TECHNICAL WORKING PAPER SERIES

ASYMPTOTIC FILTERING THEORY FOR MULTIVARIATE ARCH MODELS

Daniel B. Nelson

Technical Working Paper No. 162

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1994

This is a revision of parts of an earlier working paper, "Asymptotic Filtering and Smoothing Theory for Multivariate ARCH Models." I would like to thank Andrea Bertozzi, Tim Bollerslev, George Easton, Dean Foster, Boaz Schwartz, and seminar participants at Brigham Young, Chicago, the Econometric Society Summer Meetings, Harvard/M.I.T., the Multivariate Financial Time Series Conference, the NBER Asset Pricing Group, Northwestern, Princeton, Utah, Yale, and Wisconsin (Madison) for helpful discussions. This material is based on work supported by the National Science Foundation under grants #SES-9110131 and #SES-9310683. The Center for Research in Security Prices provided additional research support. This paper is part of NBER's research program in Asset Pricing. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #162
August 1994

ASYMPTOTIC FILTERING THEORY FOR MULTIVARIATE ARCH MODELS

ABSTRACT

ARCH models are widely used to estimate conditional variances and covariances in financial time series models. How successfully can ARCH models carry out this estimation when they are misspecified? How can ARCH models be optimally constructed? Nelson and Foster (1994) employed continuous record asymptotics to answer these questions in the univariate case. This paper considers the general multivariate case. Our results allow us, for example, to construct an asymptotically optimal ARCH model for estimating the conditional variance or conditional beta of a stock return given lagged returns on the stock, volume, market returns, implicit volatility from options contracts, and other relevant data. We also allow for time-varying shapes of conditional densities (e.g., "heteroskewticity" and "heterokurticity"). Examples are provided.

Daniel B. Nelson
Graduate School of Business
University of Chicago
1101 East 58th Street
Chicago, IL 60637
and NBER

1. Introduction

It is widely understood that conditional moments of asset returns are time varying. Understanding this variation is crucial in all areas of asset pricing theory. Accordingly, an enormous literature has developed on estimating conditional variances and covariances in returns data. Many techniques have been employed for this purpose, but perhaps the most widely used are the class of ARCH (autoregressive conditionally heteroskedastic) models introduced by Engle (1982) and expanded in many ways since. (See the survey papers of Bollerslev, Chou, and Kroner (1992), Bera and Higgins (1993), and Bollerslev, Engle, and Nelson (1993).)

In practice, researchers using ARCH models typically assume that the model is "true" and that apart from errors in parameter estimates in finite samples, the conditional variances produced by the model are "true." Another interpretation of ARCH models (see Nelson (1988, 1992), Foster and Nelson (1994), Nelson and Foster (1992,1994), Harvey, Ruiz, and Shephard (1994), Watanabe (1992)) is that they are not "true" per se, but instead are *filters* through which returns data can be passed to produce *estimates* of conditional variances and covariances. Under this interpretation, it is reasonable to ask how well misspecified ARCH models can estimate conditional variances and covariances, and how ARCH models can be constructed to optimally carry out this filtering. In particular, Nelson and Foster (1994) develop a continuous record asymptotic distribution theory for the conditional variance estimates generated by a univariate ARCH model, and also develop a class of asymptotically optimal ARCH conditional variance estimators. This paper extends these results to the multivariate case.

The extension from the univariate to the multivariate case is important for several reasons. Most obviously, models of asset pricing often involve conditional covariances, which

a univariate model cannot accommodate. Even if we are interested in estimating the volatility of a single variable, however, the multivariate extension is still important. Suppose, for example, that we are interested in estimating the conditional variance of a particular stock's return. While a univariate ARCH model would use only lagged returns and deterministic calendar effects, many other sources of information are also relevant: for example, implied volatilities from options prices, (e.g., Chiras and Manaster (1978), Day and Lewis (1992), trading volume, (e.g., Karpoff (1987), Gallant, Rossi, and Tauchen (1992)) high-low spreads, (e.g., Parkinson (1980), Garman and Klass (1980), Wiggins (1991)) interest rates, (e.g., Christie (1982), Glosten, Jagannathan, and Runkle (1993)), volatilities of domestic and foreign stock indices, (e.g., Braun, Nelson, and Sunier (1991) and King and Wadhwani (1990)) and volatilities of other stocks (e.g., Cox and Rubinstein (1985, pp. 280-285)). Utilizing such information requires a multivariate model. Finally, although it is conventional in ARCH modelling to keep the form of the conditional distribution constant (e.g., conditionally normal or conditionally Student's t with fixed degrees of freedom), it may be better to allow the conditional skewness or conditional kurtosis to be time varying, as has been suggested by Bera and Lee (1992) and Hansen (1992). If we require more than one unobservable state variable to describe the conditional distribution, we need a multivariate filtering theory.

The outline for the rest of the paper is as follows: in Section 2, we present and interpret the main filtering theorems. Section 3 considers the special case in which the data are generated by a diffusion. Section 4 develops examples of the main theorems. Section 5 is a brief conclusion.

2. Filtering theory for near-diffusions

The setup which we consider below is a multivariate generalization of the case considered in Nelson and Foster (1994) (henceforward *NF*). We will keep similar notation wherever possible. We refer the reader to *NF* for further discussion of the assumptions.

As in *NF*, there are two leading cases for the data generating processes, discrete time stochastic volatility and discretely observed diffusion. We present the results for near-diffusions in this section and handle the diffusion case in section 3.

In the stochastic volatility case we assume that for t = 0, h, 2h, ...,[1]

$$\begin{bmatrix} X_{t+h} \\ Y_{t+h} \end{bmatrix} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} + \begin{bmatrix} \mu(X_t, Y_t, t, h) \\ \kappa(X_t, Y_t, t, h) \end{bmatrix} h^{\delta} + \begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} h^{1/2} \tag{2.1}$$

$$E_t \begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} = \begin{bmatrix} 0_{n\times 1} \\ 0_{m\times 1} \end{bmatrix}, \quad Cov_t \begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} = \Omega(X_t, Y_t, t) \tag{2.2}$$

where δ equals either 3/4 or 1. We assume that for each $h > 0$, $\{X_t, Y_t\}_{t=0,h,2h,...}$ is Markovian, and that $[\xi_{X,t+h}', \xi_{Y,t+h}']'$ possess conditional densities $f(\xi_{X,t+h}, \xi_{Y,t+h} | X_t, Y_t, t)$, $f(\xi_{Y,t+h} | \xi_{X,t+h}, X_t, Y_t, t)$, and $f_h(\xi_{X,t+h} | X_t, Y_t, t)$. $\mu(\cdot)$ and $\kappa(\cdot)$ are continuous. (X_0, Y_0) may be fixed or random. X_t is an observable $n\times 1$ process, while Y_t is an unobservable $m\times 1$ process. Our interest is in estimating $\{Y_t\}$.

The analysis is asymptotic, in that h approaches 0. In (2.1)-(2.2), "t" is assumed to be a discrete multiple of h. To define (X_t, Y_t) for general t, set $(X_t, Y_t) \equiv (X_{h[t/h]}, Y_{h[t/h]})$, where [t/h]

---

[1] Notice that the scale factor $h^{1/2}$ on the ξ terms in (2.1) and (2.1') below is missing in the univariate case presented in NF equations (5.1) and (5.1'). These scale terms are correct here, but were inadvertently omitted in NF. In addition, the "h^{-1}" terms in NF (5.8)-(5.9) should be deleted.

is the integer part of t/h. As in *NF*, "E_t" and "Cov_t" denote, respectively, the expectation vector and covariances matrix conditional on time t information—i.e., the σ-algebra generated by $\{X_\tau, Y_\tau\}_{0 \le \tau \le t}$ or, equivalently for our purposes, by $(X_t, Y_t, \hat{Y}_t)$. ($\hat{Y}_t$ is defined below.)

We call (2.1)-(2.2) a near-diffusion, since when $\delta = 1$ and some mild regularity conditions are satisfied (see, e.g., Ethier and Kurtz (1986, Chapter 7, Theorem 4.1) it converges weakly as $h \downarrow 0$ to the diffusion process

$$d\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \mu(X_t, Y_t, t, 0) \\ \kappa(X_t, Y_t, t, 0) \end{bmatrix} dt + \Omega(X_t, Y_t, t)^{1/2} dW_t \tag{2.3}$$

where W_t is an $(n+m)\times 1$ standard Brownian motion.

As in *NF*, μ and κ drop out of the asymptotic distribution of the measurement error as $h \downarrow 0$ when $\delta = 1$. It is possible, however, to keep μ and κ in the asymptotic, provided we are willing to take $\delta = 3/4$, which we call the 'fast drift' case. Unfortunately, $\delta = 3/4$ is not, in general, compatible with a diffusion limit such as (2.3). Nevertheless, we will continue to call it a near-diffusion.

The Multivariate ARCH Model

We consider ARCH models[2] which generate estimates $\hat{Y}_t$ (for t = 0, h, 2h, 3h, ...) of the unobservable state variables by the recursion

$$\hat{Y}_{t+h} = \hat{Y}_t + h^\delta \cdot \hat{\kappa}(X_t, \hat{Y}_t, t, h) + h^{1/2} \cdot G(\hat{\xi}_{X,t+h}, X_t, \hat{Y}_t, t, h), \tag{2.4}$$

$$\hat{\xi}_{X,t+h} \equiv h^{-1/2}[X_{t+h} - X_t - h^\delta \cdot \hat{\mu}(X_t, \hat{Y}_t, t, h)]. \tag{2.5}$$

As in *NF*, $\hat{\kappa}(\cdot)$, $\hat{\mu}(\cdot)$, and $G(\cdot)$ are functions selected by the econometrician. G, $\hat{\kappa}$, and $\hat{\mu}$ are

[2] We call these ARCH models because the models (considered as data generating processes) make volatility a function of lagged residuals. A number of multivariate ARCH models take this form (see section 4 below.) The model in (2.4)-(2.5) also encompasses the extended Kalman filter, and the Autoregressive Conditional Density models of Hansen (1992).

continuous in all arguments and G is differentiable in $\hat{Y}_t$, $\xi_{X,t+h}$, and h almost everywhere and must possess one-sided derivatives everywhere. Just as μ and κ are the true drifts in X_t and Y_t, $\hat{\mu}$ and $\hat{\kappa}$ are the econometrician's (possibly misspecified) specifications of these drifts. $\hat{\xi}_{X,t+h}$ is a residual obtained using $\hat{\mu}$ in place of μ. The ARCH model treats this fitted residual $\hat{\xi}_{X,t+h}$ as if it were the true residual $\xi_{X,t+h}$. In ARCH models (considered for the moment as data generating processes rather than as filters) $\xi_{Y,t+h}$, the innovation in Y_{t+h}, is a function of $\xi_{X,t+h}$, X_t, Y_t, and t. $G(\cdot)$ is the econometrician's specification of this function. The ARCH model treats the fitted $\hat{Y}_t$ as if it were the true Y_t. Accordingly, we make the normalizing assumption that for all (X_t,Y_t,t,h), $E_t[G(\xi_{X,t+h},X_t,Y_t,t,h)\ ] = 0_{m\times 1}$.

We present results for the $\delta=3/4$, fast-drift case only. Why focus on the fast drift case? Admittedly, this experiment is a bit unnatural, since $\delta=3/4$ is generally not compatible with a diffusion limit for $\{X_t,Y_t\}$ as $h\downarrow 0$. In the standard drift case, however, the drift terms drop out of the asymptotic distribution of the ARCH model's measurement error in estimating Y_t. However, as Lo and Wang (1993) have argued, misspecification in conditional means may have an economically significant effect on volatility estimates, so it seems useful to develop an asymptotic in which the drifts do not drop out. This requires rescaling the drifts. By doing so, we adopt the view that asymptotics needn't correspond to a natural data experiment to be useful.[3] Finally, nothing is lost by using fast drifts, since the results for the $\delta=1$ case are recovered by setting $\mu(\cdot) = \hat{\mu}(\cdot) = 0_{n\times 1}$ and $\kappa(\cdot) = \hat{\kappa}(\cdot) = 0_{m\times 1}$.

---

[3] Others have taken this view: see, e.g., Phillips' (1988) "near integrated" processes, (e.g., Phillips (1988)) in which an AR root approaches 1 as the sample size grows, or the analysis of asymptotic local power, in which a sequence of alternative hypotheses collapse to the null at an appropriate rate as the sample size grows (see, e.g. Serfling (1980, Chapter 10)).

The Asymptotic Distribution of the Measurement Error

We now define the vector of normalized measurement errors

$$Q_t \equiv h^{-1/4}[\hat{Y}_t - Y_t] \tag{2.6}$$

As in *NF*, we may substitute from (2.6) for $\hat{Y}_t$ and expand $G(\hat{\xi}_{X,t+h}, X_t, \hat{Y}_t, t, h)$ in a Taylor series in h around h=0. When $\delta = 3/4$, we obtain[4]

$$\begin{aligned} Q_{t+h} = Q_t &+ h^{1/2}E_t[\partial G(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial Y]\cdot Q_t + h^{1/2}[\hat{\kappa}(X_t, Y_t, t, h) - \kappa(X_t, Y_t, t, h)] \\ &+ h^{1/2}E_t[\partial G(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial \xi_X]\cdot(\mu(X_t, Y_t, t, h) - \hat{\mu}(X_t, Y_t, t, h)) \\ &+ h^{1/4}[G(\xi_{X,t+h}, X_t, Y_t, t, 0) - \xi_{Y,t+h}] + O(h^{3/4}). \end{aligned} \tag{2.7}$$

In (2.7) $\partial G(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial Y$ is an m×m matrix of partial derivatives, with $i-j^{th}$ element $\partial G_i(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial Y_j$, and $\partial G(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial \xi_X$ is an m×n matrix with $i-j^{th}$ element $\partial G_i(\xi_{X,t+h}, X_t, Y_t, t, 0)/\partial \xi_{X_j}$.

Like the results in *NF*, we obtain our asymptotic results via passage to a continuous time (diffusion) limit process. This requires us to make assumptions on the behavior of the conditional moments of Q_t as we pass to continuous time:

Assumption 1: The following functions are well defined, continuous in t, and twice differentiable in X_t and Y_t.

$$\begin{aligned} A(X,Y,t) \equiv lim_{h\downarrow 0}\{ &[\hat{\kappa}(X,Y,t,h) - \kappa(X,Y,t)] \\ &+ E[\partial G(\xi_{X,t+h}, X_t, Y_t, t, h)/\partial \xi_X \mid (X_t, Y_t) = (X,Y)][\mu(X,Y,t) - \hat{\mu}(X,Y,t,h)]\} \end{aligned} \tag{2.8}$$

$$B(X,Y,t) \equiv - lim_{h\downarrow 0} E[\partial G(\xi_{X,t+h}, X_t, Y_t, t, h)/\partial Y \mid (X_t, Y_t) = (X,Y)], \tag{2.9}$$

$$C(X,Y,t) \equiv$$

$$lim_{h\downarrow 0} E[(G(\xi_{X,t+h}, X_t, Y_t, t, h) - \xi_{Y,t+h})(G(\xi_{X,t+h}, X_t, Y_t, t, h) - \xi_{Y,t+h})' \mid (X_t, Y_t) = (X,Y)]. \tag{2.10}$$

---

[4] For more details on the derivation of (2.7), see NF equations (3.7)-(3.9').

Further,

$$h^{-1/2}E[Q_{t+h}-Q_t \mid (X_t, Y_t, Q_t)=(X,Y,Q)] \to A(X,Y,t) - B(X,Y,t)\cdot Q, \tag{2.11}$$

$$h^{-1/2}Cov[Q_{t+h}-Q_t \mid (X_t, Y_t, Q_t)=(X,Y,Q)] \to C(X,Y,t) \tag{2.12}$$

as $h \downarrow 0$ uniformly on every bounded (X,Y,Q,t) set.[5]

We will often write 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. Define the matrix norm $\|A\| \equiv [Trace(AA')]^{1/2}$. For some $\delta > 0$

$$E[\| h^{-1/2}(Y_{t+h}-Y_t)\|^{2+\delta} \mid (X_t, Y_t)=(X,Y)], \tag{2.13}$$

$$E[\| h^{-1/2}(X_{t+h}-X_t)\|^{2+\delta} \mid (X_t, Y_t)=(X,Y)], \tag{2.14}$$

$$E[\| G(\hat{\xi}_{X,t+h}, X_t, \hat{Y}_t, t, h)\|^{2+\delta} \mid (X_t, Y_t, Q_t)=(X,Y,Q)] \tag{2.15}$$

are bounded as $h \downarrow 0$, uniformly on every bounded (X,Y,Q,t) set.

The first two conditional moments of $Q_{t+h}-Q_t$ are $O(h^{1/2})$, in contrast to the first two conditional moments of $X_{t+h}-X_t$ and $Y_{t+h}-Y_t$ in (2.1) and (2.2) when $\delta=1$,[6] which are O(h). $\{Q_t\}$ therefore oscillates much more rapidly as $h \downarrow 0$ than X_t or Y_t. In fact, $\{Q_t\}$ acts like heteroskedastic white noise (not at all like a diffusion) as $h \downarrow 0$. Our asymptotic results depend on our being able to derive a diffusion limit for Q_t. To do this, we must resort to a change in the time scales. Specifically, we choose a time T, a large positive number M, and a point in the state space (X,Y,Q), and condition on the event $(X_T,Y_T,Q_T) = (X,Y,Q)$. We then take the asymptotically vanishing time interval of calendar time $[T,T+M\cdot h^{1/2}]$ and stretch it into a time

---

[5] i.e., on every set of the form $\{(X,Y,Q,t): \|(X,Y,Q,t)\| < \Lambda\}$ for some finite, positive Λ. We could, for example, write (2.12) more formally as: For every Λ, $0 < \Lambda < \infty$,

$$(2.12') \lim_{h \downarrow 0} \sup_{\|(X,Y,Q,t)\| < \Lambda} \| h^{-1/2}Cov[Q_{t+h}-Q_t \mid (X_t,Y_t,Q_t)=(X,Y,Q)] - C(X,Y,t)\| = 0.$$

[6] When $\delta=3/4$ the first conditional moments are $O(h^{3/4})$.

interval [0,M] on a new, "fast" time scale. On this fast time scale, X_t and Y_t move more and more slowly as $h \downarrow 0$, becoming constant at the values X_T and Y_T. Q_t, on the other hand, converges to a diffusion limit. Essentially, our asymptotic analysis analyzes the behavior of Q_t in the neighborhood of given values of X_t, Y_t. This change of time scales is described in more detail and is graphically illustrated in *NF* figure 1, to which we refer the reader.

