ADAPTIVE ESTIMATION OF AUTOREGRESSIVE MODELS WITH TIME-VARYING VARIANCES

By

Ke-Li Xu and Peter C.B. Phillips

October 2006

COWLES FOUNDATION DISCUSSION PAPER NO. 1585

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

March 21, 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 the 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

*Cowles Foundation for Research in Economics, Yale University, New Haven, Connecticut USA 06520-8281. Telephone: (203) 432-3695. Fax: (203) 432-6167. E-mail addresses: peter.phillips@yale.edu, keli.xu@yale.edu. Phillips gratefully acknowledges support from the Kelly Foundation at the Business School, University of Auckland, and the NSF under NSF Grant No. SES 04-142254. The authors gratefully acknowledge valuable comments from participants at the second meeting of Singapore Econometric Study Group (SESG) at Singapore Management University.

1 Introduction

Recently robust estimation and inference methods have been developed in autoregressions to account for for potentially conditional heteroskedasticity in the innovation process. In this spirit, Kuersteiner (2001, 2002) developed efficient instrumental variables estimators for autoregressive and moving average (ARMA) models and autoregressive models of finite (p-th) order (AR(p)). Goncalves and Kilian (2004a, 2004b) used bootstrap methods to robustify inference in AR(p) and AR(∞) models with unknown conditional heteroskedasticity. These methods and results rely on the assumption that the unconditional variance of errors is constant over time.

Unconditional homoskedasticity seems unrealistic in practice, especially in view of the recent emphasis in the empirical literature on structural change modeling for economic time series. To accommodate models with error variance changes, 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 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).

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 which can adapt for a wide range of possibilities. Accordingly, this paper seeks to develop an efficient estimation procedure which adapts for the presence of different and unknown forms of variance dynamics. We focus on the stable AR(p) model whose 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 transition functions of time.

Efficient estimation of linear models with heteroskedasticity under *iid* assumptions 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 setting. In the time series context, Harvey and Robinson (1988) considered a regression model with deterministically trending regressors, whose error is an AR(p) process scaled by a continuous function of time. 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. In this case, the nearly integrated process is scaled by the factor $T^{-1/2}$, where T is the sample size, to obtain a nondegenerate limit theory. For nearly integrated covariates with deterministic drift, the corresponding normalization would be T^{-1} and Hansen's model be analogous to the model considered here. Regression models in which the conditional variance of the error is an unscaled function of an integrated time series has recently been investigated by Chung and Park (2004) using Brownian local time limit methods developed in Park and Phillips (1999, 2001).

Recently, increasing attention has been paid to potential structural error variance changes in integrated process models. The effects of 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 has been done on stable autoregressions, most of the attention in the literature being concerned with the case of step changes 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. Robust inference in such models is dealt with in another paper (Phillips and Xu, 2005).

The remainder of the paper proceeds as follows. Section 2 introduces the model and assumptions and develops a limit theory 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 3 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 to assess the finite sample performance of the adaptive estimator in Section 4. Section 5 concludes. Proofs of the main results are collected in two appendices.

2 The Model

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, $A(L) = 1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p, \quad \beta_p \neq 0$, 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 taken as 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 $\sup_t \mathbb{E}{\varepsilon_t}^{4\mu} < C < \infty$.

Remarks. (1) 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. Similar formulations have been widely used in the econometric literature, for example by Robinson (1989,1991) in the estimation of time-varying parameter of linear and nonlinear regression, and by Harvey and Robinson (1988) in the efficient estimation of regressions with deterministic trending regressors. In recent work, Cavaliere (2004b) analyzes the effects of heteroskedasticity on unit root tests using this specification of the error variance.

(2) 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.

Under the stated assumptions, the process Y_t has Wold 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, 2005):

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

where

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

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

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

satisfies

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

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 $w_t^2 = \sigma_t^{-2}$(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[1] 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

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

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 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 4.

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 ($\delta > 1$).

3 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^*. Let $K(z)$ be a kernel function defined on the real line such that $K(z)$ is continuous at all but a finite number of points, $0 \leq \sup\limits_{-\infty<z<\infty} K(z) < C$ for some finite real number C and $\int_{-\infty}^{\infty} K(z)dz = 1$. Let $\widehat{u}_t = Y_t - X_{t-1}'\widehat{\beta}$ be the OLS residuals and define the weighted squared residuals

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

where

$$w_{ti} = \frac{K\left(\frac{t-i}{Tb}\right)}{\sum\limits_{i=1}^{T} K\left(\frac{t-i}{Tb}\right)} := \frac{K_{ti}}{\sum\limits_{i=1}^{T} K_{ti}}$$

with $K_{ti} := K\left(\frac{t-i}{Tb}\right)$ and b is the bandwidth parameter, dependent on T. The implementation of the estimator $\widehat{\sigma}_t^2$ depends on the choice of kernel function K and the bandwidth b. Consider the uniform kernel $K(z) = 0.5$ for $|z| \leq 1$, and $K(z) = 0$ otherwise. Then

$$\widehat{\sigma}_t^2 = \frac{1}{Tb} \sum_{|i-t| \leq Tb} \widehat{u}_i^2$$

is the average of $\widehat{u}_i^2$ for i falling into the bin with the center t and length $2Tb$. Kernel functions with infinite support are also possible, such as the Gaussian kernel, $K(z) = (2\pi)^{-1/2} \exp(-t^2/2)$ for $-\infty < z < \infty$. In this case, w_{ti} assigns smaller weights to those $\widehat{u}_i^2$'s whose i is far from t.

Define the adaptive least squares (ALS) estimator of β

$$\widetilde{\beta} = \left(\sum_{t=1}^{T} X_{t-1} X'_{t-1} \widehat{\sigma}_t^{-2} \right)^{-1} \left(\sum_{t=1}^{T} X_{t-1} Y_t \widehat{\sigma}_t^{-2} \right) \tag{11}$$

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). $\mathbb{E}(\varepsilon_t^3 | \mathcal{F}_{t-1}) = 0$, *a.s.*;

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

Remarks. 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 (v) is a rate condition that requires $b \to 0$ at a slower rate than $T^{-1/2}$. Assumption (iv) is satisfied if ε_t has a symmetric distribution conditional on the lagged observations, which is somewhat restrictive. This assumption could be avoided and the main theorem below (Theorem 2) would still hold, if we replaced the estimator $\widehat{\sigma}_t^2$ by

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

We note in the simulations that the performance of the ALS estimator based on (12) is dominated by that based on (11), so that we do not pursue this estimator further here.

The main result is as follows.

Theorem 2 *Under Assumptions (i)-(v) with (iii') instead of (iii), as* $T \to \infty$,

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

where Γ^{-1} *is estimated by (6).*

Remarks.

