NBER WORKING PAPER SERIES

ASYMPTOTIC FILTERING THEORY FOR UNIVARIATE ARCH MODELS

Daniel B. Nelson

Dean P. Foster

Technical Working Paper No. 129

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 1992

This paper is a revision of the working paper, "Estimating Conditional Variances with Univariate ARCH Models: Asymptotic Theory." We would like to thank two referees, an associate editor, Phillip Braun, John Cochrane, Enrique Sentana, reviewers for the NSF, and seminar participants at the University of Chicago, Harvard/M.I.T., Michigan State, Minnesota, the SAS Institute, Washington University, Yale, the 1991 Midwest Econometrics Group meetings, the 1991 NBER Summer Institute, and the 1992 ASA meetings for helpful comments. This material is based on work supported by the National Science Foundation under grant #SES-9110131. We thank the Center for Research in Security Prices, the University of Chicago Graduate School of Business, and the William S. Fishman Research Scholarship for additional research support, and Boaz Schwartz for research assistance. This paper is part of NBER's Research Program in Asset Pricing. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

1. INTRODUCTION

Most asset pricing theories relate expected returns on assets to their conditional variances and covariances. An enormous literature in empirical finance has documented that these conditional moments change over time. Practical experience (as in the 1929 and 1987 stock market crashes) reinforces this conclusion. Unfortunately, conditional variances and covariances are not directly observable, and researchers and market participants must use estimates of conditional second moments. To create these estimates, they rely on models which are, no doubt, misspecified. How accurate are these estimated variances and covariances? How can researchers estimate them more accurately?

Since their introduction by Engle (1982), ARCH models have become a widely used tool for estimating conditional variances and covariances. (See the survey of Bollerslev, Chou, and Kroner (1992).) Suppose that for each t, ξ_t is a (scalar) innovation in a time series model. Interpreted as a data generating mechanism, a univariate ARCH model assumes that

(1.1) $E_{t-1}[\xi_t] = 0$ and $Var_{t-1}[\xi_t] = \sigma_t^2$, with

(1.2) $\sigma_t^2 \equiv \sigma^2(\xi_{t-1}, \xi_{t-2}, \ldots, t)$.

That is, σ_t^2 is the conditional variance of ξ_t given time t-1 information, and is a function of time and past ξ_t's.

As are all statistical and economic models, ARCH models are at best a rough approximation to reality: it is too much to hope that the models are "true." As we will see,

however, we needn't think of (1.1)-(1.2) as the true data generating mechanism in order for ARCH models to be useful in extracting conditional variances from data. Given an arbitrary sequence $\{\xi_t\}_{t=-\infty,\infty}$, we can use (1.2) to create a corresponding $\{\sigma_t^2\}_{t=-\infty,\infty}$ sequence: under conditions developed below, this sequence may provide a good estimate of the true conditional variance of $\{\xi_t\}_{t=-\infty,\infty}$, even when the model (1.1)-(1.2) is misspecified. One can think of (1.2) as a *filter* though which we pass the data to produce an *estimate* of the conditional variance. We should note, however, that we use the term "estimation" as it is used in the filtering literature rather than as it is used in the statistics literature—i.e., the ARCH model "estimates" the true conditional variance in the same sense that a Kalman filter estimates unobserved state variables in a linear system.[1]

A previous paper, Nelson (1992), gave one likely reason for the empirical success of ARCH: when both observable variables and conditional variances change "slowly" relative to the sampling interval (in particular, when the data generating process is well approximated by a diffusion and the data are observed at high frequencies) then broad classes of ARCH models—even when misspecified—provide continuous-record consistent estimates of the conditional variances. That is, as the observable variables are recorded at finer and finer intervals, the conditional variance estimates produced by the (misspecified) ARCH model converge in probability to the true conditional variances.

This paper builds on this earlier work by deriving the asymptotic *distribution* of the

1) See, e.g., the use of the term in Anderson and Moore (1979) Chapter 2, or Arnold (1973) Chapter 12.

measurement error. This allows us to approximate the measurement accuracy of ARCH conditional variance estimates and compare the efficiency achieved by different ARCH models. We are also able to characterize the relative importance of different kinds of misspecification; for example, we show that misspecifying conditional means adds only trivially (at least asymptotically) to measurement error, while other factors (for example, capturing the "leverage effect," accommodating thick tailed residuals, and correctly modelling the variability of the conditional variance process) are potentially much more important. Third, we are able to characterize a class of asymptotically optimal ARCH conditional variance estimates.

In Section 2, we state the basic functional limit theorem we employ throughout the paper. In Section 3, we use this theorem to develop an asymptotic approximation for the measurement errors in an ARCH model's estimate of the conditional variances when the data are generated by a diffusion. The class of ARCH models considered is fairly broad, encompassing, for example, the GARCH(1,1) model of Bollerslev (1986), the EGARCH model of Nelson (1991), and the model of Taylor (1986) and Schwert (1989). Section 4 derives asymptotically optimal ARCH conditional variance estimates in the diffusion case. Examples are provided. Section 5 expands the analysis of Sections 3 and 4 to the case when the data are generated by a stochastic *difference* equation rather than a stochastic *differential* equation. Surprisingly, this change makes a considerable difference in the limit theorems and optimality theory. Section 6 compares the filtering properties of several

We assume that x_t is observable at discrete time intervals of length h. y_t is the *unobservable* process controlling the instantaneous conditional variance of x_t, $\sigma(y_t)^2$. Our interest is in using an ARCH model to create an estimate $\hat{y}_t$ of y_t given the discrete observations $(x_0, x_h, x_{2h}, \ldots x_{[t/h]h})$. Our ARCH filtering theory is asymptotic in that we let h approach zero.

For reasons developed below, we also consider a variant of (3.1) in which the drifts in $\{x_t, y_t\}$ explode as $h \downarrow 0$:

$$(3.1') \quad d\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu(x_t, y_t, t) \\ \kappa(x_t, y_t, t) \end{bmatrix} h^{-1/4} dt + \begin{bmatrix} \sigma(y_t)^2 & \rho(x_t, y_t, t)\Lambda(x_t, y_t, t)\sigma(y_t) \\ \rho(x_t, y_t, t)\Lambda(x_t, y_t, t)\sigma(y_t) & \Lambda(x_t, y_t, t)^2 \end{bmatrix}^{1/2} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}.$$

Define $E_t[\cdot]$ to be the conditional expectation given time t information (i.e., given the σ-algebra generated by $\{x_\tau, y_\tau\}_{0 \le \tau \le t}$). We take expectations with respect to this information set, rather than with respect to the smaller information set observed by the econometrician, the σ-algebra generated by x_0, x_h, $x_{2h}, \ldots x_{h[t/h]}$. The $\{x_t, y_t\}$ process is Markovian, so if f in $E_t[f]$ depends only on $\{x_\tau, y_\tau\}_{\tau \ge t}$, then $E_t[f] = E[f | x_t, y_t]$. We will similarly use $SD_t[\cdot]$, $Corr_t[\cdot]$, and $Var_t[\cdot]$ for the conditional standard deviation, correlation and variance.

We now define the normalized innovations in x_t and y_t:

$$(3.2) \quad \xi_{x,t+h} \equiv h^{-1/2}[x_{t+h} - x_t - E_t[x_{t+h} - x_t]], \text{ and}$$

$$(3.3) \quad \xi_{y,t+h} \equiv h^{-1/2}[y_{t+h} - y_t - E_t[y_{t+h} - y_t]].$$

We consider the class of ARCH models which generate $\hat{y}_t$ (for t an integer multiple

of h) by the recursion

(3.4) $\hat{y}_{t+h} = \hat{y}_t + h\cdot\hat{\kappa}(x_t,\hat{y}_t,t,h) + h^{1/2}\cdot g(\hat{\xi}_{x,t+h},x_t,\hat{y}_t,t,h)$,

(3.5) $\hat{\xi}_{x,t+h} \equiv h^{-1/2}[x_{t+h} - x_t - h\cdot\hat{\mu}(x_t,\hat{y}_t,t,h)]$ under (3.1), and

(3.4') $\hat{y}_{t+h} = \hat{y}_t + h^{3/4}\cdot\hat{\kappa}(x_t,\hat{y}_t,t,h) + h^{1/2}\cdot g(\hat{\xi}_{x,t+h},x_t,\hat{y}_t,t,h)$,

(3.5') $\hat{\xi}_{x,t+h} \equiv h^{-1/2}[x_{t+h} - x_t - h^{3/4}\cdot\hat{\mu}(x_t,\hat{y}_t,t,h)]$ under (3.1').

$\hat{\xi}_{x,t+h}$ is a residual obtained using the (in general, incorrect) drift $\hat{\mu}$. The ARCH model (3.4) treats this fitted residual $\hat{\xi}_x$ and fitted $\hat{y}$ *as if* they were the true residual ξ_x and the true y. The ARCH model also treats $h^{1/2}\cdot g(\hat{\xi}_x,x,\hat{y},t,h)$ as if it were the true innovation in the y_t process. Accordingly, we assume that

(3.6) $h^{-1}E[g(\xi_{x,t+h},x_t,y_t,t,h)|x_t=x,y_t=y] \to 0$ as $h\downarrow 0$

uniformly on bounded (x,y,t) sets.

The class of ARCH models encompassed by (3.4)-(3.5) is fairly broad, encompassing, for example, GARCH(1,1), which sets $y \equiv \sigma^2$, $\hat{\kappa} \equiv \omega-\theta\sigma^2$, and $g \equiv \alpha\cdot(\hat{\xi}_x^2-\hat{\sigma}^2)$, AR(1) EGARCH, which sets $y \equiv \ln(\sigma^2)$, $\hat{\kappa} \equiv \beta\cdot[\ln(\hat{\sigma}^2)-\alpha]$, and $g \equiv \theta\hat{\xi}_x/\hat{\sigma}+\gamma[|\hat{\xi}_x/\hat{\sigma}|-E_t|\xi_x/\sigma|]$, and the model of Taylor (1986, Chapter 4) and Schwert (1990), which sets $y \equiv \sigma$, $\hat{\kappa} \equiv \omega-\theta\hat{\sigma}$, and $g \equiv \alpha\hat{\sigma}\cdot(|\hat{\xi}_x/\hat{\sigma}|-E_t|\xi_x/\sigma|)$.

We next define q_t, the normalized measurement error in y. For t an integer multiple of h, we set

(3.7) $q_t \equiv h^{-1/4}(\hat{y}_t - y_t)$.

For general t, we take $q_t \equiv q_{[t/h]h}$, making $\{q_t\}$ a random step function with jumps at time

intervals of length h. Under (3.1), the conditional mean and standard deviation of the increments in (x_t,y_t) over intervals of length h are O(h) and $O(h^{1/2})$, respectively. (3.4)-(3.6) impose this on the increments of $\hat{y}_t$ *when* $\hat{y}_t = y_t$. When $\hat{y}_t = y_t$, $h^{1/2}g(\cdot)$ is the noise term in the increment of $\hat{y}_t$, and $h\cdot\hat{\kappa}$ is the instantaneously predictable component of $\hat{y}_t$ As is standard in ARCH models, $g(\cdot)$ is driven by the fitted residuals $\{\hat{\xi}_t\}$. We assume that $\hat{\kappa}(\cdot)$ and $\hat{\mu}(\cdot)$ are continuous, and that $g(\xi,x,y,t,h)$ is differentiable almost everywhere, possessing one-sided derivatives everywhere.

Why use such an ARCH model to extract estimates of the state rather than, say, a nonlinear Kalman filter? (See, e.g., Kitagawa (1987), Maybeck (1982, Chapter 13).) First, since ARCH models are so widely used in practice, it seems reasonable to investigate their properties. Second, ARCH models are much more computationally tractable than standard nonlinear filters, which are typically infinite dimensional, and which involve extensive numerical integration. When the ARCH model is assumed to be the true data generating process, it is easily fit using maximum-likelihood methods. Finally, we will also see that the optimal ARCH filters are readily interpretable, and an explicit asymptotic distribution theory can be derived for these filters.

Expanding $g(\hat{\xi}_x,x,\hat{y},t,h)$ in a Taylor series around $\hat{\xi} = \xi$, $\hat{y} = y$, and $h = 0$, we obtain under (3.1) and (3.1') respectively

(3.8) $g(\hat{\xi}_{x,t+h},x_t,\hat{y}_t,t,h) \approx g(\xi_{x,t+h},x_t,y_t,t,0) + h^{1/4}q_t\cdot E_t[\partial g(\xi_{x,t+h},x_t,y_t,t,0)/\partial y]$, and

(3.8') $g(\hat{\xi}_{x,t+h}, x_t, \hat{y}_t, t, h) \approx g(\xi_{x,t+h}, x_t, y_t, t, 0) + h^{1/4} q_t \cdot E_t[\partial g(\xi_{x,t+h}, x_t, y_t, t, 0)/\partial y]$

$+ h^{1/4} \cdot (\hat{\mu}(x_t, y_t, t, h) - \mu(x_t, y_t, t)) \cdot E_t[\partial g(\xi_{x,t+h}, x_t, y_t, t, 0)/\partial \xi]$.

Substituting into (3.7), we have under (3.1) and (3.1')

(3.9) $q_{t+h} \approx q_t + h^{1/2} q_t \cdot E_t[\partial g(\xi_{x,t+h}, x_t, y_t, t, 0)/\partial y]$

$+ h^{1/4}[g(\xi_{x,t+h}, x_t, y_t, t, 0) - \xi_{y,t+h}]$ and

(3.9') $q_{t+h} \approx q_t + h^{1/2} q_t \cdot E_t[\partial g(\xi_{x,t+h}, x_t, y_t, t, 0)/\partial y] + h^{1/2}[\kappa(x_t, y_t, t) - \hat{\kappa}(x_t, y_t, t, h)]$

$+ h^{1/2}(\hat{\mu}(x_t, y_t, t, 0) - \mu(x_t, y_t, t, 0)) \cdot E_t[\partial g(\xi_{x,t+h}, x_t, y_t, t, 0)/\partial \xi]$

$+ h^{1/4}[g(\xi_{x,t+h}, x_t, y_t, t, 0) - \xi_{y,t+h}]$.

We also have

(3.10) $y_{t+h} \approx y_t + h \cdot \kappa(x_t, y_t, t) + h^{1/2} \xi_{y,t+h}$ and

(3.11) $x_{t+h} \approx x_t + h \cdot \mu(x_t, y_t, t) + h^{1/2} \xi_{x,t+h}$

under (3.1). (Under (3.1'), replace the $h \cdot \kappa$ and $h \cdot \mu$ by $h^{3/4}\kappa$ and $h^{3/4}\mu$.) Recall that Theorem 2.1 characterized the limit diffusion using the first two conditional moments of the state variables—in our case x_t, y_t, and q_t. Using the approximations (3.9)–(3.11) to characterize these conditional moments requires the following assumptions:

ASSUMPTION 1: *The following functions are well-defined and twice continuously differentiable in x_t and y_t:*

(3.12) $A(x,y,t) \equiv 0$ under (3.1), and

$$A(x,y,t) \equiv [\kappa(x,y,t) - \hat{\kappa}(x,y,t,0)]$$

$$+ \lim_{h\downarrow 0} [\hat{\mu}(x,y,t,h) - \mu(x,y,t)] \cdot E[\partial g(\xi_{x,t+h}, x_t, y_t, t, h)/\partial \xi_x | x_t = x, y_t = y]$$

under (3.1'),

(3.13) $B(x,y,t) \equiv - \lim_{h\downarrow 0} E[\partial g(\xi_{x,t+h}, x_t, y_t, t, h)/\partial y | x_t = x, y_t = y]$,

(3.14) $C(x,y,t) \equiv \lim_{h\downarrow 0} E[g(\xi_{x,t+h}, x_t, y_t, t, h) - \xi_{y,t+h} | x_t = x, y_t = y]^2$.

Further,

(3.15) $h^{-1/2} E[q_{t+h} - q_t | x_t = x, y_t = y, q_t = q] \rightarrow (A(x,y,t) - q \cdot B(x,y,t))$, *and*

(3.16) $h^{-1/2} Var[q_{t+h} - q_t | x_t = x, y_t = y, q_t = q] \rightarrow C(x,y,t)$

as $h\downarrow 0$, where the convergence in (3.15)-(3.16) is uniform on every bounded (x,y,q,t) set.

To simplify notation we shall often substitute A_t, B_t, and C_t for $A(x_t,y_t,t)$, $B(x_t,y_t,t)$, and $C(x_t,y_t,t)$.

ASSUMPTION 2: *For some $\delta > 0$*

(3.17) $E[|h^{-1/2}(y_{t+h} - y_t)|^{2+\delta} | x_t = x, y_t = y] \rightarrow 0$

(3.18) $E[|h^{-1/2}(x_{t+h} - x_t)|^{2+\delta} | x_t = x, y_t = y] \rightarrow 0$

as $h\downarrow 0$, uniformly on every bounded (x,y,t) set, and

(3.19) $\lim\sup_{h\downarrow 0} E[|h^{-1/2} g(\hat{\xi}_{x,t+h}, x_t, \hat{y}_t, t, h)|^{2+\delta} | x_t = x, y_t = y, q_t = q]$ *is bounded uniformly on every bounded (x,y,q,t) set.*

These assumptions are written in the most natural form for applying Theorem 2.1. As we will see in Section 4, they can be verified in many applications.