Another bit of notation first: we will make frequent use of two different types of matrix square roots. When we write $F^{1/2}$ for a positive semidefinte $d \times d$ matrix F, we mean the unique positive semidefinite (and therefore symmetric) matrix with $F^{1/2}F^{1/2} = F$. When we write $\underline{F^{1/2}}$, we mean any $d \times d$ matrix satisfying $\underline{F^{1/2}}\underline{F^{1/2}}' = F$.[7]

Formally, we define $Q_{T,\tau} \equiv Q_{T+\tau h^{1/2}}$, and obtain the multivariate version of the limit diffusion for $Q_{T,\tau}$ on the "fast" time scale

$$dQ_{T,\tau} = (A_T - B_T Q_{T,\tau})d\tau + C_T \underline{^{1/2}} dW_\tau^*, \tag{2.16}$$

where W_τ^* is an $m \times 1$ standard Brownian motion. $Q_{T,\tau}$ requires two time subscripts because of the change in the time scale: τ is the subscript on the fast time scale, while T is the starting point of the interval $[T, T+h^{1/2}M]$, over which the asymptotic analysis is conducted. In *NF*, the measurement error process followed an Ornstein-Uhlenbeck process, the continuous time version of a gaussian AR(1), on the fast time scale. Here, $Q_{T,\tau}$ follows the continuous time version of a gaussian VAR(1) (see, e.g., Karatzas and Shreve (1988, Section 5.6) or Arnold (1973, Section 8.2)). As in *NF*, we initialize the stochastic differential equation (2.16) by conditioning on X_T, Y_T, and $Q_{T,0}$.

---

[7] For proof that $F^{1/2}$ is unique, see Horn and Johnson (1985, Theorem 7.2.6). $\underline{F^{1/2}}$ need not be positive definite or symmetric, and is not in general unique: if U is any orthonormal real matrix, $(\underline{F^{1/2}}U)(\underline{F^{1/2}}U)' = F$.

Theorem 2.1. Let Assumptions 1-2 be satisfied, and let $0 < T < \infty$, $0 < \tau < \infty$. *Let* Θ *be a bounded, open subset of* R^{n+2m+1} *on which for some* $\epsilon > 0$

The real parts of all the eigenvalues of $B(X,Y,T)$ *are bounded below by* ϵ, (2.17)

$$\| A(X,Y,T) \| < 1/\epsilon, \ \| B(X,Y,T) \| < 1/\epsilon \ \text{and} \ \| C(X,Y,T) \| < 1/\epsilon \tag{2.18}$$

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 (2.16) as* $h \downarrow 0$. *This convergence is uniform on* Θ.

Proof: see Appendix.

Corollary. Under the conditions of Theorem 2.1, for every $(X_T, Y_T, Q_T, T) \in \Theta$ *and every* $\tau > 0$,

$$[Q_{T,\tau} \mid (X_T, Y_T, Q_T) = (X,Y,Q)] \Rightarrow N[b_{T,\tau}, V_{T,\tau}] \tag{2.19}$$

where "$\Rightarrow$*" denotes weak convergence as* $h \downarrow 0$, *and*

$$b_{T,\tau} = \exp[-B_T \cdot \tau] \cdot [Q + \int_0^\tau \exp[-B_T \cdot s] A_T ds] \tag{2.22}$$

$$= \exp[-B_T \cdot \tau] \cdot [Q - B_T^{-1} A_T] + B_T^{-1} A_T, \ \text{and}$$

$$V_{T,\tau} = \exp[-B_T \cdot \tau] \cdot [\int_0^\tau \exp[B_T \cdot s] C_T \exp[B_T' \cdot s] ds] \cdot \exp[-B_T' \cdot \tau], \tag{2.23}$$

or equivalently to (2.23)

$$vec(V_{T,\tau}) = -\exp[-(I_{m\times m} \otimes B_T + B_T \otimes I_{m\times m})\tau][[I_{m\times m} \otimes B_T + B_T \otimes I_{m\times m}]^{-1} vec(C_T)] \tag{2.24}$$

$$+ [I_{m\times m} \otimes B_T + B_T \otimes I_{m\times m}]^{-1} vec(C_T),$$

where "exp" is the matrix exponential, "$\otimes$*" is the Kronecker product and "vec" is the operator which stacks the columns of a matrix into a column vector.* $V_{T,\tau}$ *is positive semidefinite and symmetric for all* $\tau > 0$.

Proof: see Appendix.

Note that as in *NF*, the theorem yields weak convergence for $\{Q_{T,\tau}\}_{[0,M]}$—i.e., for the

time interval $[T,T+h^{1/2}M]$ on the standard time scale or $[0,M]$ on the "fast" time scale. Since this holds uniformly for every finite M, Lemma 5.2 of Helland (1982) guarantees that it also holds uniformly for M_h, where $M_h \to \infty$ sufficiently slowly as $h \downarrow 0$.[8] We then have $Q_{T,M_h} \Rightarrow N(b_T^*, V_T^*)$ as $h \downarrow 0$, where $\text{vec}(V_T^*) \equiv [I_{m\times m} \otimes B_T + B_T \otimes I_{m\times m}]^{-1}\text{vec}(C_T)$ and $b_T^* \equiv B_T^{-1}A_T$. Note that in general, deviation from zero of any element of either $\hat{\kappa}(X,Y,t,0) - \kappa(X,Y,t,0)$ or $\hat{\mu}(X,Y,t,0) - \mu(X,Y,t,0)$ can create asymptotic bias in other elements of $\hat{Y}$. V_T^* has two other representations which will later prove useful (see Karatzas and Shreve (1988, pp. 355-358)).

$$B_T V_T^* + V_T^* B_T' = C_T \tag{2.25}$$

$$V_T^* = \int_0^\infty \exp[-B_T \cdot s] C_T \exp[-B_T' \cdot s] ds. \tag{2.26}$$

Because the matrix exponential in (2.26) is never singular (Bellman (1970 p. 170)) and the sum or integral of positive definite matrices is positive definite (e.g., Magnus and Neudecker (1988, Chapter 11, Theorem 9)), V_T is singular if and only if C_T is. B_T and C_T are continuous functions of X_t, Y_t, and t. Since X_t, Y_t, and t are (asymptotically) constant at their time T values on the 'fast' time scale, B_T and C_T are asymptotically equivalent to B_t and C_t for $T \le t \le T+h^{1/2}M_h$. This local constancy of X_t, Y_t, and t makes possible the transformation of the filtering problem from a highly nonlinear (and completely intractable) problem to a solvable problem.

To interpret the Corollary, recall that $Q_{T,\tau} \equiv Q_{T+\tau h^{1/2}}$, which we combine with (2.22)-(2.23) to obtain

$$Y_{T+\tau h^{1/2}} \approx \hat{Y}_{T+\tau h^{1/2}} - h^{1/4}\tilde{Q}_{T+\tau h^{1/2}}, \text{ where } [\tilde{Q}_{T+\tau h^{1/2}} | \hat{Y}_T, Y_T, X_T] \sim N(b_{T+\tau h^{1/2}}, V_{T+\tau h^{1/2}}) \tag{2.27}$$

Of course, Y_T and Q_T are unobservable. If, however, the filter has been running a "long" time

---

[8] Helland's Lemma 5.2 does not specify the rate at which $M_h \to \infty$, only that there is such a rate. It may be very slow, e.g., $\ln[\ln(h^{-1})]$. In our example, M_h must surely increase at a slower rate than $h^{-1/2}$, or else (X_t, Y_t, t) would not be asymptotically constant on $[T,T+h^{1/2}M_h]$.

(i.e., it was initialized at some time t-$M_h h^{1/2}$, where M_h goes to infinity sufficiently slowly as $h \downarrow 0$), then we have

$$Y_t \approx \hat{Y}_t - h^{1/4} Q_t, \text{ where } Q_t \sim N(b_t, V_t), \tag{2.28}$$

where b_t and V_t can be evaluated at $(X_t, \hat{Y}_t, t)$, allowing us to characterize the uncertainty in Y_T given the information in X_t and $\hat{Y}_t$. This allows us to draw confidence bands and so forth. Note, however, that the information on Y_t in X_t and $\hat{Y}_t$ does not fully summarize all information in the sample path of $\{X_t\}_{0 \le t \le T}$, since our filter is not an implementation of Bayes' theorem. Implementing Bayes' theorem analytically seems impossible in our problem—if it were possible, we would naturally prefer to do so. Even numerical implementation of Bayes' theorem is rarely possible with current methods: see, however, the recent work of Jacquier, Polson and Rossi (1992) for a successful implementation in an important special case.

Optimality

In *NF*, optimality was defined in terms of minimizing V_T^* while eliminating b_T^*. Since there is no bias-variance tradeoff, this is equivalent to minimizing the asymptotic mean-squared error. A natural multivariate generalization is to minimize Trace$[b_T^* b_T^{*\prime} + V_T^*]$. We could just as easily minimize $u'[b_T^* b_T^{*\prime} + V_T^*]u$ for an arbitrarily selected $m \times 1$ vector u. None of the optimality results would be affected, because the proofs of the optimality results show that if b_0 and V_0 are the bias vector and error covariance matrix achieved by optimal filters proposed below and b_1 and V_1 are the bias and covariance matrix achieved by any other filter then $[b_1 b_1' + V_1] - [b_0 b_0' + V_0]$ is positive semidefinite.

As in *NF*, there are two sources of uncertainty at time t in estimating the value of Y_{t+h}: first, there is uncertainty about the first difference $Y_{t+h} - Y_t$. Second, there is uncertainty about

the level of Y_t. Asymptotically, these two sources of uncertainty are of the same order ($O_p(h^{1/4})$). Correspondingly, the optimal filter turns out to have two terms. The first, which we call P, is the *prediction component* which extracts the information in $\xi_{X,t+h}$ regarding $Y_{t+h}-Y_t$. Formally,

$$P(\xi_X,X,Y,t) \equiv E[\xi_{Y,t+h}|(\xi_{X,t+h},X_t,Y_t)=(\xi_X,X,Y)]. \tag{2.29}$$

The second term, which we call S, is the estimation or *score component*, the $m\times 1$ vector

$$S(\xi_X,X,Y,t) \equiv \partial \ln[f(\xi_{X,t+h}|(X_t,Y_t)=(X,Y))]/\partial Y, \tag{2.30}$$

where in (2.30) $\xi_{X,t+h}$ is evaluated at ξ_X. This term extracts, in a manner analogous to maximum likelihood estimation with Y_t treated as a parameter, information in $\xi_{X,t+h}$ on the *level* of Y_t. We will further interpret these terms in the theorems below.

To simplify notation, we write S_{t+h} for $S(\xi_{X,t+h},X_t,Y_t,t)$ and P_{t+h} for $P(\xi_{X,t+h},X_t,Y_t,t)$. When we take conditional expectations we will drop more time subscripts and write, for example, $E_t[SS']$ for $E_t[S_{t+h}S_{t+h}']$ and $E_t[(\xi_Y-P)(\xi_Y-P)']$ for $E_t[(\xi_{Y,t+h}-P_{t+h})(\xi_{Y,t+h}-P_{t+h})']$.

Assumption 3. For every h, the conditional densities $f(\xi_X,\xi_Y|X,Y,t)$ and $f(\xi_X|X,Y,t,h)$ are well defined and continuous in X, t, and h, and $f(\xi_X|X,Y,t,h)$ is continuously differentiable in Y almost everywhere, with one sided partial derivatives with respect to Y everywhere. Further, for some $\delta > 0$

$$E[\|P_{t+h}\|^{2+\delta}|X_t=X,Y_t=Y], \text{ and} \tag{2.31}$$

$$E[\|S_{t+h}\|^{2+\delta}|X_t=X,Y_t=Y] \tag{2.32}$$

are bounded uniformly on every bounded (X,Y,t) set as $h\downarrow 0$.

Assumption 4. Let there exist a unique, positive semidefinite solution ω_T to the matrix Riccati equation:

$$E_T[PS']\omega_T + \omega_T E_T[SP'] + \omega_T E_T[SS']\omega_T = E_T[(\xi_Y-P)(\xi_Y-P)']. \tag{2.33}$$

Perhaps the simplest sufficient condition for a unique positive semidefinite solution to (2.33) is that $E_T[SS']$ is positive definite (this is easily verified using Kučera (1972, Condition 3)). This condition can sometimes be weakened—see Kučera (1972) and Lancaster and Rodman (1980). Using (2.26), the definitions in (2.9)-(2.10), and Theorem 2.2 below, it is also easy to check that positive definiteness of $E_T[SS']$ and $E_T[(\xi_Y - P)(\xi_Y - P)']$ implies that ω_T is positive definite.

Theorem 2.2. Let Assumptions 1, 3, and 4 be satisfied. A set of sufficient conditions for $Trace[b_T^* b_T^{*\prime} + V_T^*]$ *to be minimized is that*

$$\lim_{h \downarrow 0} \hat{\kappa}(X,Y,T,h) - \kappa(X,Y,T,h) = 0, \tag{2.34}$$

$$\lim_{h \downarrow 0} \hat{\mu}(X,Y,T,h) - \mu(X,Y,T,h) = 0, \text{ and} \tag{2.35}$$

$$G(\xi_X, X, Y, T, h) = P(\xi_X, X, Y, T) + \omega_T S(\xi_X, X, Y, T), \tag{2.36}$$

where ω_T *is the positive semidefinite solution to (2.33). The minimized* $V_T^* = \omega_T$ *and the minimized* $b_T^* b_T^{*\prime} = 0_{m \times m}$. *Let* $\tilde{G}$ *satisfy the regularity conditions of Theorem 2.1 and let* $\tilde{V}_T$ *be the asymptotic error covariance matrix delivered by Theorem 2.1 using* $\tilde{G}$ *in place of the G in (2.36). Then* $\tilde{V}_T = \omega_T$ *if and only if* $\tilde{G}(\xi_X, X, Y, T, h) = G(\xi_X, X, Y, T, h)$ *almost surely.*

Proof: see Appendix.

In *NF*, the optimal G_{T+h} was equal to P_{T+h} plus ω_T times S_{T+h}. This might lead one to expect that for i=1 to m, $G_{i,T+h}$ (i.e., the i^{th} element of the vector G_{T+h}) would be a linear combination of $P_{i,T+h}$ and $S_{i,T+h}$. Perhaps surprisingly, this is not the case when ω_T is non-diagonal: $G_{i,T+h}$ equals $P_{i,T+h}$ plus a matrix-weighted average of the elements of S_{T+h}.

$E_t[SP']$, $E_t[SS']$, and $E_t[(\xi_Y - P)(\xi_Y - P)']$ are matrix functions of X_t, Y_t, and t. Through the matrix Riccati equation, ω_t is as well. For $t \in [T, T+h^{1/2}M_h]$, $(X_t, Y_t, t) \to (X_T, Y_T, T)$ as $h \downarrow 0$.

On the diffusion limit on the transformed time scale, (X_t, Y_t, t) are constants evaluated at (X_T, Y_T, T). For our purposes, $E_t[SP']$, $E_t[SS']$, $E_t[(\xi_Y - P)(\xi_Y - P)']$, and ω_t are asymptotically equivalent to $E_T[SP']$, $E_T[SS']$, $E_T[(\xi_Y - P)(\xi_Y - P)']$, and ω_T when $T \leq t \leq T + h^{1/2}M_h$.

The assumptions of theorems 2.1 and 2.2 are not as general as we would like. The first-order Markov structure immediately rules out, for example, fractionally integrated models (see, e.g., Baillie, Bollerslev, and Mikkelsen (1993)). As is well-known, finite-order markov models can written in first order Markov form. This, however, does not usually help in our setup. Suppose, for example, that X_{t+h} is a scalar with conditional variance σ_t^2, and that $\ln(\sigma_t^2)$ is a linear ARMA(2,1). To write it in first-order markov form, suppose we are able to decompose $\ln(\sigma_t^2)$ into the sum of two linear AR(1) components $y_{1,t}$ and $y_{2,t}$. $\ln(\sigma_t^2) = y_{1,t} + y_{2,t}$ appears in the score term S_{t+h}, but $y_{1,t}$ and $y_{2,t}$ do not appear individually, and so $E_t[SS']$ is singular. In many cases, this prevents existence (let alone uniqueness) of a solution to the Riccati equation. The B_T matrix of theorem 2.1 is generally singular in this case, so both theorems 2.1 and 2.2 break down. Similar problems arise for many higher order markov processes: in general to avoid singular $E_t[SS']$, we need *all* the elements of Y_t to enter the score S_{t+h} directly in a nondegenerate way. Extending this paper's results to higher-order models is left for further research.