(1) In practice, the bandwidth parameter b, when estimating the function g, can be chosen using cross-validation on the average squared error – see Wong (1983). Let $\widehat{\widehat{\sigma}}_t^2$ be defined in (12). The cross-validatory choice of b is the value b^* which minimizes

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

(2) Alternative estimators include the one employed by Harvey and Robinson (1988), who deal with the time series regression with trending regressors. Rather than estimating each σ_t^2 separately, they split the data into K blocks and estimate 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[2] 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 is the minimum and maximum length of the K blocks. Compared to our estimator, this estimator is cheaper 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 to the block. Furthermore, unreported simulation results show that its performance is dominated by our kernel-based estimator in most cases.

[2] The proof is available from the authors upon request.

4 Simulations

This section examines the finite sample performance of the ALS efficient procedure proposed in Section 3 using simulations of the heteroskedastic AR(1) model

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

where $\sigma_t = g\left(\frac{t}{T}\right)$. The following values of β are used $\{-0.5, 0.1, 0.9\}$, and $\varepsilon_t \sim iid\mathcal{N}(0,1)$.

Our simulation design basically follows Cavaliere (2004) and Cavaliere and Taylor (2004). The g functions generating heteroskedasticity are taken as the step function and polynomial function used in Examples 1 and 2, viz.,

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

$$\text{Model 2: } g(r)^2 = \sigma_0^2 + (\sigma_1^2 - \sigma_0^2)r^m, r \in [0,1].$$

In Model 1, the break date is chosen from $\{0.1, 0.5, 0.9\}$ and the ratio of post-break and pre-break standard deviations $\delta = \sigma_1/\sigma_0$ is set to the values $\{0.2, 5\}$. In Model 2, the order of polynomial function is taken from $\{1, 2, 6\}$, and $\delta \in \{0.2, 5\}$. Without loss of generality, we let $\sigma_0 = 1$. The estimates of β are obtained with sample size $T = 60$ and $T = 200$, and the number of replications is set to 10,000.

We report estimates for β obtained by OLS, infeasible GLS and ALS. The label "ALS1" denotes the kernel-based ALS estimator (11) using the fixed bandwidth parameter b, $b = 0.1333$ when $T = 60$, and $b = 0.040$ when $T = 200$. The label "ALS2" refers to the ALS estimator with the bandwidth parameter chosen by the cross-validation procedure suggested in Section 3.

Table 1 reports the ratio of the root mean squared errors (RMSE) of estimators considered relative to the RMSE of GLS in Model 1. OLS is clearly inefficient and the ALS estimator works reasonably well in all cases considered. The largest inefficiency of 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. The densities of the OLS and ALS estimators (after cross validation) in the cases mentioned above are plotted in Figure 3. In Panel (a) and (b), the significant improvement of ALS estimator upon OLS can be seen, while in Panel (c) and (d), we observe little difference between two estimators.

We also note that the cross-validation procedure to choose the bandwidth of the ALS estimator works satisfactorily, but seems to be dominated by the one using the specified fixed bandwidth. When the sample size is increased from $T = 60$ to $T = 200$, the ALS estimators have the smaller ratio of RSME, while no improvement is observed for OLS.

Table 2 reports the ratio of the RMSE's of estimators considered relative to the RMSE of GLS in Model 2. The RMSE of the OLS estimator is more steady across the parameters in the heteroskedasticity function than in Model 1. The ALS estimator works remarkably well. Its ratio of RMSE, relative to GLS is below 10% in all cases considered, especially when the sample size is large. The densities of the OLS and ALS estimators (after cross validation) when $m \in \{2, 6\}$, and $\delta \in \{0.2, 5\}$, are plotted in Figure 4.

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 checked 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.

In summary, the kernel-based ALS estimator and cross-validation procedure appear to perform very well, at least within the simulation design considered. Its advantages are clear, is convenient for practical use and has uniformly good performance over the parameter space.

5 Further Remarks

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

6 Appendix A: Proofs of the Theorems.

This section gives the proofs of Theorem 1 and Theorem 2.

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

$$\sqrt{T}(\widehat{\beta}_{WLS} - \beta) = \left(\tfrac{1}{T} \sum_{t=1}^{T} w_t^2 X_{t-1} X_{t-1}' \right)^{-1} \left(\tfrac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t^2 X_{t-1} u_t \right). \tag{13}$$

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

$$\frac{1}{T} \sum_t \left(w_t^2 Y_{t-h} Y_{t-h-k} - w_t^2 \mathbb{E}(Y_{t-h} Y_{t-h-k}) \right) \xrightarrow{p} 0. \tag{14}$$

Lemma A(ii) of Phillips and Xu (2005) shows that for every continuous point r of $g(\cdot)$

$$\lim_{T\to\infty} \mathbb{E} Y_{[Tr]-h} Y_{[Tr]-h-k} \quad = g^2(r)\gamma_k, \tag{15}$$

where $[\cdot]$ refers to the integer part. So by (14)

$$\frac{1}{T} \sum_t w_t^2 Y_{t-h} Y_{t-h-k} = \frac{1}{T} \sum_t w_t^2 \mathbb{E}(Y_{t-h} Y_{t-h-k}) + o_p(1)$$

$$= \sum_{t=1}^{T} \int_{\frac{t}{T}}^{\frac{t+1}{T}} w_{[Tr]}^2 \mathbb{E} Y_{[Tr]-h} Y_{[Tr]-h-k} dr + o_p(1)$$

$$= \int_{\frac{1}{T}}^{\frac{T+1}{T}} w_{[Tr]}^2 \mathbb{E} Y_{[Tr]-h} Y_{[Tr]-h-k} dr + o_p(1) \xrightarrow{p} \left(\int w^2 g^2 \right) \gamma_k. \tag{16}$$

So we have $\frac{1}{T}\sum_{t=1}^{T} X_{t-1}X_{t-1}'\sigma_t^{-2} \xrightarrow{p} \left(\int w^2 g^2 \right) \Gamma$. Next we show that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t^4 X_{t-1} X_{t-1}' u_t^2 \xrightarrow{p} \left(\int w^4 g^4 \right) \Gamma, \tag{17}$$

which holds if $\frac{1}{T}\sum_{t=1}^{T} w_t^4 Y_{t-h} Y_{t-h-k} u_t^2 \xrightarrow{p} \gamma_k$ for $1 \leq h \leq p$, $0 \leq k \leq p-h$. Indeed, since $\{w_t^4 Y_{t-h} Y_{t-h-k} u_t^2 - w_t^4 \sigma_t^2 \mathbb{E} Y_{t-h} Y_{t-h-k}, \mathcal{F}_t\}$ are martingale differences, so $\frac{1}{T}\sum_{t=1}^{T} w_t^4 Y_{t-h} Y_{t-h-k} u_t^2 = \frac{1}{T}\sum_{t=1}^{T} w_t^4 \sigma_t^2 \mathbb{E} Y_{t-h} Y_{t-h-k} + o_p(1) \xrightarrow{p} \left(\int w^4 g^4 \right) \gamma_k$ by similar arguments to (16). Furthermore, $\mathbb{E}\|w_t^2 X_{t-1} u_t\|^4 < \infty$ by Lemma 1 (b) with $\mu = 2$. By the central limit theorem for vector martingale differences, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t^2 X_{t-1} u_t \xrightarrow{d} N\left(0, \left(\int w^4 g^4 \right) \Gamma \right)$. Then Theorem 1 follows from (13).

