Continuous Record Asymptotics for Rolling Sample Variance Estimators*

Dean P. Foster
University of Pennsylvania

Dan B. Nelson
University of Chicago

October 27, 1995
First draft: December 91

Abstract

It is widely known that conditional covariances of asset returns change over time. Researchers doing empirical work have adopted many strategies for accommodating conditional heteroskedasticity. Among the popular strategies are: (a) chopping the available data into short blocks of time and assuming homoskedasticity within the blocks, (b) performing one-sided rolling regressions, in which only data from, say, the preceding five year period is used to estimate the conditional covariance of returns at a given date, and (c) performing two-sided rolling regressions, in which covariances are estimated for each date using, say, five years of lags and five years of leads. Another model–GARCH–amounts to a one-sided weighted rolling regression. We develop continuous record asymptotic approximations for the measurement error in conditional variances and covariances when using these methods. We derive asymptotically optimal window lengths for standard rolling regressions and optimal weights for weighted rolling regressions. As an empirical example, we estimate volatility on the S&P 500 stock index using daily data from 1928 to 1990.

Keywords: Stochastic volatility, ARCH, continuous record.

JEL Category: Multiple/Simultaneous Equations Models, Time series (C:32).

*We would like to thank Phillip Braun, John Cochrane, Gene Fama, Wayne Ferson, Ken French, Gerard Gennotte, Lars Hansen, Robin Lumsdaine, two referees, and the seminar participants at the University of Chicago, Northwestern, the 1991 NBER Program on Asset Pricing Meetings, and at the 1991 N.B.E.R./N.S.F. Time Series Meeting for helpful comments. This material is based on work supported by the National Science Foundation under grants # SES 9110131 and # SES-9310683. We thank the University of Chicago Graduate School of Business, the Center for Research in Security Prices, and William S. Fishman Research Scholarship for additional support.

1 Introduction

Most asset pricing theories relate expected returns on assets to their conditional variances and covariances. See, for example, the review of the ARCH literature in Bollerslev, Chou, and Kroner (1992). It is widely recognized that these conditional moments change over time. Unfortunately, conditional covariances are not directly observable, so in tests of asset pricing theories researchers must use estimates of conditional second moments. Similarly, market participants use estimates of conditional variances and covariances in hedging, option pricing, and in many other aspects of portfolio selection. How accurate are these estimated variances and covariances? How can they be estimated more accurately?

If conditional variances and covariances were *constant* over time, then standard statistical techniques would yield the answer to these questions. When conditional heteroskedasticity is present, these techniques will not suffice. In fact, as we see in Section 2 below, statistical methods that assume constant variances and covariances even over short time intervals present a misleadingly optimistic picture of how accurate the measurement is.

Though there are many strategies for estimating time-varying variances and covariances, among the most popular have been (a) chopping the returns data into blocks of time and treating conditional variances and covariances as constant within each block (e.g., Merton (1980), Poterba and Summers (1986), French, Schwert, and Stambaugh (1987)), and (b) the rolling regression approach of Officer (1973) and Fama and MacBeth (1973).

The appeal of such strategies is clear: on the one hand, they allow for the possibility (almost a certainty in economic applications!) that the parameters of the process evolve randomly over time. On the other hand, they impose little structure on the precise *way* in which the parameters evolve. All of these strategies accommodate random evolution in parameters by estimating the value of the parameters at time t using only data "near" t. For example, Fama and MacBeth (1973) estimated conditional betas at date t using only the returns data for a period of five to eight years prior to date t—a "rolling regression."[1]

As Fama and MacBeth explain it, this estimation strategy "reflects a desire to balance the statistical power obtained with a large sample from a stationary process against potential problems caused by any non-constancy of the β_i." The more important "the statistical

[1] These estimation strategies are also popular on Wall Street: see, for example, the Merrill Lynch (1986) beta book, which uses a five-year rolling regression with monthly data to estimate betas. Rolling regressions are also used in estimating conditional means (see, for example, Banerjee, Lumsdaine, and Stock (1991)), although our results do not apply directly to this case.

power obtained with a large sample" is, the more inclined a researcher should be to use a *long* string of data in the rolling regression. On the other hand, minimizing the "potential problems caused by any non-constancy of the β_i" points toward using a *short* period for the rolling regression.

Fama and MacBeth's choice of a 5-7 year window was motivated by the work of Fisher (1970) and Gonedes (1973), who found that this window length gave the best out-of-sample forecasting performance for individual stocks. In related work, Fisher (1970), and Fisher and Kamin (1985) develop approximate distributions for measurement errors in betas and optimal weighting schemes under the assumption that conditional betas are random walks independent of market returns.[2]

In this paper, we extend these theoretical results to a much broader class of data generating processes. In Section II we show how, under weak assumptions, to approximate the distribution of measurement errors in estimated conditional variances and covariances. These results are broad enough to accommodate not only one and two-sided rolling regressions, but also more general weighting schemes such as the ARCH(p) model of Engle (1982) and one of the multivariate extensions proposed by Bollerslev, Engle, and Wooldridge (1988).[3] In Section 3, we characterize optimal window lengths and optimal weights to use in rolling regressions. Section 4 considers estimation of conditional betas. In Section 5, we provide an empirical example. Section 6 is a brief conclusion. The proofs are collected in the Appendix.

2 Asymptotic distributions

To illustrate the intuition behind our approximation method, consider the following simple case; suppose the data are generated by the diffusion

$$dX_t = \mu(X_t, \sigma_t)dt + \sigma_t \cdot dW_{1,t} \tag{1}$$

$$d\sigma_t^2 = \lambda(X_t, \sigma_t)dt + \Lambda(X_t, \sigma_t) \cdot dW_{2,t} \tag{2}$$

[2] There is a large literature on random coefficient regression, of which the work of Fisher (1970) and Fisher and Kamin (1985) is an application. See, for example, Chow (1984) and the references therein.

[3] Asymptotic measurement error distributions for conditional variances generated by other ARCH models (which cannot be accommodated by the methods in this paper) are given in Nelson and Foster (1994), Nelson (1994).

where $W_{1,t}$ and $W_{2,t}$ are (possibly correlated) standard Brownian motions, X_t and σ_t^2 are scalars, and $\Lambda(\cdot,\cdot)$, $\lambda(\cdot,\cdot,\cdot)$, and $\mu(\cdot,\cdot,\cdot)$ are continuous, with $\Lambda(\cdot,\cdot)$ strictly positive.

Our assumption that $\Lambda(\cdot,\cdot)$ is strictly positive separates our approach from that of the non-parametric liturature.

Suppose that the $\{X_t\}$ process is observable but $\{\sigma_t^2\}$ is not. How can we use the information in the sample path of $\{X_t\}$ to estimate the path of $\{\sigma_t^2\}$? It is well known that as a diffusion is observed at finer and finer time intervals (say of length h), its conditional variance at any instant can be approximated with ever greater accuracy, until in the limit as $h \to 0$, it is known exactly. To understand why, note first that because σ_t^2 in (1)-(2) is generated by a diffusion, it is continuous (with probability one) as a function of time. This implies that for every $\epsilon > 0$ and every $t > 0$ there exists, with probability one, a random $\delta(t) > 0$ such that

$$\sup_{t-\delta(t)\leq s\leq t} |\sigma_s^2 - \sigma_t^2| < \epsilon. \tag{3}$$

That is, over suitably small time intervals, the change in σ_t^2 can be made as small as we like. Now choose a small constant $\delta > 0$ and chop the interval $[t-\delta, t]$ into M equal pieces. We then estimate σ_t^2 by

$$\hat{\sigma}_t^2(\delta, M) \equiv \delta^{-1}\sum_{j=1}^{M}(X_{t-(j-1)\delta/M} - X_{t-j\delta/M})^2 \tag{4}$$

(4) is a standard one-sided rolling regression in which we act as if μ_t were identically zero. When δ is small, μ_t and σ_t^2 are effectively constant, so when we condition on $\mu_{t-\delta}$ and $\sigma_{t-\delta}^2$, the normalized increments $(M/\delta)^{1/2}[X_{t-(j-1)\delta/M} - X_{t-j\delta/M}]$ are approximately i.i.d. $N(0, \sigma_{t-\delta}^2)$. Under suitable moment conditions, the tails of these normalized increments are well-behaved (i.e., not too thick), allowing us to apply a law of large numbers yielding $[\hat{\sigma}_t^2(\delta, M) - \sigma_t^2] \to 0$ in probability as $\delta \to 0$ and $M \to \infty$. Failing to correct for the non-zero drifts in X_t and σ_t^2 does not interfere with consistency–the effect of the drift terms on $\hat{\sigma}_t^2(\delta, M)$ vanishes as $M \to \infty$ and $\delta \to 0$.

Though quite a special case, (1)−(4) illustrate the basic intuition underlying our results: as $M \to \infty$ and $\delta \to 0$, the normalized increments in X_t become approximately i.i.d. with zero conditional mean, finite conditional variance, and sufficiently thin tails, allowing us to apply a law of large numbers to estimate σ_t^2. As we see below, it is possible–in a far more general setting–to apply a central limit theorem to develop an asymptotic normal distribution for the measurement error $[\hat{\sigma}_t^2(\delta, M) - \sigma_t^2]$.

We will now introduce the notation need for our theorems. For each $h > 0$, consider a random vector step function ${}_hX_t \in \mathbf{R}^k$ which makes jumps only at times $0, h, 2h$, and so on. Assume that ${}_hX_t$ is a random process with an (almost surely) finite conditional covariance matrix. Formally, ${}_hX_t$ is a locally square integrable semimartingale– see e.g., Jacod and Shiryaev (1987) chapters $1-2$. We take ${}_hX_t$ to be adapted to the filtration $\{{}_h\mathcal{F}_t\}$ where $\{{}_h\mathcal{F}_t\}$ is increasing and right continuous. ${}_hX_t \in \mathbf{R}^k$ can be decomposed into a "predictable" part and a martingale part, i.e., the Doob-Meyer decomposition.

$$ {}_h\Delta X_\tau \equiv {}_hX_\tau - {}_hX_{\tau-h} = {}_h\mu_\tau h + ({}_hM_{\tau+h} - {}_hM_\tau) = {}_h\mu_\tau \Delta\tau + \Delta_h M_\tau $$

where ${}_h\mu_t \in \mathbf{R}^k$ is ${}_h\mathcal{F}_{t-h}$ measurable, and $\Delta_h M_t \in \mathbf{R}^k$ is a local martingale difference array with an (almost surely) finite conditional covariance matrix. Further, to make our sums look like integrals, we set $\Delta\tau = h$, and $\Delta_h M_\tau \equiv {}_hM_\tau - {}_hM_{\tau-h}$.

The conditional covariance matrix of ${}_h\Delta X_\tau$ per unit of time is the $k \times k$ matrix ${}_h\Omega_\tau = [{}_h\Omega_{(ij)\tau}]$. In other words,

$$ E({}_h\Delta M_\tau \cdot {}_h\Delta M_\tau^T | {}_h\mathcal{F}_{\tau-h}) = {}_h\Omega_\tau \Delta\tau. $$

${}_h\Omega_\tau$ is ${}_h\mathcal{F}_{\tau-h}$ measurable.

Our interest is in estimating ${}_h\Omega_t$ when it randomly evolves over time. Just as the change in ${}_hX_\tau$ can be decomposed into a drift component (i.e., a component that is predictable one step ahead) and a martingale component, so, we assume, can the change in ${}_h\Omega_\tau$:

$$ \Delta_h \Omega_\tau = {}_h\lambda_\tau \Delta\tau + {}_h\Delta M_\tau^* $$

where ${}_h\lambda_\tau$, the instantaneous drift in ${}_h\Omega_\tau$, is ${}_h\mathcal{F}_{\tau-2h}$ measurable, and ${}_hM_\tau^*$ is a $k \times k$ matrix-valued local martingale with respect to the filtration ${}_h\mathcal{F}_{\tau-h}$. Further,

$$ E({}_h\Delta M_{(ij)\tau}^* \cdot {}_h\Delta M_{(kl)\tau}^* | {}_h\mathcal{F}_{\tau-2h}) = {}_h\Lambda_{(ijkl)\tau} \Delta\tau $$

So ${}_h\Lambda_\tau$ is ${}_h\mathcal{F}_{\tau-2h}$ measurable. ${}_h\lambda_t$ and ${}_h\Lambda_t$ are, respectively, the drift and variance per unit of time in the conditional variance process ${}_h\Omega_t$. Since ${}_h\Omega_t$ is a $k \times k$ matrix, its drift ${}_h\lambda_t$ is as well. The "variance of the variance" process Λ_τ is a $k \times k \times k \times k$ tensor. As we see below, the more variable the ${}_h\Omega_t$ process is (as measured by ${}_h\Lambda_t$) the less accurately it can be measured.

The class of data generating processes encompassed in this setup is very large, including, for example, discrete time stochastic volatility models (e.g., Melino and Turnbull (1990)), diffusions observed at discrete intervals of length h, (e.g., Wiggins (1987), Hull and White (1987)), ARCH models, (e.g., Bollerslev, Chou and Kroner (1990)) and many random coefficient models (Chow (1984)).