Changing the Time Scales

In (3.1), $\{x_t\}$, $\{y_t\}$, and $\{q_t\}$ are all scaled to be $O_p(1)$, while the first two conditional moments of $x_{t+h}-x_t$ and $y_{t+h}-y_t$ are $O_p(h)$ as $h\downarrow 0$. In (3.15)-(3.16), on the other hand, the first two conditional moments of $q_{t+h}-q_t$ are $O_p(h^{1/2})$. As $h\downarrow 0$, $\{q_t\}$ oscillates much more rapidly than $\{x_t,y_t\}$. If $\{q_t\}$ mean-reverts (which it will if $E_t[\partial g(\xi_{x,t+h},x_t,y_t,t,0)/\partial y] < 0$), it does so more and more rapidly as $h\downarrow 0$. As we pass from annual observations of x_t to monthly to daily to hourly (and etc.), the rescaled measurement error q_t looks more and more like heteroskedastic white noise (i.e., *not* like a diffusion). In other words, in the limit as $h\downarrow 0$, $\{q_t\}$ *operates on a faster natural time scale* than $\{x_t,y_t\}$.

To use the Stroock-Varadhan results to approximate the behavior of $\{q_t\}$ requires that we change the time scales, which Theorem 2.1 allows via our choice of Δ. Specifically, we choose a time T, a large positive number M, and a point in the state space (x,y,q,t), and condition on the event $(x_T,y_T,q_T) = (x,y,q)$. We then take the vanishingly small time interval $[T,T+M\cdot h^{1/2}]$ on our old time scale (i.e., calendar time) and stretch it into a time interval [0,M] on a new, "fast" time scale. Formally, this involves using $\Delta = 1/2$ in place of $\Delta = 1$ in applying Theorem 2.1. On the usual calendar time scale, $\{x_t,y_t\}$ are a diffusion and $\{q_t\}$ is (asymptotically) white noise. On the new "fast" time scale, the $\{x_t,y_t,t\}$ process moves more and more slowly as $h\downarrow 0$, becoming constant at the values (x_T,y_T,T) in the limit as $h\downarrow 0$ while $\{q_t\}$ is (asymptotically) a diffusion. On the new time scale, therefore, we require two

time subscripts for the $\{q_t\}$ process, one giving the time T on the standard time scale and one giving the time elapsed since T on the "fast" time scale. We therefore write

(3.20) $q_{T,\tau} \equiv q_{T+\tau h^{1/2}}$,

where the τ is the time index on the fast time scale.[5] Our analysis of the measurement error process is therefore *local* in character: in a sense it treats the more slowly varying $\{x_t,y_t,t\}$ as constant at the values (x_T,y_T,T) and examines the behavior of the measurement error in the neighborhood of (x_T,y_T,T).

Figure 1 illustrates this changing of the time scales with artificially generated data. The upper panel in Figure 1A plots a simulated $\{y_t\}_{t=0,h,2h,3h...}$ series from a diffusion model and the corresponding $\{\hat{y}_t\}$ generated by an EGARCH model based on monthly observations of the observable $\{x_t\}_{t=0,h,2h,3h...}$ series.[6] Time is measured in annual units so h = 1/12. In the upper panel of Figure 1B the $\{y_t\}_{t=0,h,2h,3h...}$ and $\{\hat{y}_t\}_{t=0,h,2h,3h...}$ based on daily observations (h = 1/264) are plotted. The measurement error $\{\hat{y}_t - y_t\}$ is smaller for daily than for monthly data (the empirical variances in the simulation are .144 and 1.230 respectively) and is also less autocorrelated at fixed lags of calendar time (e.g., the autocorrelation at a 1 month lag is .219 with daily observations and .477 with monthly

---

5) $q_{T,\tau}$ also depends on the time T startup point (x,y,q,t), though we suppress this in the notation.

6) We used the Wiggins (1987) model, (4.20)-(4.22) below. The model was simulated using the Euler stochastic difference equation approximation with daily increments. We assumed 22 (trading) days per month. The parameter values used were $\mu = 0$, $\Lambda = 2.1$, $\alpha = -3.9$, $\beta = .825$, $\rho = -.69$. $y \equiv \ln(\sigma_t^2)$. The EGARCH models estimated were the asymptotically optimal EGARCH models for this diffusion--see Section 6 below.

observations). We allow for the shrinking variance with the $h^{-1/4}$ term in the definition of q_t. The time deformation allows us to handle the changing serial correlation: for example, set $M \equiv 12^{1/2}$, $T = 20$ and $T+M \cdot h^{1/2} = 21$ for monthly data and call these $\tau = 0$ and $\tau = 1$ on the "fast" time scale. As the upper panels indicate, the interval of *calendar* time between $\tau=0$ and $\tau=1$ shrinks with h, from one year with monthly data to about 2½ months with daily data. The lower panels plot the corresponding $q_{20,\tau}$ processes. The diffusion approximation we derive is for $q_{20,\tau}$ as $h \downarrow 0$.

Making this change of time scale, and conditioning on $(x_T, y_T, q_{T,0}, T)$ it is straightforward to derive the limit diffusion of $\{q_{T,\tau}\}_{h \downarrow 0}$:

$$(3.21) \quad dq_{T,\tau} = (A_T - B_T q_{T,\tau})d\tau + C_T^{1/2} dW_\tau,$$

where W_τ is a standard Brownian motion (on the "fast" time scale). Since the distributions of $\xi_{x,T+h}$ and $\xi_{y,T+h}$ are functions of x_t, y_t, t, and h, B_T and C_T are functions only of y_T x_T, T, and h, and are constant (conditional on y_T, x_T, and T) in the diffusion limit on the fast time scale. On the fast time scale, $\{q_{T,\tau}\}$ follows an Ornstein-Uhlenbeck process, the continuous time equivalent of a Gaussian AR(1).

THEOREM 3.1: *Let Assumptions 1-2 be satisfied, and let T and τ be strictly positive, finite numbers. Let Θ be any bounded, open subset of R^4 on which for some $\epsilon > 0$ and all $(x,y,q,T) \in \Theta$, $|A(x,y,T)| < 1/\epsilon$, $\epsilon < B(x,y,T) < 1/\epsilon$ and $C(x,y,T) < 1/\epsilon$. Then for every $(x,y,q,T) \in \Theta$, $\{q_{T,\tau}\}_{[0,M]}$ (conditional on $(x_T, y_T, q_T)=(x,y,q)$) converges weakly to the*

diffusion (3.21) as h↓0. This convergence is uniform on Θ.

Proof: see the Appendix.

COROLLARY: *Under the conditions of Theorem 3.1, for every* $(x_T, y_T, q_T, T) \in \Theta$,

(3.22) $[q_{T,M} \mid (x_T, y_T, q_{T,0}) = (x,y,q)] \xrightarrow{d}$

$$N[A_T/B_T + e^{-M \cdot B_T}(q_{T,0} - A_T/B_T), C_T(1 - e^{-2M \cdot B_T})/2B_T],$$

where "$\xrightarrow{d}$" *denotes convergence in distribution as h↓0.*

Interpretation

If $B_T > 0$, then for large M (recall that we can make M as large as we like as long as it is finite), $q_{T,M}$ *given* x_T *and* y_T, is approximately $N[A_T/B_T, C_T/2B_T]$. (Though $\lim_{M\to\infty} \lim_{h\downarrow 0} [q_{T,M} \mid (x_T, y_T, q_{T,0}) = (x,y,q)] \stackrel{d}{=} N(A_T/B_T, C_T/2B_T)$, Theorem 2.1 does not allow us to interchange the limits or to make M a function of h such that $M(h) \to \infty$ as h↓0.) Four comments are in order:

First, under the conditions of Theorem 3.1, $[\hat{y}_t - y_t]$ is $O_p(h^{1/4})$. Although this *rate* of convergence ($O_p(h^{1/4})$) is the same throughout the state space (whenever the conditions of the theorem are satisfied), the asymptotic variance of the measurement error is a function of B_T and C_T and therefore of x_T, y_T, and T. The $O_p(h^{1/4})$ rate seems fairly slow, implying, for example, that in going from annual to daily returns data the standard deviation of the measurement error falls by about a factor of four. If the variance per unit of time were *constant*, we could achieve an $O_p(h^{1/2})$ rate of convergence (see, e.g., Merton (1980)),

and the standard deviation of the measurement error would fall by a factor of 16. Our slower convergence rate results from the fact that the ARCH variance estimators are shooting at a rapidly oscillating target.

Second, Theorem 3.1 analyzes the local behavior of $\{q_{T,\tau}\}$ in the neighborhood of the time T state values (x_T, y_T, q_T). This analysis may not be particularly useful if q_T is exploding as $h \downarrow 0$, as it would if $[\hat{y}_T - y_T]$ converges to zero at a rate slower that $h^{1/4}$. A technical device gets us around this problem: in related work (Foster and Nelson (1991)), we show that under mild regularity conditions, rolling regression estimators achieve the $O_p(h^{1/4})$ convergence rate for $[\sigma(\hat{y}_t)^2 - \sigma(y_t)^2]$. A rolling regression at the beginning of a sample—say from dates $T\text{-}h^{1/2}$ to T—can be used to initialize the ARCH filter at time T.[7]

Third, Theorem 3.1 also allows us to characterize the asymptotic autocorrelation: using the autocorrelation function of the Ornstein-Uhlenbeck (see Arnold (1973 Section 8.3)) we have for small h and large positive τ and τ'

$$(3.23)\quad \mathrm{Corr}(q_{T+\tau h^{1/2}}, q_{T+\tau' h^{1/2}}) \approx \exp[-|\tau-\tau'| \cdot B_T].$$

According to (3.23), $\{q_t\}$ is asymptotically white noise on the standard (calendar, slow) time scale, since the (asymptotic) serial correlation in the measurement errors vanishes except at lag lengths shrinking to zero at rate $O(h^{1/2})$. It is conditionally heteroskedastic white noise, however, since the asymptotic variance of q_t depends on x_t, y_t

---

7) Formally, to guarantee that $q_t = O_p(h^{1/4})$ for all t, choose a large positive N and define $\hat{y}_T$ by $\hat{y}_T \equiv \hat{y}_T(\mathrm{ARCH})$ whenever $|\hat{\sigma}_T^2(\mathrm{ARCH}) - \hat{\sigma}_T^2(\mathrm{Rolling\ Regression})| < h^{1/4}N$, and $\hat{y}_T \equiv \hat{\sigma}_T^2(\mathrm{Rolling\ Regression})$ otherwise.

and t.

Fourth, under (3.1), the asymptotic distribution of the measurement error process $\{q_t\}$ depends on $\rho(\cdot)$, $\sigma(\cdot)$, $\Lambda(\cdot)$ and $g(\cdot)$, but not on $\hat{\kappa}$, λ, μ, or $\hat{\mu}$. While errors in the drift terms $\hat{\kappa}$, λ, μ, or $\hat{\mu}$ affect $\{q_t\}$ for fixed $h > 0$, they are asymptotically negligible as $h \downarrow 0$. (3.1') blows up these drift terms at an appropriate rate to keep them from dropping out of the asymptotic distribution. In this "large drift" asymptotic, nonzero $[\hat{\kappa} - \kappa]$ and $[\mu - \hat{\mu}]$ create an asymptotic bias in $q_t \equiv h^{-1/4}[\hat{y} - y]$, but do not affect its asymptotic variance. This confirms the intuition given in Nelson (1992) that seriously misspecified ARCH models may consistently extract conditional variances from data observed at higher and higher frequencies—i.e., consistency is not incompatible with (moderately) explosive (as $h \downarrow 0$) misspecification in the drifts.

To gauge the size of the bias, note that blowing up $[\hat{\kappa} - \kappa]$ and $[\mu - \hat{\mu}]$ at an $h^{-1/4}$ rate introduced an $O(h^{1/4})$ bias in the measurement error $[\hat{y} - y]$, suggesting that the effect of *non*-exploding drifts introduces an $O(h^{1/2})$ bias.

4. ASYMPTOTIC OPTIMALITY

In discussing optimal ARCH model selection, several warnings are in order:

First, we consider optimality only within the class of ARCH models given by (3.4)-(3.5) and subject to the regularity conditions in the Assumptions.

Second, we evaluate optimality in terms of the approximate asymptotic bias A_T/B_T.

and approximate asymptotic variance $C_T/2B_T$. As is well-known in standard large-sample asymptotics, minimizing asymptotic variance need not be the same as minimizing the limit of the variances. More importantly, Theorems 3.1 allows us to make M arbitrarily large, but does not allow us to take M to infinity as $h \downarrow 0$, so A_T/B_T and $C_T/2B_T$ do not exactly correspond to the bias and variance delivered by Theorem 3.1, though they become arbitrarily close for large M.

Third, much the same difficulty arises in defining *globally* optimal ARCH filters as arises in defining globally optimal estimators in statistics: just as the "optimal" estimate of a parameter Ψ is Ψ itself, the "optimal" ARCH model when $(y_T,x_T,T) = (y,x,t)$ is the model in which $\hat{y}_{T,\tau}$ is held *constant* at y. Even if y_t is randomly changing, the estimation error would be $O_p(h^{\Delta})$ (note the faster rate of convergence when $\Delta > 1/4$) in an $O(h^{\Delta})$ neighborhood of $y_t = y$. Obviously, in other regions of the state space such an estimate would perform disastrously. We call the ARCH model globally optimal if it eliminates the (approximate) asymptotic bias A_T/B_T—even in the "fast drift" case—and minimizes the (approximate) asymptotic variance $C_T/2B_T$ for every (x,y,t). Hence our optimality concept is patterned on the UMVUE (uniform minimum variance unbiased estimator) criterion.

Eliminating Asymptotic Bias

From (3.12) and Theorem 3.1, it is clear that our first condition for global optimality, elimination of the asymptotic bias A_T/B_T, can be achieved by setting $\hat{\kappa}(x,y,t,h) = \kappa(x,y,t)$

and $\hat{\mu}(x,y,t,h)=\mu(x,y,t)$ for all (x,y,t). Though this choice of $\hat{\kappa}(x,y,t,h)$ and $\hat{\mu}(x,y,t,h)$ is *sufficient* to eliminate asymptotic bias, it need not be *necessary*, since it is possible that bias from $\hat{\mu}(x,y,t,h) \neq \mu(x,y,t)$ exactly offsets bias from $\hat{\kappa}(x,y,t,h) \neq \kappa(x,y,t)$.

Minimizing Asymptotic Variance

Suppose for the moment that the conditional density of $\xi_{x,T+h}$, say $f(\xi_{x,T+h}|x_T,y_T)$, is well-defined and is differentiable in y. Integrating by parts then allows us to write B_T as

(4.1) $B_T = \lim_{h\downarrow 0} -E_T[\partial g(\xi_{x,T+h},x_T,y_T,T,h)/\partial y]$

$= \lim_{h\downarrow 0} E_T[g(\xi_{x,T+h},x_T,y_T,T,h)\cdot \partial \ln[f(\xi_{x,T+h}|x_T,y_T,T,h)]/\partial y]$.

Under (3.1) or (3.1') the increments in $\{x_{t+h}-x_t, y_{t+h}-y_t\}$ approach conditional normality as $h\downarrow 0$ (see, e.g., Stroock and Varadhan (1979, pp. 2-4)), and $(\xi_{x,t+h},\xi_{y,t+h})$ is approximately (conditionally) bivariate normal with mean 0 and variances $\sigma(y_t)^2$ and $\Lambda(x_t\, y_t\, ,t)^2$, and correlation $\rho(x_t,y_t,t)$. If $f(\xi_{x,T+h}|x_T,y_T,T)$, $g(\xi_x,x,y,t)$ and their partial derivatives with respect to y are sufficiently well-behaved, (4.2)-(4.4) will hold: define

$$(4.2)\quad \begin{bmatrix} \epsilon_x \\ \epsilon_y \end{bmatrix} \overset{d}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma(y)^2 & \rho(x,y,t)\Lambda(x,y,t)\sigma(y) \\ \rho(x,y,t)\Lambda(x,y,t)\sigma(y) & \Lambda(x,y,t)^2 \end{bmatrix} \right).$$

(i.e., ϵ_x and ϵ_y are bivariate normal with mean and covariances given by (4.2).) Then

(4.3) $B(x,y,t) = \sigma(y)^{-3}\sigma'(y)E[\epsilon_x^2 g(\epsilon_x,x,y,t,0)]$, and

(4.4) $C(x,y,t) = E[(g(\epsilon_x,x,y,t,0)-\epsilon_y)^2]$.