The Proof of the Theorem 2. We follow closely the proof of the theorem in Robinson (1987) using some of his notation. First, note that $\widetilde{\beta}$ satisfies

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

Define $a(f) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{t-1} u_t f_t^{-2}$ 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

$$\begin{aligned}
\sqrt{T}(\widetilde{\beta} - \beta) &= A(\widehat{\sigma})^{-1} a(\widehat{\sigma}) \\
&= A(\sigma)^{-1} a(\sigma) + A(\widehat{\sigma})^{-1}(a(\widehat{\sigma}) - a(\sigma)) - A(\sigma)^{-1}(A(\widehat{\sigma}) - A(\sigma)) A(\widehat{\sigma})^{-1} a(\sigma).
\end{aligned}$$

We have $A(\sigma) \xrightarrow{p} \Gamma$ which is positive definite, and $a(\sigma) = O_p(1)$, which follows from Markov's inequality and

$$\mathbb{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_{t-h} u_t \sigma_t^{-2}\right)^2 = \frac{1}{T}\sum_{t=1}^{T} \sigma_t^{-4} \mathbb{E} Y_{t-h}^2 u_t^2 \leq C \frac{1}{T}\sum_{t=1}^{T} \mathbb{E} Y_{t-h}^2 u_t^2 < \infty,$$

by Lemma 1 (b) and Assumption (i). Hence Theorem 2 follows if we prove

$$A(\widehat{\sigma}) - A(\sigma) \xrightarrow{p} 0, \ a(\widehat{\sigma}) - a(\sigma) \xrightarrow{p} 0. \tag{18}$$

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 (18) 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

$$\left\| a(\widehat{\sigma}) - a(\widetilde{\sigma}) \right\| \overset{TI}{\leq} \left(\min_{1\leq t\leq T} \widetilde{\sigma}_t^2\right)^{-1} \left(\min_{1\leq t\leq T} \widehat{\sigma}_t^2\right)^{-1} \sum_{t=1}^{T} \frac{\left\| X_{t-1} u_t \right\|}{\sqrt{T}} \left| \widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2 \right|$$

$$\overset{CS}{\leq} \left(\min_{1\leq t\leq T} \widetilde{\sigma}_t^2\right)^{-1} \left(\min_{1\leq t\leq T} \widehat{\sigma}_t^2\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \left\| X_{t-1} u_t \right\|^2\right)^{1/2} \left(\sum_{t=1}^{T} \left| \widetilde{\sigma}_t^2 - \widehat{\sigma}_t^2 \right|^2\right)^{1/2} = O_p(\frac{1}{Tb}) \xrightarrow{p} 0,$$

by Lemma 1, 7, 9, 10.

(b) We write

$$\begin{aligned} a(\widetilde{\sigma}) - a(\overline{\sigma}) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{t-1} u_t \left(\widetilde{\sigma}_t^{-2} - \overline{\sigma}_t^{-2}\right) \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{t-1} u_t (\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2) \overline{\sigma}_t^{-4} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_{t-1} u_t (\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)^2 \widetilde{\sigma}_t^{-2} \overline{\sigma}_t^{-4}, \end{aligned} \tag{19}$$

which holds since for two any nonzero real numbers p and q we have the following equality $p^{-1} - q^{-1} = (q-p)q^{-2} + (q-p)^2 p^{-1} q^{-2}$. We will show the two terms of (19) 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

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

$$= \overline{\sigma}_t^{-2}\mathbb{E}(X_{t-1}u_t|\mathcal{F}_{t-1}) - \overline{\sigma}_t^{-4}\left(\mathbb{E}\left(X_{t-1}u_t\left(\sum_{i=1,i\neq t}^{T} w_{ti}u_i^2\right)|\mathcal{F}_{t-1}\right) + w_{tt}\mathbb{E}(X_{t-1}u_t^3|\mathcal{F}_{t-1})\right). \quad (20)$$

By Assumption (iv), $\mathbb{E}(X_{t-1}u_t^3|\mathcal{F}_{t-1}) = X_{t-1}\mathbb{E}(u_t^3|\mathcal{F}_{t-1}) = 0$. Further, we have

$$\mathbb{E}\left(X_{t-1}u_t\left(\sum_{i=1,i\neq t}^{T} w_{ti}u_i^2\right)|\mathcal{F}_{t-1}\right) = 0,$$

which holds since for the term $i > t$,

$$\begin{aligned}\mathbb{E}\left(X_{t-1}u_tu_i^2|\mathcal{F}_{t-1}\right) &= X_{t-1}\mathbb{E}\left(u_tu_i^2|\mathcal{F}_{t-1}\right) = X_{t-1}\mathbb{E}\left(u_t\mathbb{E}\left(u_i^2|\mathcal{F}_{i-1}\right)|\mathcal{F}_{t-1}\right)\\ &= X_{t-1}\mathbb{E}\left(u_t|\mathcal{F}_{t-1}\right) = 0,\end{aligned}$$

and for the term $i < t$,

$$\mathbb{E}\left(X_{t-1}u_tu_i^2|\mathcal{F}_{t-1}\right) = X_{t-1}u_i^2\mathbb{E}\left(u_t|\mathcal{F}_{t-1}\right) = 0.$$

Thus, by (20) $\mathbb{E}(X_{t-1}u_t(\overline{\sigma}_t^2 - \widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}|\mathcal{F}_{t-1}) = 0$. So the first term of (19) converges to zero in probability by the Markov inequality and