An Important Special Case

In *NF*, a closed form was available for ω_T. In the multivariate case, unfortunately, there often is not, though there is a large literature on numerical and other techniques for solving equations such as (2.33), which arise in linear control and Kalman filtering problems—see, e.g., Anderson and Moore (1971, Chapter 15). Theorem 2.3 considers the important special case in

which $E_T[SS']$ is positive definite and $E_T[PS'] = 0_{m\times m}$:

Theorem 2.3. (a) Let $E_T[SS']$ be positive definite and let $E_T[PS'] = 0_{m\times m}$. Then the unique positive semidefinite solution to the Riccati equation of Assumption 4 is[9]

$$\omega_T = (E_T[SS'])^{-1/2}[(E_T[SS'])^{1/2}(E_T[(\xi_Y-P)(\xi_Y-P)'])(E_T[SS'])^{1/2}]^{1/2}(E_T[SS'])^{-1/2}, \qquad (2.37)$$

(b) In addition to the assumptions of Theorem 2.2, let the joint conditional density of $[\xi_X, \xi_Y]$ be elliptically symmetric. Then $P_{t+h} = E_t[\xi_Y\xi_X'](E_t[\xi_X\xi_X'])^{-1}\xi_{X,t+h}$ (i.e., given X_t and Y_t, the prediction component P_{t+h} is linear in $\xi_{X,t+h}$) and $E_t[PS'] = 0_{m\times m}$.

Proof: see Appendix.

The intuition behind (b) is straightforward. Let the $k\times 1$ vector $Z \sim N(0,\Omega)$. As is well known, $E[Z_i|Z_k\ldots Z_m]$ is computed as a linear regression, and the differential with respect to Ω of the log of the density is $\frac{1}{2}\text{Trace}(d\Omega)\Omega^{-1}(ZZ'-\Omega)\Omega^{-1}$,[10] which is, of course, orthogonal to any linear combination of the Z_i's. While the functional form of the score is different for other elliptically symmetric distributions, the orthogonality between the Z's and the score for parameters of the covariance matrix still holds, as does the linear regression structure of the conditional expectations—see e.g., Cambanis, Huang, and Simons (1981) and Mitchell (1989).

The elliptically symmetric case includes, for example, the multivariate normal, multivariate t, as well as the case in which $\{X_t,Y_t\}$ is generated by a discretely observed diffusion (see Theorem 3.1 below.)

---

[9] $E_T[SS']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$ can typically be computed analytically as functions of the state variables. Given numerical values of $E_T[SS']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$, fast algorithms are available to compute ω_T in (2.37)—for example, using the MatrixPower command in version 2.2 of *Mathematica* (Wolfram Research, Inc. (1992)). Given access to routines for computing eigenvalues and eigenvectors (for example the eigrg2 and eigrs2 commands in *Gauss* (Aptech Systems Inc. (1992))), it is also easy to write code to compute matrix square roots of real symmetric matrices.

[10] see, e.g., Magnus and Neudecker (1988, Section 15.3).

Fisher Information and Unpredictable Components

To interpret the matrix Riccati equation of Assumption 4, consider first the simplest case, in which $E_T[SS']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$ are diagonal and $E_T[PS'] = 0_{m\times m}$. (2.37) then simplifies to

$$\omega_T = (E_T[SS'])^{-1/2}(E_T[(\xi_Y-P)(\xi_Y-P)'])^{1/2}. \tag{2.38}$$

$E_T[SS']$ is the filtering analogue of the Fisher information: the larger the Fisher information, the smaller ω_T. $E_T[(\xi_Y-P)(\xi_Y-P)']$, on the other hand, is a measure of the variance in the innovations variance in Y_{t+h} *after the predictable component* $P(\xi_{X,t+h}, X_t, Y_t, t)$ *has been removed.* Again, the larger this residual variance, the more unpredictably variable $\{Y_t\}$ is, and consequently the larger the measurement error covariance matrix ω_T is. This argument can be extended to the general case in which $E_T[SS']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$ may or may not be diagonal and $E_T[PS']$ may not be a matrix of zeros:

Theorem 2.4. Let $E_T[SS']$ *and* $E_T[(\xi_Y-P)(\xi_Y-P)']$ *be positive definite. Let* Δ *be* $m\times m$ *with* Δ *positive semidefinite and* $\Delta \neq 0_{m\times m}$, *let* ζ *be a scalar, and define* $\omega_T(\zeta)$ *implicitly by*

$$E_T[PS']\omega_T(\zeta) + \omega_T(\zeta)E_T[SP'] + \omega_T(\zeta)E_T[SS']\omega_T(\zeta) = E_T[(\xi_Y-P)(\xi_Y-P)'] + \zeta\Delta. \tag{2.33'}$$

Then $d\omega_T(\zeta)/d\zeta$ *evaluated at* $\zeta = 0$ *equals*

$$\int_0^\infty \exp[-s(E_T[PS']+\omega_T(0)E_T[SS'])]\Delta\exp[-s(E_T[SP']+E_T[SS']\omega_T(0))]ds \tag{2.39}$$

which is positive semidefinite and non-null. Similarly, let β *be* $m\times m$ *with* β *positive semidefinite and* $\beta \neq 0_{m\times m}$, *and let* δ *be a scalar. Define* $\omega_T(\delta)$ *implicitly by*

$$E_T[PS']\omega_T(\delta) + \omega_T(\delta)E_T[SP'] + \omega_T(\delta)(E_T[SS']+\delta\cdot\beta)\omega_T(\delta) = E_T[(\xi_Y-P)(\xi_Y-P)']. \tag{2.33''}$$

Then $d\omega_T(\delta)/d\delta$ *evaluated at* $\delta = 0$ *equals*

$$-\int_0^\infty e^{-s(E_t[PS']+\omega_t(0)E_t[SS'])}\beta e^{-s(E_t[SP']+E_t[SS']\omega_t(0))}ds \tag{2.40}$$

which is negative semidefinite and non-null. Finally, let α *be an* $m\times m$ *matrix, not necessarily symmetric, and let* η *be a scalar. Define* $\omega_T(\eta)$ *implicitly by*

$$(E_t[PS']+\eta\cdot\alpha)\omega_T(\eta) + \omega_T(\eta)(\eta\cdot\alpha'+E_t[SP']) + \omega_T(\eta)E_t[SS']\omega(\eta) \tag{2.33'''}$$
$$= E_t[(\xi_Y-P)(\xi_Y-P)']$$

Then $d\omega_T(\eta)/d\eta$ *evaluated at* $\eta = 0$ *equals*

$$-\int_0^\infty e^{-s(E_t[PS']+\omega_t(0)E_t[SS'])}(\alpha\omega_T(0) + \omega_T(0)\alpha')\, e^{-s(E_t[SP']+E_t[SS']\omega_t(0))}ds \tag{2.41}$$

Proof: see Appendix.

$d\omega_T(\eta)/d\eta$ is less readily interpretable than are $d\omega_T(\zeta)/d\zeta$ and $d\omega_T(\delta)/d\delta$, since we cannot easily identify it as positive or negative semidefinite using simple assumptions on α. Clearly $d\omega_T(\eta)/d\eta$ is negative semidefinite if $\alpha\omega_T + \omega_T\alpha'$ is positive semidefinite, but even if we are willing to assume, say, that $\alpha + \alpha'$ is positive semidefinite, $\alpha\omega_T + \omega_T\alpha'$ may not be. If α is positive semidefinite, $\alpha\omega_T + \omega_T\alpha'$ is as well (Taussky (1968, p. 177)). But there is no particular reason why α should even be symmetric.

Conditional Moment Matching

One important aspect of the *NF* filtering results which is preserved in the multivariate setting is moment matching—i.e., for both the true data generating process and the ARCH model *interpreted as a data generating process*, the first two conditional moments are functions of the time t and the state variables X_t and Y_t. In *NF*, efficiency required that these functions be identical in the ARCH model and the true data generating process. Under mild regularity conditions, the same is true here. Theorem 2.2 requires matching the first conditional

moments—i.e., $\hat{\kappa}(X,Y,T,0) \equiv \kappa(X,Y,T,0)$ and $\hat{\mu}(X,Y,T,0) \equiv \mu(X,Y,T,0)$. It also matches the second moments:

Theorem 2.5. Let the conditions of Theorem 2.2 be satisfied, and let $G(\cdot)$ be given by (2.36). Then

$$E_\mu[G(\xi_{X,t+h}, X_t, Y_t, t, 0)G(\xi_{X,t+h}, X_t, Y_t, t, 0)'] = E_\mu[\xi_{Y,t+h}\,\xi_{Y,t+h}']. \tag{2.42}$$

In addition, given X and t, let $\xi_X[f(\xi_X|X,Y,t)]$ be uniformly continuous in Y on $(\xi_X, Y) \in R^{n+m}$. Then

$$E_\mu[G(\xi_{X,t+h}, X_t, Y_t, t, 0)\xi_{X,t+h}'] = E_\mu[\xi_{Y,t+h}\,\xi_{X,t+h}']. \tag{2.43}$$

Proof: see Appendix.

In a sense therefore, the asymptotically optimal ARCH model makes itself as much like the true data generating process as possible. In filtering, as we have noted, the first moment terms μ, $\hat{\mu}$, κ, and $\hat{\kappa}$ are only second-order important. Because it matches both of the first two conditional moments, the optimal filter will perform well at both filtering and forecasting—see Nelson and Foster (1994).

Moment matching also has a useful practical application as a shortcut in computing the optimal filter. Once the functional forms of S and P are known, ω can often be computed via the moment matching condition, which is frequently easier than solving the matrix Riccati equation. The model of Bollerslev, Engle, and Wooldridge (1988) analyzed in section 4 below provides an example.

Partially Optimal Filtering and Nuisance State Variables

Consideration of nuisance state variables illustrates a further parallel between optimal ARCH filters and maximum likelihood parameter estimation. In particular, suppose that only a

subset of the unobservable state variables Y_t—say the first m_1 elements of Y_t, are of direct interest—that is, we are interested in minimizing $u'[b_T^* b_T^{*\prime} + V_T^*]u$, where $u' = [1_{1\times m_1}, 0_{1\times(m-m_1)}]$. Is it necessary for *all* of the elements of $G(\cdot)$ to equal the optimal G of Theorem 2.2? Under what circumstances can we use suboptimal filters to estimate the nuisance state variables (the last $m-m_1$ elements of Y_t) while retaining efficiency in estimating the first m_1 elements of Y_t?

Theorem 2.6. Let ω_T and G be as in Theorem 2.2. Let $E_T[SS']$ and $E_T[(\xi_Y - P)(\xi_Y - P)']$ be positive definite, and let $E_T[PS']$, $E_T[SS']$ and $E_T[(\xi_Y - P)(\xi_Y - P)']$ be block diagonal—i.e., of the form

$$\begin{bmatrix} K & 0 \\ 0' & J \end{bmatrix}, \tag{2.44}$$

where K is $m_1 \times m_1$, J is $(m-m_1)\times(m-m_1)$, "0" is an $m_1\times(m-m_1)$ matrix of zeros. Then ω_T is block diagonal of the same form. Now define an $m\times 1$ function $\tilde{G}(\xi_{X,t+h}, X_t, Y_t, t, h)$ which satisfies Assumptions 1 and 2, and whose first m_1 elements equal the corresponding elements of G. Call $\tilde{V}_T$ the steady state error covariance matrix delivered by Theorem 2.1 using this $\tilde{G}$. Write $\tilde{V}_T$ and ω_T in partitioned form as

$$\tilde{V}_T = \begin{bmatrix} \tilde{V}_{T,1} & \tilde{V}_{T,2} \\ \tilde{V}_{T,2}' & \tilde{V}_{T,3} \end{bmatrix}, \quad \omega_T = \begin{bmatrix} \omega_{T,1} & 0 \\ 0' & \omega_{T,3} \end{bmatrix}, \tag{2.45}$$

where $\tilde{V}_{T,1}$ and $\omega_{T,1}$ are $m_1\times m_1$, $\tilde{V}_{T,2}$ and "0" are $m_1\times(m-m_1)$, and $\tilde{V}_{T,3}$ and $\omega_{T,3}$ are $(m-m_1)\times(m-m_1)$. Then $\tilde{V}_{T,1} = \omega_{T,1}$.

Proof: see Appendix.

Note that the conditions (2.44) and (2.45) allow all the elements of P and S to depend

on any or all of the elements of Y_t—i.e., we do not require that the first m_1 elements of Y_t be independent (or conditionally independent) of the last $(m-m_1)$ elements.

Theorem 2.6 is closely connected to the familiar result from maximum likelihood estimation that if the information matrix $E_T[SS']$ is block diagonal between the parameters of interest and the nuisance parameters, then consistent but inefficient estimates of the nuisance parameters may be employed without preventing asymptotically efficient estimation of the parameters of interest (see, e.g., Cox and Reid (1987) or the application in Engle (1982, sections 5-6)). Our case, however, is made more complicated by the presence of the terms $E_T[SP']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$, which also must be block diagonal.

Change of Variables

So far, we have considered only optimality *given* the definitions of the state variables X_t and Y_t. Is there an optimal way to define the state variables? Suppose we define the functions

$$\Upsilon_t \equiv \Upsilon(X_t,Y_t,t), \quad \chi_t \equiv \chi(X_t,t), \tag{2.46}$$

where $\Upsilon(X,Y,t)$ and $\chi(X,t)$ are twice continuously differentiable with $\partial\Upsilon(X,Y,t)/\partial Y$ and $\partial\chi(X,t)/\partial X$ nonsingular. Given t, there is a one-to-one mapping from χ_t to X_t and, given t and X_t, there is a one-to-one mapping from Υ_t to Y_t, so the σ-algebra generated by $\{X_t,Y_t\}$ is the same as that generated by $\{\chi_t,\Upsilon_t\}$, and the σ-algebra generated by $\{X_t\}$ is the same as that generated by $\{\chi_t\}$. Like X_t, χ_t is $n\times 1$ and observable. Like Y_t, Υ_t is $m\times 1$ and unobservable.

If we first construct asymptotically optimal estimates of Y_t using Theorem 2.2 and then apply the delta method (e.g., Serfling (1980, Section 3.3 Theorem A)), we derive the asymptotic

distribution of $\Xi_t \equiv h^{-1/4}[\Upsilon(X_t,\hat{Y}_t,t)-\Upsilon(X_t,Y_t,t)]$ as

$$\Xi_t \rightarrow N\left[0_{m\times 1}, \left[\frac{\partial \Upsilon}{\partial Y}\,\omega\,\frac{\partial \Upsilon}{\partial Y'}\right]\right], \tag{2.47}$$

where the i-j^{th} element of $\partial\Upsilon/\partial Y$ is $\partial\Upsilon_i/\partial Y_j$.

Suppose, however, that we change variables *first*, and then compute the optimal filter for the $\{\chi_t,\Upsilon_t\}$ process using Theorem 2.2. Is it possible, by a judicious choice of $\Upsilon(X,Y,t)$ and $\chi(X,t)$ to achieve an asymptotic covariance matrix of $h^{-1/4}[\hat{\Upsilon}_t-\Upsilon_t]$ smaller than the matrix in (2.47)? In yet another parallel to maximum likelihood estimation, the answer is, in general, no.

Theorem 2.7. Define the state variables Υ_t and χ_t as in (2.46). In addition, let $\xi_X f(\xi_X|X,Y,t)$ be uniformly continuous in Y on $(\xi_X,Y)\in R^{n+m}$, and let there be a $\delta > 0$ such that

$$E[\,\|\,h^{-1/2}(\Upsilon_{t+h}-\Upsilon_t)\,\|^{\,2+\delta}\,|\,(\chi_t,\Upsilon_t)=(\chi,\Upsilon)], \text{ and} \tag{2.48}$$

$$E[\,\|\,h^{-1/2}(\chi_{t+h}-\chi_t)\,\|^{\,2+\delta}\,|\,(\chi_t,\Upsilon_t)=(\chi,\Upsilon)] \tag{2.49}$$

are bounded as $h\downarrow 0$, uniformly on every bounded (χ,Υ,t) set. Then the asymptotic distribution of $h^{-1/4}[\hat{\Upsilon}_t-\Upsilon_t]$ achieved by the asymptotically optimal filter for this system is given by (2.47).

Proof: see Appendix.

So we are not able to improve the asymptotic performance of the filter via a change of variables. For nonzero h, however, such transformations may be important. Nelson and Schwartz (1992) and Schwartz (1994) show that in monte carlo experiments, transformations of the state variables which reduce or eliminate the dependence of ω_Υ on Y_t substantially improve the asymptotic approximation for $h > 0$. This, of course, has many parallels in the literature on maximum likelihood going back at least to Fisher (1921).

If the assumptions of Theorem 2.1 are not satisfied, (if, for example, the limit data

generating process is a jump diffusion rather than a diffusion) transformations may be important for another reason: the "near-diffusion" assumption guaranteed that the increments in the state variables $\{X_t, Y_t\}$ would be small over small time intervals. This allowed us to approximate the increments in the transformed process $\{T_t, \chi_t\}$ using a two-term Taylor series expansion of the functions $T(\cdot)$ and $\chi(\cdot)$. If jumps are present asymptotically as $h \downarrow 0$ (i.e., if the conditional distribution of ξ_X and ξ_Y is too thick tailed) such an expansion is invalid, and the global, rather than just the local, properties of $T(\cdot)$ and $\chi(\cdot)$ become relevant—for example, NF show that under our regularity conditions, the filtering performance of the EGARCH model of Nelson (1991) is relatively robust to the presence of thick tailed errors. However, as Engle and Ng (1993) point out, EGARCH arrives at $\hat{\sigma}_t$ by *exponentiating* a function of rescaled lagged residuals, so when the normalized residual is huge (e.g., October 19, 1987) the two term taylor series approximation may break down, and because of the tail behavior of $\exp(\cdot)$, the 'robust' EGARCH model may become highly non-robust.