As is well known for standard regressions, the efficiency of least squares covariance matrix estimates depends to a considerable extent on tail thickness of the noise terms (see, e.g., Davidian and Carroll (1987)). This is true for rolling regressions as well. To motivate our next bit of notation, suppose for the moment that the $\Delta_h X_t$'s were i.i.d., scalar draws from a distribution with mean zero and variance Ω. If we estimate Ω using T observations by $\hat{\Omega} = T^{-1} \sum_{t=1}^{T} (\Delta_h X_t)^2$, the variance of $\hat{\Omega}$ is T^{-1} Var$[(\Delta_h X_t)^2]$. That is, the sample variance of $\hat{\Omega}$ depends on the *fourth* moments of the $\Delta_h X_t'$s. When ${}_h\Omega_t$ randomly evolves over time, we require an analogous measure of the *conditional* tail thickness of $\Delta_h X_t$. Accordingly, we define ${}_hB_\tau$, a $k \times k$ matrix-valued martingale by the following martingale difference array:[4]

$$ {}_h\Delta B_\tau = h^{-1/2}({}_h\Delta M_\tau \cdot {}_h\Delta M_\tau^T - {}_h\Omega_\tau \Delta\tau). $$

${}_hB_\tau$ is essentially an empirical second moment process with its conditional mean removed each period to make it a martingale. We next define the conditional variance process for ${}_hB_\tau$, the $k \times k \times k \times k$ tensor process ${}_h\theta_\tau$ with

$$ {}_h\theta_{(ijkl)\tau}\Delta\tau = E({}_h\Delta B_{(ij)\tau} \cdot {}_h\Delta B_{(kl)\tau} | {}_h\mathcal{F}_{\tau-h}). $$

$\theta_{(ijkl)t}$ is closely related to the multivariate conditional fourth moment of $\Delta_h M_t$:

$$ \begin{aligned} \theta_{(iiii)t} &= E[({}_h\Delta X_{i,t} - {}_h\mu_{i,t} \cdot h)^4 - {}_h\Omega^2_{(ii)t} | {}_h\mathcal{F}_{t-h}] \\ &= E[\Delta_h M^4_{i,t} - {}_h\Omega^2_{(ii)t} | {}_h\mathcal{F}_{t-h}]. \end{aligned} $$

$\theta_{(iiii)\tau}/\Omega^2_{(ii)\tau}$ is the conditional coefficient of kurtosis less one of the i^{th} variable at time τ. If $\Delta X_{i,t}$ is conditionally normal, then $\theta_{(iiii)\tau} = 2\Omega^2_{(ii)\tau}$.

We next define

$$ {}_h\rho_{(ijkl)\tau} \equiv \operatorname{corr}(\Delta_h B_{(ij)\tau}, \Delta_h M^*_{(kl)\tau-\Delta\tau} | {}_h\mathcal{F}_{\tau-h}). $$

[4]The reason for the $h^{-1/2}$ in the definition of B is to keep ${}_hB = O_p(1)$. Thus, the notation will remind us the size of various integrals. In other words, for M and M^*, we have the usual "size" condition that ${}_h\Delta M^2 = O(\Delta\tau)$, and ${}_h\Delta M^{*2} = O(\Delta\tau)$, and now this also holds for the B process: ${}_h\Delta B^2 = {}_h\theta\Delta\tau = O(\Delta\tau)$.

${}_h\rho_t$ is the conditional correlation between the innovations in the empirical second moment process ${}_hB_\tau$ and the innovations in the conditional variance process Ω_τ. The behavior of ${}_h\rho_{(ijkl)\tau}$ is an important determinant of our ability to measure ${}_h\Omega_t$ accurately. To see why, suppose that ${}_h\Omega_t$ is generated by a diagonal multivariate GARCH model as in Bollerslev, Engle, and Wooldridge (1988). In this case ${}_h\Omega_t$ equals a distributed lag of the outer product of residual vectors and therefore ${}_h\rho_{(iiii)t} = 1$. In this case, rolling regressions can estimate ${}_h\Omega_t$ arbitrarily well, since $\Delta_h\Omega_t$ is *perfectly* correlated with elements of $\Delta_h X_t \Delta_h X_t^T$. I.e., when we see $\Delta_h X_t$ this tells us all we need to know about the change in ${}_h\Omega_t$. On the other hand, suppose that ${}_h\Omega_t$ is generated by a diffusion observable at intervals of length h. In this case ${}_h\rho_{(ijkl)t} = 0$, and though $\Delta_h X_t \Delta_h X_t^T$ contains information about the *level* ${}_h\Omega_t$, it in general contains no information about *changes* in ${}_h\Omega_t$. The case where ${}_h\rho < 0$ is a sort of "reverse GARCH" case, in which larger than expected residuals cause variance to drop. Our results are able to accommodate this case, though it seems unlikely to be practically relevant. In general, however, the higher $|{}_h\rho_{(ijkl)t}|$, the more accurately measurable is ${}_h\Omega_{(ij)t}$. Unfortunately, we will have to assume a value for ρ because we will see that it is not identifiable.

The estimator we will study is

$$ {}_h\hat{\Omega}_{(ij)T} \equiv \sum_\tau {}_hw_{(ij)(\tau-T)}[{}_h\Delta X_{(i)\tau} - h \cdot {}_h\hat{\mu}_{(i)\tau}][{}_h\Delta X_{(j)\tau} - h \cdot {}_h\hat{\mu}_{(j)\tau}], \tag{5} $$

where ${}_h\hat{\mu}$ is a estimate of ${}_h\mu$, and ${}_h\Omega_{(ij)}$ is the ij^{th} component of ${}_h\Omega$, ${}_h\hat{\mu}_{(i)}$ is the i^{th} component of ${}_h\hat{\mu}$, and ${}_hw_{\tau-T}$ is a $k \times k$ weighting matrix for which $\sum {}_hw_{(ij)(\tau-T)}\Delta\tau = 1$. For now both the conditional mean estimate ${}_h\hat{\mu}_t$ and the weights ${}_hw_{(ij)t}$ as exogenously given, though below we consider data-dependent selection of ${}_hw_{(ij)t}$.

A special case of the above is the standard flat–weight rolling regression motivated by the following argument. $E(\Delta M)^2/\Delta\tau = \Omega$, so it seems reasonable that if we average terms like $\Delta\hat{M}^2/\Delta\tau$, we should get a good approximation to Ω. So, the rolling regression estimator of Ω is defined as:

$$ {}_h\hat{\Omega}_T \equiv [(n+m)h]^{-1} \sum_{\tau=T-nh}^{\tau=T+(m-1)h} [{}_h\Delta X_\tau - h_h\hat{\mu}_\tau][{}_h\Delta X_\tau - h_h\hat{\mu}_\tau]^T $$

Thus the weights are equal over some region. So,

$$ {}_hw_{(ij)\tau} = \begin{cases} \frac{1}{(n+m)h} & -nh \le \tau < mh \\ 0 & \text{otherwise} \end{cases} \tag{6} $$

So, for example, when $n = m = kh^{-1/2}$ for some constant k, (which when $\rho = 0$ will turn out to be the asymptotically optimal way of choosing a rolling regression) we see that ${}_h w_{\tau - T} \cong h^{-1/2} k^{-1}$ near T, and 0 far away from T, with $\sum w \Delta\tau = 1$. Here m is the number of leads and n is the number of lags. In a standard one-sided rolling regression, m is set equal to zero and ${}_h w_{(ij)t-T} = 1/nh$ for $T - nh \leq t < T$ and zero otherwise.

When $m = 0$ and the weights are non-negative but otherwise unconstrained in (5), we have a special case of the multivariate GARCH model of Bollerslev, Engle, and Wooldridge (1988). The method of treating conditional covariances as constant over blocks of time (e.g., Merton (1980), Poterba and Summers (1986), French, Schwert, and Stambaugh (1987)) is also easily accommodated: here $w = 1/hK$ whenever $t - T$ is in the same time block as time T and equals zero otherwise. K is the number of observations within the block.

2.1 Assumptions

The first assumption requires the first few conditional moments of ${}_h X_t$ and ${}_h \Omega_t$ remain bounded with small changes over small time intervals as $h \to 0$: This assumption essentially allows us to apply the central limit theorem locally in time.

Assumption A *The following 8 expressions are all $O_p(1)$:*

(i) $\sup_{s,t \in [T, T+h^{1/2}]} |\hat{\mu}_s - \mu_t|$

(ii) $\sup_{t \in [T, T+h^{1/2}]} |{}_h\lambda_{(i)t} - {}_h\lambda_{(i)T}|$

(iii) ${}_h\lambda_T$

(iv) ${}_h\Omega_T$

(v) ${}_h\theta_T$

(vi) ${}_h\Lambda_T$

(vii) *for some* $\epsilon > 0$, $E\left(|h^{-1/2}{}_h\Delta M_T^*|^{2+\epsilon} \middle| \mathcal{F}_{T-2h}\right)$

(viii) *for some* $\epsilon > 0$, $E\left(|h^{-1/2}{}_h\Delta B_T|^{2+\epsilon} \middle| \mathcal{F}_{T-h}\right)$.

Assumption A is not as formidable as its 8 parts appear. For example, if all these processes are actually continuous semi-martingales, then Assumption A will hold with only non-explosiveness conditions. This is made precise in the following definition and following restatement of Assumption A.

Definition *We will call ${}_h X_\tau$ a discretized continuous semi-martingale if there exists a process ${}_0 X_\tau$, such that ${}_h X_{ih} = {}_0 X_{ih}$ and ${}_0 X_\tau$ is a continuous semi-martingale with differential* *representation of* $d_0 X_\tau = {}_0\mu_\tau d\tau + {}_0\Omega_\tau^{1/2} d_0 W_\tau$*, where both* ${}_0\mu_\tau$ *and* ${}_0\Omega_\tau$ *are continuous semi-martingales with* Ω *positive definite* $[a.s]$*. Further,* $d_0\Omega_\tau = {}_0\lambda_\tau d\tau + {}_0\Lambda_\tau d_0 W'_\tau$*, where both* ${}_0\lambda_\tau$ *and* ${}_0\Lambda_\tau$ *are continuous semi-martingales, and* W_τ *and* W'_τ *are multivariate Brownian motions.*

Assumption (A') ${}_hX_\tau$ *is a discretized continuous semi-martingale for which there exists a random variable* M *with finite mean such that for all* $\tau \leq K$ *(*K *finite) the following holds [almost surely]:* $|{}_0\mu_\tau| + |{}_0\Omega_\tau| + |{}_0\theta_\tau| + |{}_0\lambda_\tau| + |{}_0\Lambda_\tau| \leq M$*. Also assume* ${}_h\hat{\mu}_\tau \equiv 0$*.*

From standard arguments Assumption A' can be shown to imply Assumption A. Thus, we see that Assumption A is more of a regularity condition rather than a restrictive assumption.

Assumption B ${}_h\theta_\tau, {}_h\Lambda_\tau$ *and* ${}_h\rho_\tau$ *change slowly over time. That is to say*

$$\begin{aligned}
\sup_{T \leq \tau \leq T + h^{1/2}} |{}_h\theta_\tau - {}_h\theta_T| &= o_p(1), \\
\sup_{T \leq \tau \leq T + h^{1/2}} |{}_h\Lambda_\tau - {}_h\Lambda_T| &= o_p(1), and \\
\sup_{T \leq \tau \leq T + h^{1/2}} |{}_h\rho_{(ijkl)\tau} - {}_h\rho_{(ijkl)T}| &= o_p(1).
\end{aligned}$$

Assumption B tells us that the “hyper-parameters” are regular enough that they can be estimated. Again this isn't a very restrictive assumption in the sense that these terms would naturally be $O_p(h^{1/2})$ if θ, Λ, and ρ followed SDEs.

Assumption C *The diagonal elements of* ${}_h\theta_\tau$ *and* ${}_h\Lambda_\tau$ *are non-vanishing. That is to say* $\forall i\ \forall j$*:* $1/{}_h\theta_{(ijij)T} = O_p(1)$*, and* $1/{}_h\Lambda_{(ijij)T} = O_p(1)$*.*

Assumption C tells us that we can get a non-degenerate asymptotic distribution at the natural rate of convergence. If assumption C were dropped, our asymptotic variance calculation would still hold. But the results might be trivial in the sense that we get an asymptotic normal with zero variance. Assumption C avoids this.

${}_h\mu_t, {}_h\hat{\mu}_t$, and ${}_h\lambda_t$ drop out of the asymptotic distribution of the measurement error in the conditional covariance estimate produced by the rolling regression–i.e., these terms are of only second order importance in determining the measurement error. In fact, if we explode ${}_h\mu_t, {}_h\hat{\mu}_t$, and ${}_h\lambda_t$to infinity as $h \to 0$ at a sufficiently slow rate, these conditional moments *still* drop out of the asymptotic distribution of the measurement error.

Definition *${}_hT_*$ and ${}_hT^*$ are the "start" and "end" times of the rolling regression. That means ${}_hw_{\tau-T}=0$ for $\tau<{}_hT_*$ or $\tau>{}_hT^*$.*

Note it is not required that ${}_hw_{\tau-T}$ be non-zero between T_* and T^*. This will be useful when considering two different weights. T_* will then typically be the earlier of the starting times and T^* the later of the ending times. The next assumption restricts the behavior of the weights ${}_hw_{\tau-T}$:

Assumption D

$$\begin{aligned} {}_hT^*-{}_hT_* &= O(h^{1/2}), \\ \sum_{\tau=T_*,T_*+h,\ldots}^{T^*} {}_hw_{(ij)\tau-T}\Delta\tau &= 1, \ \textit{and} \\ \sup_\tau(|{}_hw_{(ij)T}|) &= O(h^{-1/2}). \end{aligned}$$

Assumption D requires that the total number of lags and leads used in the rolling regression is going to infinity at rate $h^{-1/2}$, though the time interval over which the weights are nonzero is shrinking to 0 at rate $h^{1/2}$. Assumption A guarantees that changes in ${}_h\Omega_t$ are small over small time intervals: As in the illustration at the beginning of this section, as $h\to 0$ the rolling regression generates its conditional covariance estimate ${}_h\Omega_t$ using a *growing* number of residuals generated over a *shrinking* period of time. Unfortunately, however, Assumption D also *requires* that the number of residuals assigned nonzero weights is bounded for each h. This accommodates the ARCH(p) model of Engle (1982) with p growing at rate $h^{-1/2}$ as $h\to 0$, but formally excludes the GARCH(p,q) model of Bollerslev (1986). We can, however, *approximate* GARCH models to arbitrary accuracy by considering ARCH(p) models for arbitrarily large but finite (for each h) order.