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_{t-1}u_t(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)\overline{\sigma}_t^{-4}\right\|^2 &\leq \frac{C}{T}\sum_{t=1}^{T}\mathbb{E}\left\|X_{t-1}u_t\right\|^2(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)^2\\
&\leq \frac{C}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left\|X_{t-1}u_t\right\|^4\right)^{1/2}\cdot\left(\mathbb{E}(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)^4\right)^{1/2}\\
&\leq \left(\max_t\mathbb{E}(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)^4\right)^{1/2}\cdot\frac{C}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left\|X_{t-1}u_t\right\|^4\right)^{1/2}\\
&= O_p(\frac{1}{Tb})\xrightarrow{p}0,
\end{aligned}
$$

by Lemma 1 and 5. For the second term of (19),

$$
\begin{aligned}
\left\|\sum_{t=1}^{T}\frac{X_{t-1}u_t}{\sqrt{T}}(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)^2\widetilde{\sigma}_t^{-2}\overline{\sigma}_t^{-4}\right\| &\leq C\left(\frac{1}{T}\sum_{t=1}^{T}\left\|X_{t-1}u_t\right\|^2\right)^{1/2}\left(\sum_{t=1}^{T}(\overline{\sigma}_t^2-\widetilde{\sigma}_t^2)^4\right)^{1/2}\\
&= O_p(\frac{1}{T^{1/2}b})\xrightarrow{p}0,
\end{aligned}
$$

by Lemma 1 and 5. This completes the proof of (b).

(c) First we note

$$
\sigma_t^2\left(\overline{\sigma}_t^{-2}-\sigma_t^{-2}\right)^2\leq\overline{\sigma}_t^{-4}\sigma_t^{-2}\left|\overline{\sigma}_t^2+\sigma_t^2\right|\cdot\left|\overline{\sigma}_t^2-\sigma_t^2\right|\leq C\left|\overline{\sigma}_t^2-\sigma_t^2\right|. \tag{21}
$$

Since $\{X_{t-1}u_t\}$ is an m.d. sequence, we get

$$
\begin{aligned}
\mathbb{E}\left\|a(\overline{\sigma})-a(\sigma)\right\|^2 &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\left\|X_{t-1}\right\|^2 u_t^2\right)\left(\overline{\sigma}_t^{-2}-\sigma_t^{-2}\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\left\|X_{t-1}\right\|^2 \mathbb{E}(u_t^2|\mathcal{F}_{t-1})\right)\left(\overline{\sigma}_t^{-2}-\sigma_t^{-2}\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|X_{t-1}\right\|^2 \sigma_t^2\left|\overline{\sigma}_t^{-2}-\sigma_t^{-2}\right|^2 \\
&\le \frac{C}{T}\sum_{t=1}^{T}\mathbb{E}\left\|X_{t-1}\right\|^2\left|\overline{\sigma}_t^{2}-\sigma_t^{2}\right| \\
&\le C\max_t\mathbb{E}\left\|X_{t-1}\right\|^2\cdot\frac{1}{T}\sum_{t=1}^{T}\left|\overline{\sigma}_t^{2}-\sigma_t^{2}\right| = o_p(1),
\end{aligned}
$$

by Lemma 1 and 11.

(d) It follows from

$$
\begin{aligned}
\left\|A(\widehat{\sigma})-A(\widetilde{\sigma})\right\| &\le \left(\min_{1\le t\le T}\widetilde{\sigma}_t^2\right)^{-1}\left(\min_{1\le t\le T}\widehat{\sigma}_t^2\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\left\|X_{t-1}\right\|^2\left|\widetilde{\sigma}_t^2-\widehat{\sigma}_t^2\right| \\
&\le C\max_t\left|\widetilde{\sigma}_t^2-\widehat{\sigma}_t^2\right|\cdot\frac{1}{T}\sum_{t=1}^{T}\left\|X_{t-1}\right\|^2 = O_p(\frac{1}{\sqrt{Tb}}),
\end{aligned}
$$

by Lemma 1, 7, 8, 9.

(e) This can be proved in the same way as (d) by employing Lemma 6.

(f) It follows from

$$
\begin{aligned}
\left\|A(\overline{\sigma})-A(\sigma)\right\| &\le \left(\min_{1\le t\le T}\overline{\sigma}_t^2\right)^{-1}\left(\min_{1\le t\le T}\sigma_t^2\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\left\|X_{t-1}\right\|^2\left|\overline{\sigma}_t^2-\sigma_t^2\right| \\
&\le \left(\min_{1\le t\le T}\overline{\sigma}_t^2\right)^{-1}\left(\min_{1\le t\le T}\sigma_t^2\right)^{-1}\max_t\left\|X_{t-1}\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left|\overline{\sigma}_t^2-\sigma_t^2\right| = o_p(1),
\end{aligned}
$$

by Lemma 1, 4, 11.

7 Appendix B: Supplementary Lemmas and Proofs.

This section states and proves some results (Lemma 1-Lemma 11) used in the proofs of the theorems.

Lemma 1 (a) For $1 \leq \mu < \infty$ and $1 \leq h \leq p$,

$$\sup_{1 \leq t \leq T} \mathbb{E} Y_{t-h}^{2\mu} < \infty$$

holds if $\sup_{1 \leq t \leq T} \mathbb{E}\varepsilon_t^{2\mu} < \infty$;and

$$\sup_{1 \leq t \leq T} \mathbb{E}\left(Y_{t-h} u_t\right)^{2\mu} < \infty$$

holds if $\sup_{1 \leq t \leq T} \mathbb{E}\varepsilon_t^{4\mu} < \infty$.

Proof. (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

$$\mathbb{E}\left|u_{t-h-k} u_{t-h-l}\right|^{\mu} \leq \left(\mathbb{E}u_{t-h-k}^{2\mu} \mathbb{E}u_{t-h-l}^{2\mu}\right)^{1/2} < \infty.$$

So we have

$$\begin{aligned} \mathbb{E}\left(Y_{t-h}\right)^{2\mu} &= \left\|Y_{t-h}^2\right\|_{\mu}^{\mu} \leq \left(\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \alpha_k \alpha_l \left\|u_{t-h-k} u_{t-h-l}\right\|_{\mu}\right)^{\mu} \\ &\leq C\left(\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \alpha_k \alpha_l\right)^{\mu} = C\left(\sum_{k=0}^{\infty} \alpha_k\right)^{2\mu} < \infty. \end{aligned}$$

(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

$$\mathbb{E}\left|u_{t-h-k} u_{t-h-l} u_t^2\right|^{\mu} \leq \left(\mathbb{E}u_{t-h-k}^{4\mu} \mathbb{E}u_{t-h-l}^{4\mu}\right)^{1/4} \left(\mathbb{E}u_t^{4\mu}\right)^{1/2} < \infty,$$

so

$$
\begin{aligned}
\mathbb{E}\left(Y_{t-h}u_t\right)^{2\mu} &= \left\|Y_{t-h}^2u_t^2\right\|_\mu^\mu \le \left(\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\alpha_k\alpha_l\left\|u_{t-h-k}u_{t-h-l}u_t^2\right\|_\mu\right)^\mu \\
&\le C\left(\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\alpha_k\alpha_l\right)^\mu < \infty.
\end{aligned}
$$

Lemma 2. For any $1 \le t \le T$, $\frac{1}{Tb}\sum_{i=1}^{T}K_{ti} \to \int_{-\infty}^{\infty}K(z)dz = 1$, where $K_{ti} = K(\frac{t-i}{Tb})$.