(4.2)-(4.4) turn out to be important technical conditions in deriving the asymptotic variance

minimizer, as does a stronger version of (3.18):

ASSUMPTION 3: *For every (x,y,t), (4.2)-(4.4) hold, and for some $\delta > 0$,*

(4.5) lim sup$_{h\downarrow 0}$ $E[|h^{-1/2}(x_{t+h}-x_t)|^{4+\delta}|x_t=x,y_t=y]$ is uniformly bounded on every bounded (x,y,t) set.

The conditional density of $\xi_{x,t+h}$ can, in principle, be derived from the Kolmogorov backward or forward equations (see, e.g., Stroock and Varadhan (1979, Chapters 2-3, 9-10)). In practice, it is easier to check Assumption 3 directly than to establish convergence for the derivatives of the conditional density.

THEOREM 4.1: *Let Assumptions 1-3 hold. Under either (3.1) or (3.1'), the approximate asymptotic variance $C_T/2B_T$ is minimized by setting*

$$(4.6)\quad g(\xi_x,x,y,t,h) = \left[\frac{\rho(x,y,t)\Lambda(x,y,t)\xi_x}{\sigma(y)}\right] + \left[\frac{\Lambda(x,y,t)[1-\rho(x,y,t)^2]^{1/2}}{2^{1/2}}\right]\left[\frac{\xi_{x_2}}{\sigma(y)^2}-1\right].$$

The minimized $C_T/2B_T = \dfrac{[1-\rho(x_T,y_T,T)^2]^{1/2}\Lambda(x_T,y_T,T)\sigma(y_T)}{[2^{1/2}\sigma'(y_T)]}$,

and the corresponding approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2-\sigma(y)^2]$ *equals* $[1-\rho(x_T,y_T,T)^2]^{1/2}\Lambda(x_T,y_T,T)\sigma(y_T)^3\sigma'(y_T)8^{1/2}$.

Proof: see Appendix.

Sometimes it is of interest to minimize $C_T/2B_T$ within a particular class of ARCH

models (e.g., to find the optimal GARCH(1,1) model). Accordingly, we consider models in which

(4.7) $g(\xi_x,x,y,t,h) = a(x,y,t,h)\cdot g^*(\xi_x,x,y,t,h)$, where

(4.8) $E_t[g^*(\xi_{x,t+h},x_t,y_t,t,h)] = 0$, $Var_t[g^*(\xi_{x,t+h},x_t,y_t,t,h)] = 1$,

and $E_t[\partial g^*(\xi_{x,t+h},x_t,y_t,t,0)/\partial y] < 0$.

We now treat $g^*(\cdot)$ as given and optimize over $a(\cdot)$.

THEOREM 4.2: *The approximate asymptotic variance $C_T/2B_T$ is minimized subject to the constraints (4.7)-(4.8) by setting*

(4.9) $a(x,y,t,h) \equiv \Lambda(x,y,t)$.

The minimized $C_T/2B_T$ equals

$$(4.10) \qquad \frac{\sigma(y_T)^3[\Lambda(x_T,y_T,T) - Cov_T[\xi_{y,T+h},g^*(\xi_{x,T+h},x_T,y_T,T,0)]]}{\sigma'(y_T)E_T[\xi_{x,T+h}^2 g^*(\xi_{x,T+h},x_T,y_T,T,0)]},$$

and the corresponding approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y}_T)^2-\sigma(y_T)^2]$ is

$$4\sigma'(y_T)\sigma(y_T)^5 \frac{[\Lambda(x_T,y_T,T)-Cov_T[\xi_{y,T+h},g^*(\xi_{x,T+h},x_T,y_T,T,0)]]}{E_T[\xi_{x,T+h}^2 g^*(\xi_{x,T+h},x_T,y_T,T,0)]}.$$

Proof: see Appendix.

Interpretations

Since $\{x_t,y_t\}$ is generated by a diffusion, the increments $x_{t+h}-x_t$ and $y_{t+h}-y_t$ are approximately conditionally normal for small h. The first term in (4.6), $\rho(x,y,t)\Lambda(x,y,t)\xi_x/\sigma(y)$

is $E[\xi_{y,t+h}|\xi_{x,t+h},x_t,y_t]$ using the (limiting) conditionally normal distribution. Given the information in $\xi_{x,t+h}$, (4.6) optimally forecasts the innovation in y_{t+h}, $h^{1/2}\xi_{y,t+h}$.

To understand the second term in (4.6), consider y_t as an unknown parameter in the conditional distribution of $\xi_{x,t+h}$. Given y_t, $\xi_{x,t+h}$ is approximately $N[0,\sigma(y_t)^2]$. We may write the (limiting) loglikelihood as

$$(4.11)\quad \ln(f(\xi_{x,t+h}|y_t)) = -.5\cdot\ln(2\pi) - \ln[\sigma(y_t)] - .5\cdot\xi_{x,t+h}^2/\sigma(y_t)^2,$$

so the score is

$$(4.12)\quad \frac{\partial \ln(f(\xi_{x,t+h}|x_t,y_t,y))}{\partial y} = \left[\frac{\xi_{x,t+h}^2}{\sigma(y_t)^2}-1\right]\frac{\sigma'(y_t)}{\sigma(y_t)^2}.$$

For a given x, y, and t, the second term on the right hand side of (4.6) is proportional to the score. As in maximum likelihood estimation, $\hat{y}_t$ is moved up when the score is positive and is moved down when the score is negative.[8]

Consider the problem of predicting y_{t+h} given x_t, x_{t-h}, x_{t-2h},.... There are two sources of uncertainty about y_{t+h}: first, uncertainty about $y_{t+h}-y_t$, i.e., uncertainty about *changes* in y_t. Second, there is uncertainty about the *level* of y_t. These two sources are asymptotically of the same order, $O_p(h^{1/4})$. The first term on the right side of (4.6) optimally extracts information about $y_{t+h}-y_t$ contained in $x_{t+h}-x_t$. The second term, in a manner analogous to maximum likelihood estimation, extracts information about y_t itself.

---

8) This is merely a heuristic: convergence in distribution of $[\xi_x,\xi_y]$ does not, of course, imply convergence of the conditional density or of its derivative, and we do not need to prove such convergence to verify Assumptions 1-3.

Conditional Moment Matching and the Connection with Consistent Forecasting

Nelson and Foster (1991) develop conditions under which the forecasts generated by a misspecified ARCH model approach the forecasts generated by the true model as a continuous time limit is approached. For example, suppose the diffusion (3.1)-(3.3) generates the data and the misspecified ARCH model (3.4)-(3.5) is used to estimate y_t and to make probabalistic forecasts about the future path of $\{x_t,y_t\}$. In particular, suppose the ARCH model is used in forecasting *as if* it were the true model—i.e., as if, instead of (3.1) we had

$$(4.13)\quad x_{t+h} = x_t + h\cdot \hat{\mu}(x_t,y_t,t,h) + h^{1/2}\cdot \xi_{x,t+h}$$

$$(4.14)\quad y_{t+h} = y_t + h\cdot \hat{\kappa}(x_t,y_t,t,h) + h^{1/2}\cdot g(\xi_{x,t+h},x_t,y_t,t,h),$$

where $\xi_{x,t+h}|x_t,y_t,t \sim N[0,\sigma(y_t)^2]$. Under what circumstances do forecasts generated by this misspecified model approach forecasts generated by (3.1)-(3.3) as $h\downarrow 0$? It turns out that these conditions are closely related to the conditions for asymptotically efficient filtering. In particular, the conditions for consistent forecasting include the first-moment-matching requirement that $\hat{\kappa}(x,y,t,h) = \kappa(x,y,t)$ and $\hat{\mu}(x,y,t,h) = \mu(x,y,t)$ for all (x,y,t). As we saw earlier, this is a sufficient condition to eliminate asymptotic bias in the "large drift" case. We also require that the second moments are matched in the limit as $h\downarrow 0$—i.e., that

$$(4.15)\quad Cov_t\begin{bmatrix}\xi_x\\ \xi_y\end{bmatrix} = Cov_t\begin{bmatrix}\xi_x\\ g(\xi_x,x,y,t,0)\end{bmatrix} = \begin{bmatrix}\sigma(y)^2 & \rho(x,y,t)\Lambda(x,y,t)\sigma(y)\\ \rho(x,y,t)\Lambda(x,y,t)\sigma(y) & \Lambda(x,y,t)^2\end{bmatrix}.$$

Using the approximate bivariate normality of ξ_x and ξ_y, it is easy to show that the

asymptotically optimal $g(\cdot)$ of Theorem 4.1 satisfies (4.15).

What accounts for this moment matching property of optimal ARCH filters? Recall that the first two conditional moments of $\{x_t,y_t\}$ (along with properties (a) and (d) of Section 2) characterize the distribution of the process. Optimal ARCH filters make themselves *as much like the true data generating process as possible.* Since the first two conditional moments characterize the true data generating process in the continuous time limit, the optimal ARCH filters match these two moments.[9]

Interestingly, misspecification in the drifts κ and μ has only a second-order ($O_p(h^{1/2})$) affect on filtering (and hence on one-step ahead forecasting) but has a first-order affect on many-step-ahead forecasting performance. Over short time intervals, diffusions act like driftless Brownian motions, with the noise swamping the drift. In the medium and long-term, however, the drift exerts a crucial impact on the process.

Invariance to the Definition of y_t

There is considerable arbitrariness in our definition of y_t. Suppose, for example, that we define $\tilde{y}_t \equiv \tilde{y}(y_t)$ for some monotone increasing, twice continuously differentiable function $\tilde{y}(\cdot)$. We could then apply Ito's Lemma to (3.1), re-writing it as a stochastic integral equation in x_t and $\tilde{y}_t$. If the regularity conditions of Theorem 4.1 are satisfied, the Theorem yields an asymptotically optimal filter for the new system. Is it possible to reduce

9) This is ignoring the (pathological) case when biases from $\hat{\kappa} \neq \kappa$ and from $\hat{\mu} \neq \mu$ exactly cancel.

the asymptotic variance of $[\hat{\sigma}_t^2-\sigma_t^2]$ by a judicious choice of $\tilde{y}(\cdot)$? Using Ito's Lemma it is easy to verify that the answer is no, provided that the regularity conditions are satisfied for both the (x_t,y_t) and $(x_t,\tilde{y}_t)$ systems: the $\sigma'(y)$ in the asymptotic variance of $[\hat{\sigma}_t^2-\sigma_t^2]$ in Theorem 4.1 is replaced by $\sigma'(y)/[\partial\tilde{y}(y_t)/\partial y]$ in the $\tilde{y}$ system, but the $\Lambda(x,y,t)$ is replaced by $\Lambda(x,y,t)\cdot[\partial\tilde{y}(y_t)/\partial y]$, leaving the asymptotic variance of $[\hat{\sigma}_t^2-\sigma_t^2]$ unchanged. Within the limits of the regularity conditions, the definition of y_t is arbitrary.

Examples

The asymptotically optimal ARCH filter of Theorem 4.1 looks unlike ARCH models commonly used in the literature. GARCH(1,1) is an exception. Suppose the data are generated by the diffusion

(4.16) $dx_t = \mu\cdot dt + \sigma_t dW_{1,t}$

(4.17) $d\sigma_t^2 = (\omega - \theta\sigma_t^2)dt + 2^{1/2}\alpha\sigma_t^2 dW_{2,t}$,

where $W_{1,t}$ and $W_{2,t}$ are independent standard Brownian motions. If we set $y_t\equiv\sigma_t^2$, this is in the form of (3.1). The asymptotically optimal filter of Theorem 4.1 sets $\hat{\mu} = \mu$ and

(4.18) $\hat{\sigma}_{t+h}^2 = \hat{\sigma}_t^2 + (\omega - \theta\hat{\sigma}_t^2)\cdot h + h^{1/2}\alpha(\hat{\xi}_{x,t+h}^2 - \hat{\sigma}_t^2)$,

which is recognizable as a GARCH(1,1) when we re-write (4.18) as

(4.19) $\hat{\sigma}_{t+h}^2 = \omega\cdot h + (1 - \theta\cdot h - \alpha\cdot h^{1/2})\hat{\sigma}_t^2 + h^{1/2}\alpha\hat{\xi}_{x,t+h}^2$.

THEOREM 4.3: *(4.16)-(4.18) satisfy the conditions of Theorems 3.1 and 4.1.*

Proof: see Appendix.

If $dW_{1,t}dW_{2,t} = \rho dt$, $\rho \neq 0$ (i.e., $W_{1,t}$ and $W_{2,t}$ are correlated) GARCH(1,1) is no longer optimal: the second moment matching condition (4.15) fails, since $Corr_t[\alpha(\xi_{x,t+h}^2 - \sigma_t^2), \xi_{x,t+h}]$ is zero not ρ. A modification of GARCH(1,1) proposed by Engle and Ng (1991, equation (11)) can be shown to be optimal in this case.

To further illustrate the construction of globally optimal ARCH models, we next consider two models from the option pricing literature. In each model, S_t is a stock price and σ_t is its instantaneous returns volatility. We observe $\{S_t\}$ at discrete intervals of length h. In each model, we have

$$(4.20) \quad dS_t = \mu S_t dt + S_t \sigma_t dW_{1,t}.$$

The first model (see Wiggins (1987), Hull and White (1987), Melino and Turnbull (1990), and Scott (1987)) sets

$$(4.21) \quad d[\ln(\sigma_t^2)] = -\beta[\ln(\sigma_t^2) - \alpha]dt + \psi \cdot dW_{2,t},$$

where $W_{1,t}$ and $W_{2,t}$ are standard Brownian motions independent of (S_0, σ_0^2) with

$$(4.22) \quad \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix} \left[dW_{1,t} \; dW_{2,t} \right] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} dt.$$

μ, ψ, β, and α are constants.

Bates and Pennacchi (1990), Gennotte and Marsh (1991), and Heston (1991) propose a second model, which replaces (4.21) with

$$(4.23) \quad d\sigma_t^2 = -\beta[\sigma_t^2 - \alpha]dt + \psi \cdot \sigma_t \cdot dW_{2,t}.$$

Here σ_t^2 is generated by the "square root" diffusion popularized as a model of the short-

term interest rate by Cox, Ingersoll, and Ross (1985).

Now consider ARCH filtering of these models. Suppose we define $x_t \equiv \ln(S_t)$ and $y_t \equiv \ln(\sigma_t^2)$. We may then re-write (4.20)-(4.21) as

(4.20') $dx_t = (\mu - \exp(y_t)/2)dt + \exp(y_t/2)dW_{1,t}$,

(4.21') $dy_t = -\beta[y_t - \alpha]dt + \psi \cdot dW_{2,t}$.

THEOREM 4.4: *The asymptotically optimal ARCH model for the model (4.20')-(4.21') and (4.22) is*

(4.24) $\hat{\mu}(x, \hat{y}) \equiv \mu - exp(\hat{y}/2)$

(4.25) $\hat{y}_{t+h} = \hat{y}_t - \beta[\hat{y}_t - \alpha] \cdot h$

$+ h^{1/2}\psi[\rho\, \hat{\xi}_{x,t+h} \cdot exp(-\hat{y}_t/2) + [(1-\rho^2)/2]^{1/2} \cdot (\hat{\xi}^2_{x,t+h} \cdot exp(-\hat{y}_t) - 1)]$.

(4.20')-(4.21') and (4.24)-(4.25) satisfy Assumptions 1-3. The resulting (minimized) approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2 - \sigma(y)^2]$ *is* $[2(1-\rho^2)]^{1/2}h^{1/2}\psi\sigma^4$.

Proof: see Appendix.

Next consider the model given by (4.20) and (4.22)-(4.23). Using Ito's Lemma and $y \equiv \ln(\sigma^2)$, (4.23) becomes

(4.23') $dy_t = (-\beta + \exp(-y_t)[\beta\alpha - \psi^2/2])dt + \psi \cdot \exp(-y_t/2) \cdot dW_{2,t}$.

The asymptotically optimal filter suggested by Theorem 4.1 is

(4.26) $\hat{y}_{t+h} = \hat{y}_t + (-\beta + \exp(-\hat{y}_t)[\beta\alpha - \psi^2/2]) \cdot h$

$+ h^{1/2}\psi\exp(-\hat{y}_t/2)[\rho \cdot \hat{\xi}_{x,t+h} \cdot \exp(-\hat{y}_t/2) + [(1-\rho^2)/2]^{1/2} \cdot (\hat{\xi}^2_{x,t+h} \cdot \exp(-\hat{y}_t) - 1)]$.

Unfortunately, the regularity conditions break down at $y = -\infty$ (i.e., at $\sigma^2 = 0$). (In fact, the

stochastic differential equation (4.23') is not well-defined in this case, although (4.23) is.) This is not a problem in the theorem as long as the boundary $y_t=-\infty$ is unattainable in finite time. When $2\beta\alpha \leq \psi^2$, however, this boundary is attained in finite time with positive probability (see Cox, Ingersoll and Ross (1985).) We therefore exclude this case.