Typically $w_{(ij)\tau-T}\geq 0$ but this is not required. Assumption D also requires $\sum w_{(ij)\tau-T}\Delta\tau=1$. Interpreting the rolling regression as a multivariate GARCH model, this corresponds to an IGARCH ("Integrated GARCH") model– see Engle and Bollerslev (1986). For the theorems we can relax this condition to only assume that $\sum w_{(ij)\tau-T}\Delta\tau=1+o(h^{1/4})$. For intuition on why IGARCH is approached as $h\to 0$, see Nelson (1992).

Definition

$${}_h\Psi_{(ij)x}\equiv\begin{cases} \displaystyle\sum_{\tau=x+h,x+2h\ldots}^{\infty} {}_hw_{(ij)\tau}\Delta\tau & \textit{if } x\geq 0 \\ \displaystyle-\sum_{\tau=-\infty}^{x} {}_hw_{(ij)\tau}\Delta\tau & \textit{if } x<0 \end{cases}.$$

Note: ${}_h\Psi_x$ is only defined if x/h is an integer. This is like an integral of ${}_hw_\tau$ in the sense that $\Delta\Psi(x)/\Delta x = -{}_hw_x$. For example, in the case of the flat weight rolling regression for a univariate process

$$ {}_h\Psi_{s-T} = \frac{({}_hT^* - s)I_{s\geq T} - (s - {}_hT_*)I_{s<T}}{{}_hT^* - {}_hT_*}, $$

where ${}_hT^*$ is the right end point of the rolling regression and ${}_hT_*$ is the left end point. Define the following sums,

$$ \begin{aligned} {}_hS_{ww} &\equiv h^{1/2}\sum_\tau {}_hw_\tau^2\Delta\tau \\ {}_hS_{\Psi\Psi} &\equiv h^{-1/2}\sum_\tau {}_h\Psi_\tau^2\Delta\tau \\ {}_hS_{w\Psi} &\equiv \sum_\tau {}_hw_\tau \cdot {}_h\Psi_\tau\Delta\tau \end{aligned} $$

In the multivariate case, w_τ is replaced by $w_{\tau(ij)}$. So, ${}_hS_{ww}$is $k \times k \times k$×k. So, these sums are actually tensors. For example ${}_hS_{w_{ij}\Psi_{kl}} \equiv \sum_\tau {}_hw_{\tau(ij)}\cdot {}_h\Psi_{\tau(kl)}\Delta\tau$.

Finally, define the normalized measurement error process

$$ {}_hQ_t \equiv h^{-1/4}({}_h\hat{\Omega}_t - {}_h\Omega_t) $$

Its conditional covariances are asymptotically the $k \times k \times k \times k$ tensor process ${}_hC_\tau$ with elements given by

$$ \begin{aligned} {}_hC_{(ijkl)t} \equiv\ & {}_hS_{w_{ij}w_{kl}} \cdot {}_h\theta_{(ijkl)t} + {}_hS_{\Psi_{ij}\Psi_{kl}} \cdot {}_h\Lambda_{(ijkl)t} + \\ & + {}_hS_{w_{ij}\Psi_{kl}} \cdot {}_h\rho_{(ijkl)t}\sqrt{{}_h\theta_{(ijij)t} \cdot {}_h\Lambda_{(klkl)t}} + \\ & + {}_hS_{w_{kl}\Psi_{ij}} \cdot {}_h\rho_{(klij)t}\sqrt{{}_h\theta_{(klkl)t} \cdot {}_h\Lambda_{(ijij)t}} \end{aligned} $$

Which in the scalar case is just (where the 'h' has been deleted from $C, \theta, \Lambda, \rho, \Psi$, and S)

$$ C_t \equiv S_{ww}\theta_t + 2S_{w\Psi}\rho_t\sqrt{\theta_t\Lambda_t} + S_{\Psi\Psi}\Lambda_t \tag{7} $$

2.2 Main Convergence Theorems

Theorem 1 (Representation:) *If Assumptions A & D hold, then*

$$ \begin{aligned} {}_hQ_{(ij)T} &\equiv h^{-1/4}({}_h\hat{\Omega}_{(ij)T} - {}_h\Omega_{(ij)T}) \\ &= h^{1/4}\sum_\tau {}_hw_{\tau-T}\, {}_h\Delta B_{(ij)\tau} + h^{-1/4}\sum_\tau {}_h\Psi_{\tau-T}\, {}_h\Delta M^*_{(ij)\tau} + o_p(1) \end{aligned} $$

Theorem 2 (Asymptotic Distribution:) *If Assumptions A–D hold, then*

$$ {}_hQ_T|\mathcal{F}_{T_*} \text{ is asymptotically distributed } N(0, {}_hC_{T_*}). \tag{8}$$

PROOFS: See the Appendix.

The matrix normal distribution in Theorem 2 has the obvious interpretation –i.e., the asymptotic covariance of ${}_hQ_{(ij)t}$ and ${}_hQ_{(kl)t}$ given $\mathcal{F}_{T_*}$is $C_{(ijkl)T_*}$. Alternatively, using an appropriate sense of a tensor square-root, equation (8) says, $C^{-1/2}Q \stackrel{D}{\to} N(0,1)$ where 1 is the tensor identity.

To illustrate the application of Theorem 2, consider a multivariate rolling regression with flat weights. Assume that ${}_hn_{ij} = n_0h^{-1/2}$, and ${}_hm_{ij} = m_0h^{-1/2}$. This is a restricted form of a rolling regression in which all of the windows are the same size. For all i and j the same weighting is then used. In other words, ${}_hw_\tau = h^{1/2}(n_0+m_0)^{-1}I\{\tau \in [-n_0h^{1/2}, m_0h^{1/2}]\}$. Thus, Assumption D is satisfied. So, in this case, (the following approximations are easy to see if one thinks of each sum as being approximated by an integral):

$$\begin{aligned} h^{-1/2}\Psi_{(h^{1/2}s)} &= \frac{(m_0-s)I_{s\geq T}+(s-n_0)I_{s<T}}{m_0+n_0} \\ S_{ww} &= \sum w_s^2 h \cong \frac{1}{m_0+n_0} \\ S_{\Psi\Psi} &= \sum \Psi_s^2 h \cong \frac{m_0^3+n_0^3}{3(m_0+n_0)^2} \\ S_{w\Psi} &= \sum w_s\Psi_s h \cong \frac{m_0-n_0}{2(m_0+n_0)} \end{aligned}$$

Because of our assumption that all w_{ij} are the same, we don't need to distinguish between $S_{w_{ij}w_{kl}}$ and just call all of them S_{ww}. Likewise for $S_{w\Phi}$ and $S_{\Psi\Psi}$. We can now compute the variance of $\hat{\Omega}_{ij}$. Then, (where to simplify the equations we have taken: ${}_h\theta_{(ijij)T_*} = \theta, {}_h\Lambda_{(ijij)T_*} = \Lambda, {}_h\rho_{(ijij)T_*} = \rho$)

$$\begin{aligned} C_{(ijij)T_*} &= \theta S_{ww} + 2\sqrt{\theta\Lambda}\rho S_{w\Psi} + \Lambda S_{\Psi\Psi} \\ &= \frac{\theta}{m_0+n_0} + \sqrt{\theta\Lambda}\rho\frac{m_0-n_0}{n_0+m_0} + \Lambda\frac{m_0^3+n_0^3}{3(n_0+m_0)^2} \end{aligned} \tag{9}$$

Consider the three components of the asymptotic covariances in (9): the first term, θS_{ww}, would be present even in the i.i.d. case. This term reflects sampling error, and can be made arbitrarily small by making $n_0 + m_0$ sufficiently large. Indeed, if the conditional covariance matrix ${}_h\Omega_t$ were constant, the other terms in $C_{(ijkl)T_*}$ would vanish, and letting

n_0+m_0 be infinite would be optimal. The third term, $\Lambda S_{\Psi\Psi}$, reflects the variability in ${}_h\Omega_t$. This term can be made arbitrarily small by making n_0+m_0 sufficiently small: the smaller the window over which the rolling regression is conducted, the more like a constant ${}_h\Omega_t$ is within the window. As indicated in our discussion of ${}_h\rho$, the second term, $\sqrt{\theta\Lambda}\rho S_{w\Psi}$, comes from the covariance between the first and last terms. This term drops out when the data are generated by a diffusion but not, for example, when the data are generated by a GARCH model. This term also controls how much information about ${}_h\Omega_\tau$ is in the "past" residuals as opposed to the future residuals.

2.3 Consistent Estimation of Nuisance Parameters

To construct correct asymptotic confidence intervals, we must have consistent estimates of the components of the conditional covariance of the measurement error ${}_hQ_t$, namely ${}_h\theta_t$, ${}_h\Lambda_t$, and ${}_h\rho_t$. Sometimes some of these are known a priori: for example, when $\{{}_hX_t, {}_h\Omega_t\}$ is generated by a diffusion process, ${}_h\rho_{(ijkl)t} \to 0$, ${}_h\theta_{(iiii)t}/{}_h\Omega^2_{(ii)t} \to 2$ and $\theta_{ijkl} \to 0$ otherwise as $h \to 0$, thus leaving only Λ_t to estimate. In more general circumstances, however, they all must be estimated.

We next consider estimation of ${}_h\theta_\tau$and ${}_h\Lambda_\tau$.

Since we have only the most indirect methods of obtaining information about these parameters, we will need to assume that the processes under consideration are "regular" over a slightly longer interval. To do this we will use the following uniform convergence idea. We will say that $X_T = o_p(1)$ holds uniformly over $T \in [T', T' + K_h h^{1/2}]$ if for all $\epsilon > 0$,

$$\sup_{T \in [T', T' + K_h h^{1/2}]} P(|X_T| > \epsilon) \to 0 \text{ as } h \to 0.$$

Assumption E *Assume there exists a function $K_h$such that $K_h \to \infty$ as $h \to 0$ such that Assumption A holds uniformly over $T \in [T', T' + K_h h^{1/2}]$.*

By way of example, consider assumption A part (iii). It tells us that λ_T is small: $|{}_h\lambda_T| = O_p(1)$. In other words, Assumption A-iii by itself says: $\forall\epsilon > 0$, $\exists M$ such that $P(|{}_h\lambda_T| > M) < \epsilon$ for sufficiently small h. Under Assumption E, we have the following stronger statement:$\forall\epsilon > 0$, $\exists M$ such that

$$\sup_{T\in[T',T'+K_h h^{1/2}]} P(|{}_h\lambda_T| > M) < \epsilon$$

for sufficiently small h. We now need to assume that our "targets" don't change very much over short time intervals. In other words, we need a stronger version of Assumption B.

Assumption F *For the K_h in assumption E,*

$$\sup_{\tau \in [T', T' + K_h h^{1/2}]} |{}_h\theta_\tau - {}_h\theta_T| = o_p(1)$$

$$\sup_{\tau \in [T', T' + K_h h^{1/2}]} |{}_h\Lambda_\tau - {}_h\Lambda_T| = o_p(1)$$

$$\sup_{\tau \in [T', T' + K_h h^{1/2}]} |{}_h\rho_\tau - {}_h\rho_T| = o_p(1)$$

Assumption (F') *For the K_h in Assumption E, (in the univariate case only)*

$$\sup_{\tau\in[T',T'+K_h h^{1/2}]} |{}_h\theta_\tau / {}_h\Omega_\tau^2 - {}_h\theta_T / {}_h\Omega_T^2| = o_p(1)$$

$$\sup_{\tau\in[T',T'+K_h h^{1/2}]} |{}_h\Lambda_\tau / {}_h\Omega_\tau^2 - {}_h\Lambda_T / {}_h\Omega_T^2| = o_p(1)$$

$$\sup_{\tau\in[T',T'+K_h h^{1/2}]} |{}_h\rho_\tau - {}_h\rho_T| = o_p(1)$$

Assumption F trivially implies Assumption B. That F' implies Assumption B follows from the "near" constancy of Ω_T over intervals of length $h^{1/2}$. Assumption F is more natural for the proof of our convergence theorem, and is easily understood in the multivariate setting.

In the univariate case, the advantage of using θ/Ω^2 and Λ/Ω^2 instead of θ and Λ respectively, is that it may be more believable that the "shape" parameters are constant than the parameters themselves: Constant θ/Ω^2 is equivalent to constant conditional kurtosis of the increments in ${}_hX_t$. When ${}_hX_t$ is generated by a diffusion, for example, $\theta/\Omega^2 = 2$. Constant Λ/Ω^2 is equivalent to $\ln({}_h\Omega_t)$ being conditionally homoskedastic. Many ARCH and stochastic volatility models effectively assume this (see Nelson and Foster (1994)) and, as we see in the empirical application below, this homoskedastic $\ln({}_h\Omega_t)$ seems a reasonable approximation for U.S. stock prices.

In the univariate case, these assumptions are equivalent in the sense that a process that satisfies F for some K_h will satisfy F' for some other K_h (and visa versa). But if one of these $K_h's$ is significantly larger than the other, it will allow the use of more data in estimating θ and Λ.

We will now outline estimators for θ and Λ. First define a matrix $f_\epsilon(\tau)$ which can be thought of as being $\hat{\Omega}_\tau - \hat{\Omega}_{\tau-\epsilon}$ for an appropriately defined $\hat{\Omega}$:

$$f_\epsilon(\tau) = \sum_{s=\tau,\tau+h,\ldots}^{\tau+\epsilon h^{1/2}} h^{-1/2}(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)'/\epsilon - $$
$$- \sum_{s=\tau,\tau-h,\tau-2h,\ldots}^{\tau-\epsilon h^{1/2}} h^{-1/2}(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)'/\epsilon$$

$\hat{\theta}$ and $\hat{\Lambda}$ can now be defined as:

$$\hat{\theta}_{(ijkl)T} = \frac{\epsilon}{2K_h} \sum_{\tau=T,T+h,\ldots}^{T+K_n h^{1/2}} f_{(ij)\epsilon}(\tau) f_{(kl)\epsilon}(\tau) - \hat{\Lambda}_{(ijkl)T}\epsilon^2/3 \tag{10}$$

$$\hat{\Lambda}_{(ijkl)T} = \frac{3}{2\delta K_h} \sum_{\tau=T,T+h,\ldots}^{T+K_n h^{1/2}} f_{(ij)\delta}(\tau) f_{(kl)\delta}(\tau) - 3\hat{\theta}_{(ijkl)T}/\delta^2 \tag{11}$$

In the case where $\epsilon \ll \delta$ these estimators are more intuitive because the "corrections" are small and only the sums themselves need be considered. To actually get the estimators, we have to solve the simultaneous equations (10) and (11). These estimators are designed to work with Assumption F. The following theorem shows they achieve this goal.