Proof. Let $t-i = [Tx]$, where x is a real number, $|x| < 1$. Then

$$
\begin{aligned}
\frac{1}{Tb}\sum_{i=1}^{T}K_{ti} &= \sum_{i=1}^{T}\int_{(t-i)/T}^{(t-i+1)/T}K(\frac{[Tx]}{Tb})d\left(\frac{x}{b}\right) \overset{z=x/b}{=} \sum_{i=1}^{T}\int_{(t-i)/Tb}^{(t-i+1)/Tb}K(\frac{[Tbz]}{Tb})dz \\
&= \int_{(t-T)/Tb}^{t/Tb}K(\frac{[Tbz]}{Tb})dz \to \int_{-\infty}^{\infty}K(z)dz = 1.
\end{aligned}
$$

Lemma 3. $\max_{t,i} w_{ti} = O(\frac{1}{Tb})$.

Proof. It follows from $w_{ti} = \left(\frac{1}{Tb}\sum_{i=1}^{T}K_{ti}\right)^{-1}\frac{K_{ti}}{Tb}$ and Lemma 2.

Lemma 4. $\min_{1\le t\le T}\overline{\sigma}_t^2 \ge c > 0$.

Proof. It follows from $\min_{1\le t\le T}\overline{\sigma}_t^2 \ge \min_{1\le i\le T}\sigma_i^2\cdot\left(\sum_{i=1}^{T}w_{ti}\right) \ge \inf_{s\in[0,1]}g^2(s) \ge c > 0$.

Lemma 5. $\max_{1\le t\le T}\mathbb{E}\left|\widetilde{\sigma}_t^2 - \overline{\sigma}_t^2\right|^4 = O\left(\frac{1}{(Tb)^2}\right)$.

Proof. 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

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

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

$$\mathbb{E}\left(\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right)^4 = \mathbb{E}\left(\sum_{i=1}^{T} w_{ti}a_i\right)^4 \overset{BI(p=4)}{\leq} \mathbb{E}\left(\sum_{i=1}^{T} w_{ti}^2 a_i^2\right)^2$$
$$\leq \frac{1}{(Tb)^2}\mathbb{E}\left(\sum_{i=1}^{T} w_{ti}a_i^2\right)^2 \leq \frac{1}{(Tb)^2}\sum_{i=1}^{T} w_{ti}\mathbb{E}a_i^4 = O\left(\tfrac{1}{(Tb)^2}\right),$$

where the second-to-last line is by Jensen inequality $f(\sum_{i=1}^{T} w_{ti}a_i^2) \leq \sum_{i=1}^{T} w_{ti}f(a_i^2)$ with convex function $f(x)=x^2$.

Lemma 6. $\max_t \left|\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right|^\delta = O_p(T^{-\delta/4}b^{-\delta/2})$, for $\delta=1,2$.

Proof. It holds since

$$P\left(\max_t \left|\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right|^\delta > CT^{-\delta/4}b^{-\delta/2}\right) \leq \sum_{t=1}^{T} P\left(\left|\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right|^\delta > CT^{-\delta/4}b^{-\delta/2}\right)$$
$$\leq C^{-4}Tb^2\sum_{t=1}^{T}\mathbb{E}\left|\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right|^4 \leq C^{-4}Tb^2\cdot T\cdot O\left(\tfrac{1}{(Tb)^2}\right) \leq O(1).$$

Lemma 7. $\left(\min_{1\leq t\leq T}\widetilde{\sigma}_t^2\right)^{-1} = O_p(1)$, as $T\to\infty$.

Proof. It follows from Lemma 4 and

$$\min_{1\leq t\leq T}\overline{\sigma}_t^2 \leq \min_{1\leq t\leq T}\widetilde{\sigma}_t^2 + \max_t\left|\widetilde{\sigma}_t^2-\overline{\sigma}_t^2\right| = \min_{1\leq t\leq T}\widetilde{\sigma}_t^2 + o_p(1).$$

Lemma 8. $\max_{1\leq t\leq T}\left|\widehat{\sigma}_t^2-\widetilde{\sigma}_t^2\right| = O_p(\frac{1}{\sqrt{Tb}})$.

Proof. Note that

$$\begin{aligned}
\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 &= \sum_{i=1}^{T} w_{ti}\left(\widehat{u}_i^2 - u_i^2\right) \\
&= \sum_{i=1}^{T} w_{ti}\left((\widehat{\beta}-\beta)' X_{i-1}X_{i-1}'(\widehat{\beta}-\beta) - 2u_i X_{i-1}'(\widehat{\beta}-\beta)\right),
\end{aligned}$$

and $\max\limits_{t,i}\sum\limits_{i=1}^{T} w_{ti}^2 \le \max\limits_{t,i} w_{ti} \cdot \sum\limits_{i=1}^{T} w_{ti} = O(\frac{1}{Tb})$. Thus

$$\begin{aligned}
\max_{1\le t\le T}\left|\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2\right| &\le \max_{1\le t\le T}\sum_{i=1}^{T} w_{ti}\left|(\widehat{\beta}-\beta)' X_{i-1}X_{i-1}'(\widehat{\beta}-\beta) - 2u_i X_{i-1}'(\widehat{\beta}-\beta)\right| \\
&\le \max_{1\le t\le T}\sum_{i=1}^{T} w_{ti}\left\|\widehat{\beta}-\beta\right\|^2\left\|X_{i-1}\right\|^2 + 2\max_{1\le t\le T}\sum_{i=1}^{T} w_{ti}\left\|u_i X_{i-1}'\right\|\left\|\widehat{\beta}-\beta\right\| \\
&\le \max_{t,i} w_{ti}\cdot\left\|\widehat{\beta}-\beta\right\|^2\sum_{i=1}^{T}\left\|X_{i-1}\right\|^2 + 2\left\|\widehat{\beta}-\beta\right\|\cdot\left(\max_{t,i}\sum_{i=1}^{T} w_{ti}^2\right)^{1/2}\cdot\left(\sum_{i=1}^{T}\left\|u_i X_{i-1}'\right\|\right)^{1/2} \\
&= O_p(\frac{1}{Tb}) + O_p(\frac{1}{\sqrt{Tb}}) = O_p(\frac{1}{\sqrt{Tb}}).
\end{aligned}$$

Lemma 9. $\left(\min\limits_{1\le t\le T}\widehat{\sigma}_t^2\right)^{-1} = O_p(1)$, as $T\to\infty$.