3. Diffusions

We now consider the case in which the data are generated by a diffusion:

$$d\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \mu(X_t, Y_t, t) \\ \kappa(X_t, Y_t, t) \end{bmatrix} h^{\delta - 1} dt + \Omega(X_t, Y_t, t)^{1/2} dW_t \tag{3.1}$$

where X_t and $\mu(X_t, Y_t, t)$ are $n \times 1$ and Y_t and $\kappa(X_t, Y_t, t)$ are $m \times 1$, W_t is an $(n+m) \times 1$ standard Brownian motion, and $\Omega(X_t, Y_t, t)$ is $(n+m) \times (n+m)$. (X_0, Y_0) are assumed random but independent of $\{W_t\}_{0 \le t < \infty}$. We also assume that (3.1) has a unique weak-sense solution for each h, $0 < h < 1$.

We assume that $\{X_t\}$ is observable at discrete intervals of length h. As in the near-

diffusion case, our interest is in estimating $\{Y_t\}$ using an ARCH model of the form (2.4)-(2.5). As we will see, the results for the diffusion case are identical to those for the near-diffusion case when $[\xi_{X,t+h}', \xi_{Y,t+h}']'$ is conditionally multivariate normal, i.e., for the model

$$\begin{bmatrix} X_{t+h} \\ Y_{t+h} \end{bmatrix} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} + \begin{bmatrix} \mu(X_t, Y_t, t) \\ \kappa(X_t, Y_t, t) \end{bmatrix} h^{\delta} + \begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} h^{1/2}, \text{ where} \tag{3.2}$$

$$\begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} | X_t, Y_t \sim N[\, 0_{(m+n)\times 1}, \Omega(X_t, Y_t, t) \,] \tag{3.3}$$

Theorem 3.1. For each h, $0 < h < 1$, let the diffusion (3.1) possess a unique weak-sense solution, and let $\mu(\cdot)$, $\kappa(\cdot)$, and $\Omega(\cdot)$ be continuous. Define $G(\cdot)$, $\hat{\mu}$, and $\hat{\kappa}$ as in Theorem 2.1, where G has two partial derivatives with respect to ξ_X and Y almost everywhere, and let the absolute values of G and these two partial derivatives be bounded above by some polynomial in ξ_X with the coefficients of the polynomial continuous in X_t, Y_t, and t. Then all the results of Section 2 that are valid for the conditionally normal stochastic volatility model (3.2)-(3.3) hold for the diffusion model (3.1).

Proof: See Appendix.

Essentially, Theorem 3.1 allows us to treat a discretely observed diffusion as if it were generated by a conditionally normal stochastic difference equation. For example, since the multivariate normal is elliptically symmetric, the results of Theorems 2.3 and 2.5 are satisfied for diffusions.

The major complication that arises in the Proof of Theorem 3.1 is that the moments required in Assumptions 1-3 may not exist. Even when these moments do exist, proving so can be quite tedious (see, e.g., the proofs of Theorems 4.3-4.5 in NF), since most of the stochastic

volatility diffusion models fail the usual "growth condition" (see, e.g., Arnold (1973, Chapter 6, 6.2.5)) that guarantees the existence of arbitrary finite conditional moments.

A better approach is to use a diffusion approximation result that does not require bounded conditional moments. In the proof of Theorem 3.1, we adopt this approach, based on Ethier and Kurtz (1986, Chapter 7, Corollary to Theorem 4.1)) in place of NF, Theorem 2.1. NF's conditions are easier to check in the near-diffusion case, while Ethier and Kurtz's conditions are easier to check in the diffusion case.

4. Examples

We next turn to selected examples of the results of the first three sections.

4.1 Conditional Heterokurticity

ARCH models generally assume a constant shape to the conditional distribution—e.g., a conditional Student's t distribution with constant degrees of freedom. There is evidence, however, that the shapes of the conditional distributions of asset returns may be time-varying. This is illustrated in Figure 1, which plots standardized residuals $\hat{\xi}_{X,t}/\hat{\sigma}_t$ exceeding four in absolute value, where the $\hat{\sigma}_t$'s are generated by a univariate EGARCH model fit to S&P 500 daily returns from January 1928 through December 1990. If the EGARCH model is correctly specified (or if is a relatively efficient filter), the standardized residuals should be approximately iid. Figure 1 captures the outliers. It is quite clear that the large residuals clump together over time—i.e., there are many more outliers in the 1940's, 1950's and late 1980's than during other periods. The conditional skewness may also be changing over time.[11]

[11] For evidence of time-varying higher-order conditional moments found using other methodologies, see Bates (1991,1993) and Turner and Weigel (1992).

To keep things relatively simple, we will next consider a model with changing conditional kurtosis but constant (zero) conditional skewness. In particular, we consider a model in which the (scalar) observable state variable x_t exhibits both time varying conditional variance (governed by the unobservable state variable y_t) and time varying conditional tail-thickness (governed by the unobservable state variable ν_t). In particular, we assume that $[X_{t+h}', Y_{t+h}']'$ is conditionally multivariate t[12] with 2 + exp(ν_t) degrees of freedom:

$$\begin{bmatrix} x_{t+h} \\ y_{t+h} \\ \nu_{t+h} \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \\ \nu_t \end{bmatrix} + \begin{bmatrix} \mu_1(x_t, y_t, \nu_t) \\ \mu_2(x_t, y_t, \nu_t) \\ \mu_3(x_t, y_t, \nu_t) \end{bmatrix} h + \begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \\ \xi_{\nu,t+h} \end{bmatrix} h^{\frac{1}{2}} \tag{4.1}$$

$$\begin{bmatrix} \xi_{x,t+h} \\ \xi_{y,t+h} \\ \xi_{\nu,t+h} \end{bmatrix} \sim MVT_{2+e^{\nu_t}} \left[\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} e^{y_t} & e^{y_t/2}\rho_1\Lambda_1 & e^{y_t/2}\rho_2\Lambda_2 \\ e^{y_t/2}\rho_1\Lambda_1 & \Lambda_1^2 & \rho_3\Lambda_1\Lambda_2 \\ e^{y_t/2}\rho_2\Lambda_2 & \rho_3\Lambda_1\Lambda_2 & \Lambda_2^2 \end{bmatrix} \right] \tag{4.2}$$

To satisfy the regularity conditions and to ensure that the conditional variances are well defined, the degrees of freedom are bounded below by 2. To enforce positive semidefiniteness on the conditional covariance matrix, we also require $|\rho_j| \leq 1$ for $j=1,2,3$ and $1-\rho_1^2-\rho_2^2-\rho_3^2+2\rho_1\rho_2\rho_3 \geq 0$.[13]

Our interest is in estimating both y_t and ν_t. Without the aid of Theorem 2.2, it is by no means intuitively clear how to proceed: a GARCH estimate of σ_t^2, for example, has an immediate and intuitive interpretation as a smoothed empirical variance, but how does one

---

[12] There are a number of "multivariate t" distributions—here we mean the multivariate t with a common denominator and common degrees of freedom. See Johnson and Kotz (1972, Sections 37.3-37.4).

[13] Note that this is an ARCH model according to the definition of Bollerslev, Engle and Nelson (1993), though they did not focus on higher-order moments. Certainly, time varying higher-order moments are "heteroskedasticity" in the sense of McCulloch (1985). Hansen (1992) would call this an "autoregressive conditional density" model.

empirically smooth conditional kurtosis? A straightforward application of Theorem 2.2, however, yields

Theorem 4.1. The asymptotically optimal filter for (4.1)-(4.2) is:

$$G_{t+h} \equiv P_{t+h} + \omega_t S_{t+h}, \text{ where} \tag{4.3}$$

$$\omega_t = (E_t[SS'])^{-1/2}[(E_t[SS'])^{1/2}(E_t[(\xi_Y - P)(\xi_Y - P)'])(E_t[SS'])^{1/2}]^{1/2}(E_t[SS'])^{-1/2}, \tag{4.4}$$

$$P_{t+h} \equiv E_t\left[\begin{bmatrix}\xi_{y,t+h}\\ \xi_{v,t+h}\end{bmatrix} \mid \xi_{x,t+h}\right] = \begin{bmatrix}\rho_1\Lambda_1\\ \rho_2\Lambda_2\end{bmatrix}\xi_{x,t+h}e^{-\gamma_t/2} \tag{4.5}$$

$$S_{t+h} = \begin{bmatrix} \dfrac{(3+e^{\nu_t})\xi_{x,t+h}^2}{2(\xi_{x,t+h}^2+e^{\gamma_t+\nu_t})} - \dfrac{1}{2} \\ \dfrac{e^{\nu_t}}{2}[\psi(\dfrac{3+e^{\nu_t}}{2}) - \psi(1+\dfrac{e^{\nu_t}}{2})] - \dfrac{e^{\nu_t}}{2}\ln(1+\dfrac{\xi_{x,t+h}^2}{e^{\gamma_t+\nu_t}}) + \dfrac{(3+e^{\nu_t})\xi_{x,t+h}^2}{2(\xi_{x,t+h}^2+e^{\gamma_t+\nu_t})} - \dfrac{1}{2} \end{bmatrix} \tag{4.6}$$

$$E_t[(\xi_Y - P)(\xi_Y - P)'] = \begin{bmatrix} (1-\rho_1^2)\Lambda_1^2 & (\rho_3-\rho_1\rho_2)\Lambda_1\Lambda_2 \\ (\rho_3-\rho_1\rho_2)\Lambda_1\Lambda_2 & (1-\rho_2^2)\Lambda_2^2 \end{bmatrix} \tag{4.7}$$

$$E_t[SS'] = \begin{bmatrix} \dfrac{2+e^{\nu_t}}{2(5+e^{\nu_t})} & \dfrac{3}{e^{2\nu_t}+8e^{\nu_t}+15} \\ \dfrac{3}{e^{2\nu_t}+8e^{\nu_t}+15} & \dfrac{e^{2\nu_t}}{4}[\psi'(1+\dfrac{e^{\nu_t}}{2}) - \psi'(\dfrac{3+e^{\nu_t}}{2})] - \dfrac{(e^{\nu_t}-1)(e^{\nu_t}+6)}{2(e^{\nu_t}+3)(e^{\nu_t}+5)} \end{bmatrix} \tag{4.8}$$

$\psi(\cdot)$ is the Euler Psi (or Digamma[14]) function $\psi(x) \equiv d[\ln \Gamma(x)]/dx$ for $x > 0$.

Proof: see appendix.

It seems clear that (4.3)-(4.8) could not have been arrived at by ad hoc modelling

[14] For given values of ν, $\psi(\cdot)$ and its derivative $\psi'(\cdot)$ can be easily computed using, for example, the PolyGamma function in *Mathematica* (Wolfram Research, Inc. (1992)). It is also easy to write code for computing $\psi(\cdot)$ and $\psi'(\cdot)$ using the asymptotic approximations and recurrence relations given in Davis (1964, formulas 6.3.6, 6.3.18, 6.4.6, and 6.4.12).

strategies. Ignoring the existence of conditional heterokurticity can have important consequences. Suppose, for example, that we ignore ν_t, and estimate y_t using the filter corresponding to a conditionally normal distribution. When the conditional degrees of freedom $2 + \exp(\nu_t)$ exceed four, the filter will still achieve a $h^{1/4}$ rate of convergence for $(\hat{y}_t - y_t)$, but its efficiency can be very bad (see NF figure 2 and pp. 27-28), a conclusion reinforced by the monte carlo experiments reported in Schwartz (1994)). When the conditional degrees of freedom drop below four, the filter breaks down altogether.

It may seem surprising that higher-order conditional moments can be consistently estimated in a diffusion limit framework, since distributionally, diffusions are characterized by their first two conditional moments.[15] To see why higher-order conditional moments can be extracted, note that if the conditional degrees of freedom process has a diffusion limit, it has (in the limit) continuous sample paths almost surely and so is asymptotically constant (as is the conditional variance process) over a vanishingly small time interval $[T, T+M_h h^{1/2}]$. Yet over that vanishingly small interval we see a growing number $[M_h h^{-1/2}]$ of realizations of $h^{-1/2}(x_{t+h} - x_t)$, each of which is (asymptotically as $h \downarrow 0$) conditionally Student's t with a constant variance and constant degrees of freedom. We can estimate the variance and degrees of freedom in this case just as we can in the iid case. The asymptotically optimal filter carries this estimation out.

4.2 Multivariate GARCH Models

A number of multivariate extensions of GARCH have appeared in the literature. In all of these models we begin with an $n \times 1$ observable process $\{X_t\}_{t=0,h,2h,\ldots}$ and its associated

[15] More precisely, diffusions with unique weak-sense solutions are characterized by their first two conditional moments, sample path continuity, and the distribution of the initial starting point.

innovations process $\{\xi_{X,t}\}_{t=0,h,2h,\ldots}$. We assume that $E_t[\xi_{X,t+h}] = 0_{m\times 1}$ and $Cov_t[\xi_{X,t+h}] = \Omega_t$, where Ω_t is determined by a vector of unobservable state variables Y_t. For simplicity, we assume that $\{X_t, Y_t\}$ is conditionally multivariate normal. We will consider a number of GARCH models, first considering them as data generating process, and then as filters.

The first model we consider is due to Bollerslev, Engle, and Wooldridge (1988). Considered as a data generating process, this model sets

$$\Omega_{t+h} = \gamma_h + \beta_h \odot \Omega_t + \alpha_h \odot (\xi_{X,t+h}\xi_{X,t+h}') \tag{4.9}$$

where "$\odot$" is the Hadamard (i.e., element-by-element) product. If Ω_t, γ_h, β_h, α_h are all nonnegative definite, then as a consequence of the Schur product theorem (e.g. Horn and Johnson (1985, Theorem 7.5.3)), Ω_{t+h} is nonnegative definite. In this model, the ijth element of Ω_t is a function only of the ijth elements of lagged $\xi_{X,t}\xi_{X,t}'$. Now set $\gamma_h \equiv h^{\delta}\cdot\gamma$, $\alpha_h \equiv h^{1/2}\alpha$, and $\beta_h \equiv I - h^{1/2}\alpha - h^{\delta}\cdot\theta$, where $\delta = 3/4$ or 1, θ is symmetric (but not necessarily positive semidefinite) and α and γ are symmetric and nonnegative definite. Let "vech" be the operator that stacks the upper triangle of a matrix into a vector. Applying this operator to (4.9),[16]

$$\begin{aligned}\mathrm{vech}(\Omega_{t+h}) &= \mathrm{vech}(\Omega_t) + h\cdot\mathrm{vech}(\gamma) - h\cdot\mathrm{vech}(\theta)\odot\mathrm{vech}(\Omega_t)\\ &\quad + h^{1/2}\mathrm{vech}(\alpha)\odot\mathrm{vech}(\xi_{X,t+h}\xi_{X,t+h}'-\Omega_t)\end{aligned} \tag{4.9'}$$

For simplicity, let $E_t[X_{t+h}-X_t] = 0_{m\times 1}$ for all $h > 0$. If $\delta=1$, we obtain the diffusion limit:

$$\begin{aligned}dX_t &= \Omega_t^{1/2}dW_{X,t},\\ d\mathrm{vech}(\Omega_t) &= [\mathrm{vech}(\gamma)-\mathrm{vech}(\theta)\odot\mathrm{vech}(\Omega_t)]dt + \Lambda(\Omega_t)^{1/2}dW_{\Omega,t},\end{aligned} \tag{4.10}$$

where $W_{X,t}$ and $W_{\Omega,t}$ are independent $n\times 1$ and $m\times 1$ brownian motions, and $m = n(n+1)/2$.

---

[16] The constant correlations model of Bollerslev (1990) is a variant on this model, obtained by replacing the vech operator with the operator that stacks the diagonal elements of a square matrix into a vector. The conditional correlations are assumed constant.

$\Lambda(\Omega)$ is the covariance matrix of vech(α)$\odot$vech($ZZ'-\Omega$), where $Z \sim MVN(0,\Omega)$. Using subscripts to denote matrix or vector elements, we can evaluate the terms in $\Lambda(\Omega)$ using the relation $E[Z_i \cdot Z_j \cdot Z_k \cdot Z_m] = \Omega_{ij}\Omega_{km} + \Omega_{ik}\Omega_{jm} + \Omega_{im}\Omega_{jk}$ (Anderson (1985, Section 2.6, (26))). Note that the diffusion limit is on the standard (not the fast) time scale.