THEOREM 4.5: *Let $2\beta\alpha > \psi^2$. The asymptotically optimal model for (4.20') and (4.22')-(4.23') is given by (4.26). (4.20'), (4.22'), (4.23) and (4.26) satisfy Assumptions 1-3. The resulting (minimized) approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2-\sigma(y)^2]$ is $[2(1-\rho^2)]^{1/2}h^{1/2}\psi\sigma^3$.*

Proof: see Appendix.

The differences in the optimal filters for the two models are most easily understood in terms of the moment matching conditions: In (4.21) and its associated optimal ARCH model, the conditional variance of σ_t^2 rises linearly with σ_t^4, while in (4.23) the conditional variance of σ_t^2 rises linearly with σ_t^2. As we will see in Section 6, most commonly-used ARCH models effectively assume that the "variance of the variance" rises linearly with σ_t^4. If (4.23) generates the data, GARCH, EGARCH, and other such models will be very inefficient filters when σ^2 is very low or very high, since the $g(\cdot)$ functions in these ARCH models cannot match the ARCH and true "variance of the variance" everywhere in the state space.

5. NEAR DIFFUSIONS

In this section we consider the case in which the data are generated in discrete time by the stochastic volatility model

$$(5.1)\quad \begin{bmatrix} x_{t+h} \\ y_{t+h} \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} \mu(x_t,y_t,t,h) \\ \kappa(x_t,y_t,t,h) \end{bmatrix} h + \begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \end{bmatrix},$$

or, in the "fast drift" case analogous to (3.1')

$$(5.1')\quad \begin{bmatrix} x_{t+h} \\ y_{t+h} \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} \mu(x_t,y_t,t,h) \\ \kappa(x_t,y_t,t,h) \end{bmatrix} h^{3/4} + \begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \end{bmatrix},$$

for some (small) h > 0, where

$$(5.2)\quad E_t\begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad Cov_t\begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \end{bmatrix} = \begin{bmatrix} \sigma(y_t)^2 & \rho(x_t,y_t,t)\Lambda(x_t,y_t,t)\sigma(y_t) \\ \rho(x_t,y_t,t)\Lambda(x_t,y_t,t)\sigma(y_t) & \Lambda(x_t,y_t,t)^2 \end{bmatrix}.$$

We assume further that the process $\{x_t,y_t\}_{t=0,h,2h...}$ is Markovian. Again, in (5.1)-(5.2), "t" is assumed to be a discrete multiple of h. To define the process for general t, set $(x_t,y_t) \equiv (x_{h[t/h]},y_{h[t/h]})$. This makes each $\{x_t,y_t\}$ process a step function with jumps at discrete intervals of length h. "E_t" and "Cov_t" denote, respectively, expectation and covariances conditional on time t information——i.e., the σ-algebra generated by $\{x_\tau\,_\tau\}_{0\le\tau\le t}$ or, equivalently for our purposes, by $(x_t,y_t,\hat{y}_t,t)$. Note that the structure of the first two conditional moments of x_t and y_t in (5.1)-(5.2) are the same as in (3.1)-(3.2). In fact, under (5.1) and the regularity conditions assumed below, $\{x_t,y_t\}$ converges weakly to the diffusion

(3.1)-(3.2) as $h \downarrow 0$. We therefore call such a $\{x_t, y_t\}$ processes a *near diffusion*.

Why the Near Diffusion Case is Important

At first glance, it would seem that there is little gain to generalizing the results of Section 3 to the near diffusion case. The intuition is this: the estimated conditional variance process $\{\sigma(\hat{y}_t)^2\}$ is a functional of the sample path of the $\{x_t\}$ of (5.1)-(5.2). The $\{x_t\}$ of (5.1)-(5.2) converges weakly to the $\{x_t\}$ of (3.1)-(3.2) as $h \downarrow 0$. If the mapping from $\{x_t\}$ to $\{\sigma(\hat{y}_t)^2\}$ is sufficiently well-behaved, the continuous mapping theorem should guarantee that $\{\sigma(\hat{y}_t)^2\}$ converges to the limit derived in Section 3 as $h \downarrow 0$, yielding the same results on efficiency and etc. as in the diffusion case.

Unfortunately, this intuition is wrong, since the mapping from $\{x_t\}$ to $\{\sigma(\hat{y}_t)^2\}$ is not at all well behaved in the sense required by the continuous mapping theorem. To see why, consider the case of iid residuals. Let

$$x_{t+h} = x_t + h^{1/2}\xi_{t+h} \tag{5.3}$$

where $x_0 = 0$ and for all t and all h, ξ_t is iid with mean 0 and variance σ^2. Consider the least squares estimator of σ^2, given at time t by

$$\hat{\sigma}_{t+h}^2 \equiv (t/h)^{-1} \sum_{j=0,h,2h...}^{t/h} h^{-1}[x_{(j+1)h} - x_{jh}]^2 = (t/h)^{-1} \sum_{j=0,h,2h...}^{t/h} \xi_j^2. \tag{5.4}$$

Standard invariance arguments (e.g., using Donsker's theorem: see Jacod and Shiryaev (1987)) show that as $h \downarrow 0$, $\{x_t\}$ converges weakly to the limit process given by

(5.5) $x_t = \sigma W_{1,t}$,

where $W_{1,t}$ is a standard Brownian motion. *This holds regardless of the distribution of* ξ_t, *provided* ξ_t *is iid with mean 0 and variance* σ^2. If $E[\xi_t^4] < \infty$, $\{(t/h)^{1/2}(\hat{\sigma}_{t+h}^2 - \sigma^2)\}$ also converges weakly, to the process $\{\psi_t\}$ with

(5.6) $\psi_t \equiv \sigma^2(E[\xi_t^4/\sigma^4] - 1)^{1/2}W_{2,t}$,

where $W_{2,t}$ is a second standard Brownian motion. *Note that the diffusion limit of* $\{x_t\}$ *does not depend on the distribution of* ξ_t/σ, *but the diffusion limit of* $\{\psi_t\}$ *does, through the fourth moment of* ξ_t/σ.

Suppose, for example, that $\{x_t\}$ is a Brownian motion observed at time intervals of length h, so $\xi_t \sim N(0,\sigma^2)$. Here $(E[\xi_t^4/\sigma^4] - 1) = 2$. (This is the case analogous to (3.1)-(3.2).) Moving from the diffusion to the near diffusion case by changing the distribution of ξ_t can have drastic consequences for $\hat{\sigma}_t$. For example, let ξ_t equal σ with probability 1/2 and equal $-\sigma$ with probability 1/2. In this case, $(E[\xi_t^4/\sigma^4] - 1) = 0$. In fact, a single observation is sufficient to recover σ with no error. On the other hand, suppose ξ_t is iid Student's t with 3 degrees of freedom and variance σ^2. Now $E[\xi_t^4] = \infty$ and $\{\psi_t\}$ fails to converge. In each of these cases, the diffusion limit of $\{x_t\}$ is the same, while the limit of $\{\psi_t\}$ is not. Even in this iid case, there is a crucial difference between the behavior of the variance estimate for the limit diffusion and the variance estimate for a sequence of process converging to the diffusion. As we will see below, this remains true in the more general

ARCH case as well.[10]

Many empirical studies of asset market volatility have found that returns remain somewhat thick tailed even after conditional heteroskedasticity is accounted for (e.g., Baillie and Bollerslev (1989), Nelson (1989,1991)). Near diffusions easily accommodate this. The diffusion case examined in Section 3, on the other hand, effectively assumes conditional normality for sufficiently small h. The near diffusion case is therefore likely to be practically important, and will allow us to consider optimality for different conditional distributions and the robustness of different ARCH models to the presence of conditionally thick tailed residuals.

Main Results

THEOREM 5.1: *Let Assumptions 1-2 hold with (5.1) and (5.1') replacing (3.1) and (3.1'). Then the statement of Theorem 3.1 holds.*

Proof: see Appendix.

The interpretation of the optimal filter in terms of an estimation component and a forecasting component applies in the near diffusion context also. We accordingly define the prediction component

---

10) Here is another heuristic: for Brownian motions, the instantaneous variance is the stochastic derivative of the quadratic variation of the process. Since Brownian motions are continuous with probability one, the Weierstrass theorem guarantees that they can be approximated to arbitrary accuracy with finite-order polynomials. Yet the quadratic variation of such a polynomial is always zero.

(5.6) $P(\xi_x,x,y,t,h) \equiv E[\xi_{y,t+h}|(\xi_{x,t+h},x_t,y_t,t,h)=(\xi_x,x,y,t,h)]$,

and the estimation (or score) component

(5.7) $S(\xi_x,x,y,t,h) \equiv \partial \ln[f(\xi_{x,t+h}|x,y,t,h)]/\partial y$,

where $f(\xi_{x,t+h}|x,y,t,h)$ is the conditional density of $\xi_{t,t+h}$ given $(x_t,y_t) = (x,y)$. In the diffusion case of Sections 3 and 4, $P(\cdot)$ is proportional to ξ_x and $S(\cdot)$ is proportional to $\xi_x^2-\sigma^2$. This is *not* generally true in the near diffusion case unless ξ_x and ξ_y are conditionally bivariate normal.

ASSUMPTION 4: *For every h, the conditional densities $f(\xi_x,\xi_y|x,y,t,h)$ and $f(\xi_x|x,y,t,h)$ are well-defined and continuous in x, t, and h and continuously differentiable in y. Further, for some $\delta > 0$*

(5.8) $h^{-1}E[|P(\xi_{x,t+h},x_t,y_t,t,h)|^{2+\delta}|x_t=x,y_t=y] \to 0$, *and*

(5.9) $h^{-1}E[|S(\xi_{x,t+h},x_t,y_t,t,h)|^{2+\delta}|x_t=x,y_t=y] \to 0$

as $h\downarrow 0$, uniformly on every bounded (x,y,t) set.

THEOREM 5.2: *$C_T/2B_T$ is minimized by setting*

(5.10) $g(\xi_x,x,y,t,h) = P(\xi_x,x,y,t) + \omega(x,y,t)\cdot S(\xi_x,x,y,t)$, *where*

$$(5.11)\quad \omega(x,y,t) \equiv \frac{[Cov_t(S,P)^2 + (\Lambda^2 - Var_t(P))\cdot Var_t(S)]^{1/2} - Cov_t(P,S)}{Var_t(S)}.$$

(The arguments have been dropped in (5.11) to simplify notation.) The approximate asymptotic

variance achieved by any other $g(\cdot)$ function (say $\tilde{g}$) satisfying the constraints $E_t[\tilde{g}] = 0$ and $B_T > 0$ is strictly higher unless $\tilde{g}(\xi_x,x,y,t,0) = g(\xi_x,x,y,t,0)$ with probability one.

The minimized $C_T/2B_T = \omega(x_T,y_T,T)$, and the corresponding approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2 - \sigma(y)^2]$ equals $4 \cdot \omega(x_T,y_T,T) \cdot \sigma(y_T)^2[\sigma'(y_T)]^2$.

Proof: see Appendix.

THEOREM 5.3: *The approximate asymptotic variance $C_T/2B_T$ is minimized subject to the constraints (4.7)-(4.8) by setting*

(5.12) $a(x,y,t,h) \equiv \Lambda(x,y,t)$.

The minimized $C_T/2B_T$ is $\Lambda[1 - Corr_T(\xi_y,g^*)]/E_T[g^* \cdot S]$, and the approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2 - \sigma(y)^2]$ is $4\sigma^2(\sigma')^2\Lambda[1 - Corr_T(\xi_y,g^*)]/E_T[g^* \cdot S]$.

Proof: see Appendix.

The interpretations of the optimal filter given in Section 4—i.e., moment matching, asymptotic irrelevance of transformations $\tilde{y}(y_t)$, and the prediction and estimation components of the optimal filter—continue to hold. To further understand the distinction between prediction component $P(\xi_x,x,y,t)$ and the estimation component $S(\xi_x,x,y,t)$, consider first the case in which an ARCH model is the true data generating process—i.e., in which the innovation in the y process, ξ_y, is a function of the innovation in x, ξ_x, and possibly the other state variables x, y, and t, say $\xi_y = g(\xi_x,x,y,t)$. Now $\xi_y = P(\cdot)$, and $\omega(\cdot) = 0$. The innovations in y_t are observable, so it is not surprising that the asymptotic variance of the

measurement error is zero.

Another polar case arises when $P(\cdot) = 0$ with probability one, so ξ_x contains no information that helps predict ξ_y. This was true, for example, in (4.16)-(4.17), when the x_t and y_t were driven by independent Brownian motions. In this case, the asymptotic variance is $\Lambda/SD_t(S)$. This is easily interpretable: Λ is the conditional standard deviation of y_t—the more locally variable y_t is, the less accurately it can be estimated. $Var_t(S)$, on the other hand, is the filtering analogue of the Fisher information—the smaller the Fisher information, the higher the asymptotic variance of the parameter estimates.

6. ANALYSIS OF SOME COMMONLY USED ARCH MODELS

GARCH(1,1)

In the GARCH(1,1) model of Bollerslev (1986), we have $y \equiv \sigma^2$, $\hat{\kappa} \equiv \omega - \theta\sigma^2$, and $g \equiv \alpha \cdot [\xi_x^2 - \sigma^2]$. As we saw in Section 4, GARCH(1,1) is asymptotically optimal for the diffusion (4.16)-(4.17). More generally, suppose the data are generated by either (3.1), (3.1'), (5.1)-(5.2) or (5.1')-(5.2) with $y \equiv \sigma^2$. By Theorem 5.2, GARCH(1,1) is optimal when for some α and all ξ_x, x, σ^2, and t, $\hat{k}(x,\sigma^2,t) = \omega - \theta\sigma^2$ and $P(\xi_x,x,\sigma^2,t) + \omega(x,\sigma^2,t)S(\xi_x,\sigma^2,t) = \alpha[\xi_x^2 - \sigma^2]$. When GARCH(1,1) is the true data generating process, the GARCH(1,1) filter is (trivially) optimal and $P(\xi_x,x,\sigma^2,t) = \alpha[\xi_x^2 - \sigma^2]$ while $\omega(x,\sigma^2,t) = 0$. Even when $P(\xi_x,\sigma^2,t) = 0$, GARCH is optimal provided ξ_x is conditionally normal.

Next consider minimizing $C_T/2B_T$ with a model of the form

(6.1) $g \equiv \alpha(x,\sigma^2,t)\cdot(\xi_x^2-\sigma^2)$.

By Theorems 4.2 and 5.3, the optimal $\alpha(x,\sigma^2,t)$ equals $\Lambda(x,\sigma^2,t)/SD_t(\xi_x^2)$. Suppose the conditional kurtosis of ξ_x is a constant K (this is always satisfied in the diffusion case, where K = 3, and is satisfied in many discrete stochastic volatility models as well). We may then write the (constrained) optimal g as

(6.2) $g \equiv \Lambda(x,\sigma^2,t)\sigma^{-2}(K-1)^{-1/2}\cdot(\xi_x^2-\sigma^2)$.

GARCH(1,1) further constrains α to be constant, so clearly if GARCH(1,1) is to be the constrained optimum in the class of models (6.1), Λ^2, the conditional variance of σ^2, must be linear in σ^4. (Or, equivalently, the conditional variance of $\ln(\sigma^2)$ is constant.) As we will see, many other commonly used ARCH models effectively make the same assumption. The resulting (locally) minimized asymptotic variance of $h^{-1/4}[\hat{\sigma}^2-\sigma^2]$ is

(6.3) $h^{1/2}SD_t(\xi_x^2/\sigma^2)\cdot[1-Corr_t(\xi_x^2,\xi_y)]\cdot\sigma^2\cdot\Lambda$.

That is, GARCH(1,1) can more accurately measure σ_t^2: (1) the less locally variable σ_t^2 is (as reflected by Λ), (2) the lower the conditional kurtosis of ξ_x, (3) the lower σ_t^2, and (4) the more the true data generating mechanism resembles GARCH(1,1) (e.g., if GARCH(1,1) is the data-generating process, $Corr_t(\xi_x^2,\xi_y) = 1$).

The Taylor/Schwert Model

Davidian and Carroll (1987) argue (though not explicitly in a time series or ARCH context) that scale estimates based on absolute residuals are more robust to the presence

of thick tailed residuals than scale estimates based on squared residuals. Schwert (1989) applied the Davidian and Carroll intuition to ARCH models, conjecturing that estimating σ_t^2 with the square of a distributed lag of absolute residuals (as opposed to estimating it with a distributed lag of squared residuals, as in GARCH) would be more robust to ξ_x's with thick tailed distributions. Taylor (1986, Chapter 4) proposed a similar method. Hence, we consider the model $y \equiv \sigma$, $\hat{\kappa} \equiv \omega - \theta\hat{\sigma}$, and $g \equiv \alpha \cdot [|\hat{\xi}_x| - E_t|\xi_x|]$, with the data generated by either (3.1), (3.1'), (5.1)-(5.2) or (5.1')-(5.2). We also assume that the distribution of ξ_x/σ does not vary with x, σ, t, or h—i.e., that σ enters the distribution of ξ_x only as a scale parameter.

