta = 5$ and the sample size $T = \{50, 200\}$.

error dist.	T	OLS	ALS h_1	ALS h_2	ALS h_3	ALS h_4	ALS CV	GLS
				$\beta = 0.1$				
$\frac{\chi^2(5)-5}{\sqrt{10}}$	50	2.0441	1.3597	1.3298	1.3983	1.4721	1.4277	[.1375]
	200	2.1478	1.1170	1.1022	1.1148	1.1364	1.1157	[.0701]
$\sqrt{0.6}t_5$	50	1.9072	1.4207	1.3863	1.4259	1.4834	1.4405	[.1394]
	200	2.1648	1.1687	1.1477	1.1583	1.1767	1.1545	[.0704]
				$\beta = 0.9$				
$\frac{\chi^2(5)-5}{\sqrt{10}}$	50	2.0241	1.3419	1.3424	1.3853	1.4286	1.4208	[.0902]
	200	2.2665	1.1729	1.1345	1.1278	1.1364	1.1457	[.0327]
$\sqrt{0.6}t_5$	50	2.0371	1.3108	1.3060	1.3605	1.4243	1.3851	[.0850]
	200	2.1579	1.1515	1.1233	1.1216	1.1337	1.1321	[.0364]