We now consider the model as a filter: let us reintroduce the possibility of 'fast' drift by replacing (4.10) by

$$\begin{aligned} dX_t &= \Omega_t^{1/2} dW_{X,t}, \\ d\mathrm{vech}(\Omega_t) &= [\mathrm{vech}(\gamma) - \mathrm{vech}(\theta) \odot \mathrm{vech}(\Omega_t)] h^{\delta-1} dt + \Lambda(\Omega_t)^{1/2} dW_{\Omega,t}. \end{aligned} \tag{4.10'}$$

We observe $X_0, X_h, X_{2h}, \ldots$, and wish to estimate $\{Y_t\}$, where $Y_t \equiv \mathrm{vech}(\Omega_t)$. Clearly if $\delta = 3/4$, the optimal filter sets $\hat{\mu} = \mu = 0_{m\times 1}$ and $\hat{\kappa} = \kappa = \mathrm{vech}(\gamma) - \mathrm{vech}(\theta) \odot \mathrm{vech}(\Omega_t)$. (If $\delta = 1$, it doesn't matter what $\hat{\mu}$ and $\hat{\kappa}$ are set to.) In the proof of 4.2 below, we show that (4.10') has a unique weak-sense solution for every $h > 0$. The other conditions of Theorem 3.1 are easily checked as well, so the filtering results will be the same as if the data were generated by the system

$$\begin{aligned} X_{t+h} &= X_t + h^{1/2} \xi_{X,t+h} \\ Y_{t+h} &= Y_t + [vech(\gamma) - vech(\theta) \odot Y_t] h^{\delta} + h^{1/2} \xi_{Y,t+h}, \text{ where} \\ \begin{bmatrix} \xi_{X,t+h} \\ \xi_{Y,t+h} \end{bmatrix} \mid X_t, Y_t &\sim N\left(\begin{bmatrix} 0_{n\times 1} \\ 0_{m\times 1} \end{bmatrix}, \begin{bmatrix} \Omega(Y_t) & 0_{n\times m} \\ 0_{m\times n} & \Lambda(Y_t) \end{bmatrix} \right) \end{aligned} \tag{4.10''}$$

Since $\xi_{X,t+h}$ and $\xi_{Y,t+h}$ are conditionally independent, $P_{t+h} = 0_{m\times 1}$. To compute the score S_{t+h}, we follow Magnus and Neudecker (1988, Section 15.4):

$$d\ln(f(\xi_{X,t+h} \mid X_t, Y_t)) = \frac{1}{2} vec(d\Omega(Y_t))' [\Omega(Y_t)^{-1} \otimes \Omega(Y_t)^{-1}] vec(\xi_{X,t+h}\xi_{X,t+h}' - \Omega(Y_t)) \tag{4.11}$$

Since the elements of the $m = n(n+1)/2$ dimensional vector Y include all the n^2 elements of Ω, we may write, $DY = Vec(\Omega)$, where D is the $n^2 \times m$ duplication matrix (see Magnus and

Neudecker (1988, Section 3.8)). This allows us to write

$$d\ln(f(\xi_{X,t+h} \mid X_t, Y_t)) = \frac{1}{2}(dY_t)'D'[\Omega(Y_t)^{-1}) \otimes \Omega(Y_t)^{-1}]D\,vech(\xi_{X,t+h}\xi'_{X,t+h} - Y_t) \tag{4.12}$$

$$\begin{aligned} &so\ S_{t+h} = \frac{1}{2}D'[\Omega(Y_t)^{-1}) \otimes \Omega(Y_t)^{-1}]D\,vech(\xi_{X,t+h}\xi'_{X,t+h} - Y_t) \\ &and\ G_{t+h} = \omega_t S_{t+h} \end{aligned} \tag{4.13}$$

Now, invoking the moment matching theorem 2.5, we recognize that the optimal filter is the GARCH Model (4.9'):

Theorem 4.2. Let the data be generated by (4.10') with X_t observed at intervals of length $h > 0$. X_0 and Y_0 are fixed. Let $Y_t \equiv vech(\Omega_t)$. Then the asymptotically optimal filter is

$$\hat{Y}_{t+h} = \hat{Y}_t + h^\delta \cdot vech(\gamma) - h^\delta \cdot vech(\theta) \odot \hat{Y}_t + h^{1/2} vech(\alpha) \odot [vech(\xi_{X,t+h}\xi_{X,t+h}') - \hat{Y}_t], \tag{4.14}$$

where $\xi_{X,t+h} \equiv h^{-1/2}[X_{t+h} - X_t]$.

Proof: see Appendix.

Next, consider the GARCH(1,1) version of the BEKK model of Engle and Kroner (1994), in which

$$\Omega_{t+h} = \gamma_h + \beta_h \odot \Omega_t + \sum_{k=1}^{K} A'_{h,k}\xi_{X,t+h}\xi'_{X,t+h}A_{h,k} + \sum_{k=1}^{K} B'_{h,k}\Omega_t B_{h,k} \tag{4.15}$$

where γ_h and β_h are nonnegative definite, and the $A_{h,k}$ and $B_{h,k}$ are arbitrary $n \times n$ matrices.[17] To obtain a diffusion limit (on the standard time scale) for this model, we need to put it in the form "state variables$_{t+h}$ = state variables$_t$ + h·drift + $h^{1/2}$ noise." The only terms on the right of (4.15) not known at time t are the $\xi_{X,t+h}\xi_{X,t+h}'$ terms, which we may convert to $O(h^{1/2})$ noise

[17] For notational convenience, we have added the $\beta_h \odot \Omega_t$ term to the BEKK model as presented in Engle and Kroner (1994). As Engle and Kroner show, this term can be effectively added to their equation (2.2) by a suitable choice of the $B_{h,k}$'s.

terms by subtracting Ω_t and setting $A_{h,k} \equiv h^{1/4} A_k$:

$$\Omega_{t+h} = \gamma_h + \beta_h \odot \Omega_t + \sum_{k=1}^{K} [B'_{h,k}\Omega_t B_{h,k} + h^{1/2} A'_k \Omega_t A_k] + h^{1/2} \sum_{k=1}^{K} A'_k (\xi_{X,t+h}\xi'_{X,t+h} - \Omega_t) A_k \qquad (4.16)$$

To obtain a diffusion limit, we need the terms known at time t to satisfy

$$\gamma_h + \beta_h \odot \Omega_t + \sum_{k=1}^{K} [B'_{h,k}\Omega_t B_{h,k} + h^{1/2} A'_k \Omega_t A_k] = \Omega_t + O(h). \qquad (4.16)$$

Unfortunately, there is in general no way to do this unless the A_k matrices are all diagonal, since the off-diagonal elements create drift terms of order $h^{1/2}$, not the required order h. "Fast" drift in the sense of section 2 above won't help, since this would allow drift of order $h^{3/4}$, but not of order $h^{1/2}$. If we are willing to assume diagonal A_k, then we may set $\gamma_h = h \cdot \gamma$, $\beta_h = I - \alpha \cdot h^{1/2} - \theta \cdot h$, $\alpha \equiv \Sigma_{k=1,K} A_k^2$, $B_{h,k} = h^{1/2} B_k$, θ symmetric, and γ positive semidefinite. Note that the term $\alpha \cdot h^{1/2}$ in the drift, induced by the diagonal elements of the A_k, does not cause any problem, since it is dominated by the "I" term in β_h, so β_h is positive definite for sufficiently small h. We now obtain

$$\Omega_{t+h} = \Omega_t + h \cdot \gamma - h \cdot \theta \odot \Omega_t + h \sum_{k=1}^{K} B'_k \Omega_t B_k + h^{1/2} \alpha \odot (\xi_{X,t+h}\xi'_{X,t+h} - \Omega_t) \qquad (4.17)$$

which, apart from the $B_k'\Omega B_k$ terms, is the same as the Bollerslev, Engle and Wooldridge model of Theorem 4.2, and has the same diffusion limit, with the addition of the $[\Sigma_{k=1,K} B_k'\Omega_t B_k]dt$ term. The $G(\cdot)$ term of the optimal filter is also the same, though the $\hat{\kappa}$ term of the optimal filter must now accommodate the new component of the drift. The analysis of Kroner and Ng's (1993) "General Dynamic Covariance Model" is similar.

Finally, consider the Factor ARCH model of Engle, Ng, and Rothschild (1990), in which

$$\Omega_t = \bar{\Omega} + \sum_{k=1}^{K} \beta_k \beta_k' \lambda_{k,t}, \tag{4.18}$$

where the $\lambda_{k,t}$ terms are the conditional variances of certain linear combinations of the X_t's. The $\lambda_{k,t}$ terms follow univariate ARCH models. It is clear that the existence of a diffusion limit for this process depends completely on the existence of diffusion limits for the $\lambda_{k,t}$ terms. If such diffusion limits exist, we have

$$\begin{aligned} dX_t &= \Omega_t^{1/2} dW_{X,t} \\ d\Omega_t &= \Sigma_{k=1,K}\, \beta_k' \beta_k\, d\lambda_{k,t}, \end{aligned} \tag{4.19}$$

where $\Omega_0 = \bar{\Omega} + \Sigma_{k=1,K}\, \beta_k' \beta_k\, \lambda_{k,0}$. It should be easy to see that if the data are generated by the diffusion (4.19) with X_t observable at intervals h, solving the optimal filtering problem is equivalent to solving the univariate optimal filtering problems for each of the $\{\lambda_{k,t}\}$ processes.

Finally, it is important to note that the form of the optimal filter would change if we did not assume conditionally normal errors. For example, a stochastic volatility model without conditionally normal errors could easily have the diffusion limit (4.10). However, the optimal filters for the sequence of stochastic volatility models would not, in general, correspond to the optimal filters for the discretely observed diffusion. For example, if $(\xi_{X,t+h}, \xi_{Y,t+h})$ are conditionally student's *t* with fixed finite degrees of freedom $d > 2$, the score term will not correspond to (4.13). For discussion in the univariate case, see NF, Sections 5-6.

5. Conclusion

While we have made progress in the theory of multivariate ARCH filtering, much remains to be done, particularly consideration of higher-order markov models and the effect of estimated parameters. In addition, the importance of matrix Riccati equations in our analysis

strongly suggests a link between our results and the theory of optimal control and linear filtering, in which such equations also figure prominently. The exact nature of this connection is not yet clear: unlike the terms in the linear or the extended Kalman filter (see, e.g., Anderson and Moore (1979)) the $G(\cdot)$ functions that form the basis of our filters are in general highly nonlinear in both the state variables X and Y and in the innovations ξ_X. In related work, Nelson (1994) develops asymptotic smoothing theory for ARCH models, and Nelson and Schwartz (1992) and Schwartz (1994) present monte carlo evidence on the effect of parameter estimation, and on filter performance for small but not infinitesimal h.

Perhaps most importantly, specifying the optimal filter requires specifying an underlying stochastic volatility model. Even in the context of linear models, parsimonious specification of multivariate systems is nontrivial. It is not likely to be any easier in the non-linear context.

Appendix

Proof of Theorem 2.1. The proof is substantially identical to the proof of Theorem 3.1 in NF.

Proof of the Corollary. The results are taken from Karatzas and Shreve (1988, Section 5.6). Their (6.12)-(6.13) give the differential equations for $b_{T,\tau}$ and $V_{T,\tau}$:

$$db_{T,\tau}/d\tau = -B_T b_{T,\tau} + A_T, \tag{A.1}$$

$$dV_{T,\tau}/d\tau = -B_T V_{T,\tau} - V_{T,\tau} B_T' + C_T \tag{A.2}$$

with initial conditions $b_{T,0} = Q$ and $V_{T,0} = 0_{m\times m}$. The unique solutions are given in (2.22)-(2.23). (2.22) is their (6.10) and (2.23) is their (6.14'). To go from (A.2) to (2.24), take the vec of both sides of (A.2) using the rule for evaluating the vec of a product of two matrices (see Magnus and Neudecker (1988, Chapter 2, Section 4, equation (7)). This yields

$$(d/dt)\mathrm{vec}(V_{T,\tau}) = -[I_{m\times m}\otimes B_T + B_T\otimes I_{m\times m}]^{-1}\mathrm{vec}(V_{T,\tau}) + \mathrm{vec}(C_T), \tag{A.3}$$

which is a standard linear vector o.d.e. with solution (2.24).

Proof of Theorem 2.2. Since there is no variance-bias tradeoff we may consider $b_T^* b_T^{*\prime}$ and V^* separately. Clearly (2.34)-(2.35) are sufficient to eliminate $b_T^* b_T^{*\prime}$, so we turn our attention to V^*. Our strategy is to guess a solution to the minimization problem and then to verify it. To arrive at the guess in (2.36), we differentiate $\mathrm{Trace}[b_T b_T' + V_T^*]$ with respect to $G(\xi_X,\cdot)$, drop the $d\xi_X$, $d\xi_y$ terms, (naively) treating $G(\xi_X,\cdot)$ as a separate choice variable for each ξ_X (so there are an uncountable number of choice variables). This yields (2.36). To verify global optimality of this $G(\cdot)$, we will need the following lemma, (basically a consequence of the law of iterated expectations).

Lemma A.1. Let $\theta(\cdot)$ be an $k\times 1$ integrable vector function of ξ_X, X, Y, and t. Then $E_t[\theta'(\xi_X,X,Y,t)\xi_Y'] = E_t[\theta'(\xi_X,X,Y,T,\tau)P']$.

Proof of Lemma A.1.

$$E_t[\theta\xi_Y] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta'\ \xi_Y\, f(\xi_X,\xi_Y|X,Y,T,\tau)d\xi_Y d\xi_X \tag{A.4}$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta'\ \left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\xi_Y f(\xi_Y|\xi_X,X,Y,t)d\xi_Y\right] f(\xi_X|X,Y,t)d\xi_X \tag{A.5}$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta'\ P\, f(\xi_X|X,Y,t)\,d\xi_X = E_t[\theta' P]. \tag{A.6}$$

Now we continue with the proof of Theorem 2.2. We will drop time subscripts when it should cause no confusion. It will be convenient to have a simpler expression for B than (2.9). Recall that the ij^{th} element of B is $-E_t[\partial G_i/\partial Y_j]$. Integrating by parts and using the fact that $E_t[G] = 0_{m\times 1}$ yields $E_t[\partial G_i/\partial Y_j] = -E_t[G_i S_j]$, so

$$B = E_t[GS']. \tag{A.7}$$

Now consider another choice of G, say $\tilde{G} \equiv P + \omega S + H$, where H is a function of ξ_X, X, Y,

T, and t satisfying the conditions Theorem 2.1 put on admissible G functions. If the real parts of the eigenvalues of $\tilde{B} \equiv E_t[\tilde{G}S'] = E_t[PS'] + \omega E_t[SS'] + E_t[HS']$ are all positive (as required by Theorem 2.1) there will be a bounded $\tilde{V}$ that is the asymptotic covariance matrix of the measurement error process (see Lancaster and Tismenetsky (1985, Chapter 12.3, Theorem 3)). If $\tilde{V}$ is unbounded its trace is clearly larger than that of ω, so we assume that the real parts of the eigenvalues of $\tilde{B}$ are strictly positive. By (2.25),

$$\tilde{B}\tilde{V} + \tilde{V}\tilde{B}' - \tilde{C} = 0, \tag{A.8}$$

where by (2.10), $\tilde{C} = E_t[(P + \omega_T S + H - \xi_Y)(P + \omega_T S + H - \xi_Y)']$. Similarly, ω satisfies

$$B\omega + \omega B' - C = 0, \tag{A.9}$$

where $B \equiv E_t[PS'] + \omega_T E_t[SS']$ and $C \equiv E_t[(P + \omega S - \xi_Y)(P + \omega S - \xi_Y)']$. Subtracting (A.9) from (A.8), substituting for B, C, $\tilde{B}$, and $\tilde{C}$, and simplifying terms (employing Lemma A.1) leads to

$$\tilde{B}(\tilde{V}-\omega) + (\tilde{V}-\omega)\tilde{B}' = E_t[HH'], \tag{A.10}$$

which is an equation of the same form as (2.25), and has (see Lancaster and Tismenetsky (1985, Chapter 12.3, Theorem 3)) a solution of the same form as (2.26):

$$(\tilde{V}-\omega) = \int_0^\infty \exp[-\tilde{B}\cdot s]E_t[HH']\exp[-\tilde{B}'\cdot s]ds. \tag{A.11}$$

The right-hand side of (A.11) is clearly positive semidefinite, since a sum or integral of positive semidefinite matrices is positive semidefinite. $\tilde{V}$ therefore exceeds ω by a positive semidefinite matrix. Because of the non-singularity of $\exp[-\tilde{B}\cdot s]$, $\tilde{V}=\omega$ if and only if $E_t[HH']$ is a matrix of zeros. Finally, the matrix Riccati equation follows by substituting for B and C in (A.9).

Proof of Theorem 2.3 (a). (2.33) becomes

$$E_T[(\xi_Y-P)(\xi_Y-P)'] = \omega_T E_T[SS']\omega_T. \tag{2.32'}$$

Pre and post multiplying by $E_T[SS']^{1/2}$, we obtain

$$E_T[SS']^{1/2}E_T[(\xi_Y-P)(\xi_Y-P)']E_T[SS']^{1/2} = E_T[SS']^{1/2}\omega_T E_T[SS']^{1/2}E_T[SS']^{1/2}\omega_T E_T[SS']^{1/2}. \quad (A.12)$$

Taking symmetric matrix square roots,

$$[E_T[SS']^{1/2}E_T[(\xi_Y-P)(\xi_Y-P)']E_T[SS']^{1/2}]^{1/2} = E_T[SS']^{1/2}\omega_T E_T[SS']^{1/2}. \quad (A.13)$$

Pre and post multiplying by $E_T[SS']^{-1/2}$, yields (2.37).

(b). That $P_{t+h} = E_T[\xi_Y\xi_X'](E_T[\xi_X\xi_X'])^{-1}\xi_{X,t+h}$ is Cambanis, Huang, and Simons (1981, Corollary 5). $E_T[PS'] = 0_{m\times m}$ follows from Mitchell (1989, (2.4) and (2.7)).

To prove Theorem 2.4 we need the following lemma:

Lemma A.2. Let $E_T[SS']$ and $E_T[(\xi_Y-P)(\xi_Y-P)']$ be positive definite. Then the matrix $-(E_T[PS'] + \omega_T E_T[SS'])$ is stable—i.e., the real parts of its eigenvalues are strictly negative.

Proof of Lemma A.2. By Kučera (1972, Theorem 3), a sufficient condition for $-(E_T[PS'] + \omega_T E_T[SS'])$ to be stable is that there exist real matrices M_1 and M_2 such that

$$-E_T[SP'] + E_T[SS']\cdot M_1 \text{ and} \quad (A.14)$$

$$-E_T[PS'] + E_T[(\xi_Y-P)(\xi_Y-P)']\cdot M_2 \quad (A.15)$$

are stable. For $i=1,2$, we choose $M_i = -I_{m\times m}/k_i$, where k_i is a small positive number. (A.14)-(A.15) are now equivalent to the stability of

$$-E_T[SP']\cdot k_1 - E_T[SS'] \text{ and} \quad (A.16)$$

$$-E_T[PS']\cdot k_2 - E_T[(\xi_Y-P)(\xi_Y-P)']. \quad (A.17)$$

$-E_T[(\xi_Y-P)(\xi_Y-P)']$ and $-E_T[SS']$ are stable by assumption. Now consider (A.16)-(A.17) as functions of k_i. Clearly, these functions are analytic, implying that their eigenvalues are continuous functions of k_i (see Lancaster and Tismenetsky (1985, Chapter 11.7 Theorem 1 and Exercise 6)). This, in turn, implies the stability of (A.16) and (A.17) for sufficiently small k_i.