If we allow α to be a function of x, y, and t, the optimal α is $\Lambda/[SD_t(|\xi_x|)]$. The minimized asymptotic variance of $h^{-1/4}[\hat{\sigma}_t^2 - \sigma_t^2]$ is

(6.4) $4 \cdot h^{1/2} \cdot SD_t(|\xi_x|) \cdot \sigma^3 \cdot [1 - Corr_t(|\xi_x|, \xi_y)] \cdot \Lambda/E_t|\xi_x|$

As in the GARCH case, *global* optimality of the Taylor/Schwert model in this class of models requires that α is constant. Again, when $E_t[|\xi_x/\sigma|]$ is constant, this is equivalent to $\ln(\sigma^2)$ being conditionally homoskedastic in the diffusion limit.

Schwert's conjecture that this model is more robust than GARCH to conditionally thick tailed ξ_x's can be rigorously justified. For example, compare the relative efficiencies of GARCH(1,1), the Taylor/Schwert model, and the asymptotically optimal filter, supposing for the moment that $Corr_t(\xi_x^2, \xi_y) = Corr_t(|\xi_x|, \xi_y) = 0$, and that ξ_x is conditionally Student's t with $K > 2$ degrees of freedom. It is important to note that (6.4) is not directly

comparable to (6.3), since in (6.3) $y \equiv \sigma^2$ and Λ is the instantaneous standard deviation of σ^2, whereas in (6.4) $y \equiv \sigma$ and Λ is the instantaneous standard deviation of σ. In general the Λ in (6.3) equals 2σ times the Λ in (6.4). Making this adjustment, we can compare the variances in (6.4), (6.5), and the variance achieved by the optimal filter. Figure 2 compares the minimized asymptotic variances of $h^{-1/4}[\hat{\sigma}_t^2 - \sigma_t^2]$ achieved by these filters, plotting both the ratio of the GARCH to the Taylor/Schwert variance and the ratio of GARCH to the optimal filter error variances over a range of K values. The GARCH error variance approaches ∞ as $K \downarrow 4$, and is infinite for $K \leq 4$. The Taylor/Schwert model's error variance approaches infinity as $K \downarrow 2$ and is infinite for $K \leq 2$, but is finite for $K > 2$. The optimal filter's variance is bounded as $K \downarrow 2$. The horizontal line at 1 in Figure 1 divides the region in which the Taylor/Schwert error variance is lower (above the line) and the region in which the GARCH error variance is lower (below the line). For low degrees of freedom, K, the Taylor/Schwert model is much more efficient than GARCH, and remains more efficient until $K > 15.56$. Perhaps surprisingly, the variance of GARCH is only about 6.5% lower even as $K \rightarrow \infty$, so even under the most favorable circumstances, the efficiency gain from using GARCH is slight. When the ξ_x's are conditionally thick tailed, the efficiency loss with GARCH can be dramatic.[11] As $K \rightarrow \infty$, the GARCH and Optimal error variances

11) When ξ_x is conditionally distributed GED (Harvey (1981)), the analysis is similar: Taylor/Schwert performs much better than GARCH when ξ_x is conditionally thick tailed. On the other hand, GARCH may perform substantially (up to 22.5%) better than Taylor/Schwert when ξ_x is much *thinner* tailed than the normal. In the range of parameter estimates found for asset prices in the empirical literature (e.g., Nelson (1989,1991), Sentana (1992)) GARCH and Taylor/Schwert perform about equally.

converge, but fairly slowly: even for K = 20 the GARCH variance is about 13% higher than the optimal.

Our robustness results closely parallel those of Davidian and Carroll (1987). To see why, consider the optimal filter of Theorem 5.2, assuming for now that the prediction component $E[\xi_{y,t+h} | \xi_{x,t+h}, x_t, y_t] = 0$ almost surely for all t. The optimal g is then proportional to the score $\partial \ln[f(\xi_x | x,y,t)]/\partial y$. A necessary condition in this case for GARCH to be optimal is that ξ_x is conditionally normal. Optimality of the Taylor/Schwert model in this case requires ξ_x to be conditionally double exponential, and hence thicker tailed than the normal. Abandoning the assumption of optimality for either model, we see that GARCH uses a normal quasi-likelihood in estimating the level of y_t, while Taylor-Schwert uses a double exponential quasi-likelihood. More generally, the choice of an ARCH model embodies a choice of quasi-likelihood, a choice which can be analyzed in much the same way that Davidian and Carroll did in the case in which the conditional variance of the error term was a function (of known form) of observable variables.

Several papers in the ARCH literature have assumed conditionally Student's *t* errors and have treated the degrees of freedom as a parameter to be estimated. In modelling daily exchange rates, for example, Baillie and Bollerslev (1989) estimated degrees of freedom parameters ranging from 6.3 to 18.5, while Hsieh's (1989) estimates ranged from 3.1 to 6.5. In modelling daily stock price indices, Baillie and DeGennaro's (1990) estimated degrees

of freedom ranged from 9.2 to 10.2.[12] More broadly, thick tailed standardized residuals are the norm in empirical applications of ARCH (see Bollerslev et al (1992)). This range of estimates, along with Figure 2, suggests the use of absolute residuals as opposed to squared residuals in estimating time varying volatilities in asset returns.

Exponential ARCH (EGARCH)

The Exponential ARCH (EGARCH) model of Nelson (1991) was largely motivated by Black's (1976) empirical observation that stock volatility tends to rise following negative returns and to drop following positive returns. The EGARCH model exploits this empirical regularity by making the conditional variance estimate a function of both the size and the sign of lagged residuals. AR(1) EGARCH sets $y \equiv \ln(\sigma^2)$, $\hat{\kappa} \equiv \beta \cdot [y - \alpha]$, and $g \equiv \theta \hat{\xi}_x/\hat{\sigma} + \gamma[|\hat{\xi}_x/\hat{\sigma}| - E_t|\xi_x/\sigma|]$.

Again we assume the data are generated by either (3.1), (3.1)', (5.1)-(5.2) or (5.1')-(5.2). If we make the simplifying assumptions that $E_t[\xi_x \cdot |\xi_x|] = 0$ and that the distribution of ξ_x/σ is independent of σ, it is easy to solve for the (locally) optimal γ and θ:

(6.5) $\gamma^* \equiv \Lambda[1-\mathrm{Corr}_t^2(\xi_x,\xi_y)]^{1/2}/\mathrm{SD}_t(|\xi_x/\sigma|)$, and

12) These models were estimated under the assumption that the estimated ARCH model was correctly specified. If this is literally true, there is clearly no efficiency gain in abandoning the true model. If the model is not correctly specified the robustness of the conditional variance estimates is important.

(6.6) $\theta^* \equiv \Lambda \cdot \mathrm{Corr}_t(\xi_x, \xi_y)$.

For EGARCH to be optimal in this class requires that γ^* and θ^* are constant. It is straightforward to check that when the first two moments of ξ_x/σ and $|\xi_x/\sigma|$ are constant, constant γ^* and θ^* is equivalent to conditionally homoskedastic $\ln(\sigma_t^2)$ and constant conditional correlation between x_t and σ_t^2.

The minimized asymptotic variance of $h^{-1/4}[\hat{\sigma}_t^2 - \sigma_t^2]$ for this case is

(6.7) $2 \cdot \mathrm{SD}_t(|\xi_x|) \cdot \Lambda \cdot \sigma^4 \cdot [\,(1-\mathrm{Corr}_t^2(\xi_x, \xi_y))^{1/2} - \mathrm{Corr}_t(\xi_y, |\xi_x|)]/\mathrm{E}_t|\xi_x|$.

If we adjust for the changed definition of y and therefore of $\Lambda(\cdot)$, we see that apart from the replacement of "1" with "$(1-\mathrm{Corr}^2(\xi_x, \xi_y))^{1/2}$," this is the minimized variance for the Taylor/Schwert model. The only difference between the two models in the $O_p(h^{1/4})$ error components is that the EGARCH takes advantage of conditional correlation in ξ_x and ξ_y (e.g., "leverage effects"). This suggests that we can significantly improve the variance estimators by exploiting correlations between changes in observable variables and changes in σ^2—the effects are "first order" appearing in the dominant component of the measurement error, the $O_p(h^{1/4})$ term—and so are likely to be important in practice. This, along with its relative robustness to conditionally thick tailed ξ_x's, probably accounts for much of the empirical success of EGARCH in applications to stock market returns data (e.g., Pagan and Schwert (1990), Engle and Ng (1991)).

Related Variants

Higgins and Bera (1992) nested GARCH and the Taylor/Schwert model in a class of "NARCH" (nonlinear ARCH) models, which set $\hat{\sigma}_t^{2\delta}$ equal to a distributed lag of past absolute residuals each raised to the 2δ power. Using a geometric lag, this corresponds to $y \equiv \sigma^{2\delta}$, $\hat{\kappa} \equiv \omega - \theta \cdot \hat{y}$, and $g \equiv \alpha \cdot [|\hat{\xi}_x|^{2\delta} - E_t[|\xi_x|^{2\delta}]]$.

The chief appeal of NARCH (as with Taylor/Schwert) is that when $\delta < 1$, it is more robust to conditionally thick tailed ξ_x's than GARCH: NARCH limits the influence of large residuals essentially the same way that the ℓ_p estimators employed in the robust statistics literature (see, e.g., Davidian and Carroll (1987)). While GARCH and Taylor/Schwert use normal and double exponential likelihoods, respectively, in their estimation components, NARCH uses a GED quasi-likelihood.[13]

Another variant is the Threshold ARCH ("TARCH") model of Zakoian (1990). The locally minimized asymptotic measurement error variance (up to the $O_p(h^{1/4})$ terms) of this model are the same as EGARCH.

Similarities and Differences in $\hat{\kappa}(\cdot)$ and $a(\cdot)$ in these Models

Though the GARCH, EGARCH, Taylor/Schwert, Higgins/Bera, and Zakoian models

---

13) Davidian and Carroll (1987) considered the cases δ = 1, 1/2, 1/3, 1/4, 1/6, and lim $\delta \downarrow 0$, finding as we do that scale estimates using $\delta < 1$ are more robust in the presence of thick tailed residuals than estimates using $\delta = 1$. Scale estimates using δ's close to 0 were sensitive to "inliers" rather than outliers.

have important differences, they have at least one similarity and potential limitation: for global asymptotic optimality of $a(\cdot)$ given $g^*(\cdot)$ (as in Theorem 5.3), each requires $\{\ln(\sigma_t^2)\}$ to be conditionally homoskedastic in the diffusion limit. This is clearest when we examine the diffusion limits of these models considered as data generating processes. The diffusion limit for $\{\ln(\sigma_t^2)\}$ in the Higgins/Bera model as a data generating process takes the form[14]

$$(6.8)\quad d[\ln(\sigma_t^2)] = \delta^{-1}(\omega\sigma_t^{-2\delta} - \theta^*)dt - \alpha^* dW_t$$

where W_t is a standard Brownian motion and θ^* and α^* are constants. The Taylor/Schwert and Zakoian models are a special case of (6.8) with $\delta=1/2$, and GARCH is a special case with $\delta=1$. The AR(1) EGARCH diffusion limit for $\{\ln(\sigma_t^2)\}$ is given by (4.21). In (4.21), and for any $\delta > 0$ in (6.8), the conditional variance of the increments in $\{\ln(\sigma_t^2)\}$ is constant. A few ARCH models have been proposed that do not make this assumption (e.g., (4.26), the QARCH model of Sentana (1992), and the model of Friedman and Laibson (1989)). Unfortunately, these models make similarly restrictive assumptions on the conditional second moments. Since the second moment matching condition (4.15) has a first order effect on the measurement error variance, practitioners should probably parameterize $g(\cdot)$ in a way that allows (but does not force) the conditional variance of $\{\ln(\sigma_t^2)\}$ and its instantaneous correlation with $\{x_t\}$ to vary with the level of σ_t^2 and x_t.

---

14) The limits for $\{\sigma_t^2\}$ in GARCH(1,1) and AR(1) EGARCH as data generating processes are given in Nelson (1990). The diffusion limit for $\{\sigma_t\}$ in the Taylor/Schwert model is given in Nelson (1992). We applied Ito's Lemma to convert the diffusion limits into stochastic differential equations for $\{\ln(\sigma_t^2)\}$.

Jump Diffusions and the Friedman/Laibson Model

Friedman and Laibson (1989) argue that stock movements have "ordinary" and "extraordinary" components. This motivated their modified ARCH ("MARCH") model bounds $g(\cdot)$ to keep the "extraordinary" component from being too influential in determining $\hat{\sigma}_t^2$. The MARCH model is similar to GARCH, but with

(6.9) $\hat{\sigma}_{t+h}^2 = \omega \cdot h + \beta_h \cdot \hat{\sigma}_t^2 + h^{1/2} g(\hat{\xi}_x)$, where

(6.10) $g(\hat{\xi}_x) \equiv \alpha > 0$ if $\gamma \cdot \hat{\xi}_x^2 \geq \pi/2$

$\equiv \alpha \cdot \sin[\gamma \cdot \hat{\xi}_x^2]$ if $\gamma \cdot \hat{\xi}_x^2 < \pi/2$.

Friedman and Laibson's model can also be understood as a robust filtering procedure: if ξ_x has occasional large outliers, least-squares based procedures such as GARCH will not estimate σ_t^2 efficiently. Much the same intuition comes from the near diffusion results, in which the conditional distribution of ξ_x is allowed to be considerably thicker tailed than the normal. In accord with Friedman and Laibson's (and Davidian and Carroll's) intuition, the thicker tailed the conditional distribution of ξ_x, the less weight should be given to "large" observations, at least in the score component S of the optimal filter. For example, if $\{\xi_x, \xi_y\}$ is conditionally bivariate t with $K > 2$ degrees of freedom, we have (in the notation of Theorem 5.2)

(6.11) $P \equiv \rho(x,y,t)\Lambda(x,y,t)\xi_x/\sigma(y)$, and

$$(6.12)\quad S \equiv \frac{\sigma'(y)}{\sigma(y)}\left[\frac{K+1}{K-2}\cdot\frac{\xi_x^2/\sigma(y)^2}{1+\xi_x^2/[(K-2)\sigma(y)^2]}\right].$$

Given y, the score is bounded above by $K\cdot\sigma'(y)/\sigma(y)$. The lower the degrees of freedom, the tighter the bound. P, however, remains linear in ξ_x, as in the conditionally normal case.

Unfortunately, the near diffusion assumption does not allow ξ_x and ξ_y to be *too* thick tailed. In particular, ξ_x and ξ_y are assumed to have (conditionally) bounded $2+\delta$ absolute moments, which is why we assumed $K > 2$ in the Student's *t* case. In the limit as $h\downarrow 0$, this effectively rules out the possibility that "lumpy" information arrival causes occasional large jumps in x_t or y_t. Such occasional large jumps may well be a feature of some financial time series. For example, using daily stock returns data, Nelson (1989, 1991) generated $\{\hat{\sigma}_t^2\}$ using an EGARCH model, but found occasional large outliers (i.e., more than five or six times $\hat{\sigma}_t$).[15]

It would be interesting to extend our results to allow the data to be generated by a jump-diffusion (or a near jump-diffusion). We suspect that this would lead to bounds on both P and S in the optimal $g(\cdot)$ function: If we fail to impose such a bound, $g(\hat{\xi}_x, x, \hat{y}, t)$ will be enormous when a jump occurs, which may well make such jumps too influential in

---

15) For example, the market drop on September 26, 1955 (in response to Eisenhower's heart attack) was eleven estimated conditional standard deviations. There were drops of about seven estimated conditional standard deviations on November 3, 1948 and June 26, 1950 in response to Truman's surprise reelection and the beginning of the Korean war respectively. The Crash of October 19, 1987 was almost eight estimated conditional standard deviations, and the market rose seven estimated conditional standard deviations on July 6, 1955.

determining $\{\hat{y}_t\}$.

7. CONCLUSION

One widely voiced criticism of ARCH models (see, e.g., Campbell and Hentschel (1991) and Andersen (1992)) is that they are ad hoc—i.e., though they have been successful in empirical applications, they are *statistical* models, not *economic* models. This criticism, though correct, does not go far enough; even as purely *statistical* models, ARCH models are ad hoc. In applied work, there has been considerable arbitrariness in the choice of ARCH models, despite (perhaps because of) the plethora of proposed ARCH specifications. Many models have been proposed, but few compared to the infinite potential number of ARCH models. How can we choose between these models? How do we design new models? We summarize the main implications of our results for the design of ARCH filters for near diffusions as follows:

Rule 1: Asymptotically optimal ARCH models are as "similar" to the true data generating process as possible, in the sense that the first two conditional moments (as functions of the state variables and time) implied by the ARCH model considered as a data generating process have the same functional form as in the true data generating process.