Proof. It follows from Lemma 7 and

$$\min_{1\le t\le T}\widetilde{\sigma}_t^2 \le \min_{1\le t\le T}\widehat{\sigma}_t^2 + \max_t\left|\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2\right| = \min_{1\le t\le T}\widehat{\sigma}_t^2 + o_p(1).$$

Lemma 10. $\sum\limits_{t=1}^{T}\left(\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2\right)^2 = O_p(\frac{1}{(Tb)^2})$.

Proof. Since

$$\widehat{\sigma}_t^2 - \widetilde{\sigma}_t^2 = \sum_{i=1}^{T} w_{ti}\left(\widehat{u}_i^2 - u_i^2\right) = (\widehat{\beta}-\beta)'\left(\sum_{i=1}^{T} w_{ti}^2 X_{i-1}X_{i-1}'\right)(\widehat{\beta}-\beta) - 2\left(\sum_{i=1}^{T} w_{ti}^2 u_i X_{i-1}'\right)(\widehat{\beta}-\beta),$$

we have

$$\sum_{t=1}^{T}\left(\widehat{\sigma}_t^2-\widetilde{\sigma}_t^2\right)^2$$

$$\overset{T}{\leq}\sum_{t=1}^{T}C\left(\left\|\widehat{\beta}-\beta\right\|^4\left\|\sum_{i=1}^{T}w_{ti}^2X_{i-1}X_{i-1}'\right\|^2+\left\|\sum_{i=1}^{T}w_{ti}^2u_iX_{i-1}'\right\|^2\left\|\widehat{\beta}-\beta\right\|^2\right)$$

$$\leq\left\|\widehat{\beta}-\beta\right\|^4\sum_{t=1}^{T}C\left(\sum_{i=1}^{T}w_{ti}^2\left\|X_{i-1}\right\|^2\right)^2+\left\|\widehat{\beta}-\beta\right\|^2\sum_{t=1}^{T}C\left(\sum_{i=1}^{T}w_{ti}^2\left\|u_iX_{i-1}'\right\|\right)^2 \tag{22}$$

The first term of (22) is bounded by

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

and similarly the second term of (22) is $O_p(\frac{1}{T^2b^2})$. So Lemma 10 follows.

Lemma 11. $\frac{1}{T}\sum_{t=1}^{T}|\overline{\sigma}_t^2-\sigma_t^2|=o(1)$.

Proof. Without loss of generality and given the sample size T, suppose $\frac{t_1}{T},\cdots,\frac{t_D}{T}$ happen to be the discontinuous points of g, then D is a finite number (independent of T). For $M>0$ define $\sum_i{}'=\sum_{|i-t|\leq MTb}$ and $\sum_i{}''=\sum_{|i-t|>MTb}$.Then for $t\neq t_1,\cdots,t_D$

$$\begin{aligned}\frac{1}{Tb}\sum_{i=1}^{T}K_{ti}\left|\overline{\sigma}_t^2-\sigma_t^2\right| &= \frac{1}{Tb}\sum_{i=1}^{T}K_{ti}\left|\sigma_i^2-\sigma_t^2\right|\\ &\leq \frac{1}{Tb}\sum_i{}'K_{ti}\left|g^2(\tfrac{i}{T})-g^2(\tfrac{t}{T})\right|+\frac{C}{Tb}\sum_i{}''K_{ti}\end{aligned} \tag{23}$$

The first term of (23) is bounded by

$$C \max_{|i-t|\le MTb}\left|\frac{i-t}{T}\right|\frac{1}{Tb}\cdot\sum_i{}' K_{ti}\le C\frac{MTb}{T}\frac{1}{Tb}\cdot 2MTb.$$

For the second term, similarly to the proof of Lemma 2 we can get

$$\frac{C}{Tb}\sum_i{}'' K_{ti}\to C\int_{|z|\ge M}K(z)dz.$$

Thus (23) converges to zero by letting $T\to\infty$ then $M\to\infty$. In view of Lemma 2, we establish $\max_{t\neq t_1,\cdots,t_D}|\overline{\sigma}_t^2-\sigma_t^2| = o(1)$. Thus

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}|\overline{\sigma}_t^2-\sigma_t^2| &= \frac{1}{T}\sum_{t=t_1,\cdots,t_D}|\overline{\sigma}_t^2-\sigma_t^2|+\frac{1}{T}\sum_{t\neq t_1,\cdots,t_D,t=1}^{T}|\overline{\sigma}_t^2-\sigma_t^2|\\ &\le \frac{D}{T}C+\frac{1}{T}(T-D)\max_{t\neq t_1,\cdots,t_D}|\overline{\sigma}_t^2-\sigma_t^2| = o(1).\end{aligned}$$

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

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

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

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

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

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

[9] Chung, H., Park, J. Y., 2004. Nonstationary nonlinear heteroskedasticity in regression. Working paper, Rice University.

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

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

[12] 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.

[13] Goncalves, S., Kilian, L., 2004a. Bootstrapping autoregression with conditional heteroskedasticity of unknown form. Journal of Econometrics 123, 89-120.

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

[15] Goncalves, S., Kilian, L., 2004b. Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. CIRANO Working paper.

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

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

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

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

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

[21] Kuersteiner, G. M., 2001. Optimal instrumental variables estimation for ARMA models. Journal of Econometrics 104, 359-405.

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

[23] 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.

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

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

[26] 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.

[27] Phillips, P. C. B., Xu, K.-L., 2005. Inference in autoregression under heteroskedasticity. mimeo, Yale University.

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

[29] 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.

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

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

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

[33] 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.

Table 1: The ratio of the RMSE relative to that of GLS in Model 1 (The levels of RMSE are reported for GLS)