Theorem 3 (Consistency) *Under Assumption D, E, and F, both $\hat{\theta}_T$ and $\hat{\Lambda}_T$ are consistent pointwise in T.*

Proof: See the Appendix.

For the scalar case, Assumption F' should hold over a longer interval and so "better" estimates of θ and Λ should be available. Estimators appropriate for this situation will now be given. The definition of f_ϵ is notationally simpler in the scalar case:

$$f_\epsilon(\tau) = \sum_{s=\tau,\tau+h,\ldots}^{\tau+\epsilon h^{1/2}} h^{-1/2}(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)^2/\epsilon - $$
$$\sum_{s=\tau,\tau-h,\ldots}^{\tau-\epsilon h^{1/2}} h^{-1/2}(\Delta_h X_s - {}_h\hat{\mu}_s \Delta s)^2/\epsilon$$

Now modify (10) and (11) as follows:

$$\hat{\theta}_T = \frac{\hat{\Omega}_T^2 \epsilon}{2K_h} \sum_{\tau=T,T+h,\ldots}^{T+K_n h^{1/2}} f_\epsilon(\tau)^2/\hat{\Omega}_\tau^2 - \hat{\Lambda}_T \epsilon^2/3 \tag{10$'$}$$

$$\hat{\Lambda}_T = \frac{3\hat{\Omega}_T^2}{2\delta K_h} \sum_{\tau=T,T+h,\ldots}^{T+K_n h^{1/2}} f_\delta(\tau)^2/\hat{\Omega}_\tau^2 - 3\hat{\theta}_T/\delta^2 \tag{11$'$}$$

These are the estimators that we actually use in the empirical example.

We will see from the simulations that the following estimator appears to do somewhat better for $\hat{\Lambda}$ in the scalar case:

$$\hat{\Lambda}_T = \frac{3\hat{\Omega}_T^2}{2\delta K_h} \sum_{\tau=T}^{T+K_n h^{1/2}} \log(\hat{\Omega}(T+\delta h^{1/2})/\hat{\Omega}_T)^2 - 3\hat{\theta}_T/\delta^2, \tag{12}$$

where $\hat{\Omega}$ is taken to be a 1 sided rolling regression of length $\delta h^{1/2}$. Eq. (12) can be seen to be close to (11') if a one term Taylor series for the log is used.

Without further assumptions on the processes X_t and Ω_t it is impossible to estimate ${}_h\rho_{(ij\,kl)t}$. In order to prove this we have to find two models which have identical observable random variables (i.e. the distributions for the X_t's are the same) but have different values for the parameter ${}_h\rho_{(ij\,kl)t}$. Luckily Shephard (1994) does exactly this. He starts out with a stochastic volitility model which by construction has a ρ of zero. In other words, $\Omega_t = \mathrm{Var}(X_t|\mathcal{F}_{t-h})$ where $\mathcal{F}$ is the σ-field generated by both the observed state variable X and the unobservable latent variable Ω. He integrates out this latent state variable and generates a GARCH model. This is equivalent to looking at $\Omega'_t = \mathrm{Var}(X_t|\mathcal{G}_{t-h})$ where $\mathcal{G}$ is the σ-field generated only by the observable state variable X. In this new model, $\rho = 1$. Since only the process X is observed, it is impossible to distinguish between these two models. Thus, ρ is unidentifiable since it changes with the definition of the σ-field.

3 Efficiency and Optimality

Throughout this section, we will use various techniques of estimating a particular Ω_{ij}. Thus, we will think of i, j as fixed. We will call ${}_h\theta_{(ijij)T_*} = \theta$, ${}_h\Lambda_{(ijij)T_*} = \Lambda$, ${}_h\rho_{(ijij)T_*} = \rho$. Further, because we will want to compare windows of different lengths, we will take our conditioning time to be $T_* = T - kh^{-1/2}$ for some sufficiently large k.

3.1 Optimal Lead and Lag Lengths in Standard (flat-weight) Rolling Regressions

Consider again the example of the standard rolling regression in the previous Section, in which, for some nonnegative n_0 and m_0 the weights are given by ${}_hw_t = h^{1/2}(n_0+m_0) \cdot I(t \in [-n_0 h^{1/2}, m_0 h^{1/2}])$. This weighting scheme is of special interest, since it is most frequently encountered in practice. The asymptotic standard error for the ij^{th} element of the measurement error in the conditional covariance matrix is given in (9). In other words,

$$\begin{aligned} SE(\hat{\Omega}_{ij})^2 &= (\text{bias})^2 + \ \text{Var}(\hat{\Omega}_{ij}) \cong 0^2 + C_{ijij} \\ &\cong \frac{\theta}{m_0 + n_0} + \sqrt{\theta\Lambda}\rho\frac{m_0 - n_0}{n_0 + m_0} + \Lambda\frac{m_0^3 + n_0^3}{3(n_0 + m_0)^2}. \end{aligned}$$

Where the first equality follows by definition of the mean squared error, and the approximation follows from our Theorem 2.

Theorem 4 (flat-weight)

- *The asymptotic variance-minimizing backward looking flat-weight rolling regression (i.e., $m_0 = 0$) is given by setting $n_0 = \sqrt{\frac{3\theta}{\Lambda}}$. The asymptotic measurement error variance (see (9)) achieved by this choice of m_0 and n_0 is $(2-\rho)\sqrt{\frac{\Lambda\theta}{3}}$.*
- *The asymptotic variance-minimizing forward looking flat-weight rolling regression (i.e., $n_0 = 0$) is given by setting $m_0 = \sqrt{\frac{3\theta}{\Lambda}}$. The asymptotic measurement error variance achieved with this choice of m_0 and n_0 is $(2+\rho)[\Lambda\theta/3]^{1/2}$.*
- *When $\rho > (3/4)^{1/2}$, the one-sided backward-looking flat-weight rolling regression is asymptotically optimal in the class of flat-weight rolling regressions. When $\rho < -(3/4)^{1/2}$, the optimum is a one-sided forward-looking rolling regression. When $|\rho| \leq (3/4)^{1/2}$, the asymptotic optimum is a two-sided rolling regression with*

$$n_0 = \sqrt{3(1-\rho^2)\theta/\Lambda} + \rho\sqrt{\theta/\Lambda}, \text{ and} \tag{13}$$

$$m_0 = \sqrt{3(1-\rho^2)\theta/\Lambda} - \rho\sqrt{\theta/\Lambda} \tag{14}$$

The minimized asymptotic variance when $|\rho| \leq (3/4)^{1/2}$ is $\sqrt{\Lambda\theta(1-\rho^2)/3}$.

Proof: See the Appendix.

Note the role of ${}_h\rho$ in determining the optimal weighting scheme: when GARCH generates the data, ${}_h\rho = 1$ and all information used by the rolling regression about ${}_h\Omega_t$ is in the *lagged* residuals. The closer ${}_h\rho$ is to 1 therefore, the more weight is optimally put on lagged (as opposed to led) residuals.

The ${}_h\rho = 0$ case is also instructive: here the optimal weighting scheme is two-sided with equal window lengths on each side. This cuts the asymptotic variance exactly in half compared with the optimal one-sided rolling regression.

3.2 Optimal Weighted Rolling Regressions

Although flat-weight rolling regressions are widely used, they are generally non-optimal:

Theorem 5 (Optimal weights) *Define θ and Λ as in (7) and let $\alpha \equiv \sqrt{\Lambda/\theta}$.*

- *The asymptotic variance-minimizing backward looking (i.e., all the weight is on lagged residuals) weight function ${}_0w_t$ is given by $I_{\{t<0\}}\alpha e^{\alpha t}$. This achieves an asymptotic measurement error variance of $\sqrt{\Lambda\theta}(1-\rho)$.*
- *The asymptotic variance-minimizing forward looking weight function ${}_0w_t$ is given by $I_{\{t>0\}}\alpha e^{-\alpha t}$. This achieves an asymptotic measurement error variance of $\sqrt{\Lambda\theta}(1+\rho)$.*
- *The asymptotic variance-minimizing weight function ${}_0w_t$ is given by*

$$
{}_0w_s = \begin{cases} p\alpha e^{-\alpha s} & \text{for } s \geq 0 \\ (1-p)\alpha e^{\alpha s} & \text{for } s < 0, \end{cases} \tag{15}
$$

where $p = (1-\rho)/2$. This achieves an asymptotic measurement error variance of $(1/2)\sqrt{\Lambda\theta}(1-\rho^2)$.

Proof: See the Appendix.

`shorten, kill, etc!`

Others (Geno-Catalot, Laredo and Picard (1992), Corradi and White (1994), Banon (1978), Dohnal (1987) and Florens-Zmirou (1993)) have estimated Ω_t by non-parametric methods. Their estimators often achieve better rates of convergence then we do since they assume that Ω_t is much smoother than we assume it to be. On the other hand, we can often handle a more general situation than they can. So, the choice of estimator and its resulting rate of convergence depends on which assumptions are appropriate.

Note that the estimators recommended Theorem 5 violate our assumptions in the sense that ${}_0w_s$ does not have compact support. Of course the recommended ${}_0w_s$ can be arbitrarily well approximated by a w which does have compact support.

Further notice that in terms of forecasting (i.e. backwards looking) the optimal weighting is the same regardless of the value of ρ. Thus even if ρ can not be estimated, optimal forecasts for Ω are still available. Of course, we wouldn't know how accurate these forecasts in fact are!

Another popular strategy for estimating conditional covariances–chopping the data up into short blocks and estimating covariances as if they were constant within the blocks (see, e.g., Merton (1980), Poterba and Summers (1986), French, Schwert, and Stambaugh (1987))–is a special case of the two-sided flat-weight rolling regression. Suppose the block is composed of a total of K observations. At the left (right) end point of the block, the covariance matrix estimate is a one-sided rolling regression using K led (lagged) residuals. Between the two end points, the estimate is a two-sided rolling regression. If we set $K \equiv h^{-1/2}k_0$, then the asymptotic measurement error variance at a point a fraction η through the block ($0 \leq \eta \leq 1$) is obtained from (9) by setting $n_0 = k_0\eta$ and $m_0 = k_0(1-\eta)$:

$$\hat{C} = \theta/k_0 + \rho\sqrt{\theta\Lambda}(1-2\eta) + (\Lambda k_0/3)[\eta^3 + (1-\eta)^3] \tag{16}$$

which, when $|\rho\sqrt{\theta\Lambda}k_0| \leq 1/2$, is minimized when $\eta = 1/2 - \rho\sqrt{\theta/\Lambda}k_0$, lending a bow shape to the confidence intervals.

An obvious implication of Theorem 5 is that flat-weighting schemes such as one or two-sided rolling regressions or block-constant estimators are inefficient. Unfortunately, however, constructing the asymptotically efficient weights requires consistent estimates of the nuisance parameter processes $\{\rho_t\}, \{\Lambda_t\}$, and $\{\theta_t\}$. Can we construct dominating weighting schemes without knowing $\{\rho_t\}, \{\Lambda_t\}$, and $\{\theta_t\}$? The answer, it turns out, is yes:

Theorem 6 (Dominating flat weights) *For every i and j, define the weights ${}_hw_{(ij)\tau-T}$ by (6) (i.e., we use $n = n_0h^{-1/2}$ lagged residuals and $m = m_0h^{-1/2}$ led residuals). Define an alternative set of weights ${}_hw^*_{(ij)\tau-T}$ by*

$$ {}_hw^*_{(ij)\tau-T} = \begin{cases} 3^{1/2}(n_0+m_0)\exp[-3^{1/2}h^{1/2}(T-\tau)/m_0] & \textit{if } \tau > T \\ 3^{1/2}(n_0+m_0)\exp[-3^{1/2}h^{1/2}(T-\tau)/n_0] & \textit{if } \tau < \ T. \end{cases} \tag{17}$$

Then the asymptotic variance obtained using ${}_hw^*_{(ij)\tau-T}$ is lower than the asymptotic variance obtained by using ${}_hw_{(ij)\tau-T}$ for any ρ, θ, and Λ, with

$$\hat{C} - \hat{C}^* = (1-\sqrt{3}/2)\left(\ \psi(0-)^2(\theta/m_0 + \Lambda m_0/3) + \psi(0)^2(\theta/n_0 + \Lambda n_0/3)\ \right) > 0.$$

The idea behind Theorem 6 is simple: we leave the total *share* of the weight put on led and lagged residuals unchanged, but alter the *shape* of the weights on each side of time T from a block-shape to an exponential decline.

There is another natural way to dominate a block-constant estimation scheme, provided we are willing to consider average, rather than pointwise, measures of accuracy: integrate the measurement error variance (16) across the block (i.e., integrate (16) over η from 0 to

1), yielding an average measurement error variance across the block of ("b.c." is for "block constant") $\hat{C}_{\text{b.c.}} = \theta/k_0 + (\Lambda k_0/6)$ Now consider a flat-weight, two-sided rolling regression using $K/2 = .5k_o h^{-1/2}$ leads and the same number of lags. By (9), this achieves an average measurement error variance of ("t.s." is for "two sided") $\hat{C}_{\text{t.s.}} = \theta/k_0 + (\Lambda k_0/12)$, which is strictly smaller whenever $\Lambda > 0$, regardless of the values of k_0, ρ, and Ω. Of course, this two-sided rolling regression is itself dominated by an exponentially weighted rolling regression constructed as in Theorem 6.