The choice of an ARCH model therefore embodies an implicit assumption about the joint variability of the state variables x_t and y_t. GARCH(1,1), AR(1) EGARCH, the Taylor/Schwert model, NARCH, and TARCH effectively assume that the conditional

variance of σ_t^2 is linear in σ_t^4. Some ARCH models, (e.g., GARCH, NARCH, Taylor/Schwert), effectively assume that increments in x_t and σ_t^2 are uncorrelated. EGARCH assumes a constant conditional correlation. It is probably wise to relax these constraints, since specification of the conditional second moments of $\{x_t,y_t\}$ affects the $O_p(h^{1/4})$ terms, and so is likely to be important in practice.

Rule 2: The optimally selected $g(\cdot)$ has two components. The first, $E[\xi_{x,t+h} | \xi_{y,t+h}, x_t, y_t]$, forecasts *changes* in $\{y_t\}$, for example by taking advantage of "leverage effects." The second component of the optimal $g(\cdot)$ is proportional to the score of ξ_x, treating y as an unobserved parameter. This term estimates the *level* of $\{y_t\}$ in much the same way as a maximum likelihood estimator of a scale parameter in the i.i.d. case. The robustness results of Davidian and Carroll (1987) hold in the ARCH context: in particular, EGARCH and Taylor/Schwert are more robust than GARCH to conditionally thick tailed ξ_x's. It is probably wise to design ARCH models to be robust to thick tailed ξ_x's (perhaps by bounding $g(\cdot)$ as suggested by Friedman and Laibson), since conditional leptokurtosis seem to be the rule in financial applications of ARCH.

Rule 3: The asymptotic conditional mean of $[\hat{\sigma}_t^2 - \sigma_t^2]$ is zero when the drifts in $\{x_t,y_t\}$ are well specified—i.e., when $\mu(x,y,t) = \hat{\mu}(x,y,t)$ and $\kappa(x,y,t) = \hat{\kappa}(x,y,t)$. Incorrect specification of the drifts creates an $O_p(h^{1/2})$ asymptotic bias in $[\hat{y}_t - y_t]$. Such bias has a second order effect on $[\hat{y}_t - y_t]$, but a first order effect on the medium and long-term forecasts generated by the ARCH model. If filtering rather than forecasting is of primary

concern, specification of $\hat{\mu}(\cdot)$ and $\hat{\kappa}(\cdot)$ is probably less important than specification of $g(\cdot)$.

Our results could be extended in a number of interesting ways. In a sequel, we will allow x_t and y_t to be vectors, and will consider smoothing (i.e., allowing $\hat{y}_t$ to depend on leads and lags of x_t) as well as filtering. Other extensions are also possible. For example, versions of Theorem 2.1 are available which allow a jump diffusion limit for $\{x_t,y_t\}$, and which do not require $\{x_t,y_t,\hat{y}_t\}$ to be Markov (see, e.g., Jacod and Shiryaev (1987, Chapter 9)). Unfortunately, the regularity conditions are considerably more difficult to check than the conditions of Theorem 2.1.[16]

The most important extension would be to allow the parameters of the ARCH model to be estimated by quasi-maximum likelihood methods. All this paper's suggestions for ARCH model building rely on an as yet unproven assumption: namely that fitting misspecified parametric ARCH models by maximum likelihood selects (asymptotically) the 'best' available filter. Monte Carlo experiments (to be reported in a sequel) suggests that this assumption is reasonable in practice. Unfortunately, however, formal asymptotic theory for ARCH parameter estimates has proven very difficult even for well-specified ARCH models, and is not likely to be any easier for misspecified models.

---

16) Foster and Nelson (1991) are able to drop the Markov assumption in analyzing rolling regressions and GARCH by using a central limit theorem for semimartingales in place of Theorem 2.1. It isn't clear if this method can be successfully applied to broader classes of ARCH models.

APPENDIX

PROOF OF THEOREM 2.1: *This is a modification of Stroock and Varadhan (1979, Theorem 11.2.3). The version of the result we cite is from Ethier and Kurtz (1986 Chapter 7, Corollary 4.2). We set Ethier and Kurtz's "n" equal to* $h^{-\Delta}$. *There are two changes from Ethier and Kurtz's version of the Theorem: first, we define* $\Omega_{\Delta,h}(y)$ *as a covariance rather than as a conditional second moment. The conditional first and second moments are each* $O(h^{-\Delta})$, *so the difference between the conditional covariance and the conditional second moment vanishes at rate* $O(h^{-2\Delta})$ *as* $h \downarrow 0$. *Second, Ethier and Kurtz use truncated expectations in (2.2)-(2.3)—i.e.,*

(2.2') $\mu_{\Delta,h}(y) \equiv h^{-\Delta} E[({}_hY_{k+1} - {}_hY_k) \cdot I(\|{}_hY_{k+1} - {}_hY_k\| < 1) \mid {}_hY_k = y]$, *and*

(2.3') $\Omega_{\Delta,h}(y) \equiv h^{-\Delta} Cov[({}_hY_{k+1} - {}_hY_k) \cdot I(\|{}_hY_{k+1} - {}_hY_k\| < 1) \mid {}_hY_k = y]$,

where $I(\cdot)$ *is an indicator function. They then replace (d') with the requirement* $h^{-\Delta} P[\|{}_hY_{k+1} - {}_hY_k\| > \epsilon \mid {}_hY_k = y] \rightarrow 0$ *as* $h \downarrow 0$, *uniformly on every bounded y set. To see that (2.2')-(2.3') and (d") follow from (2.2)-(2.3) and (d'), see the proof of Nelson (1990, Theorem 2.2). (Our version is a little simpler to state, but Ethier and Kurtz's is more general and is sometimes easier to verify. The moment conditions in this paper could be weakened using Ethier and Kurtz's version.) For the uniformity of the weak convergence on bounded* $\{y_0\}$ *sets when* ${}_hY_0$ *is fixed, see Stroock and Varadhan (1979, Theorem 11.2.3).*

PROOFS OF THEOREMS 3.1-3.2: *The proofs are very similar, so we prove only Theorem 3.2, and leave the slightly simpler Theorem 3.1 to the reader. We employ Theorem*

2.1 with $\Delta \equiv 1/2$, treating x, y, and q as state variables and conditioning on (x_T, y_T, q_T).

First, we consider the first two conditional moments of $(x_{t+h}-x_t)$, $(y_{t+h}-y_t)$, $(q_{t+h}-q_t)$, and verify that these increments vanish to zero in probability at an appropriate rate. Under (3.1)-(3.3), the first two conditional moments of $\{x_t, y_t\}$ and the $2+\delta$ absolute moment are all $O(h)$ on bounded (x,y,t) sets. The "fast drift" assumption makes the first conditional moments $O(h^{3/4})$. This means that when we apply Theorem 2.1 using $\Delta \equiv 1/2$, the first two conditional moments of $\{x_t, y_t\}$ converge, respectively, to a vector and matrix of zeros, since the conditional moments are $o(h^{1/2})$. By Assumption 1, the first two conditional moments of $q_{t+h}-q_t$ (normalized by $h^{-1/2}$) converge to $(A_t - B_t q_t)$ and C_t. The normalized covariances of the increments in q_t, x_t, and y_t converge to 0. This convergence is uniform on bounded (x,y,q,t) sets, as required. By Assumption 2, each element of the state vector has a bounded (uniformly on bounded (x,y,q,t) sets) $2+\delta$ absolute moment. Since we used $h^{-1/2}$ as the normalizing factor in the conditional moments, we choose $\Delta = 1/2$ in applying Theorem 2.1. Since the drifts and variances of x_t and y_t are zeros in the limit, $x_{T,\tau}$ and $y_{T,\tau}$ are constant at X_T and y_T in the limit and A_t, B_t, and C_t are asymptotically constant in the diffusion limit on the fast time scale at their time T values.

All that remains is to verify that the diffusion limit has a unique weak-sense solution. Note first that the limit diffusion is clearly non-explosive, since $x_{T,\tau}$ and $y_{T,\tau}$ are constants and $q_{T,\tau}$ follows an Ornstein-Uhlenbeck process. By Assumption 1, A_t, B_t, and C_t are twice continuously differentiable, so weak-sense uniqueness follows by Stroock and Varadhan (1979,

Corollary 6.3.3 and Theorem 10.1.3). The Theorem now follows by Theorem 2.1.

PROOF OF THEOREM 4.1: *Assumption 3 allows us to solve the optimization problem as if* $\xi_{x,t+h}$ *and* $\xi_{y,t+h}$ *were conditionally bivariate normal with covariance matrix given by the last term on the right side of (4.15). The theorem then follows as a special case of Theorem 5.2 below.*

PROOF OF THEOREM 4.2: *By Assumption 3, (dropping time subscripts and function arguments to ease notation)* $C_T/2B_T = E_T[g^2 - 2g\xi_y + \xi_y^2]/2E_T[-\partial g/\partial y] = -E_T[a^2 \cdot g^{*2} - 2a \cdot g^* \cdot \xi_y + \xi_y^2]/2aE_T[\partial g^*/\partial y]$. *Minimizing this with respect to a is equivalent to minimizing* $(a + \Lambda^2/a)$. *The first and second-order conditions then yield* $a = \Lambda$ *(4.10) then follows.*

Lemmas A.1 and A.2 are needed in the proofs of Theorems 4.3-4.5:

LEMMA A.1: *Let* x_t *be an* $n \times 1$ *diffusion, generated by*

(A.1) $x_t = x_0 + \int_0^t \mu(x_s, s)ds + \int_0^t \sigma(x_s, s)dW_s$,

where $\mu(\cdot)$ *is continuous and* $n \times 1$, $\sigma(\cdot)$ *is continuous and* $n \times n$, *and* W_t *is an n-dimensional standard Brownian motion. Then* $t^{-1/2}(x_t - x_0)$ *converges in distribution to a* $N(0_{n\times 1}, \sigma(x_0, 0)\sigma(x_0, 0)')$ *as* $t \downarrow 0$. *The result still holds if the* $\int_0^t \mu(x_s, s)ds$ *in (A.1) is replaced by* $t^{-1/4}\int_0^t \mu(x_s, s)ds$.

PROOF OF LEMMA A.1: *For every t and every* τ, $0 \leq \tau \leq t$, *define* $W_{t,\tau} \equiv$

$t^{-1/2}W_{\tau t}$ *and* $x_{t,\tau} \equiv t^{-1/2}(x_{t\tau} - x_0)$. *For every t,* $W_{t,\tau}$ *is a standard Brownian motion on* $\tau \in [0,1]$. *We now re-write (A.1) as*

(A.1') $x_{t,\tau} = t^{1/2}\int_0^1 \mu(x_0+t^{1/2}x_{t,s},t\cdot s)ds + \int_0^1 \sigma(x_0+t^{1/2}x_{t,s},t\cdot s)dW_{t,s}$.

As $t\downarrow 0$, $t^{1/2}\mu(x_0+t^{1/2}x_{t,s},t\cdot s) \to 0$ *and* $\sigma(x_0+t^{1/2}x_{t,s},t\cdot s) \to \sigma(x_0,0)$ *uniformly on bounded* $(x_0,x_{t,s})$ *sets. Applying Stroock and Varadhan (1979) Theorem 11.1.4,* $x_{t,\tau}$ *converges weakly to a Brownian motion with no drift and with diffusion matrix* $\sigma(x_0,0)\sigma(x_0,0)'$. *In the case in which* $t^{-1/4}\int_0^t \mu(x_s,s)ds$ *replaces* $\int_0^t \mu(x_s,s)ds$ *in (A.1),* $t^{1/4}\int_0^1 \mu(x_0+t^{1/2}x_{t,s},t\cdot s)ds$ *replaces* $t^{1/2}\int_0^1 \mu(x_0+t^{1/2}x_{t,s},t\cdot s)ds$ *in (A.1'), and the convergence is unaffected. The Lemma now follows.*

LEMMA A.2: *Let* x_t *be as in Lemma A.1. For every* $p \geq 2$ *there is a nonnegative, finite k such that for every* x_0 *and every* $i = 1$ *to* n

(A.2) $(E_0|t^{-1/2}(x_{i,t}-x_{i,0})|^p)^{1/p} \leq$

$(t^{-1+p/2}\int_0^t E_0|\mu_i(x_s,s)|^p ds)^{1/p} + k\cdot\Sigma_{j=1,n}\,(t^{-1}\int_0^t E_0|\sigma_{i,j}(x_s,s)|^p ds)^{1/p}$,

where $x_{i,t}$ *is the* i^{th} *element of* x_t, $\sigma_{i,j,t}$ *is the* i-j^{th} *element of* $\sigma(x_t,t)$, *and k is a constant depending only on p. When* $t^{-1/4}\int_0^t \mu(x_s,s)ds$ *replaces* $\int_0^t \mu(x_s,s)ds$ *in (A.1),* $t^{-1+p/4}\int_0^t E_0|\mu_i(x_s,s)|^p ds$ *replaces* $t^{-1+p/2}\int_0^t E_0|\mu_i(x_s,s)|^p ds$ *in (A.2).*

PROOF OF LEMMA A.2: *From (A.1),*

(A.3) $|t^{-1/2}(x_{i,t}-x_{i,0})| = |t^{-1/2}\int_0^t \mu_i(x_s,s)ds + \Sigma_{j=1,n}t^{-1/2}\int_0^t \sigma_{i,j}(x_s,s)dW_{j,s}|$

$\leq t^{-1/2}\int_0^t |\mu_i(x_s,s)|ds + \Sigma_{j=1,n}t^{-1/2}\,|\int_0^t \sigma_{i,j}(x_s,s)dW_{j,s}|$.

Applying Minkowski's inequality:

(A.4) $[E_0|t^{-1/2}(x_{i,t}-x_{i,0})|^p]^{1/p} \le [E_0(t^{-1/2}\int_0^t|\mu_i(x_s,s)|ds)^p]^{1/p}$

$+ \Sigma_{j=1,n}[E_0(t^{-1/2}|\int_0^t\sigma_{i,j}(x_s,s)dW_{j,s}|)^p]^{1/p}$.

By the integral version of the means inequality (Hardy, Littlewood, and Pólya (1951, Theorem 192)) $\int_0^t|\mu_i(x_s,s)|ds \le t^{(p-1)/p}(\int_0^t|\mu_i(x_s,s)|^p ds)^{1/p}$. By Karatzas and Shreve (1988, Exercise 3.25), $E_0(|t^{-1/2}\int_0^t\sigma_{i,j}(x_s,s)dW_{j,s}|)^p \le mt^{-1}E_0\int_0^t|\sigma_{i,j}(x_s,s)|^p ds$, where m is a constant depending only on n and p. Substituting both these inequalities into (A.3) yields (A.2). Substituting $t^{-1/4}\mu_i(x_s,s)$ for $\mu_i(x_s,s)$ yields the remainder of the Lemma.

PROOF OF THEOREM 4.3: *That (4.16)-(4.17) has a unique weak-sense solution was established in Nelson (1990). Next, note that $\xi_x = \hat{\xi}_x$, since $\hat{\mu} = \mu$. Employing the definitions of y, $\hat{y}$, ξ_x, ξ_y and q, and using the "exploding drifts" convention (3.1'), we have*

(A.5) $\xi_{x,t+h} = h^{-1/2}\int_t^{t+h} y_s^{1/2}dW_{1,t}$,

(A.6) $\xi_{y,t+h} = \theta h^{-3/4}\int_t^{t+h}(E_t y_s - y_s)ds + h^{-1/2}\int_t^{t+h} y_s\alpha 2^{1/2}dW_{2,t}$,

(A.7) $q_{t+h} - q_t = h^{1/4}[\alpha(\xi_{x,t+h}^2 - y_t - h^{1/4}q_t) - \xi_{y,t+h}] - \theta h^{-1/2}\int_t^{t+h}(y_t - E_t y_s + h^{1/4}q_t)ds$, and

(A.8) $E_t[y_s] = \omega/\theta + (y_t - \theta/\omega)[\exp(-\theta(s-t))-1]$ or

$E_t[y_s] = \omega/\theta + (y_t - \theta/\omega)[\exp(-\theta h^{-1/4}(s-t))-1]$

in the fast drift case. By Lemma A.1, $(\xi_{x,t+h}, \xi_{y,t+h})$ given time t information, converges in distribution to a bivariate normal with means of zero, no correlation, and variances y_t and $2\alpha y_t^2$. To verify Assumptions 1-3, however, we need convergence of moments (up to order $4+\delta$ for ξ_x and $2+\delta$ for ξ_y) as well. These moments (conditional on $(x_t=x, y_t=y, q_t=q)$) must be uniformly bounded on every bounded (x,y,q,t) set.