			$T=60$				$T=200$			
β	τ	δ	OLS	ALS1	ALS2	GLS	OLS	ALS1	ALS2	GLS
-0.5	0.1	0.2	2.1204	1.3246	1.3405	[.0967]	2.3136	1.1564	1.2091	[.0583]
		1	1.0000	1.0101	1.0130	[.1190]	1.0000	1.0030	1.0058	[.0569]
		5	1.0329	1.0595	1.0570	[.1156]	1.0446	1.0471	1.0450	[.0613]
	0.5	0.2	1.5621	1.2714	1.3052	[.0987]	1.4704	1.1026	1.1364	[.0562]
		5	1.3140	1.1129	1.1521	[.1147]	1.3639	1.0698	1.1177	[.0608]
	0.9	0.2	1.1820	1.1767	1.1811	[.1023]	1.0915	1.1185	1.1217	[.0564]
		5	2.0619	1.2267	1.2602	[.1198]	2.4099	1.1157	1.1857	[.0601]
0.1	0.1	0.2	2.1256	1.3755	1.4076	[.1113]	2.3017	1.1224	1.1831	[.0648]
		1	1.0000	1.0197	1.0095	[.1296]	1.0000	1.0094	1.0051	[.0659]
		5	1.0324	1.0516	1.0424	[.1259]	1.0430	1.0415	1.0467	[.0732]
	0.5	0.2	1.4741	1.2324	1.2612	[.1150]	1.4650	1.1155	1.1547	[.0643]
		5	1.2784	1.1029	1.1326	[.1310]	1.3786	1.0504	1.0693	[.0698]
	0.9	0.2	1.1527	1.1665	1.1575	[.1161]	1.0970	1.1070	1.1183	[.0655]
		5	2.0710	1.2388	1.2740	[.1252]	2.2879	1.0839	1.1138	[.0690]
0.9	0.1	0.2	1.9045	1.2771	1.3360	[.0624]	2.3275	1.1754	1.2246	[.0295]
		1	1.0000	1.0044	1.0081	[.0776]	1.0000	1.0041	1.0055	[.0365]
		5	1.0352	1.0441	1.0388	[.0797]	1.0516	1.0526	1.0540	[.0337]
	0.5	0.2	1.7187	1.2607	1.3005	[.0607]	1.6318	1.1637	1.2052	[.0279]
		5	1.5026	1.1886	1.2416	[.0794]	1.3985	1.0535	1.0773	[.0358]
	0.9	0.2	1.2994	1.2706	1.2591	[.0617]	1.1829	1.1299	1.1558	[.0289]
		5	2.2604	1.2429	1.3065	[.0695]	2.3215	1.0857	1.1646	[.0346]

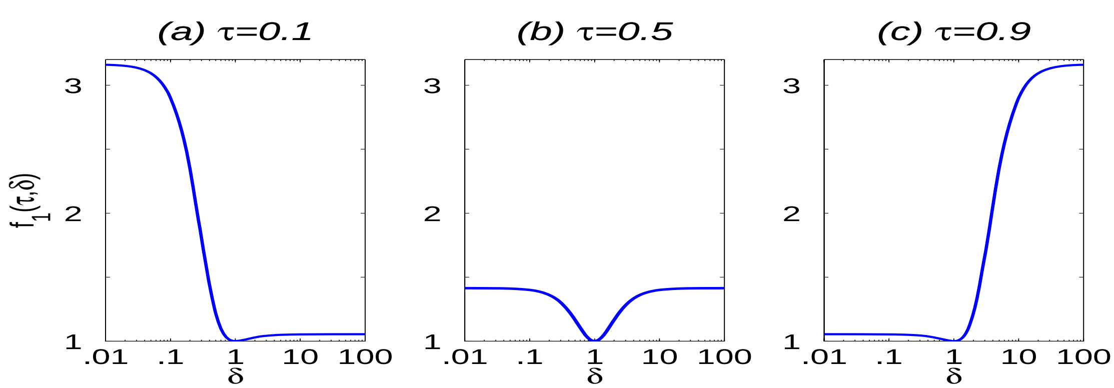

Figure 1: *The values of* $f_1(\tau,\delta)$ *(y-axis) in Example 1 across* δ *(x-axis) for different values of* τ*: (a)* $\tau=0.1$*; (b)* $\tau=0.5$*; (c)* $\tau=0.9$.

Table 2: The ratio of the RMSE relative to that of GLS in Model 2 (The levels of RMSE are reported for GLS)

			$T = 60$				$T = 200$			
β	m	δ	OLS	ALS1	ALS2	GLS	OLS	ALS1	ALS2	GLS
-0.5	1	0.2	1.1329	1.0269	1.0500	[.1151]	1.1344	1.0371	1.0370	[.0613]
		5	1.0869	1.0214	1.0471	[.1223]	1.1005	1.0245	1.0226	[.0610]
	2	0.2	1.1408	1.0739	1.0823	[.1105]	1.0781	1.0173	1.0243	[.0624]
		5	1.2286	1.0447	1.0696	[.1193]	1.2579	1.0336	1.0226	[.0587]
	6	0.2	1.0926	1.0861	1.0856	[.1095]	1.0474	1.0550	1.0400	[.0610]
		5	1.5504	1.0607	1.0994	[.1192]	1.5361	1.0251	1.0412	[.0639]
0.1	1	0.2	1.1297	1.0406	1.0608	[.1260]	1.1149	1.0343	1.0362	[.0672]
		5	1.1428	1.0364	1.0573	[.1305]	1.1251	1.0295	1.0269	[.0743]
	2	0.2	1.0887	1.0465	1.0619	[.1257]	1.0875	1.0383	1.0389	[.0678]
		5	1.1949	1.0324	1.0597	[.1332]	1.2854	1.0294	1.0287	[.0695]
	6	0.2	1.0607	1.0573	1.0573	[.1248]	1.0376	1.0258	1.0223	[.0713]
		5	1.5141	1.0553	1.0930	[.1317]	1.6076	1.0442	1.0438	[.0689]
0.9	1	0.2	1.1460	1.0378	1.0634	[.0708]	1.1552	1.0179	1.0278	[.0317]
		5	1.0962	1.0204	1.0398	[.0800]	1.1121	1.0247	1.0268	[.0352]
	2	0.2	1.1312	1.0501	1.0615	[.0702]	1.0603	1.0303	1.0249	[.0344]
		5	1.2342	1.0468	1.0820	[.0843]	1.2578	1.0194	1.0172	[.0340]
	6	0.2	1.1097	1.0933	1.0987	[.0716]	1.0302	1.0365	1.0301	[.0345]
		5	1.5187	1.0642	1.1141	[.0820]	1.6012	1.0278	1.0291	[.0339]

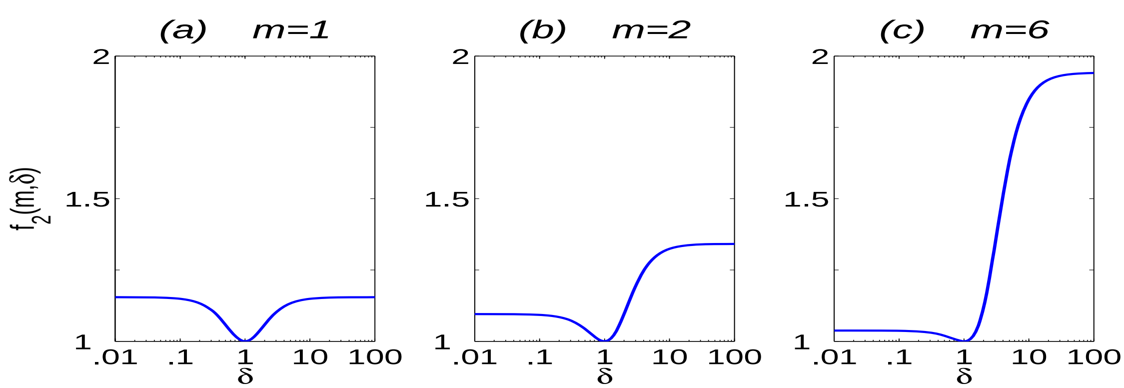

Figure 2: *The values of $f_2(m,\delta)$ (y-axis) in Example 2 across δ (x-axis) for different values of m: (a) $m = 1$; (b) $m = 2$; (c) $m = 6$.*