If we are willing to assume that $\rho = 0$, as it would be, for example, if the data are generated by a diffusion observed at discrete intervals, further dominance relations follow: in particular, a one-sided rolling regression using, say, n lags and no leads has exactly twice the asymptotic variance of a rolling regression using n lags and n leads. The resulting two-sided rolling regression is itself dominated by an exponential-weighted rolling regression constructed as in Theorem 6.

******* **Figure 1 near here** *********

Several of the dominance relations are illustrated in figure 1. Using numbers from the empirical application in Section 5, figure 1 plots the ratios of the standard deviation of measurement errors in S&P 500 volatility estimates using various estimation schemes to that obtained using the optimal two-sided exponentially weighted estimator. The graph was constructed under the assumption that $\rho = 0$. In switching from the optimal two-sided exponentially weighted estimator to the optimal flat-weight estimator, the standard deviation of the measurement error rises about 7%. In switching from the optimal-two sided to the optimal one-sided estimate, the standard deviation goes up by a factor of $\sqrt{2}$. The bow-shaped pattern attained by the block-constant scheme of French, Schwert, and Stambaugh (1987) and of Poterba and Summers (1986) is clear in figure 1: when $\rho = 0$, this estimate does relatively well mid-month but poorly at the beginning and the end of the month. Switching from this block constant scheme to using a two-sided rolling regression with the same number of residuals (as proposed above) achieves a standard error equal to the (minimized) mid-month standard error.

If standard errors are estimated for the variance estimate under the false assumption that the covariance matrix *truly* is constant within blocks, only the sampling error term θ/k_0 appears, giving an unrealistically optimistic picture of the accuracy of the estimated covariance matrix. This is illustrated in figure 2.

******* **Figure 2 near here** *********

3.3 The Relation between the Regularity Conditions and the Optimality Results

Clearly there are relaxations in the regularity conditions which would invalidate the optimality results. For example, suppose that within each month, volatility is constant, with each month's volatility an i.i.d. draw from some distribution. Presumably in this case the block-constant estimation scheme of Poterba and Summers (1986) and French, Schwert and Stambaugh (1987) would dominate two-sided exponentially declining weights. This, however, would violate our regularity conditions, which (asymptotically) ruled out discrete jumps in ${}_h\Omega_t$.

A more subtle example was suggested to us by John Campbell: suppose volatility follows a moving average process in which volatility shocks persist–with constant weight–for some period and then suddenly die out. In this case, a flat-weight rolling regression would presumably dominate an exponential weighting scheme. (This is obviously true, for example, if volatility follows Engle's (1982) ARCH(p) process with equal weights on p lagged residuals.) Here discrete jumps are not the problem, since it is easy to show that such a moving-average scheme is consistent with a continuous sample path for volatility in the limit as $h \to 0$. For example, for some $\Delta > 0$, set $\Omega_t = \exp(W_t - W_{t-\Delta})$.

Though it may not be as obvious, this scheme is also ruled out by our regularity conditions, which not only assumed that the sample paths of the state variables were (asymptotically) continuous, but also that over short time intervals, the *unpredictable component of changes in the state variables swamps the predictable component*[5]–i.e., the noise swamps the signal for sufficiently small h. In the moving average example just given, the noise and the signal are of the same stochastic order as $h \to 0$. Our regularity conditions effectively assume that shocks to the state variables decay either gradually or not at all. This means that over very short time intervals, the movements in ${}_h\Omega_t$ and ${}_hX_t$ look like random walks.

Since our estimates of Ω_t are formed over short intervals, and since X_t and Ω_t behave asymptotically like random walks over such short intervals, it should not be too surprising that our optimal weighting scheme is two-sided exponential: this is the weighting scheme obtained in the literature on random coefficient models under the assumption of a Gaussian random walk (independent of the right-hand side variables) for the regression coefficients–

[5] A continuous time semimartingale is decomposable (by definition) into the sum of a martingale (which may be of unbounded variation, and so very rapidly oscillating) and an instantaneously predictable component of bounded variation (which is much more slowly varying over short time intervals).

see, for example, Fisher and Kamin (1985).

If the regularity conditions asymptotically ruling out discrete jumps in ${}_hX_t$ are relaxed, our results are invalidated: suppose, for example, that ${}_hX_t$ is generated by a jump process, say a poisson, observed at discrete intervals of length h. For each T, the normalized residual $h^{-1/2}[{}_hX_t - {}_hX_{t-h}]$ converges in probability to zero as $h \to 0$, yet its conditional variance does not vanish to zero with h. Clearly a rolling regression using $O(h^{1/2})$ window widths cannot consistently extract this variance, since unless there is a jump within the window (which happens with vanishingly small probability as h goes to zero), the variance estimate produced by the rolling regression is 0! The problem here is that the normalized residuals $h^{-1/2}[{}_hX_t - {}_hX_{t-h}]$ are too thick tailed (i.e., they are nearly always small but are occasionally enormous – i.e. $\theta = \infty$). This prevents us from applying a law of large numbers and a central limit theorem locally in time to extract ${}_h\Omega_t$ from the squared increments in ${}_hX_t$.

We have also assumed that our variance process, Ω, does not have jumps. In this case though, the problem becomes in some sense easier instead of harder. If the variability of Ω is contained in jumps, then "most of the time" Ω is relatively constant. So, long windows can be used for the rolling regression. Unfortunately, the asymptotic variance will still be infinite, but this is now due to a few large errors. In other words, most of the time, we will be getting very accurate estimates, but when a jump occurs, we get asymptotically an infinite error.

4 Estimating conditional betas

In many applications, especially in finance, conditional betas are of greater importance than conditional variances or covariances. Suppose that $\Delta_h X_{1,t}$ is the return on some market index, while $\Delta_h X_{j,t}$ is the return on some other asset or portfolio. The true and estimated conditional betas of asset j with respect to the market index are defined respectively as

$${}_h\beta_{j,t} \equiv {}_h\Omega_{1,j,t}/{}_h\Omega_{1,1,t}, \text{ and } {}_h\hat{\beta}_{j,t} \equiv {}_h\hat{\Omega}_{1,j,t}/{}_h\hat{\Omega}_{1,1,t}.$$

Since the estimated beta is a differentiable function of the asymptotically normal covariance and variance estimates ${}_h\hat{\Omega}_{1,j,t}$ and ${}_h\hat{\Omega}_{1,1,t}$, it is also asymptotically normal (see, e.g., Serfling (1980, Section 3.3, Theorem A), with mean zero and asymptotic variance

$${}_h\hat{\Omega}^{-2}_{1,1,t}[{}_hC_{(1j1j)t} + {}_h\beta^2_{j,t}{}_hC_{(1111)t} - 2{}_h\beta_{j,t}{}_hC_{(111j)t}]. \quad (18)$$

We next consider optimality, assuming, for simplicity, that the same weights are used in forming both ${}_h\hat{\Omega}_{1,j,t}$ and ${}_h\hat{\Omega}_{1,1,t}$. This corresponds to using weighted least squares (regressing $\Delta_h X_{j,\cdot}$on $\Delta_h X_{1,\cdot}$) to estimate ${}_h\beta_{j,t}$. Substituting from (7) into (18) yields

$$\mathrm{AVAR}_t(h^{-1/4}[{}_h\hat{\beta}_{j,t} - {}_h\beta_{j,t}]) = [\theta_\beta S_{ww} + \Lambda_\beta S_{\Psi\Psi} + 2\rho_\beta\sqrt{\theta_\beta\Lambda_\beta}S_{w\Psi}] \tag{19}$$

where

$$\theta_\beta \equiv ({}_h\theta_{(1j1j)t} + {}_h\beta^2_{j,t}{}_h\theta_{(1111)t} - 2{}_h\beta_{j,t}{}_h\theta_{(111j)t})/{}_h\hat{\Omega}^2_{1,2,t}, \tag{20}$$

$$\Lambda_\beta \equiv ({}_h\Lambda_{(1j1j)t} + {}_h\beta^2_{j,t}\,{}_h\Lambda_{(1111)t} - 2\,{}_h\beta_{j,t}\,{}_h\Lambda_{(111j)t})/{}_h\Omega^2_{1,1,t}, \tag{21}$$

and (deleting the h and t subscripts to improve legibility)

$$\rho_\beta \equiv \frac{\rho_{1j1j}\sqrt{\theta_{1j1j}\Lambda_{1j1j}} + \beta^2\rho_{1111}\sqrt{\theta_{1111}\Lambda_{1111}} - \beta\rho_{111j}\sqrt{\theta_{1111}\Lambda_{1j1j}} - \beta\rho_{1j11}\sqrt{\theta_{1j1j}\Lambda_{1111}}}{\Omega^2_{11}\sqrt{\theta_\beta\Lambda_\beta}}$$

As in Section 2, the three terms are easily interpreted: θ_β is the sampling error variance, Λ_β is the instantaneous conditional variance of the increments in ${}_h\beta_{j,t}$. The $2\rho_\beta\sqrt{\theta_\beta\Lambda_\beta}$ term arises from the covariance between the other two terms. Again, this term is zero for diffusion models and many stochastic volatility models. Note that (19) has the same form as (9) if we substitute $\theta_\beta, \Lambda_\beta$, and ρ_β for θ, Λ, and ρ. Apart from these substitutions, the optimality and dominance results of Section 3 are unaffected. In particular, the asymptotically optimal weights are two-sided and exponentially declining, just as derived in the random coefficients literature under the assumption that betas follow random walks independent of returns on the market index.

5 An application: Volatility on the S & P 500

To illustrate the application of our results, we estimate the conditional variance of continuously compounded daily capital gains on the S&P 500. Our data extend from January 1928 through December 1990. Poterba and Summers (1986) and French, Schwert, and Stambaugh (1987) employed the same series (up to 1985) in their work. The series exhibits small but statistically significant serial correlation of about 6% at one lag, presumably caused by thin trading of the stocks in the underlying index–see, e.g., Scholes and Williams (1977). There is little serial correlation at longer lags. Since this serial correlation is not of interest to our application, we pre-whitened the series with an AR(1). Another 'nuisance' aspect of this data is the contribution of non-trading days to variance: i.e., stock volatility is typically higher following weekends and holidays, since the information arriving during the period of market closure must be reflected in asset prices when the market re- opens. (See, e.g., French and Roll (1986).) Nelson (1989) estimated that each non-trading day adds 22.8% to the variance of the S&P 500 on the next trading day. Accordingly, we divide each of the pre-whitened capital gains ξ_t by $(1 + .228 \cdot N_t)^{1/2}$, where N_t is the number of non-trading days preceding trading day t. The transformed series is plotted in figure 3.

******* **Figure 3 near here** *********

As noted earlier, French, Schwert, and Stambaugh (1987) employed a block-constant estimation strategy for the variance. They noted that the resulting $\hat{\Omega}_t$ series is skewed to the right, and that the variance of the innovations in $\hat{\Omega}_t$ is an increasing function of $\hat{\Omega}_t$. French, Schwert, and Stambaugh took the log of $\hat{\Omega}_t$ and found that this transformation adequately stabilized the variance. This is apparent in figure 4,

******* **Figure 4 near here** *********

which plots the log of a simple flat-weight rolling regression with a window length of 25 days on each side. We therefore make the simplifying assumption that $\ln(\Omega_t)$ is conditionally homoskedastic (i.e., $\Lambda_t = \Lambda\Omega_t^2$) . We also make the simplifying assumptions that conditional kurtosis is constant (i.e., $\theta_t = \theta\Omega_t^2$), and that $\rho_t = 0$, i.e., stochastic volatility or diffusion rather than GARCH as the data generating process. These assumptions allow us to set $K_h = \infty$ in Theorem 3. We then formed initial conditional variance estimates using two-sided flat-weight rolling regressions. From these initial variance estimates, we created estimates of θ and Λ using the method of Theorem 3. These estimates in turn implied optimal n and m values ($n = m$) for two-sided rolling regressions through formulas (13) and (14). We then iterated this procedure, at each stage using the "optimal" n and m suggested at the previous step until the procedure converged. (This occurred very rapidly, since for $m + n$ values below 52 a higher value was suggested, while for $n + m$ above 54 a lower value was suggested. We settled on a window length of 52.) The estimated θ and Λ values were 2.72 and .0120, respectively, implying through Theorem 5 an optimal exponential decay rate of $\alpha = .0665$ for a two-sided exponentially weighted rolling regression.[6]

To gauge the reliability of our asymptotic approximations, we performed 600 replications of the following experiment calibrated to the S&P 500 data: First, we generated 16885

[6]To gauge the importance of our pre-whitening and non-trading days adjustment, we repeated the estimation procedure using the raw (i.e., unadjusted) capital gains data. The results changed very little: the estimated θ and Λ were respectively, 2.668 and .0124, and the optimal $m + n$ and exponential decay rate were 51 and .068 respectively.

observations of $\ln(\Omega_t)$ and ΔM_t as

$$\ln(\Omega_t) = -.4246 + .9944 \cdot [\ln(\Omega_{t-1}) + .4246] + z_{2,t} \tag{22}$$

$$\Delta M_t = \Omega_t^{1/2} \cdot z_{1,t} \tag{23}$$

where $z_{1,t}$, and $z_{2,t}$ are mutually independent and i.i.d., with $z_{1,t}$ distributed as a Student's t with 12 degrees of freedom, mean 0 and variance 1 and $z_{2,t}$ is $N(0, .0120)$. The degrees of freedom of the Student's t distribution was selected to match the estimated conditional kurtosis from the S&P 500 data. The variance of $z_{2,t}$ was selected to match the estimate of Λ for the S&P data. The population mean of $\ln(\Omega_t)$, which was -.4246, matched the sample mean of the fitted $\ln(\hat{\Omega}_t)$. The slow mean reversion (.9944) was selected to match the unconditional variance of $\ln(\Omega_t)$ to the sample variance of the fitted $\ln(\hat{\Omega}_t)$ plus the variance of $(\ln \Omega_\tau - \ln \hat{\Omega}_\tau)$.