Proof of Theorem 2.4. To prove (2.39), define $\omega_T(\zeta)$ as in (2.33'). Differentiating, we obtain at $\zeta = 0$,

$$(E_T[PS'] + \omega_T(0)E_T[SS'])d\omega_T(\zeta)/d\zeta + (d\omega_T(\zeta)/d\zeta)(E_T[SP'] + E_T[SS']\omega_T(0)) = \Delta. \quad (A.18)$$

(2.39) now follows from Lemma A.2 and from Lancaster and Tismenetsky (1985, Chapter 12.3, Theorem 3). $d\omega_T(\zeta)/d\zeta$ is non-negative definite and non-null, since Δ is, and since matrix exponentials are non-singular. The proofs of (2.40) and (2.41) are essentially identical.

Proof of Theorem 2.5. We have

$$E_t[GG'] = E_t[(P + \omega S)(P + \omega S)'] = E_t[\, PP' + \omega SS'\omega + \omega SP' + PS'\omega]. \quad (A.19)$$

Subtracting (2.33) from (A.19) and simplifying leads to

$$E_t[GG'] = E_t[PP'] + E_t[(\xi_Y - P)(\xi_Y - P)'] = E_t[\xi_Y \xi_Y'], \quad (A.20)$$

proving (2.42). To prove (2.43), note that

$$E_t[G\xi_X'] = E_t[P\xi_X'] + \omega_T E_t[S\xi_X'], \text{ but} \quad (A.21)$$

$$E_t[S\xi_X'] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial \ln(f(\xi_X|X,Y))}{\partial Y} \xi_X' f(\xi_X|X,Y)\, d\xi_X \quad (32)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial f(\xi_X|X,Y)}{\partial Y} \xi_X'\, d\xi_X = \frac{\partial}{\partial Y} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\xi_X|X,Y)\, \xi_X'\, d\xi_X = 0. \quad (A.20)$$

The interchange of limits in (A.22) is allowed by uniform convergence. So $E_t[G\xi_X'] = E_t[P\xi_X']$ $= E_t[\xi_Y \xi_X']$ by Lemma A.1, proving (2.43).

Proof of Theorem 2.6. We adopt the same notation as in the proof of Theorem 2.2. In the case we now consider, H takes the form $[0_{1\times m_1}, \eta_{1\times(m-m_1)}]'$. (A.11) becomes

$$(\tilde{V}_T - \omega_T) = \int_0^{\infty} \begin{bmatrix} \tilde{B}_1(s) & \tilde{B}_2(s) \\ \tilde{B}_3(s) & \tilde{B}_4(s) \end{bmatrix} \begin{bmatrix} 0_{m_1\times m_1} & 0_{m_1\times(m-m_1)} \\ 0_{m-m_1\times m_1} & E_T[\eta\eta'] \end{bmatrix} \begin{bmatrix} \tilde{B}_1'(s) & \tilde{B}_3'(s) \\ \tilde{B}_2'(s) & \tilde{B}_4'(s) \end{bmatrix} ds \quad (A.23)$$

$\tilde{B}_1$ is $m_1\times m_1$, $\tilde{B}_2$ is $m_1\times(m-m_1)$, $\tilde{B}_3$ is $(m-m_1)\times m_1$, and $\tilde{B}_4$ is $(m-m_1)\times(m-m_1)$. Carrying

$$\text{where} \quad \begin{bmatrix} \tilde{B}_1(s) & \tilde{B}_2(s) \\ \tilde{B}_3(s) & \tilde{B}_4(s) \end{bmatrix} \equiv \exp[-\tilde{B}\cdot s]. \tag{A.24}$$

out the matrix multiplication in (A.23), we obtain

$$(\tilde{V}_{T,1}-\omega_{T,1}) = \int_0^\infty \tilde{B}_2(s)\, E_T[\eta\eta']\, \tilde{B}_2(s)'ds \tag{A.25}$$

which is a matrix of zeros if for all $s>0$, $\tilde{B}_2(s)$ is a matrix of zeros. But

$$\exp[-\tilde{B}\cdot s] \equiv \Sigma_{j=0,\infty}\, [(-s)^j/j!]\, \tilde{B}^j, \tag{A.26}$$

and it is easy to check that $\tilde{B}^j$ is lower block-triangular for all j whenever $\tilde{B}$ is. Therefore if $\tilde{B}$ is lower block-triangular, $\tilde{B}_2(s)$ is a matrix of zeros and $\tilde{V}_{T,1}=\omega_{T,1}$. Since $\tilde{B} \equiv E_T[\tilde{G}S'] = E_T[PS'] + \omega_T E_T[SS'] + E_T[HS']$, and $E_T[HS']$ is block lower triangular, $\tilde{B}$ is block lower-triangular if and only if B is. When $E_T[PS']$ is a matrix of zeros and ω_T and $E_T[SS']$ are block-diagonal, so is B. All that remains, therefore, is to show that ω_T is block-diagonal whenever $E_T[SS']$, $E_T[PS']$, and $E_T[(\xi_Y-P)(\xi_Y-P)']$ are. Substituting into (2.33), we obtain $[\alpha_1 + \omega_{T,1}U_1]\omega_{T,2} + \omega_{T,2}[\alpha_2' + U_2\omega_{T,3}] = 0_{m_1\times(m-m_1)}$, where

$$E_T[SS'] = \begin{bmatrix} U_1 & 0 \\ 0' & U_2 \end{bmatrix}, \quad E_T[PS'] = \begin{bmatrix} \alpha_1 & 0 \\ 0' & \alpha_2 \end{bmatrix}$$

with "0" an $m_1\times(m-m_1)$ matrix of zeros. Since $\omega_{T,1}$, U_1 , $\omega_{T,3}$, and U_2 are positive definite, the eigenvalues of $\omega_{T,1}U_1$ and $U_2\omega_{T,3}$ are real and positive (Taussky (1968, p. 177)). Applying the same reasoning as in the proof of Lemma A.2, we see that the real parts of the eigenvalues of $[\alpha_2' + U_2\omega_{T,3}]$ and $[\alpha_1 + \omega_{T,1}U_1]$ are positive. This in turn implies that $\omega_{T,2} = 0_{m_1\times(m-m_1)}$ (Lancaster and Tismenetsky (1985, Chapter 12.3, Theorem 3)).

Proof of Theorem 2.7. To simplify the presentation of the proof, we take $\mu = \hat{\mu} = 0_{n\times 1}$ and $\kappa = \hat{\kappa} = 0_{m\times 1}$. Since the asymptotically optimal filter eliminates asymptotic bias, these terms

are not of direct interest. We have

$$\Upsilon_{t+h} - \Upsilon_t = \Upsilon(X_{t+h}, Y_{t+h}, t+h) - \Upsilon(X_t, Y_t, t), \text{ and} \tag{A.27}$$

$$\chi_{t+h} - \chi_t = \chi(X_{t+h}, t+h) - \chi(X_t, t) \tag{A.28}$$

Expanding (A.27)-(A.28) in Taylor series around X_t, Y_t, and t, we have

$$\Upsilon_{t+h} = \Upsilon_t + \frac{\partial \Upsilon}{\partial t}\cdot h + \frac{\partial \Upsilon}{\partial X}(X_{t+h}-X_t) + \frac{\partial \Upsilon}{\partial Y}(Y_{t+h}-Y_t) + O(\|(X_{t+h}-X_t, Y_{t+h}-Y_t, h)\|^2) \tag{A.29}$$

$$= \Upsilon_t + \frac{\partial \Upsilon}{\partial t}\cdot h + \left[\frac{\partial \Upsilon}{\partial X}\xi_{X,t+h} + \frac{\partial \Upsilon}{\partial Y}\xi_{Y,t+h}\right] h^{1/2} + O(\|(X_{t+h}-X_t, Y_{t+h}-Y_t, h)\|^2)$$

$$\chi_{t+h} = \chi_t + \frac{\partial \chi}{\partial t}\cdot h + \frac{\partial \chi}{\partial X}(X_{t+h}-X_t) + O(\|(X_{t+h}-X_t, h)\|^2) \tag{A.30}$$

$$= \chi_t + \frac{\partial \chi}{\partial t}\cdot h + \left[\frac{\partial \chi}{\partial X}\xi_{X,t+h}\right] h^{1/2} + O(\|(X_{t+h}-X_t, h)\|^2)$$

where the partial derivatives are evaluated at (X_t, Y_t, t). Note that by Assumption 2 and (2.48)-(2.49) and the corollary to Billingsley (1986, Theorem 25.12) the moments of higher-order terms (normalized by $h^{-1/2}$) vanish to zero as $h \downarrow 0$, as do the $\partial\Upsilon/\partial t$ and $\partial\chi/\partial t$ terms.

We can now write $\xi_{\Upsilon,t+h}$ and $\xi_{\chi,t+h}$ as:

$$\xi_{\Upsilon,t+h} = \left[\frac{\partial \Upsilon}{\partial X}\xi_{X,t+h} + \frac{\partial \Upsilon}{\partial Y}\xi_{Y,t+h}\right] + h.o.t., \quad \xi_{\chi,t+h} = \frac{\partial \chi}{\partial X}\xi_{X,t+h} + h.o.t. \tag{A.31}$$

where the higher order terms along with their relevant moments, disappear as $h \downarrow 0$. So

$$P_{\Upsilon,t+h} \equiv E[\xi_{\Upsilon,t+h} \mid \xi_{\chi,t+h}, \chi_t, \Upsilon_t] = E[\xi_{\Upsilon,t+h} \mid \xi_{X,t+h}, X_t, Y_t]$$

$$as\ h \downarrow 0, \quad P_{\Upsilon,t+h} \to E[\frac{\partial \Upsilon}{\partial X}\xi_{X,t+h} + \frac{\partial \Upsilon}{\partial Y}\xi_{Y,t+h} \mid \xi_{X,t+h}, X_t, Y_t] = \frac{\partial \Upsilon}{\partial X}\xi_{X,t+h} + \frac{\partial \Upsilon}{\partial Y}P_{t+h} \tag{A.32}$$

where as in Theorem 2.2, $P_{t+h} \equiv E[\xi_{Y,t+h} \mid \xi_{X,t+h}, X_t, Y_t]$. We have as $h \downarrow 0$

$$E_t[(\xi_\Upsilon - P_\Upsilon)(\xi_\Upsilon - P_\Upsilon)'] \to \left[\frac{\partial \Upsilon}{\partial Y}\right] E_t[(\xi_Y - P)(\xi_Y - P)'] \left[\frac{\partial \Upsilon}{\partial Y'}\right]. \tag{A.33}$$

Again, the existence of $2+\delta$ absolute conditional moments guarantees the convergence of the relevant conditional moments. Next, note that since the conditions of the implicit function theorem are satisfied, we have for fixed t,

$$\begin{bmatrix} d\Upsilon \\ d\chi \end{bmatrix} = \begin{bmatrix} \frac{\partial \Upsilon}{\partial Y} & \frac{\partial \Upsilon}{\partial X} \\ 0_{n\times m} & \frac{\partial \chi}{\partial X} \end{bmatrix} \begin{bmatrix} dY \\ dX \end{bmatrix}, \text{ and } \begin{bmatrix} dY \\ dX \end{bmatrix} = \begin{bmatrix} (\frac{\partial \Upsilon}{\partial Y})^{-1} & -(\frac{\partial \Upsilon}{\partial Y})^{-1}\frac{\partial \Upsilon}{\partial X}(\frac{\partial \chi}{\partial X})^{-1} \\ 0_{n\times m} & (\frac{\partial \chi}{\partial X})^{-1} \end{bmatrix} \begin{bmatrix} d\Upsilon \\ d\chi \end{bmatrix} \tag{A.34}$$

So given t, we may treat (X_t, Y_t) as implicit functions of (χ_t, Υ_t). Similarly, given (X_t, Y_t, t) or (χ_t, Υ_t, t) we may treat $\xi_{X,t+h}$ as an implicit function of $\xi_{\chi,t+h}$. Call the conditional density of $\xi_{\chi,t+h}$ given χ_t and Υ_t $\varphi(\xi_{\chi,t+h}|\chi_t,\Upsilon_t)$. Applying the change of variables formula and the chain rule,

$$\varphi(\xi_{\chi,t+h}|\chi_t,\Upsilon_t) = f(\xi_{X,t+h}|X_t,Y_t)\,\left|\frac{\partial \chi}{\partial X}\right|^{-1}, \textit{ so the score is} \tag{A.35}$$

$$\frac{\partial \ln[\varphi(\xi_{\chi,t+h}|\chi_t,\Upsilon_t)]}{\partial \Upsilon} = \frac{\partial Y}{\partial \Upsilon'}\frac{\partial \ln[f(\xi_{X,t+h}|X_t,Y_t)]}{\partial Y} = \frac{\partial Y}{\partial \Upsilon'}S_{t+h}, \tag{A.36}$$

where S_{t+h} is as in Theorem 2.2. So the Fisher information matrix for $\{\chi_t, \Upsilon_t\}$ is

$$\frac{\partial Y}{\partial \Upsilon'}E_t[SS']\frac{\partial Y}{\partial \Upsilon} \tag{A.37}$$

where $E_t[SS']$ is the Fisher information matrix for the $\{X_t, Y_t\}$ system. We also have

$$E_t[P_\Upsilon S_\Upsilon'] \to \frac{\partial \Upsilon}{\partial X}E_t[\xi_X S']\frac{\partial Y}{\partial \Upsilon} + \frac{\partial \Upsilon}{\partial Y}E_t[PS']\frac{\partial Y}{\partial \Upsilon} = \frac{\partial \Upsilon}{\partial Y}E_t[PS']\frac{\partial Y}{\partial \Upsilon} \tag{A.38}$$

where the last equality holds since $E_t[\xi_X S'] = 0_{m\times m}$—see the proof of Theorem 2.5. Let $\bar{\omega}$ be the solution to the matrix Riccati equation for the (χ,Υ) system. From (A.33), (A.37) and

(A.38),

$$\frac{\partial \Upsilon}{\partial Y} E_t[PS'] \frac{\partial Y}{\partial \Upsilon} \bar{\omega} + \bar{\omega} \frac{\partial Y}{\partial \Upsilon'} E_t[SP'] \frac{\partial \Upsilon}{\partial Y'} + \bar{\omega} \frac{\partial Y}{\partial \Upsilon'} E_t[SS'] \frac{\partial Y}{\partial \Upsilon} \bar{\omega}$$

$$= \frac{\partial \Upsilon}{\partial Y} E_t[(\xi_Y - P)(\xi_Y - P)'] \frac{\partial \Upsilon}{\partial Y'} \tag{A.39}$$

The theorem is proved if we can show that $\bar{\omega} = [\partial \Upsilon / \partial Y] \omega [\partial \Upsilon / \partial Y']$ where ω is as in Theorem 2.2. By (A.34) $[\partial Y / \partial \Upsilon]^{-1} = \partial \Upsilon / \partial Y$. Substituting for $\bar{\omega}$ in (A.39) and simplifying leads to

$$\frac{\partial \Upsilon}{\partial Y} E_t[PS'] \omega \frac{\partial \Upsilon}{\partial Y} + \frac{\partial \Upsilon}{\partial Y} \omega E_t[SP'] \frac{\partial \Upsilon}{\partial Y'} + \frac{\partial \Upsilon}{\partial Y} \omega E_t[SS'] \omega \frac{\partial \Upsilon}{\partial Y'}$$
$$= \frac{\partial \Upsilon}{\partial Y} E_t[(\xi_Y - P)(\xi_Y - P)'] \frac{\partial \Upsilon}{\partial Y'} \tag{A.40}$$

Left multiplying by $[\partial \Upsilon / \partial Y]^{-1} = \partial Y / \partial \Upsilon$ and right multiplying by $[\partial \Upsilon / \partial Y']^{-1} = \partial Y / \partial \Upsilon'$ recovers the Riccati equation of Theorem 2.2. By assumption 4 there is a unique nonnegative definite solution to this equation, so there is for the transformed system as well.

To prove Theorem 3.1, we first need two lemmas:

Lemma A.3 (Ethier and Kurtz (1986, Chapter 7, Corollary to Theorem 4.1))). Let there be a unique weak-sense solution to the $m \times 1$ stochastic integral equation

$$Z_t = Z_0 + \int_0^t \mu(Z_s) ds + \int_0^t \Omega(Z_s)^{1/2} dW_s , \tag{A.41}$$

where $\{W_s\}$ is an $n \times 1$ standard Brownian motion independent of Z_0, $\mu(x) \in R^n$ and $\varphi(x) \in R^{n \times n}$ are continuous, and Z_0 is random with distribution function F. Now consider for each $h > 0$ a discrete time $n \times 1$ Markov process $\{{}_hY_k\}_{k=0,\infty}$ and define for each $\Delta > 0$, each $h > 0$, and each integer $k \geq 0$,

$$\mu_{\Delta,h}(y) \equiv h^{-\Delta} E[({}_hY_{k+1} - {}_hY_k) \cdot I(\| {}_hY_{k+1} - {}_hY_k \| < 1) \mid {}_hY_k = y], \text{ and} \tag{A.42}$$

$$\Omega_{\Delta,h}(y) \equiv h^{-\Delta} Cov[({}_hY_{k+1} - {}_hY_k) \cdot I(\| {}_hY_{k+1} - {}_hY_k \| < 1) \mid {}_hY_k = y], \tag{A.43}$$

where $I(\cdot)$ is the indicator function. The initial value ${}_hY_0$ is random with cumulative distribution function F_h. Let

(a') $F_h(y) \Rightarrow F(y)$ *as* $h \downarrow 0$, *and for some* $\Delta > 0$, *let*

(b') $\mu_{\Delta,h}(y) \to \mu(y)$,

(c') $\Omega_{\Delta,h}(y) \to \Omega(y)$, *and for every* ϵ, $0 < \epsilon < 1$,

(d') $h^{-\Delta}P[\| {}_hY_{k+1} - {}_hY_k \| > \epsilon \mid {}_hY_k = y] \to 0$

as $h \downarrow 0$, *uniformly on every bounded y set. For each* $h > 0$, *define the process* $\{{}_hZ_t\}$ *by* ${}_hX_t \equiv {}_hY_{[t \cdot h^{-\Delta}]}$ *for each* $t \geq 0$, *where "*$[t \cdot h^{-\Delta}]$*" is the integer part of* $t \cdot h^{-\Delta}$. *Then for any T,* $0 < T < \infty$, $\{{}_hZ_t\}_{[0,T]} \Rightarrow \{Z_t\}_{[0,T]}$ *as* $h \downarrow 0$, *where (A.41) defines* $\{Z_t\}$.