If there is a suitably bounded 2+δ moment for $|\xi_y|$, *a bounded 4+δ moment for* $|\xi_x|$ *follows using Lemma A.2. For each* $h > 0$, *the* $\{\sigma_t^2\}$ *process (i.e., without* x_t*) satisfies standard Lipshitz and Growth conditions and consequently (see Arnold (1973, Section 7.1)) for every* $\delta > 0$

(A.9) $E_t[\sigma_{t+h}^{2+\delta}] \leq (1 + \sigma_t^{2+\delta})exp(C \cdot h^{1/2})$

where C is a constant depending on α, θ, *and* ω. *Moment boundedness for* $|\xi_{x,t+h}|^{4+\delta}$ *now follows by Lemma A.2, satisfying Assumption 2 and (4.5).*

Applying Lemma A.2 and (A.8) we now have

(A.10) $h^{-1/2}E[q_{t+h} - q_t \mid x_t=x, y_t=y, q_t=q] \rightarrow -\alpha q$

(A.11) $h^{-1/2}Var[q_{t+h} - q_t \mid x_t=x, y_t=y, q_t=q] \rightarrow Var[\alpha(\epsilon_x^2 - y) - \epsilon_y] = 4\alpha^2 y^2$, *and*

(A.12) $h^{-1/2}E[|q_{t+h} - q_t|^3 \mid x_t=x, y_t=y, q_t=q] \rightarrow 0$ *uniformly on bounded (x,y,q,t) sets as required, satisfying (3.15)-(3.16) and (4.2)-(4.4).*

PROOF OF THEOREM 4.4: *That the system (4.20'), (4.21'), and (4.22) has a unique weak-sense solution is established in e.g., Nelson (1990). Under the "fast drift" convention (3.1'),*

(A.13) $\xi_{x,t+h} = -h^{-3/4}(1/2)\int_t^{t+h}(exp(y_s) - E_t exp(y_s))ds + h^{-1/2}\int_t^{t+h} exp(y_s/2)dW_{1,s}$,

(A.14) $\hat{\xi}_{x,t+h} = \xi_{x,t+h} + h^{-3/4}(1/2)\int_t^{t+h}(E_t[exp(y_s)] - exp(y_t + h^{1/4}q_t))ds$

(A.15) $\xi_{y,t+h} = -\beta \cdot h^{-3/4}\int_t^{t+h}(y_s - E_t y_s)ds + h^{-1/2}\int_t^{t+h}\psi dW_{2,s}$, *and*

(A.16) $q_{t+h} - q_t = -h^{-1/2}\beta\int_t^{t+h}(y_t + h^{1/4}q_t - E_t[y_s])ds + h^{1/4}[g(\hat{\xi}_{x,t+h}, \hat{y}_t) - \xi_{y,t+h}]$

where

(A.17) $g(\hat{\xi},\hat{y}) \equiv \psi[\rho\hat{\xi}\cdot\exp(-\hat{y}/2) + [(1-\rho^2)/2]^{1/2}\cdot(\hat{\xi}^2\cdot\exp(-\hat{y}) - 1)]$.

By Lemma A.1, given time t information $(\xi_{x,t+h}, \xi_{y,t+h})$ *converges in distribution to a bivariate normal with means of zero, correlation* ρ, *and variances of* $\exp(y_t)$ *and* ψ^2. *The same is true of* $(\hat{\xi}_{x,t+h}, \xi_{y,t+h})$. *We also require convergence of moments up to order* $4+\delta$ *for* $|\hat{\xi}_{x,t+h}|$ *and* $2+\delta$ *for* $|\xi_{y,t+h}|$. y_s *is Gaussian (see, e.g., Arnold (1973 Section 8.3)), and for* $s > t$, $y_s|y_t \sim N[(\alpha+(y_t-\alpha)e^{-\beta(s-t)}), \psi^2[1-\exp^{-2\beta(s-t)}]/2\beta]$ *and has arbitrary finite moments (each uniformly bounded on bounded y sets conditional on* $y_t=y$), *as does* $\exp(y_s)$. *In the fast drift case,* y_s *given* y_t *is normal with mean and variance* $(\alpha+(y_t-\alpha)\exp(h^{-1/4}\beta(s-t))$ *and* $\psi^2[(1-\exp(-2\beta h^{-1/4}(s-t)))/(2\beta h^{-1/4})])$. *This allows us to compute* $E_t[\exp(y_s)]$ *and* $E_t[y_s]$ *explicitly. For s with* $t \le s \le t+h$, *the conditional moments of arbitrary order of both* y_s *and* $\exp(y_s)$ *remain uniformly bounded on bounded* y_t *sets. (4.5) and Assumption 2 are therefore satisfied.*

Substituting $(y_t + h^{1/4}q_t)$ *for* $\hat{y}_t$ *in (A.16)-(A.17) and using the formulas for* $E_t[\exp(y_s)]$ *and* $E_t[y_s]$ *leads to*

(A.18) $h^{-1/2}E[q_{t+h}-q_t|x_t=x,y_t=y,q_t=q] \rightarrow -q\cdot\psi[(1-\rho^2)/2]^{1/2}$, *and*

(A.19) $h^{-1/2}Var[q_{t+h}-q_t|x_t=x,y_t=y,q_t=q] \rightarrow 2\psi^2(1-\rho^2)$

(A.20) $h^{-1/2}E[|q_{t+h}-q_t|^{2+\delta}|x_t=x,y_t=y,q_t=q] \rightarrow 0$

uniformly on bounded (x,y,q,t) sets, satisfying Assumption 1 and (4.2)-(4.4).

PROOF OF THEOREM 4.5: *To establish existence an uniqueness of a weak-sense solution to the system (4.22), (4.20'), and (4.23') we first consider the system (4.20') and*

(4.22)-(4.23). For the latter system, we apply Nelson (1990) Theorem A.1. Condition A of that theorem is clearly satisfied. For its non-explosion condition, use $\varphi(x,y) \equiv 1 + x^2 + \sigma^4$. *We conclude that (4.20') and (4.22)-(4.23) has a unique weak-sense solution. When* $\beta\alpha \geq \psi^2/2$ *(or, in the fast drift case, whenever* $\beta > 0$ *and* $\alpha > 0$*)* $\sigma_t^2 = 0$ *is inaccessible (with probability one) in finite time, so the mapping from* $\{x_t, \sigma_t^2\}$ *to* $\{x_t, ln(\sigma_t^2)\}$ *is almost surely uniformly continuous on [0,T] for all* $T < \infty$. *The continuous mapping theorem then delivers a unique weak-sense solution for (4.22), (4.20'), and (4.23').*

Convergence in distribution of $(\hat{\xi}_{x,t+h}, \xi_{y,t+h})$ *(and of* $(\xi_{x,t+h}, \xi_{y,t+h})$*) given time t information to a bivariate normal with mean (0,0)', correlation* ρ *and variances* $exp(y_t)$ *and* $\psi^2 exp(y_t)$ *follows from Lemma A.1. We next check local boundedness of the moments. The conditional distribution of* σ_s^2 *given* σ_t^2 $(s>t)$ *is given by Cox, Ingersoll and Ross (1985, pp. 391-392). Using a formula for the non-central chi-square distribution (see Johnson and Kotz (1970, Chapter 28, (1)) and the integral form of the Gamma function we obtain, for* $\nu > -a$ *and* $s > t$,

$$(A.21)\quad E_t[\sigma_s^{2a}] = c^{-a} exp(-c \cdot \sigma_t^2 \cdot e^{-\beta(s-t)}) \sum_{j=0}^{\infty} \frac{(c \cdot \sigma_t^2 \cdot e^{-\beta(s-t)})^j \, \Gamma(\nu + j + a)}{\Gamma(j+1) \cdot \Gamma(\nu + j)},$$

where $\nu \equiv 2\beta h^{-1/4}\alpha/\psi^2$ *and* $c \equiv 2\beta h^{-1/4}/[\psi^2(1-exp(-\beta h^{-1/4}(s-t)))]$ *in the fast drift case and* $\nu \equiv 2\beta\alpha/\psi^2$ *and* $c \equiv 2\beta/[\psi^2(1-exp(-\beta(s-t)))]$ *otherwise. This can be rewritten as*

(A.22) $E_t[\sigma_s^{2a}] = \frac{\Gamma(\nu + a)}{c^a \Gamma(\nu)} exp(-c \cdot \sigma_t^2 \cdot e^{-\beta(s-t)}) \cdot M(a+\nu, \nu, c \cdot \sigma_t^2 \cdot e^{-\beta(s-t)}),$

$= \sigma_t^{2a}[1 + \sigma_t^{-2} O(|c|^{-1})]$ *as* $c \to 0$,

where $M(\cdot,\cdot,\cdot)$ *is a confluent hypergeometric function. The last equality in (A.22) follows from Slater (1965, 13.1.4). Since* $t \le s \le t+h$ *for the relevant moments,* $1 - exp(-\beta(s-t)) = O(h)$ *and* $c^{-1} = O(s-t)$ *as* $h \downarrow 0$. $E[\sigma_s^{2a} | ln(\sigma_t^2) = y] - \sigma_t^{2a} \to 0$ *uniformly on bounded y sets as* $h \downarrow 0$, *provided* $\nu + a > 0$.

To bound the $4+\delta$ *conditional absolute moment of* ξ_x, *set* $a = 2 + \delta/2$. *To bound the* $2+\delta$ *conditional absolute moment of* ξ_y, *set* $a = -1 - \delta/2$. *In the fast drift case, these moments are finite for sufficiently small h whenever* $\alpha > 0$ *and* $\beta > 0$. *Otherwise, we require* $2\beta\alpha > \psi^2$. *This satisfies Assumption 2 and (4.5). We now have, for the fast drift case, (the standard case is similar)*

(A.23) $\xi_{x,t+h} = -h^{-3/4}(1/2) \int_t^{t+h} (exp(y_s) - E_t\, exp(y_s))ds + h^{-1/2} \int_t^{t+h} exp(y_s/2) dW_{1,s}$,

(A.24) $\hat{\xi}_{x,t+h} = \xi_{x,t+h} + h^{-3/4}(1/2) \int_t^{t+h} (E_t[exp(y_s)] - exp(y_t + h^{1/4} q_t))ds$,

(A.25) $\xi_{y,t+h} = h^{-3/4}(\alpha\beta - \psi^2/2) \int_t^{t+h} (exp(-y_s) - E_t\, exp(-y_s))ds + \psi h^{-1/2} \int_t^{t+h} exp(-y_s/2) dW_{2,s}$,

and

(A.26) $q_{t+h} - q_t = h^{-1/2}(\alpha\beta - \psi^2/2) \int_t^{t+h} (exp(-\hat{y}_t) - E_t[exp(-y_s)])ds$

$+ h^{1/4}[g(\hat{\xi}_{x,t+h}, \hat{y}_t) - \xi_{y,t+h}]$, *where*

(A.27) $g(\hat{\xi}, \hat{y}) \equiv \psi \cdot exp(-\hat{y}/2)[\rho\, \hat{\xi} \cdot exp(-\hat{y}/2) + [(1-\rho^2)/2]^{1/2} \cdot (\hat{\xi}^2 \cdot exp(-\hat{y}) - 1)]$.

Applying Lemmas A.1-A.2, substituting $(y_t + h^{1/4} q_t)$ *for* $\hat{y}_t$ *in (A.26)-(A.27), and using*

(A.22) leads to

(A.28) $h^{-1/2}E[q_{t+h}-q_t \mid x_t=x,y_t=y,q_t=q] \rightarrow -q \cdot exp(-y/2)\,\psi[(1-\rho^2)/2]^{1/2}$,

(A.29) $h^{-1/2}Var[q_{t+h}-q_t \mid x_t=x,y_t=y,q_t=q] \rightarrow 2\psi^2(1-\rho^2)exp(-y)$, *and*

(A.30) $h^{-1/2}E[\,|q_{t+h}-q_t|^{2+\delta} \mid x_t=x,y_t=y,q_t=q] \rightarrow 0$

uniformly on bounded (x,y,q,t) sets, satisfying Assumption 1 and (4.2)-(4.4). Finally, apply the delta method to derive the asymptotic variance of $[\hat{\sigma}_t^2-\sigma_t^2]$.

PROOF OF THEOREM 5.1: *Nearly identical to the proof of Theorem 3.1.*

Before we prove Theorem 5.2, we present a heuristic derivation of the first order condition. In the proof we verify global optimality.

Under Assumption 4 we may write $C_t/2B_t$ as

$$\text{(A.31)} \qquad \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[g(\xi_x,x,y,t)-\xi_y]^2 f(\xi_x,\xi_y|x,y,t)d\xi_x d\xi_y}{2\int_{-\infty}^{\infty} g(\xi_x,x,y,t)\cdot S(\xi_x,x,y,t)\cdot f(\xi_x|x,y,t)d\xi_x}.$$

We wish to minimize this with respect to $g(\cdot)$, subject to two constraints: first that $E_t[g] = 0$, and second, that the denominator of (A.31) is nonnegative. For now we ignore these constraints since they are not binding at the solution (5.10)-(5.11). To derive the first-order conditions, we treat $g(\xi_x^*,x,y,t)$, for each (ξ_x^*,x,y,t) as a separate choice variable. Setting the partial derivative of (A.31) with respect to $g(\xi_x^*,x,y,t)$ equal to zero, dividing by $f(\xi_x^*|x,y,t)$ and multiplying by $Cov_t[g\cdot S]$ yields

$$(A.32)\quad g(\xi_x^*,x,y,t) = \int_{-\infty}^{\infty} \xi_y\, f(\xi_y|\xi_x^*,x,y,t)d\xi_y + \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[g(\xi_x,x,y,t)-\xi_y]^2 f(\xi_x,\xi_y|x,y,t)d\xi_x d\xi_y}{2\int_{-\infty}^{\infty}[g(\xi_x,x,y,t)\cdot S(\xi_x,x,y,t)]f(\xi_x|x,y,t)d\xi_x}\cdot S(\xi_x^*,x,y,t)$$

$$(A.33)\quad = P(\xi_x^*,x,y,t) + \omega(x,y,t)\cdot S(\xi_x^*,x,y,t) = P + \omega S$$

for some function $\omega(x,y,t)$. Comparing (A.32)-(A.33) with (A.31), it is clear that $C_T/2B_T = \omega(x_T,y_T,T)$. Substituting for $P + \omega S$ for g (A.33) and solving for ω leads to a quadratic in ω with two solutions:

$$(A.34)\quad \omega = \frac{\pm[Cov_t(S,P)^2 + (\Lambda^2 - Var_t(P))\cdot Var_t(S)]^{1/2} - Cov_t(P,S)}{Var_t(S)}.$$

The "+" solution is the only solution satisfying the constraint $0 \leq \int_{-\infty}^{\infty} g(\xi_x,x,y,t)\cdot S(\xi_x,x,y,t)\cdot f(\xi_x|x,y,t)d\xi_x$, leading to (5.10)-(5.11).

PROOF OF THEOREM 5.2: *Next, we verify that this $g(\cdot)$ is globally optimal.*

Dropping subscripts, we write this $g(\cdot)$ as $g \equiv P + \omega S$. Now consider a perturbation of this function, $\tilde{g} \equiv P + \omega S + H$, where H is a function of ξ_x, x, y, and t with $E_t[H] = 0$ and $Cov_t[\tilde{g},S] > 0$ (these conditions force $\tilde{g}$ to obey the constraints $E_t[\tilde{g}] = 0$ and $B_T > 0$). Our claim is that the approximate asymptotic variance of $\tilde{g}$ is strictly higher than that of g unless $H = 0$ with probability one, or equivalently

$$(A.35)\quad \frac{E_t[(P+\omega S-\xi_y)^2]}{2\cdot Cov_t[P+\omega S,S]} < \frac{E_t[(P+\omega S-\xi_y)^2] + E_t[H^2] + 2\cdot Cov_t[H,P+\omega S-\xi_y]}{2\cdot Cov_t[P+\omega S,S] + 2\cdot Cov_t[H,S]}$$

for all such H. Recall that $\omega = C_t/2B_t = E_t[(P+\omega S-\xi_y)^2]/[2\cdot Cov_t(P+\omega S,S)]$, so the left

side of (A.35) equals ω, and the $E_t[(P + \omega S - \xi_y)^2]$ term on the right of (A.35) equals $\omega \cdot [2 \cdot Cov_t(P+\omega S,S)]$. Making these substitutions, and using the positivity of both denominators in (A.35), (A.35) becomes

(A.36) $0 < E_t[H^2] - 2 \cdot (Cov_t[H,\xi_y] - Cov_t[H,P])$.

Recall that H is a function of ξ_x but not of ξ_y, and that $P \equiv E_t[\xi_y|\xi_x]$, so

$$\begin{aligned} (A.37)\ Cov_t[H,\xi_y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(\xi_x,x,y,t)\,\xi_y\, f(\xi_x,\xi_y|x,y,t)d\xi_x d\xi_y \\ &= \int_{-\infty}^{\infty} H(\xi_x,x,y,t)\left[\int_{-\infty}^{\infty} \xi_y f(\xi_y|\xi_x,x,y,t)d\xi_y\right] f(\xi_x|x,y,t)d\xi_x \\ &= \int_{-\infty}^{\infty} H(\xi_x,x,y,t)E[\xi_y|\xi_x,x,y,t]f(\xi_x|x,y,t)d\xi_x \\ &= Cov_t[H,P]. \end{aligned}$$

(A.35) is therefore equivalent to $0 < E_t[H^2]$. The approximate asymptotic variance of $h^{-1/4}[\sigma(\hat{y})^2 - \sigma(y)^2]$ follows by the delta method.