For each replication, we repeated precisely the same estimation procedure we had applied to the S&P data. Tables 1 & 2 below report means and standard deviations of the estimated parameters in the simulations. Standard errors (i.e., sample standard deviations divided by the square root of the number of simulations) are given in parenthesis.

	Mean of Estimated Coefficient	Actual Coefficient	Sample Standard Deviation
Λ	0.01051 (0.00005)	0.0120	.0012
θ	2.480 (0.003)	2.75	.082

Table 1: Using equation (12)

The estimates for both Λ and θ are downward biased (by 1.2 and 3.3 standard deviations respectively in table 1 and 5.1 and 2.5 standard deviations in table 2). The width of the asymptotic confidence intervals, the optimal $m + n$ etc., are functions of $\sqrt{\Lambda/\theta}$. The bias in this ratio is quite small–for example the optimal $m + n$ for two-sided rolling regressions is given by (13) and (14) as $(12 \cdot \theta/\Lambda)^{1/2} = 52.4$ for the simulation. The mean estimated optimal $m + n$ was 53 with a standard deviation of 3.6 (using equation 12 it was 64 ± 4.3). Our estimates of $(\Lambda/\theta)^{1/2}$ were close despite the biases in both Λ and θ. Since Λ and θ are biased in the same directions, the biases partially offset in $(\Lambda/\theta)^{1/2}$. It is also worth noting

	Mean of Estimated Coefficient	Actual Coefficient	Sample Standard Deviation
Λ	0.007 (0.001)	0.0120	.00088
θ	2.527 (0.004)	2.75	.088

Table 2: Using equation (11')

that the asymptotic standard deviation of the measurement error achieved by the optimal flat-weight or exponentially weighted rolling regressions is proportional to $(\Lambda\theta)^{1/4}$. This means that measurement errors in Λ and θ must be quite large to have much effect on the accuracy of the confidence intervals. For example, getting θ wrong by a factor of 2 throws off the confidence intervals by only about 19%. Tables 3 and 4 compare the asymptotic versus actual coverages in the measurement error, giving the proportion of measurement errors falling between $\pm 1, \pm 2$, and ± 3 estimated asymptotic standard deviations, along with the standard errors. The asymptotic confidence bands are slightly too narrow, but not drastically so.

Standard Deviations	Mean Coverage in Simulation	Asymptotic Coverage
1	0.6393 (0.0007)	0.6827
2	0.9306 (0.0004)	0.9545
3	0.9929 (0.0001)	0.9973

Table 3: Using equation (12)

******* **Figure 5 near here** *********

Figure 5 plots 95% confidence bands. We used the delta method to transform our asymptotic distribution for $h^{-1/4}(\hat{\Omega} - \Omega)$ into an asymptotic distribution for $h^{-1/4}(\ln \hat{\Omega} - \ln \Omega)$. This, combined with our assumption that $\theta_t = \theta \cdot \Omega_t^2$ and $\Lambda_t = \Lambda \Omega_t^2$ implies that the width of the confidence bounds in a log plot is constant, so the extension from figure 5 to

Standard Deviations	Mean Coverage in Simulation	Asymptotic Coverage
1	0.5915 (0.0008)	0.6827
2	0.8991 (0.0006)	0.9545
3	0.9848 (0.0002)	0.9973

Table 4: Using equation (11')

confidence bounds for the whole sample is immediate.

******* **Figure 6 near here** *********

Figure 6 is analogous to figure 5, except that it uses simulated data, and plots the true (simulated) $\Omega_t^{1/2}$ along with the ± 2 standard deviation confidence bounds. Overall, the asymptotic approximation performs tolerably well in the simulations using equation (11') and extremely well using (12).

6 Conclusion

While this paper has, we believe, shed new light on rolling regressions as conditional variance and covariance estimators, much work remains. For example, in tests of asset pricing theories the link between conditional means and conditional covariance matrices is usually crucial. As we have seen, conditional covariances can be accurately measured using high frequency data (i.e., taking h to zero). Unfortunately, estimating conditional means requires a long *span* of data as opposed to a high observation frequency, see e.g., Merton (1980). Since the asymptotic results developed in this paper are *pointwise* in time, they do not adequately equip us to study the joint evolution of conditional means and covariances *over* time.

A second limitation is our consideration only of unconstrained linear regression to compute the estimated conditional covariance matrix. Constraints on the conditional covariance matrix (e.g., on the eigenvalues or eigenvectors) are likely to prove important in dynamic factor analysis or principle components.

Finally, as we have seen, conditionally thick-tailed processes reduce the efficiency of least squares based procedures such as rolling regressions. It should be possible to adapt the methods for robust estimation of covariance matrices developed for the i.i.d. case (see, e.g., Huber (1981)) to the rolling regression framework.[7] Extending our results in these directions may prove quite challenging, but should be worth the effort.

The Wharton School, University of Pennsylvania, Philidelphia PA 19104 Phone: 215 898 8233.

University of Chicago Graduate School of Business and N.B.E.R., Chicago IL 60637. Phone: 312 702 3231.

APPENDIX

We will drop the prefix "h" from our stochastic processes to conserve space in our proofs. Lemma's, theorem's etc., will include the "h"'s. All processes depend on h.

PROOF OF THEOREM 1: We will first divide the problem into two pieces.

Definition ${}_h\overline{\Omega}_{(ij)T} = \sum_\tau {}_h\Omega_{(ij)\tau h}w_{\tau-T}\Delta\tau$.

Lemma 1 *If Assumptions A & D hold, then*

$$h^{-1/4}(\hat{\Omega}_{(ij)T} - \overline{\Omega}_{(ij)T}) = h^{1/4}\sum_\tau w_{\tau-T}\Delta B_{(ij)\tau} + o_p(1)$$

$$h^{-1/4}(\overline{\Omega}_{(ij)T} - \Omega_{(ij)T}) = h^{-1/4}\sum_\tau \Psi_{\tau-T}\Delta M^*_{(ij)\tau} + o_p(1)$$

From lemma A.1, it is obvious that theorem 1 holds. The proof of lemma A.1 relies on some other lemmas which we will prove first.

[7] Robust conditional variance estimation methods have been employed in the ARCH literature. For example, Taylor (1986) and Schwert (1989) estimate the conditional standard deviation as a distributed lag of absolute residuals (rather than estimating the conditional variance as a distributed lag of squared residuals). Schwert was explicitly motivated by the robust variance estimation methods of Davidian and Carroll (1987). For a formal analysis of the robustness properties of these models see Nelson and Foster (1992).

Lemma 2

$$\hat{\Omega}_{(ij)T} \equiv \overline{\Omega}_{(ij)T} + \sqrt{h}\sum_{\tau} w_{\tau-T}\Delta Q_{(ij)\tau} + h(\mathcal{B}_{ij} + \mathcal{B}_{ji} + \mathcal{D})$$

where

$$\mathcal{B}_{ji} \equiv \sum_{\tau}(\mu_{(j)\tau} - \hat{\mu}_{(j)\tau})w_{\tau-T}\Delta M_{(i)\tau}$$

and

$$\mathcal{D} \equiv \sum_{\tau}(\mu_{(j)\tau} - \hat{\mu}_{(j)\tau})(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau})w_{\tau-T}\Delta\tau$$

PROOF OF LEMMA A.2: First note that

$$\Delta X_{(j)\tau} - h\hat{\mu}_{(j)\tau} = \Delta M_{(j)\tau} + h\mu_{(j)} - h\hat{\mu}_{(j)\tau}$$

$$= \Delta M_{(j)\tau} + h(\mu_{(j)\tau} - \hat{\mu}_{(j)\tau})$$

So

$$\begin{aligned} [\Delta X_{(i)\tau} - h\hat{\mu}_{(i)\tau}][\Delta X_{(j)\tau} - h\hat{\mu}_{(j)\tau}] &= (\Delta M_{(i)\tau} + h(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau}))\times \\ &\quad \left(\Delta M_{(j)\tau} + h(\mu_{(j)\tau}\hat{\mu}_{(j)\tau})\right) \\ &= \Delta M_{(i)\tau}\Delta M_{(j)\tau} + h(\mu_{(j)\tau} - \hat{\mu}_{(j)\tau})\Delta M_{(i)\tau} + \\ &\quad + h(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau})\Delta M_{(j)\tau} + \\ &\quad + h^2(\mu_{(j)\tau} - \hat{\mu}_{(j)\tau})(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau}) \end{aligned}$$

Define $\mathcal{A} = \sum_{\tau} w_{\tau-T}\Delta M_{(i)\tau}\Delta M_{(j)\tau}$ Thus, $\hat{\Omega}_{(ij)T} \equiv \mathcal{A} + h(\mathcal{B}_{ij} + \mathcal{B}_{ji} + \mathcal{D})$.

Now analyzing $\mathcal{A}$:

$$\begin{aligned} \mathcal{A} &= \sum_{\tau} w_{\tau-T}\Delta M_{(i)}\Delta M_{(j)\tau} \\ &= \sum_{\tau}\Omega_{(ij)\tau}w_{\tau-T}\Delta\tau + \sum_{\tau} w_{\tau-T}(\Delta M_{(i)\tau}\Delta M_{(j)\tau} - \Omega_{(ij)\tau}\Delta\tau) \\ &= \overline{\Omega}_{(ij)T} + \sqrt{h}\sum_{\tau} w_{\tau-T}\Delta Q_{(ij)\tau} \end{aligned}$$

□

Lemma 3

$$\overline{\Omega}_{(ij)T} - \Omega_{(ij)T}\sum w_{\tau-T}\Delta\tau = \sum_{s=0}^{\infty}\Psi_{s-T}\Delta M^{*}_{(ij)s} + \mathcal{E} + \mathcal{F}$$

where $\mathcal{E} \equiv \lambda(T)\sum_{s=0}^{\infty}\Psi_{s-T}\Delta s$ *and* $\mathcal{F} \equiv \sum_{s=0}^{\infty}\Psi_{s-T}(\lambda(s) - \lambda(T))\Delta s.$

PROOF:

$$\overline{\Omega}_{(ij)T} - \Omega_{(ij)T}\sum_{s=0}^{\infty} w_{\tau-T}\Delta\tau = \sum_{\tau}(\Omega_{(ij)\tau} - \Omega_{(ij)T})w_{\tau-T}\Delta\tau$$

$$\begin{aligned}
&= (\sum_{\tau>T}\sum_{s=T}^{\tau-h}\Delta\Omega_{(ij)s}w_{\tau-T}\Delta\tau - \sum_{\tau<T}\sum_{s=\tau}^{T-h}\Delta\Omega_{(ij)s}w_{\tau-T}\Delta\tau) \\
&= (\sum_{T\leq s<\tau} w_{\tau-T}\Delta\Omega_{(ij)s}\Delta\tau - \sum_{\tau\leq s<T} w_{\tau-T}\Delta\Omega_{(ij)s}\Delta\tau) \\
&= (\sum_{s=T}^{\infty}\sum_{\tau=s+h}^{\infty} w_{\tau-T}\Delta\tau\Delta\Omega_{(ij)s} - \sum_{s=0}^{T-h}\sum_{\tau=0}^{s} w_{\tau-T}\Delta\tau\Delta\Omega_{(ij)s}) \\
&= \sum_{s=0}^{\infty}(I_{s\geq T}\sum_{\tau=s+h}^{\infty} w_{\tau-T}\Delta\tau - I_{s<T}\sum_{\tau=0}^{s} w_{\tau-T}\Delta\tau)\Delta\Omega_{(ij)s} \\
&= \sum_{s=0}^{\infty}\Psi_{s-T}\Delta\Omega_{(ij)s}
\end{aligned}$$

Now use the Doob-Meyer decomposition of $\Delta\Omega$, and we get

$$\overline{\Omega}_{(ij)T} - \Omega_{(ij)T}\sum w_{\tau-T} = \sum_{s=0}^{\infty}\Psi_{s-T}(\lambda(s)\Delta s + \Delta M^{*}_{(ij)s})$$

$$= \lambda(T)\sum_{s=0}^{\infty}\Psi_{s-T}\Delta s + \sum_{s=0}^{\infty}\Psi_{s-T}(\lambda(s) - \lambda(T))\Delta s + \sum_{s=0}^{\infty}\Psi_{s-T}\Delta M^{*}_{(ij)s}$$

□

Lemma 4 *Under Assumptions A & D the following hold*

(A.1)	$\mathcal{B}_{ij} = o_p(h^{-3/4})$	*(A.3)*	$\mathcal{E} = O_p(h^{1/2})$
(A.2)	$\mathcal{D} = O_p(1)$	*(A.4)*	$\mathcal{F} = O_p(h^{1/2})$

PROOF OF (A.1): Because $\mu, \hat{\mu}$, and w are all predictable, and ΔM is a martingale difference array,

$$E(\mathcal{B}_{ij}) = 0$$

and

$$E(\mathcal{B}_{ij}^2) = E(\sum_{\tau}(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau})^2 w_{\tau-T}^2\Omega_{(jj)\tau}\Delta\tau).$$

But, by part i of Assumption (A) we know that

$$\sup_{T_*\leq\tau\leq T^*}(\mu_{(i)\tau} - \hat{\mu}_{(i)\tau})^2 = O_p(1)$$

By Assumption D, we know $\sup(w^2_{\tau-T}) = O_p(h^{-1})$. By parts iv and vii of Assumption A, we know that $\sup(\Omega_{(jj)\tau}) = O_p(1)$. By D, we know that there are $O(h^{-1/2})$ terms in our sum. And by definition, $\Delta\tau =$ h. Thus, $E(\mathcal{B}^2_{ij}) = O_p(h^{-1/2})$. So, by Jensen's inequality

$$P(\mathcal{B}_{ij} > Mh^{-3/4}) < O_p(h^{-1/2})/M^2h^{-3/2}$$

$$= O_p(h^1) = o_p(1)$$

PROOF OF (A.2): Using part i of Assumption A and Assumption D we see that $\mathcal{D} = O_p(1)$.
PROOF OF (A.3): Using part iii of Assumption A, we see that $\lambda_T = O_p(1)$. By Assumption D and the definition of Ψ we can therefore conclude that $\mathcal{E} = O_p(h^{1/2})$.
PROOF OF (A.4): Using part ii of Assumption A, the definition of ψ, and Assumption D, we see that $\mathcal{F} = O_p(h^{1/2})$. □
PROOF OF LEMMA A.1: Follows by substituting lemma A.4 into lemmas A.2 and A.3. □