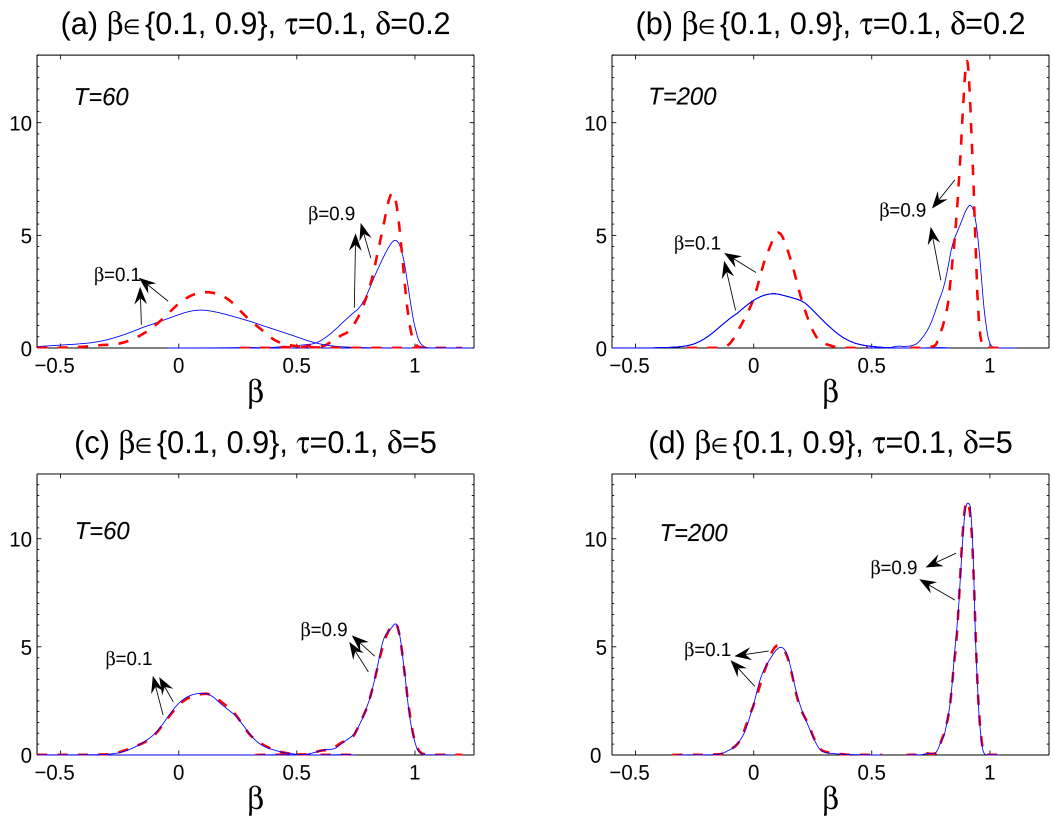

Figure 3: *Densities of the OLS (solid lines) and ALS2 (after cross-validation) estimators (dashed lines) in Model 1.*

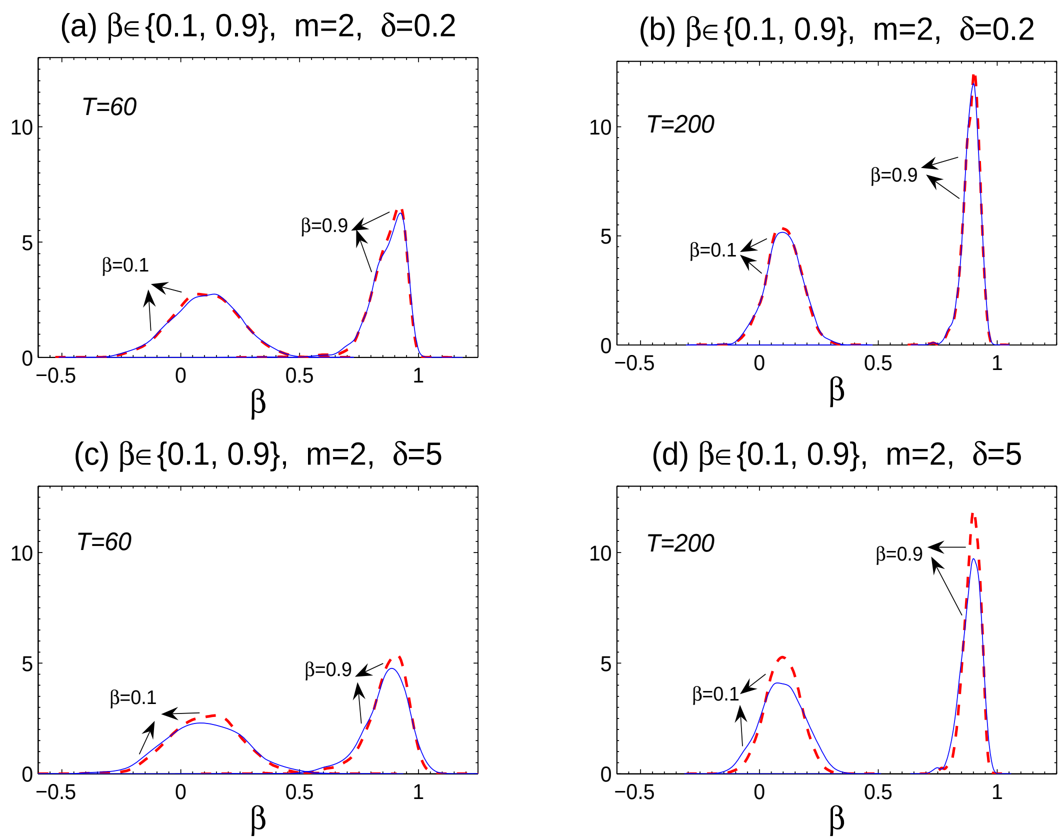

Figure 4: *Densities of the OLS (solid lines) and ALS2 (after cross-validation) estimators (dashed lines) in Model 2.*