We stated lemma A.3 for the case in which μ and Ω did not depend on t. The lemma remains true in the time inhomogeneous case, however: simply make t an element of Z_t.

We also need the following result, which adapts arguments in Friedman (1975, Chapter 5) and Arnold (1973, section 7.1) to the fast drift case. For notational simplicity, we state and prove it for the time-homogeneous case, but it too is true for the inhomogeneous case (again, make t an element of λ_t—but don't multiply its drift by $h^{-1/4}$ in the δ=3/4 case!):

Lemma A.4. For every h, $0 < h < 1$, *let there be a unique weak-sense solution to the* $n \times 1$ *equation*

$$\lambda_t = \lambda_0 + h^{-1/4} \int_0^t \nu(\lambda_s)ds + \int_0^t \varphi(\lambda_s)^{1/2}dW_s , \qquad (A.44)$$

where $\nu(\cdot)$ *and* $\varphi(\cdot)$ *are continuous,* $\{W_t\}$ *is an* $n \times 1$ *standard Brownian motion independent of* λ_0. *For each* $N > 0$, *define the stochastic integral equation*

$$\lambda_{N,t} = \lambda_{N,0} + h^{-1/4} \int_0^t \nu_N(\lambda_{N,s})ds + \int_0^t \varphi_N(\lambda_{N,s})^{1/2}dW_s , \qquad (A.45)$$

where $\lambda_{N,0} \equiv \lambda_0$ *if* $\| \lambda_0 \| \leq N$ *and* $\lambda_{N,0} \equiv 0_{n \times 1}$ *otherwise,* (A.46)

$\nu_N(\lambda) \equiv \nu(\lambda)$ *if* $\|\lambda\| \le N$, $\nu_N(\lambda) \equiv \nu(\lambda)\cdot[2-\|\lambda\|/N]$ *if* $N < \|\lambda\| \le 2N$, $\nu_N \equiv 0_{n\times 1}$ *otherwise.* (A.47)

$\varphi_N(\lambda) \equiv \varphi(\lambda)$ *if* $\|\lambda\| \le N$, $\varphi_N(\lambda) \equiv \varphi(\lambda)\cdot[2-\|\lambda\|/N]$ *if* $N < \|\lambda\| \le 2N$, $\varphi_N \equiv 0_{n\times n}$ *otherwise.* (A.48)

Then for every integer $j \ge 1$, *there is a continuous function* $M_j(\lambda,N)$ *and an* $h^* > 0$ *such that for every* h, $0 < h < h^*$, $E(\|\lambda_{N,t+h}-\lambda_{N,t}\|^{2j}|\lambda_{N,t}) \le M_j(\lambda_{N,t},N)h^j$ *almost surely.*

Proof of Lemma A.4. Note first that for every N > 0 and every h > 0, the drift and diffusion coefficients of (A.45) are bounded and continuous, so there will be a K_N such that

$$\|\nu_N(\lambda)\|^2 + \|\varphi_N(\lambda)\|^2 \le K_N^2(1 + \|\lambda\|^2) \tag{A.49}$$

Assuming h < 1, (A.49) implies

$$\|h^{-1/4}\nu_N(\lambda)\|^2 + \|\varphi_N(\lambda)\|^2 \le h^{-1/2}K_N^2(1 + \|\lambda\|^2) \tag{A.50'}$$

This is the familiar 'growth condition (e.g., Arnold (1973, Theorem 6.2.2)) which is satisfied for the $\{\lambda_{N,t}\}$ process even though it may not be satisfied for the $\{\lambda_t\}$ process. We do not need to assume the usual Lipschitz condition, since weak (as opposed to strong) existence and uniqueness of solutions to (A.44)-(A.45) is enough for our purposes. Keeping $h^{-1/4}$ outside the integral and carrying out the steps in the proof of Arnold (1973, 7.1.3) leads to

$$E_s[\|\lambda_{N,t}\|^{2j}] \le (1 + \|\lambda_{N,s}\|^{2j})\exp[2j(2j+1)K_N^2 h^{-1/4}(t-s)], \quad 0 \le s < t. \text{ Now} \tag{A.50}$$

$$\begin{aligned} E_t\|\lambda_{N,t-h}-\lambda_{N,t}\|^{2j} &= E_t\|h^{-1/4}\int_t^{t+h}\nu(\lambda_{N,s})ds + \int_t^{t+h}\varphi(\lambda_{N,s})dW_s\|^{2j} \\ &\le E_t[h^{-1/4}\int_t^{t+h}\|\nu_N(\lambda_{N,s})\|ds + \|\int_t^{t+h}\varphi(\lambda_{N,s})dW_s\|]^{2j} \end{aligned} \tag{A.51}$$

By Jensen's inequality and the convexity of x^{2j} for positive integer j,

$$\le 2^{2j-1}h^{-j/2}E_t[\int_t^{t+h}\|\nu_N(\lambda_{N,s})\|ds]^{2j} + 2^{2j-1}E_t\|\int_t^{t+h}\varphi_N(\lambda_{N,s})dW_s\|^{2j} \tag{A.52}$$

By the integral means inequality (Hardy, Littlewood and Pólya (1952, Theorem 192))

$$2^{2j-1}h^{-j/2}E_t(\int_t^{t+h}\|\nu_N(\lambda_{N,s})\|ds)^{2j} \le 2^{2j-1}h^{3j/2-1}\int_t^{t+h}\|\nu_N(\lambda_{N,s})\|^{2j}ds \tag{A.53}$$

By Jensen's inequality and (A.49)

$$\leq 2^{3j-2} h^{3j/2-1} K_N^{2j} \int_t^{t+h} E_t(1+|(\lambda_{N,s})|^{2j})\,ds \tag{A.54}$$

Substituting from (A.50) into the integrand and integrating yields, for sufficiently small h,

$$2^{2j-1} h^{-j/2} E_t(\int_t^{t+h} \| v_N(\lambda_{N,s}) \|\, ds)^{2j} \leq h^{3j/2}(1+|\lambda_{N,t}|^{2j})\, 2^{3j} K_N^{2j} \tag{A.55}$$

bounding the first term on the right-hand side of (A.52). To bound the second term, we first apply Friedman (1975, Corollary 4.6.4):

$$2^{2j-1} E_t|\int_t^{t+h} \varphi_N(\lambda_{N,s})\, dW_s|^{2j} \leq 2^{2j-1} h^{j-1} [\frac{4j^3}{2j-1}]^j \int_t^{t+h} E_t|\varphi_N(\lambda_{N,s})|^{2j}\, ds \tag{A.56}$$

Just as we did with the first term in on the right side of (A.52), we may bound $E_t|\varphi_{N,s}|^{2j}$ using (A.49), substitute into the integrand and integrate, obtaining

$$2^{2j-1} E_t|\int_t^{t+h} \varphi_{N,s}\, dW_s|^{2j} \leq 2^{2j-1} h^j K_N^{2j} [\frac{4j^3}{2j-1}]^j (3 + 2|\lambda_{N,t}|^{2j}) \tag{A.57}$$

completing the proof of the lemma.

Proof of Theorem 3.1. We will prove the theorem for $\delta=3/4$ (the $\delta=1$ case is similar but simpler.) As indicated in the text, we will use Lemma A.3 in place of Theorem 2.1 of NF. Note that the conditional moments are truncated by the $I(\| {}_hY_{k+1} - {}_hY_k \| < 1)$ term. This implies that only the local properties of the $\{{}_hY_k\}$ process enter. Suppose, for example, that ${}_hY_k \equiv \lambda_{kh}$, where λ_t is the diffusion in Lemma A.4. Then provided that $N > 1 + \|\lambda_{kh}\|$, all the conditional moments in (b')-(d') are exactly the same as if we had defined ${}_hY_k \equiv \lambda_{N,kh}$, where $\{\lambda_{N,kh}\}$ is the diffusion defined by (A.45)-(A.47). In fact, if $\|\lambda_0\| < N$, the transition probabilities for $\{\lambda_{N,t}\}$ and $\{\lambda_t\}$ are the same on the interval $[0,\tau_N]$, where the stopping time $\tau_N \equiv \inf\{ t > 0: \|\lambda_t\| > N\}$ (see Stroock and Varadhan (1979, Theorem 10.1.1).) Since the required convergence in

(b')-(d') is uniform on compact subsets of R^{n+2m} (not on all of R^{n+2m}) it will be enough to prove that the theorem holds for the diffusion $\{X_{N,t}, Y_{N,t}\}$ for all $N > 0$.

Truncating the coefficients of (3.1) in the manner described in Lemma A.4, and writing $\mu_{N,t}$ for $\mu_N(X_t, Y_t, t)$ and similarly with $\kappa_{N,t}$ and $\Omega_{N,t}$, we have

$$d\begin{bmatrix} X_{N,t} \\ Y_{N,t} \end{bmatrix} = h^{-1/4}\begin{bmatrix} \mu_{N,t} \\ \kappa_{N,t} \end{bmatrix}dt + \begin{bmatrix} c_{N11,t} & c_{N12,t} \\ c_{N21,t} & c_{N22,t} \end{bmatrix}\begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix} \tag{A.58}$$

where $\begin{bmatrix} c_{N11,t} & c_{N12,t} \\ c_{N21,t} & c_{N22,t} \end{bmatrix} \equiv \Omega_{N,t}^{1/2}$. So (A.59)

$$X_{N,t+h} = X_{N,t} + h^{-1/4}\int_t^{t+h} \mu_{N,s}\,ds + \int_t^{t+h} (c_{N11,s}\,dW_{1,s} + c_{N12,s}\,dW_{2,s}) \tag{A.60}$$

$$Y_{N,t+h} = Y_{N,t} + h^{-1/4}\int_t^{t+h} \kappa_{N,s}\,ds + \int_t^{t+h} (c_{N21,s}\,dW_{1,s} + c_{N22,s}\,dW_{2,s}) \tag{A.61}$$

That $E_t \| h^{-1/2}(X_{N,t+h}-X_t) \|^{2j}$ and $E_t \| h^{-1/2}(Y_{N,t+h}-Y_t) \|^{2j}$ are bounded for $j \geq 1$ is immediate from Lemma A.4. That $h^{-1/2}P[\| (X_{N,t+h}-X_t)', (Y_{N,t+h}-Y_t)' \| > \epsilon]$ converges to zero uniformly on compact subsets of R^{n+m} for every $\epsilon > 0$ now follows from Markov's inequality (e.g., Billingsley (1986, (5.27))). Next, take $\| (X_t', Y_t') \| < N-1$, and define $\hat{\xi}_{N,X,t+h}$ and $\xi_{N,Y,t+h}$ as

$$\hat{\xi}_{N,X,t+h} \equiv h^{-3/4}\int_t^{t+h} (\mu_{N,s} - \hat{\mu}_t)\,ds + h^{-1/2}\int_t^{t+h} (c_{N,11,s}\,dW_{1,s} + c_{N,12,s}\,dW_{2,s}) \tag{A.62}$$

$$\xi_{N,Y,t+h} \equiv h^{-1/2}\int_t^{t+h} (c_{N21,s}\,dW_{1,s} + c_{N22,s}\,dW_{2,s}) \tag{A.63}$$

Now define $\bar{\xi}_{N,X,t+h} \equiv h^{-1/2} \int_t^{t+h} (C_{N,11,t}\,dW_{1,t} + C_{N,12,t}\,dW_{2,s})$. Note that given time t information, $\bar{\xi}_{N,X,t+h}$ is gaussian, since $C_{N,11,\cdot}$ and $C_{N,12,\cdot}$ are held constant in the integrand at their time t

values. Next, expand Q_{t+h}-Q_t around $\hat{\xi}_{N,X,t+h}=\bar{\xi}_{N,X,t+h}$ and $\hat{Y}_t = Y_t$:

$$
\begin{aligned}
Q_{t+h}-Q_t &= h^{1/4}G(\bar{\xi}_{X,t+h},X_t,Y_t,t) - h^{-3/4}\int_t^{t+h}(C_{N,21,s}dW_{1,s}+C_{N,22,s}dW_{2,s}) \\
&+ h^{1/2}\frac{\partial G}{\partial Y}Q_t + h^{1/4}\frac{\partial G}{\partial \xi_X}(\hat{\xi}_{N,X,t+h}-\bar{\xi}_{X,t+h}) + h^{-1/2}\int_t^{t+h}(\hat{\kappa}_t-\kappa_{N,s})ds \\
&+ \text{higher order terms,}
\end{aligned}
\tag{A.64}
$$

where the partial derivatives are evaluated at $(\bar{\xi}_{X,t+h},X_t,Y_t,t)$. As $h\downarrow 0$, $h^{-1/2}\int_t^{t+h}(C_{N,21,s}dW_{1,s}+C_{N,22,s}dW_{2,s})$ converges weakly to the conditionally gaussian random variable $\bar{\xi}_{Y,t+h} \equiv h^{-1/2}\int_t^{t+h}(C_{21,t}dW_{1,s}+C_{22,t}dW_{2,s})$ (NF Lemma (A.1)). By Lemma A.4, for any N, $h^{-1/2}\int_t^{t+h}(C_{N,21,s}dW_{1,s}+C_{N,22,s}dW_{2,s})$ has arbitrary finite moments, so its conditional moments converge to the corresponding conditional moments of $\bar{\xi}_{Y,t+h}$ (Billingsley (1986, Corollary to Theorem 25.12)). Similarly, $\xi_{N,X,t+h}$ (and its conditional moments) converges to $\bar{\xi}_{X,t+h}$ (and its conditional moments). This moment boundedness (along with the polynomial bounds on G and its derivatives) ensures that the expectation and variances of the higher-order terms in the Taylor series expansion vanish as $h\downarrow 0$. Finally, this conditional moment boundedness also insures that $h^{-1}E_t[\int_t^{t+h}(\hat{\kappa}_t-\kappa_{N,s})ds]$ and $h^{-1}E_t[\int_t^{t+h}(\hat{\mu}_t-\mu_{N,s})ds]$ converge, respectively, to $(\hat{\kappa}_t-\kappa_t)$ and $(\hat{\mu}_t-\mu_t)$. Define the random variable $I_{\epsilon,t+h} \equiv 1$ if $\|(Q_{t+h}-Q_t)',(X_{t+h}-X_t)',(Y_{t+h}-Y_t)'\| > \epsilon$ and $\equiv 0$ otherwise. We then have

$$
\begin{aligned}
&E_t[(Q_{t+h}-Q_t)\cdot I_{1,t+h})] - (\hat{\kappa}_t-\kappa_t) + E_t[\frac{\partial G}{\partial \xi_X}](\mu_t-\hat{\mu}_t) + E_t[\frac{\partial G}{\partial Y}]Q_t \\
&E_t[(Q_{t+h}-Q_t)(Q_{t+h}-Q_t)'\cdot I_{1,t+h}] - E_t[(G(\bar{\xi}_{X,t+h},X_t,Y_t,t)-\xi_{Y,t+h})(G(\bar{\xi}_{X,t+h},X_t,Y_t,t)-\bar{\xi}_{Y,t+h})']
\end{aligned}
\tag{A.65}
$$

and for all $\epsilon > 0$, $E_t[I_{\epsilon,t+h}] \to 0$ by Markov's inequality. Because the moment bounds delivered by Lemma A.4 are uniform on compact (X,Y) sets, the convergence of these conditional moments is uniform on compacts, as required by Lemma A.3, completing the proof.

Proof of Theorem 4.1. First we need a minor modification of NF Theorem 2.1. Condition

(d') required the existence of a $\Delta > 0$ and $\delta > 0$ such that for every $\Lambda > 0$,

$$\lim_{h \downarrow 0} \sup_{|y| < \Lambda} h^{-\Delta} E[\|{}_hY_{k+1} - {}_hY_k\|^{2+\delta} | {}_hY_k = y] \to 0. \tag{A.66}$$

We can, however, replace this with the condition that there exists a $\Delta > 0$ such that for every $\Lambda > 0$ there exists a $\delta_\Lambda > 0$ such that

$$\lim_{h \downarrow 0} \sup_{|y| < \Lambda} h^{-\Delta} E[\|{}_hY_{k+1} - {}_hY_k\|^{2+\delta_\Lambda} | {}_hY_k = y] \to 0. \tag{A.67}$$

The proof of NF Theorem 2.1 is unaffected by this change. (This allows us to define the conditional degrees of freedom as $2 + e^r$ rather than as, say, $2.01 + e^r$.)