Thus, we have now completed the proof of theorem 1.
PROOF OF THEOREM 2: By Theorem 1, we need only analyze

$$h^{1/4}\sum_\tau w_{\tau-T}\Delta B_{(ij)\tau} + h^{-1/4}\sum_\tau \Psi_{\tau-T}\Delta M^*_{(ij)\tau}.$$

But since B and M^* are martingales, we know its mean to be zero and its covariance between terms ij and kl to be:

$$\begin{aligned}
&h^{1/2}\sum_\tau w_{(ij)\tau-T}w_{(kl)\tau-T}\theta_{(ijkl)\tau}\Delta\tau + h^{-1/2}\sum_\tau \psi_{(ij)\tau-T}\psi_{(kl)\tau-T}\Lambda_{(ijkl)\tau}\Delta\tau \\
&+\sum_\tau w_{(ij)\tau-T}\psi_{(kl)\tau-T}\sqrt{\theta_{(ijij)\tau}\Lambda_{(klkl)\tau}}\rho_{(ijkl)\tau}\Delta\tau \\
&+\sum_\tau w_{(kl)\tau-T}\psi_{(ij)\tau-T}\sqrt{\theta_{(klkl)\tau}\Lambda_{(ijij)\tau}}\rho_{(klij)\tau}\Delta\tau
\end{aligned}$$

which by Assumptions (A.v), (A.vi) and B is asymptotically equal to $C_{(ijkl)T}$. Now applying the standard martingale central limit (which uses Assumptions C and A.vii and A.viii) see, e.g., Liptser and Shiryayev (1980), we get the desired result. □
PROOF OF THEOREM 3: Before we begin we have to mention a detail about what we are going to prove. We will prove that trimmed-mean versions of (10) and (11) will work have the desired properties. Thus, we will replace the sum

$$\sum_\tau f_\epsilon(\tau)^2$$

by a trimmed version, namely

$$\sum_{\tau} \min(f_\epsilon(\tau)^2, M).$$

(In the multivariate case, each element of the matrix $f_\epsilon(\tau)f_\epsilon(\tau)'$ should be trimmed by the constant M.)

First we need to represent $f_\epsilon(\tau)$ as

$$f_\epsilon(\tau) = \sum_{s=\tau}^{\tau+\epsilon h^{1/2}} h^{-1/2}(\Delta X_s - \hat{\mu}_s \Delta s)^2/\epsilon - \sum_{s=\tau-\epsilon h^{1/2}}^{\tau} h^{-1/2}(\Delta X_s - \hat{\mu}_s \Delta s)^2/\epsilon$$

$$= \sum_{s} h^{-1/2} w_\epsilon((s-\tau)h^{1/2})(\Delta X_s - \hat{\mu}_s \Delta s)^2$$

where $w_\epsilon(x)$ is defined as $h^{-1/2}\frac{1}{\epsilon}\text{sgn}(x)I_{[-\epsilon,\epsilon]}(x)$, where sgn(x) is the sign of x. I.e. $\text{sgn}(x) = 1$ if $x > 0$, and $\text{sgn}(x) = -1$ if $x < 0$. Thus, we have written f_ϵ in the form of equation (5). If $\sum_s w_\epsilon(s)h = 1$, then Assumption D would hold and we could apply lemmas A.2-A.4. But, looking at the proofs of A.2-A.4 we see that this fact isn't used. Thus, from lemmas A.2-A.4 we have an asymptotic representation for $f_\epsilon(\tau)$ in terms of martingales. Using the same CLT as before, we can find the asymptotic distribution for $f_\epsilon(\tau)$. In particular $f_\epsilon(\tau)$ converges to a normal with mean zero and variance of $2\theta_\tau/\epsilon + 2\epsilon\Lambda_\tau/3$. Thus, asymptotically

$$\lim_{M\to\infty} \lim_{h\to 0} E(\min(f_\epsilon(\tau)^2, M) \to 2\theta_\tau/\epsilon + 2\epsilon\Lambda_\tau/3.$$

Now applying the law of large numbers

$$\lim_{M\to\infty} \lim_{h\to 0} 1/K \sum_{\tau} \min(f_\epsilon(\tau)^2, M)/\Omega(\tau) \to 2\theta_\tau/\epsilon + 2\epsilon\Lambda_\tau/3.$$

Substituting this into equations (10) and (11) we get the desired result.

Note: This proof also works for the multivariate problem.

Note: The importance of the truncation is that convergence in distribution will imply convergence in mean only for bounded random variables. So, we must make the $f_\epsilon(\tau)^2$ bounded to use the law of large numbers. □

PROOF OF THEOREM 4: This theorem consists of three different optimizations of equation (9). Part (a) forces m_0 to be zero, part (b) forces n_0 to be zero, and part (c) only constrains n_0 and m_0 to be non- negative. Parts (a) and (b) follow from taking derivatives and setting equal to zero. By the form of equation (9), it is obvious that there is a unique minimum. Part (c) is solved by using partial derivatives. The side constraints of non-negativity for n_0 and m_0 come into play for extreme values of ρ. Thus, we get the three part solution. □

Lemmas A.5 through A.8 set up theorem 5.

Lemma 5 *(Some calculations for exponential weights) let $c_i = \beta^i(1-\beta)$, for $i = 0, 1, 2, \ldots$. Define*

$$C_i \equiv \sum_{j=i}^{\infty} c_j = \beta^i.$$

Then,

$$\sum_{i=0}^{\infty} c_i^2 = (1-\beta)^2/(1-\beta^2)$$

and

$$\sum_{i=0}^{\infty} C_i^2 = 1/(1-\beta^2)$$

. The minimum of

$$\sum_{i=0}^{\infty} c_i^2 + A \sum_{i=0}^{\infty} C_i^2 \tag{24}$$

occurs at $\beta = 1 - \sqrt{A} + o(\sqrt{A})$, and the minimum value obtained is $\sqrt{A} + o(\sqrt{A})$.

PROOF: Note that formula (24) is equivalent to

$$(1-\beta)^2/(1-\beta^2) + A/(1-\beta^2) \tag{25}$$

The following algebra minimizes (25) to generate our result. ($\epsilon = 1 - \beta$)

$$\min_{\epsilon} (\epsilon^2 + A)/(1 - (1-\epsilon)^2)$$

$$\min_{\epsilon} (\epsilon + A/\epsilon)/(2 - \epsilon)$$

for which the minimum occurs at $\epsilon^2 = A(1-\epsilon)$, which is $\epsilon = \sqrt{A} + o(\sqrt{A})$, and the value of (25) at this point is $\sqrt{A} + o(\sqrt{A})$. □

Lemma 6 *(Discrete approximately equals continuous), Any c_i's which sum to one have the property that the value of equation (24) is at least $\sqrt{A}$. In particular, let $\mathcal{D}^+ = \{f(\cdot) | \int_0^\infty f(t)dt = 1\}$, then*

$$\begin{aligned} \sum_{i=1}^{n} c_i^2 + A \sum_{i=1}^{n} C_i^2 &\geq min_{f \in \mathcal{D}} \int_0^\infty f(t)^2 dt + A \int_0^\infty (\int_0^\infty f(s) ds)^2 dt \\ &\leq \sqrt{A}. \end{aligned} \tag{26}$$

PROOF: Taking

$$f(t) = w_i \text{ for } i \le t < i + 1$$

which is in $\mathcal{D}$, and its value is exactly the left hand side of (26). This proves the inequality part. Write

$$f(t) = \alpha e^{-\alpha t} + \eta(t),$$

with $\alpha = \sqrt{A}$. Then

$$\int_0^\infty \eta(t)dt = 0 \tag{27}$$

Because $\int f(t)dt = 1$, and $\int \alpha e^{-\alpha t} dt = 1$. The following follows by an interchange of integrals and the definition of $\eta(\cdot)$:

$$\int_0^\infty f(t)^2 dt = \alpha/2 + \int_0^\infty \eta(t)\alpha e^{-\alpha t} dt + \int_0^\infty \eta(t)^2 dt. \tag{28}$$

Obviously,

$$\int_t^\infty f(s)ds = e^{-\alpha t} + \int_t^\infty \eta(s)ds$$

Some more calculus yields:

$$\int_0^\infty \left(\int_t^\infty f(s)ds\right)^2 dt = \int_0^\infty \eta(t)dt - \int_0^\infty \eta(t)e^{-\alpha t}dt + \int_0^\infty \left(\int_t^\infty \eta(s)ds\right)^2 dt \tag{29}$$

Substituting (27) into (29) yields:

$$\int_0^\infty \left(\int_t^\infty f(s)ds\right)^2 dt = -\int_0^\infty \eta(t)e^{-\alpha t}dt + \int_0^\infty \left(\int_0^\infty \eta(s)ds\right)^2 dt \tag{30}$$

Our desired result is now A times equation (30) plus equation (28). Putting these together yields (recall $\alpha = \sqrt{A}$):

$$\text{goal } = \alpha + \int_0^\infty \eta(t)^2 dt + \int_0^\infty (\int_0^\infty \eta(s)ds)^2 dt \ge \alpha,$$

with equality holding if $f(\cdot) = \alpha e^{-\alpha t}$.

□

Lemma 7 *If $\sum_{i=0}^\infty c_i = p$, then $c_i = p\beta^i(1-\beta)$ is asymptotically (as $A \to 0$) the minimizer of equation (24) with an asymptotic value of $p^2\sqrt{A}$.*

PROOF: Lemma A.5 shows the value of (24) for these $c_i's$, and lemma A.6 shows they can't be improved upon. □

Lemma 8 *Restrict w_s such that $\int_0^\infty {}_0w_s ds = p$. Then the optimum w_s is*

$$ {}_0w_s = \begin{cases} p\alpha e^{-\alpha s} & \textit{for } s \geq 0, \\ (1-p)\alpha e^{-\alpha s} & \textit{for } s < 0, \end{cases} $$

(where $\alpha = \sqrt{\Lambda/\theta}$), which yields an asymptotic variance of

$$ \sqrt{\Lambda\theta}((2p-1)^2/2 + 1/2 - (2p-1)\rho. $$

PROOF: We will break the problem into two pieces, the positive part ($s \geq 0$) and the negative part ($s < 0$). Each will be separately minimized for each value of $p = \int_0^\infty {}_0w_s ds$. First consider

$$ \int_0^\infty w_s \Psi_s ds = \int_0^\infty w_s \int_s^\infty w_t \mathrm{dtds} $$
$$ = (1/2)(\int_0^\infty w_s \int_s^\infty w_t \mathrm{dtds} \ + \int_0^\infty w_s \int_s^\infty w_s \mathrm{dsdt}) $$
$$ = (1/2) \int_0^\infty \int_0^\infty w_s w_t \mathrm{dtds} \ = p^2/2 $$

Therefore for fixed p, minimizing the w's is the same as minimizing equation (26) above with $A = \Lambda/\theta$. Thus, the parameter of the exponential function is identical regardless of p and regardless of which side of zero we are on. So, $\alpha = \sqrt{A}$, is optimal. □

PROOF OF THEOREM 5: Lemmas A.5 through A.8 prove everything except picking the value of p. For parts (A) and (B), the value of p is determined so we are done. For part (C) we need to minimize the variance with respects to p. The variance is

$$ \sqrt{\Lambda\theta}((2p-1)^2/2 + 1/2 - (2p-1)\rho) $$

which is minimized at $(2p-1) = \rho$. Thus, the minimum occurs at $p = (1-\rho)/2$, so the optimum variance is

$$ = (1/2)\sqrt{\Lambda\theta}(1-\rho^2) $$

□

PROOF OF THEOREM 6: Equation 3.7 follows from 9 by substitution. 3.7 is obviously positive which proves our result. □

REFERENCES

BANERGEE, A., R. L. LUMSDAINE, and J. H. STOCK (1992): "Recursive and Sequential Tests of the Unit Root and Trend Break Hypothesis," *Journal of Business and Economic Statistics*, **10**, 271-288.

BANON, G. (1978): "Nonparametric Identification for Diffusion Processes," *SIAM Journal Control and Optimization*, **16**, 380-395.

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

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 J. M. WOOLDRIDGE (1988): "A Capital Asset Pricing Model with Time Varying Covariances, *Journal of Political Economy*, **96**, 116-131.

CHOW, G. C. (1984): "Random and Changing Coefficient Models," in Z. Griliches and M. D. Intriligator, eds., *The Handbook of Econometrics, Volume 2.* Amsterdam: North-Holland.

CORRADI, V. and H. WHITE (1994): "Consistent Nonparametric Estimation and Testing for the Variance of a Diffusion from Discretely Sampled Observations," UCSD.

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

DOHNAL, G. (1987): "On Estimating the Diffusion Coefficient," *Journal Applied Probability*, **24**, 105-114.

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

ENGLE, R.F. and T. BOLLERSLEV (1986): "Modelling the Persistence of Conditional Variances," *Econometric Reviews*, **5**, 1-50.

FAMA, E. F. and J. D. MACBETH (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, **81**, 607- 636.

FLORENS-ZMIROU, D. (1993): "On Estimating the Diffusion Coefficient from Discrete Observations," *Journal Applied Probability*, **30**, 790-804.

FISHER, L. (1970): "The Estimation of Systematic Risk: Some New Findings." Proceedings of the Seminar on the Analysis of Security Prices, Univeristy of Chicago.

FISHER, L. and J. H. KAMIN (1985): "Forecasting Systematic Risk: Estimates of 'Raw' Beta that Take Account of the Tendency of Beta to Change and the Heteroskedasticity of Residual Returns," *Journal of Financial and Quantitative Analysis*, 20, 127-149.