PROOF OF THEOREM 5.3: *Nearly identical to the proof of Theorem 4.2.*

REFERENCES

ANDERSEN, T. G. (1992): "A Model of Return Volatility and Trading Volume," working paper, Northwestern University.

ANDERSON, B. D. O. and J. B. MOORE (1979): *Optimal Filtering*. Englewood Cliffs, NJ: Prentice Hall.

ARNOLD, L. (1973): *Stochastic Differential Equations: Theory and Applications*. New York: Wiley.

BAILLIE, R. T., and T. BOLLERSLEV (1989): "The Message in Daily Exchange Rates: A Conditional Variance Tale." *Journal of Business and Economic Statistics*, 7, 297-305.

BAILLIE, R. T., and R. P. DeGENNARO (1990): "Stock Returns and Volatility." *Journal of Financial and Quantitative Analysis*, 25, 203-214.

BATES, D. and PENNACCHI, G. (1990): "Estimating a Heteroskedastic State Space Model of Asset Prices," Working Paper, University of Illinois at Urbana/Champaign.

BILLINGSLEY, P. (1986): *Probability and Measure*, Second ed. New York: Wiley.

BLACK, F. (1976): "Studies of Stock Market Volatility Changes." *Proceedings of the American Statistical Association, Business and Economic Statistics Section,* 177-181.

BOLLERSLEV, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31, 307-327.

BOLLERSLEV, T. (1987): "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *The Review of Economics and Statistics*, 69, 542-547.

BOLLERSLEV, T., R. Y. CHOU, and K. KRONER (1992): "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence." *Journal of Econometrics,* 52, 5-60.

CAMPBELL, J. Y. and L. HENTSCHEL (1991): "No News is Good News: An Asymmetric Model of Changing Volatility in Stock Returns." Forthcoming, *Journal of Financial Economics*.

COX, J. C., J. E. INGERSOLL, Jr., and S. A. ROSS (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica,* 53, 385-407.

DAVIDIAN, M. and R. J. CARROLL (1987): "Variance Function Estimation." *Journal of the American Statistical Association*, 82, 1079-1091.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroskedasticity withEstimates of the Variance of United Kingdom Inflation." *Econometrica,* 50, 987-1008.

ENGLE, R. F., and V. K. NG (1991): "Measuring and Testing the Impact of News on Volatility." Working Paper, U.C.S.D.

ETHIER, S. N. and T. G. KURTZ (1986): *Markov processes: Characterization and Convergence*. New York: Wiley.

FOSTER, D. P., and NELSON, D. B. (1991): "Rolling Regressions," Working Paper, University of Chicago.

FRIEDMAN, B. M. and D. I. LAIBSON (1989): "Economic Implications of Extraordinary Movements in Stock Prices." *Brookings Papers on Economic Activity,* 2:1989, 137-172.

GENNOTTE, G., and T. A. MARSH (1991): "Variations in Economic Uncertainty and Risk Premiums on Capital Assets." Working Paper, U. C. Berkeley.

HARDY, G., J. E. LITTLEWOOD, and G. PÓLYA (1951): *Inequalities*, second edition. Cambridge, UK: Cambridge University Press.

HARVEY, A. C. (1981): *The Econometric Analysis of Time Series*. Oxford: Philip Allan.

HESTON, S. L. (1991): "A Closed Form Solution for Options with Stochastic Volatility."

Working Paper, Yale University.

HIGGINS, M. L. and A. K. BERA (1992): "A Class of Nonlinear ARCH Models." *International Economic Review* 33, 137-158.

HSIEH, D. A. (1989): "Modeling Heteroskedasticity in Daily Foreign Exchange Rates." *Journal of Business and Economic Statistics*, 7 307-317.

HULL, J. and A. WHITE (1987): "The Pricing of Options on Assets with Stochastic Volatilities." *Journal of Finance,* 42, 281-300.

JACOD, J. and SHIRYAEV (1987): *Limit Theorems for Stochastic Processes*, Berlin: Springer Verlag.

JOHNSON, N. L., and S. KOTZ (1970): *Continuous Univariate Distributions, Volume 2*. New York: John Wiley and Sons.

KARATZAS, I. and S. E. SHREVE (1988): *Brownian Motion and Stochastic Calculus*. New York: Springer-Verlag.

KITAGAWA, G. (1987): "Non-Gaussian State Space Modeling of Nonstationary Time Series [with discussion]," *Journal of the American Statistical Association*, 82, 1032-1063.

MAYBECK, P. S. (1982): *Stochastic Models, Estimation, and Control, Volume 2*, New York: Academic Press.

MELINO, A. and S. TURNBULL (1990): "Pricing Foreign Currency Options with Stochastic Volatility." *Journal of Econometrics,* 45, 239-266.

MERTON, R. C. (1980): "On Estimating the Expected Return on the Market." *Journal of Financial Economics*, 8, 323-361.

NELSON, D.B. (1989): "Modeling Stock Market Volatility Changes," *1989 Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 93-98.

NELSON, D. B. (1990): "ARCH Models as Diffusion Approximations." *Journal of Econometrics*, 45, 7-39.

NELSON, D. B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica*, 59, 347-370.

NELSON, D. B. (1992): "Filtering and Forecasting with Misspecified ARCH Models I: Getting the Right Variance with the Wrong Model." *Journal of Econometrics*, 52, 61-90.

NELSON, D. B., and D. P. FOSTER (1991): "Filtering and Forecasting with Misspecified ARCH Models II: Making the Right Forecast with the Wrong Model." Working Paper, University of Chicago.

PAGAN, A. R. and G. W. SCHWERT (1990): "Alternative Models for Conditional Stock Volatility." *Journal of Econometrics*, 45, 267-290.

SCHWERT, G. W. (1989): "Why Does Stock Market Volatility Change Over Time?" *Journal of Finance*, 44, 1115-1154.

SCOTT, L. O. (1987): "Option Pricing when the Variance Changes Randomly: Theory, Estimation, and an Application." *Journal of Financial and Quantitative Analysis*, 22,

419-438.

SENTANA, E. (1992): "Quadratic ARCH models: A Potential Reinterpretation of ARCH Models as Second Order Taylor Approximations." Working Paper: LSE.

SLATER, L. J. (1965): "Confluent Hypergeometric Functions," Chapter 13 in M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

STROOCK, D. W. and S. R. S. VARADHAN (1979): *Multidimensional Diffusion Processes* Berlin: Springer Verlag.

TAYLOR, S. (1986): *Modeling Financial Time Series*. New York: Wiley.

WIGGINS, J. B. (1987): "Option Values under Stochastic Volatility: Theory and Empirical Estimates." *Journal of Financial Economics*, 19, 351-372.

ZAKOIAN, J. M. (1990): "Threshold Heteroskedastic Models." Working Paper, INSEE.

ZUCKER, R. (1965): "Elementary Transcendental Functions: Logarithmic, Exponential, Circular, and Hyperbolic," Chapter 4 in M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications.

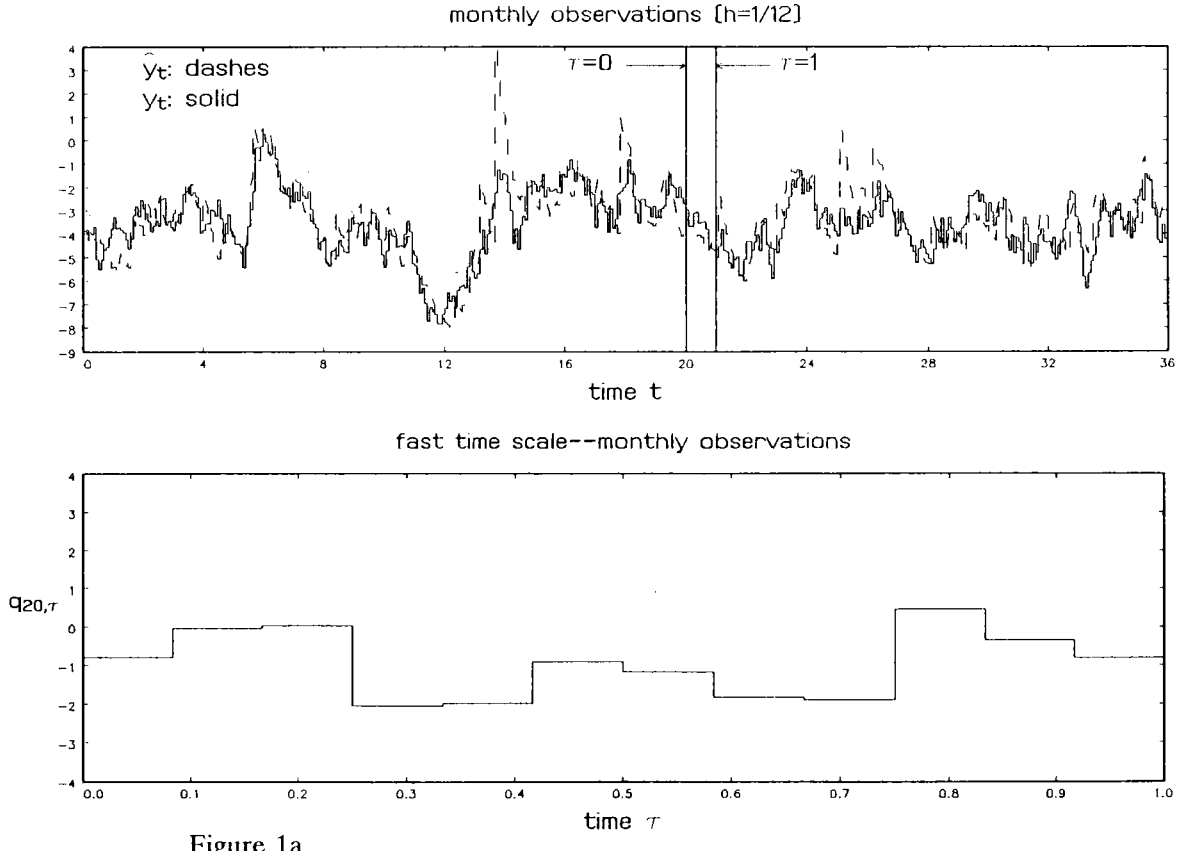
monthly observations (h=1/12)
$\hat{y}_t$: dashes
y_t: solid
τ=0
τ=1
time t
fast time scale--monthly observations
$q_{20,\tau}$
time τ

Figure 1a

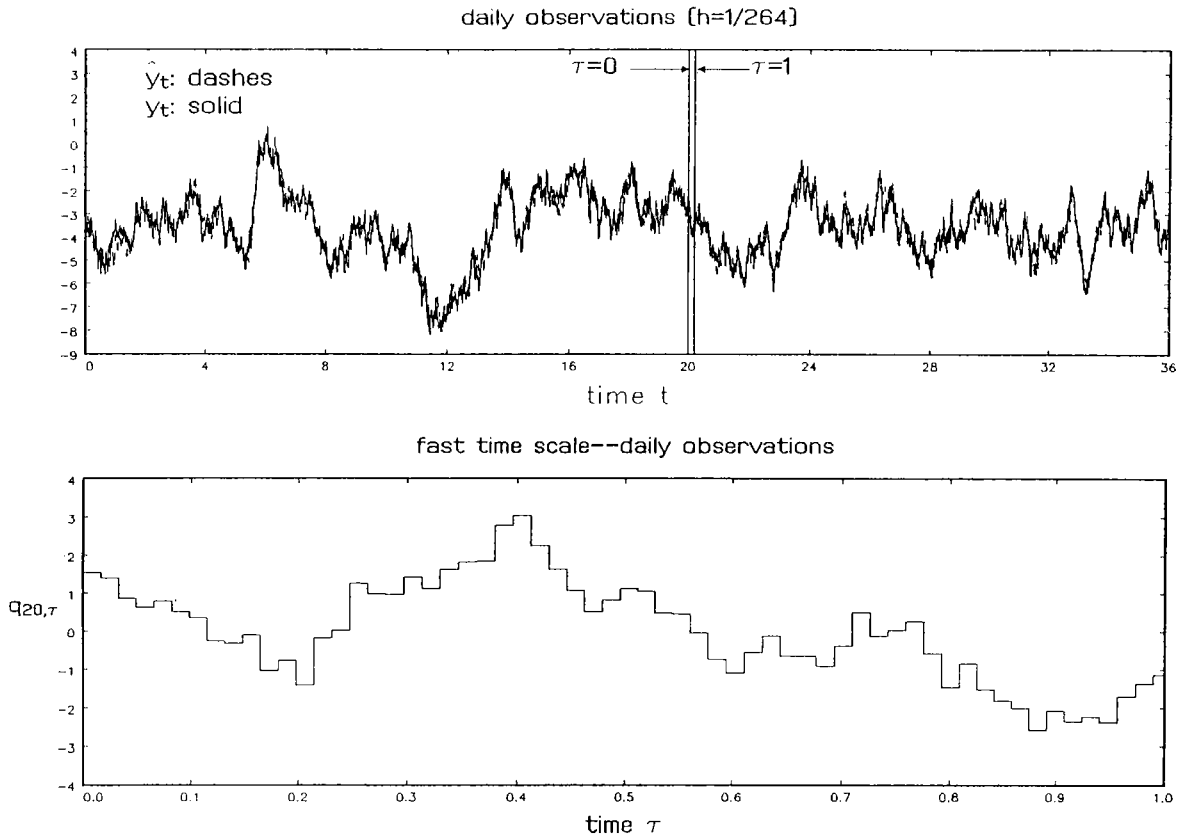
daily observations (h=1/264)
$\hat{y}_t$: dashes
y_t: solid
τ=0
τ=1
time t
fast time scale--daily observations
$q_{20,\tau}$
time τ

Figure 1b

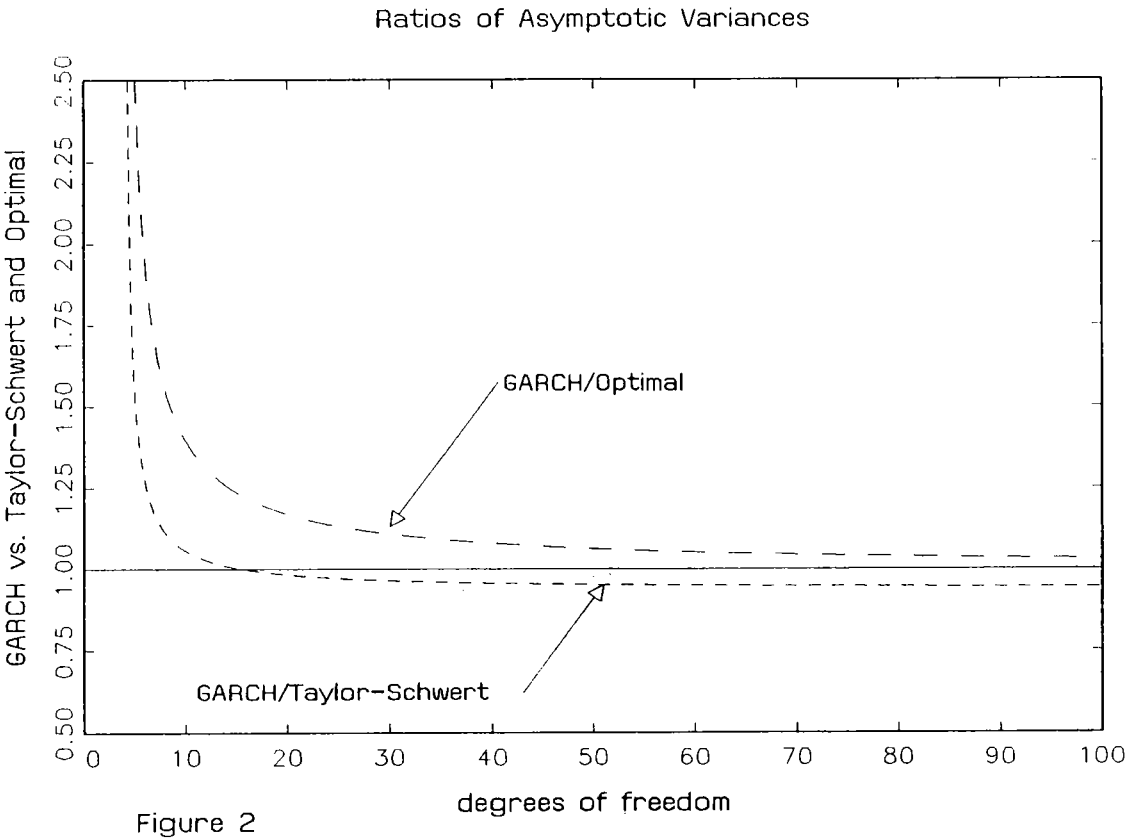
Ratios of Asymptotic Variances
GARCH vs. Taylor-Schwert and Optimal
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
0
10
20
30
40
50
60
70
80
90
100
degrees of freedom
GARCH/Optimal
GARCH/Taylor-Schwert

Figure 2