(4.5)-(4.7) are routinely derived from the multivariate t density. Once the assumptions of Theorem 2.2 are verified, (4.3)-(4.4) and Assumption 4 follow from Theorem 2.3. To derive (4.8) we made use of Prudnikov et al (1986, p. 505 formulas 53 and 56) and Davis (1964, formulas 6.2.1 and 6.3.5). The existence of the conditional densities in Assumption 3 is obvious. For bounded ν_t, the conditional degrees of freedom are bounded away from both 2 (assuring that (2.31) is satisfied) and ∞ (bounding S_{t+h} and so assuring that (2.32) is satisfied). $2+\delta$ moment boundedness of S_{t+h} and P_{t+h} in turn assures that (2.15) is satisfied. A in Assumption 1 is trivially $0_{m\times 1}$. B and C are derived for the optimal filter in the proof of Theorem 2.2. (2.11)-(2.12) follow from carrying out the Taylor series expansion in (2.7) and making use of the $2+\delta$ moment boundedness of P_{t+h} and S_{t+h}.

Proof of Theorem 4.2. Most of the work was done in the text. All remains is to demonstrate that the conditions of Theorem 3.1 hold. $\Lambda(\Omega(Y))$ and $[\text{vech}(\gamma)-\text{vech}(\theta)\odot Y]h^{\delta-1}$ are clearly continuous, so we need only demonstrate weak-sense uniqueness of the solution to (4.10') for every $h > 0$. Clearly $\Omega(Y)$ and $[\text{vech}(\gamma)-\text{vech}(\theta)\odot Y]h^{\delta-1}$ are twice differentiable in Y. Since $\Lambda(\Omega)$ is the covariance matrix of $\text{vech}(\alpha)\odot\text{vech}(ZZ'-\Omega)$, where $Z \sim N(0_{m\times 1},\Omega)$, and

$E[Z_i \cdot Z_j \cdot Z_k \cdot Z_m] = \Omega_{ij}\Omega_{km} + \Omega_{ik}\Omega_{jm} + \Omega_{im}\Omega_{jk}$, $\Lambda(\Omega(Y))$ is twice differentiable in Y as well, thus satisfying Condition A of Nelson (1990, Appendix A)). Applying Nelson (1990, Theorem A.1), the proof of weak sense uniqueness is complete if we can show that for each $h > 0$, there exists a nonnegative twice differentiable function $\varphi(X,Y,h)$ such that

$$\lim_{\|(X,Y)\| \to \infty} \varphi(X,Y,h) = \infty, \tag{A.68}$$

and for some $\lambda > 0$,

$$\sum_{i=1}^{m} (f_i - m_i Y_i)\frac{\partial \varphi}{\partial Y_i} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Omega_{ij}\frac{\partial^2 \varphi}{\partial X_i \partial X_j} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \Lambda_{ij}\frac{\partial^2 \varphi}{\partial Y_i \partial Y_j} \le \lambda \varphi(X,Y,h) \tag{A.69}$$

where $f \equiv \text{vech}(\gamma)$ and $m \equiv \text{vech}(\theta)$. The function we use is

$$\varphi(X,Y,h) \equiv [1 + \sum_{i=1}^{n} [1 - \exp(-X_i^{-2})]\,|X_i| + \sum_{j=1}^{m} [1 - \exp(-Y_i^{-2})]\,|Y_j|]\,h^{-1/4} \tag{A.70}$$

As required, $\varphi(X,Y,h)$ is twice continuously differentiable in X and Y and satisfies (A.68). $\partial^2\varphi/\partial X_i \partial X_j = \partial^2\varphi/\partial Y_i \partial Y_j = 0$ for $i \neq j$ and $\partial^2\varphi/\partial X_i \partial Y_j = 0$ for all i, j. When $|X_i|$ is large, $\partial\varphi/\partial X_i = (\pm 1 + O(X_i^{-2}))h^{-1/4}$, $\partial^2\varphi/\partial X_i^2 = O(|X_i|^{-3})$, and when $|Y_i|$ is large, $\partial\varphi/\partial Y_i = (\pm 1 + O(|Y_i|^{-2}))h^{-1/4}$, and $\partial^2\varphi/\partial Y_i^2 = O(|Y_i|^{-3})h^{-1/4}$. Note also that the elements of Ω are linear in the elements of Y, and that the elements of Λ are linear in $Y_i \cdot Y_j$ terms. It is easy to see therefore that (A.69) is satisfied for sufficiently large Y_i's. The constant term in φ ensures that for sufficiently large λ, (A.69) is satisfied for small Y_i's.

References

Anderson, B. D. O. and J. B. Moore, 1971, Linear optimal control (Prentice Hall, Englewood Cliffs, NJ).

Anderson, B. D. O. and J. B. Moore, 1979, Optimal filtering (Prentice Hall, Englewood

Cliffs, NJ).

Anderson, T. W., 1984, An introduction to multivariate statistical analysis, second edition (Wiley, New York, NY).

Aptech Systems Inc., 1992, Gauss, Version 3.0 (Aptech Systems Inc., Maple Valley, WA).

Arnold, L., 1973, Stochastic differential equations: Theory and applications (Wiley, New York, NY).

Baillie, R. T., T. Bollerslev, and H. Ole Æ Mikkelsen, 1993, Fractionally integrated generalized autoregressive conditional heteroskedasticity. Working paper (Michigan State University, East Lansing, MI.)

Bates, D. S., 1991, The crash of '87: Was it expected? The Evidence from Options Markets, Journal of Finance, 46, 1009-1044.

Bates, D. S., 1993, Jumps and stochastic volatility: Exchange rate processes implicit in PHLX Deutschemark options. Working paper (Wharton School, Philadelphia, PA).

Bellman, R., 1970, Introduction to matrix analysis, second edition (McGraw Hill, New York, NY).

Bera, A. K., and M. L. Higgins, 1993, A survey of ARCH models: Properties, estimation, and testing, Journal of Economic Surveys, 7, 305-366.

Bera, A. K. and S. Lee, 1992, Information matrix test, parameter heterogeneity, and ARCH: a synthesis. Forthcoming, Review of Economic Studies.

Billingsley, P., 1986, Probability and measure, second edition (Wiley, New York).

Bollerslev, T., 1990, Modelling the coherence in short-run nominal exchange rates: A mutivariate generalized ARCH approach, Review of Economics and Statistics, 72, 498-

505.

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.

Bollerslev, T., R. F. Engle, and D. B. Nelson, 1993, ARCH models, forthcoming in R. F. Engle and D. McFadden eds., The handbook of econometrics, Volume 4 (North Holland, Amsterdam).

Bollerslev, T., R. F. Engle, and J. M. Wooldridge, 1988, A capital asset pricing model with time-varying covariances, Journal of Political Economy, 96, 116-131.

Braun, P., D. B. Nelson, and A. Sunier, 1991, Good news, bad news, volatility, and betas. Working paper (Northwestern University, Evanston IL).

Cambanis, S., S. Huang, and G. Simons, 1981, On the theory of elliptically contoured distributions, Journal of Multivariate Analysis 11, 368-385.

Chiras, D. P., and S. Manaster, 1978, The information content of option prices and a test of market efficiency, Journal of Financial Economics 6, 213-234.

Christie, A. A., 1982, The stochastic behavior of common stock variances: value, leverage and interest rate effects, Journal of Financial Economics, 10, 407-432.

Cox, D. R., and N. Reid, 1987, Parameter orthogonality and approximate conditional inference, Journal of the Royal Statistical Society, Series B, 49, 1-39.

Cox, J. C. and M. Rubinstein, 1985, Options markets. (Prentice Hall, Englewood Cliffs, NJ).

Davis, P. J., 1964, Gamma function and related functions, in M. Abramowitz and I. A. Stegun eds., Handbook of mathematical functions (Dover, New York, NY) 253-293.

Day, T. E. and C. M. Lewis, 1992, Stock market volatility and the information content of

stock index options, Journal of Econometrics, 267-288.

Engle, R. F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50, 987-1008.

Engle, R. F., and K. F. Kroner, 1994, Multivariate simultaneous generalized ARCH, forthcoming, Econometric Theory.

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

Engle, R. F., V. Ng, and M. Rothschild, 1990, Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills, Journal of Econometrics, 45, 213-238.

Ethier, S. N., and T. G. Kurtz, 1986, Markov processes: Characterization and convergence (Wiley, New York, NY).

Fisher, R. A., 1921, On the probable error of a coefficient of correlation deduced from a small sample, Metron 1, No. 4, 1.

Foster, D. P., and Nelson, D. B., 1994, Continuous record asymptotics for rolling sample variance estimators. Working Paper (Wharton School, Philadelphia, PA)

Friedman, A., 1975, Stochastic differential equations with applications, volume 1 (Academic Press, New York, NY).

Gallant, A. R., P. E. Rossi, and G. Tauchen, 1992, Stock prices and volume, Review of Financial Studies, 5, 199-242.

Garman, M. B. and M. J. Klass, 1980, On the estimation of security price volatilities from historical data, Journal of Business 53, 67-78.

Glosten, L. R., R. Jagannathan, and D. Runkle, 1993, On the relation between the expected

value and the volatility of the nominal excess return on stocks, Journal of Finance 48, 1779-1801.

Hansen, B. E., 1992, Autoregressive conditional density estimation. Working paper (University of Rochester, Rochester, NY).

Hardy, G., J. E. Littlewood, and G. Pólya, 1952, Inequalities, second edition. (Cambridge University Press, Cambridge U.K.).

Harvey, A., E. Ruiz, and N. Shephard, 1994, Multivariate stochastic variance models, Review of Economic Studies, 61, 247-264.

Helland, I. S., 1982, Central limit theorems for martingales with discrete or continuous time. Scandinavian Journal of Statistics, 9, 79-94.

Horn, R. A., and C. R. Johnson, 1985, Matrix analysis, (Cambridge University Press, Cambridge, UK).

Hull, J. and A. White, 1987, The pricing of options on assets with stochastic volatilities. Journal of Finance, 42, 281-300.

Jacquier, E., N. G. Polson, and P. E. Rossi, 1992, Bayesian Analysis of Stochastic Volatility Models, forthcoming, Journal of Business and Economic Statistics.

Johnson, N. L., and S. Kotz, 1972, Continuous multivariate distributions. (Wiley, New York, NY).

Karatzas, I. and S. E. Shreve, 1988, Brownian motion and stochastic calculus (Springer Verlag, New York, NY).

Karpoff, J., 1987, The relation between price changes and trading volume: A survey. Journal of Financial and Quantitative Analysis 22, 109-126.

King, M. and S. Wadhwani, 1990, Transmission of volatility between stock markets, Review of Financial Studies, 3, 5-33.

Kroner, K. F., and V. K. Ng, 1993, Modelling the time varying comovement of asset returns. Working paper (University of Arizona, Tucson, AZ).

Kučera, V., 1972, A contribution to matrix quadratic equations, IEEE Transactions on Automatic Control, AC-17, 344-356.

Lancaster P., and L. Rodman, 1980, Existence and uniqueness theorems for the algebraic riccati equation, International Journal of Control, 32, 285-309.

Lancaster P., and M. Tismenetsky, 1985, The theory of matrices, second edition (Academic Press, San Diego, CA).

Lo, A. W. and J. Wang, 1993, Implementing option pricing formulas when asset returns are predictable. Working paper (Sloan School Of Management, Cambridge, MA).

Magnus, J. R. and H. Neudecker, 1988, Matrix differential calculus (Wiley, New York, NY).

McCulloch, J. H., 1985, On Heteros*edasticity, Econometrica 53, 2, 483.

Mitchell, A. F. S., 1989, The information matrix, skewness tensor and σ-connections for the general multivariate elliptic distribution, Annals of the Institute of Statistical Mathematics 41, 289-304.

Nelson, D. B., 1988, The time series behavior of stock market volatility and returns, unpublished doctoral dissertation (M.I.T. Economics Department, Cambridge, MA).

Nelson, D. B., 1990, ARCH models as diffusion approximations, Journal of Econometrics, 45, 7-38.

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, 1992, Filtering and forecasting with misspecified ARCH models II: Making the right forecast with the wrong model. forthcoming, Journal of Econometrics.

Nelson, D. B., and D. P. Foster, 1994, Asymptotic filtering theory for univariate ARCH models, Econometrica 62, 1-41.

Nelson, D. B., and B. A. Schwartz, 1992, Filtering with ARCH: A monte carlo experiment, Proceedings of the American Statistical Association Business and Economic Statistics Section, 1-6.

Phillips, P. C. B., 1988, Regression theory for near-integrated time series, Econometrica 56, 1021-1044.

Parkinson, M., 1980, The extreme value method for estimating the variance of the rate of return, Journal of Business 53, 61-65.

Prudnikov, A. P., Y. A. Brychkov and O. I. Marichev, 1986, Integrals and series: volume 1, elementary functions (Gordon and Breach Science Publishers, New York, NY).

Schwartz, B. A., 1994, Essays on ARCH filtering and estimation. Unpublished doctoral dissertation (University of Chicago GSB, Chicago, IL).

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.

Serfling, R. J., 1980, Approximation theorems of mathematical statistics (Wiley, New York,

NY).

Stroock, D. W., and S. R. S., 1979, Multidimensional diffusion processes (Springer-Verlag < Berlin).

Taussky, O., 1968, Positive-definite matrices and their role in the study of the characteristic roots of general matrices, Advances in Mathematics, 2, 175-186.

Turner, A. L., and E. J. Weigel, 1992, Daily stock market volatility: 1928-1989, Management Science, 38, 1586-1609.

Watanabe, T., 1992, Alternative approach to conditional heteroskedasticity in stock returns: approximate non-gaussian filtering. Working paper (Yale University, New Haven CT).

Wiggins, J. B., 1991, Empirical tests of the bias and efficiency of the extreme-value variance estimator for common stocks, Journal of Business, 64, 417-432.

Wolfram Research, Inc., 1992, Mathematica, Version 2.2 (Wolfram Research Inc., Champaign, IL).

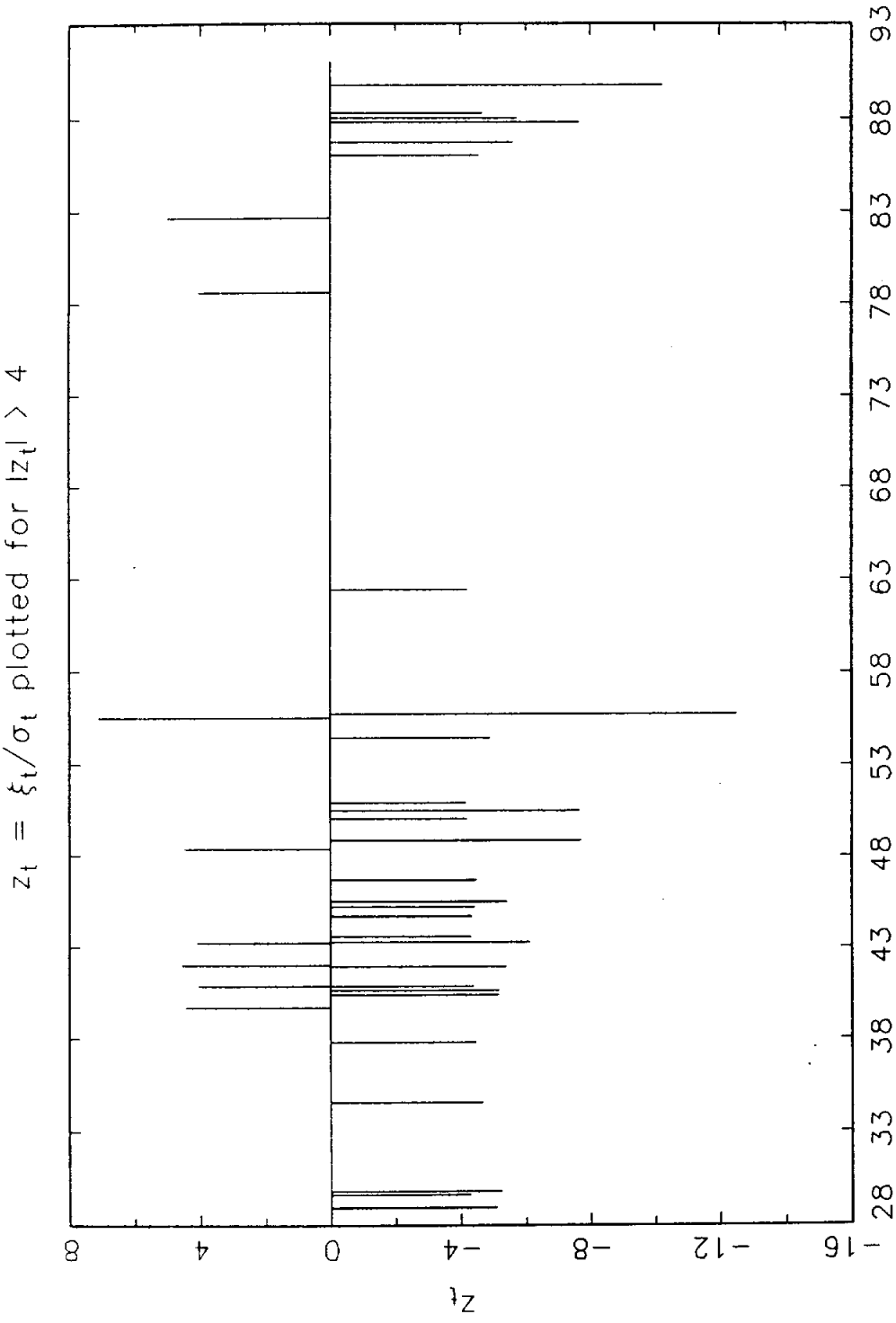
$z_t = \xi_t/\sigma_t$ plotted for $|z_t| > 4$
z_t
8
4
0
−4
−8
−12
−16
28
33
38
43
48
53
58
63
68
73
78
83
88
93

Figure 1