FRENCH, K. R., and R. ROLL (1986): "Stock Return Variances: The Arrival of Information and the Reaction of Traders," *Journal of Financial Economics*, **17**, 5-26.

FRENCH, K. R., G. W. SCHWERT, and R. F. STAMBAUGH (1987): "Expected Stock Returns and Volatility," *Journal of Financial Economics*, **19**, 3-29.

GENO-CATALOT, V., C. LAREDO, and D. PICARD (1992) "Non-parametric Estimation of the Diffusion Coefficient by Wavelets Methods," *Scandinavian Journal Statistics*, **19**, 317 - 335.

GONEDES, N. (1973): "Evidence on the Information Content of Accounting Numbers: Accounting-Based and Market Based Estimates of Systematic Risk." *Journal of Financial and Quantitative Analysis*, 8, 407-444.

HARVEY, A. C. (1989): *Forecasting, Structural Time Series Models, and the Kalman Filter.* Cambridge, UK: Cambridge University Press.

HUBER, P. J. (1981): *Robust Statistics.* New York: Wiley.

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

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

LIPTSER, R. S. and A. N. SHIRYAYEV (1980): "A Functional Central Limit Theorem for Semimartingales," *Theory of Probability and Its Applications*, **25**, 667-688.

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

MERRILL LYNCH, PIERCE, FENNER, AND SMITH, INC. (1986): *Security Risk Evaluation.*

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

NELSON, D. B. (1989): "Commentary: Price Volatility, International Market Links, and Their Implications for Regulatory Policies," *Journal of Financial Services Research* **3**, 247-254.

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

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. (1994): "Asymptotic Filtering and Smoothing Theory for Multivariate ARCH Models," University of Chicago.

NELSON, D. B., and D. P. FOSTER (1991): "Filtering and Forecasting with Misspecified ARCH Models II: Making the Right Forecast with the Wrong Model," to appear in the *Journal of Econometrics.*

NELSON, D. B., and D. P. FOSTER (1994): "Asymptotic Filtering Theory for Univariate ARCH Models," *Econometrica*, **62**, 1-41.

OFFICER, R. R. (1973): "The Variability of the Market Factor of the New York Stock Exchange," *Journal of Business*, **46**, 434-453.

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

PARKINSON, M. (1980): "The Extreme Value Method for Estimating the Variance of the Rate of Return," *Journal of Business*, **53**, 61-65.

POTERBA, J. M. and L.H. SUMMERS (1986): "The Persistence of Volatility and Stock Market Fluctuations," *American Economic Review*, **76**, 1142-1151.

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

SCHOLES, M. and J. WILLIAMS (1977): "Estimating Betas from Nonsynchronous Data," *Journal of Financial Economics*, 5, 309-327.

SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

SHEPHARD, N. (1994): "Local Scale Models: State Space Alternative to Integrated GARCH Processes," *Journal of Econometrics*, **60**, 181-202.

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

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

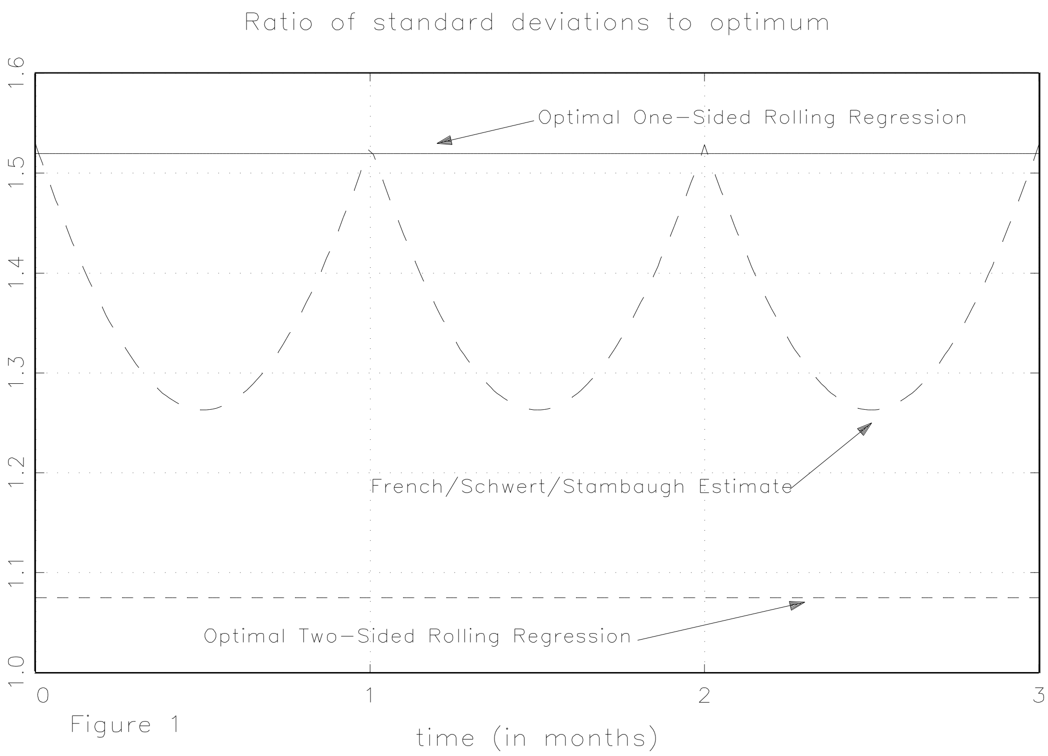

Figure 1: **Dominance relations.** This graph shows that the accuracy of various estimates of the variance vs. time. The optimal two sided exponentially weighted estimate is normalized to have a standard error of one. Relative to the accuracy of this estimator, the optimal two-sided flat weight has a standard error of 1.07. The French/Schwert/Stambaugh assumes a fixed variance over the month so its performance changes over the month. The optimal one-sided flat weight estimator uses only historical data and so does worse than the other estimators which use both the past and the future to estimate Ω_t.

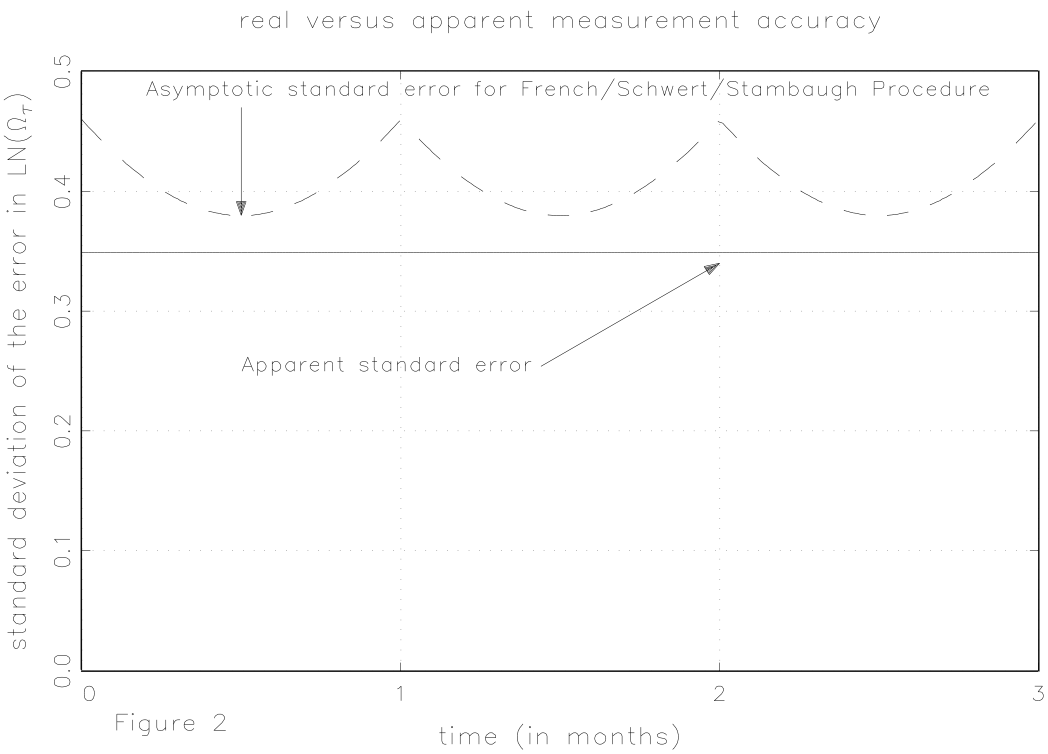

Figure 2: **Real vs apparent measurement accuracy.**

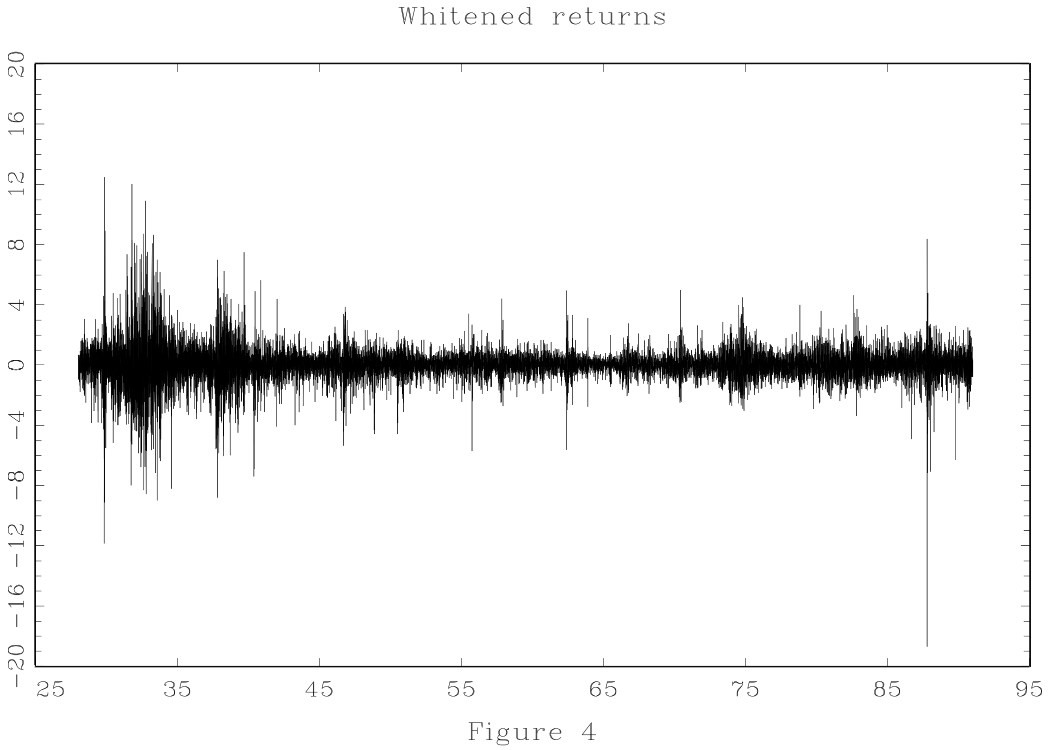

Figure 3: **Whitened returns of the S&P 500.**

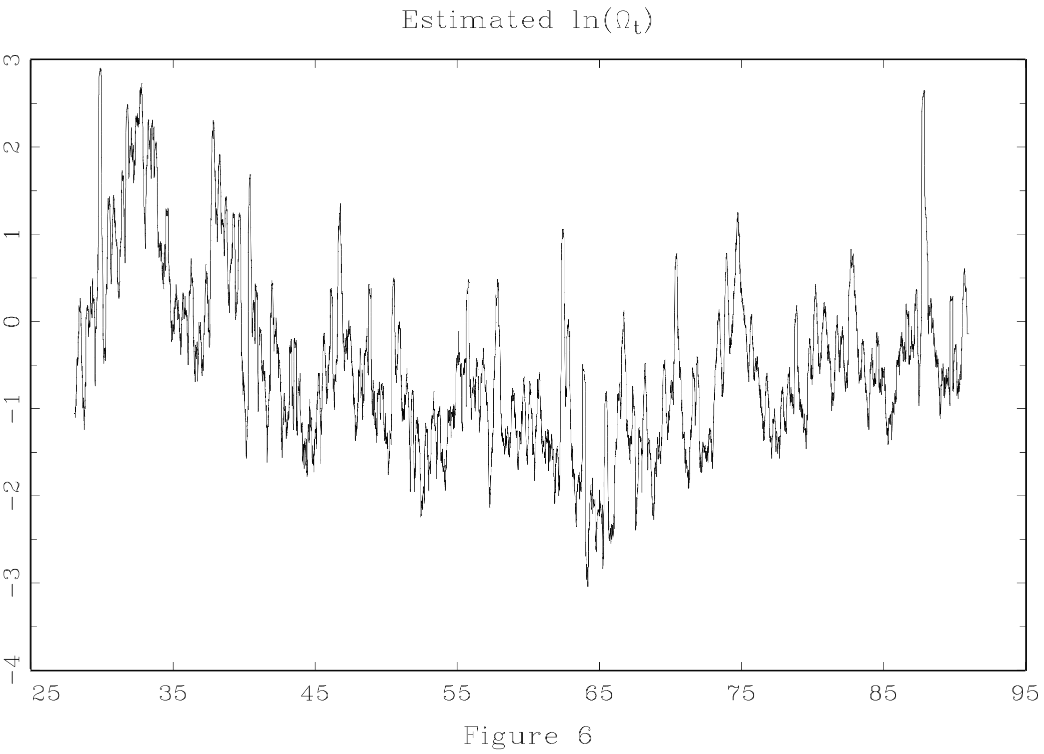

Figure 4: **Estimated log of the variance of the S&P 500:** A 25 day two-sided flat-weight rolling regression was used to estimate the variance of the S&P 500. (50 days in total.) There are 240 non-overlapping estimates of the variance (60 years $\times$ 200 trading days per year / 50 days per estimate). The graph shows that assuming conditional homoskedasticity is reasonable.

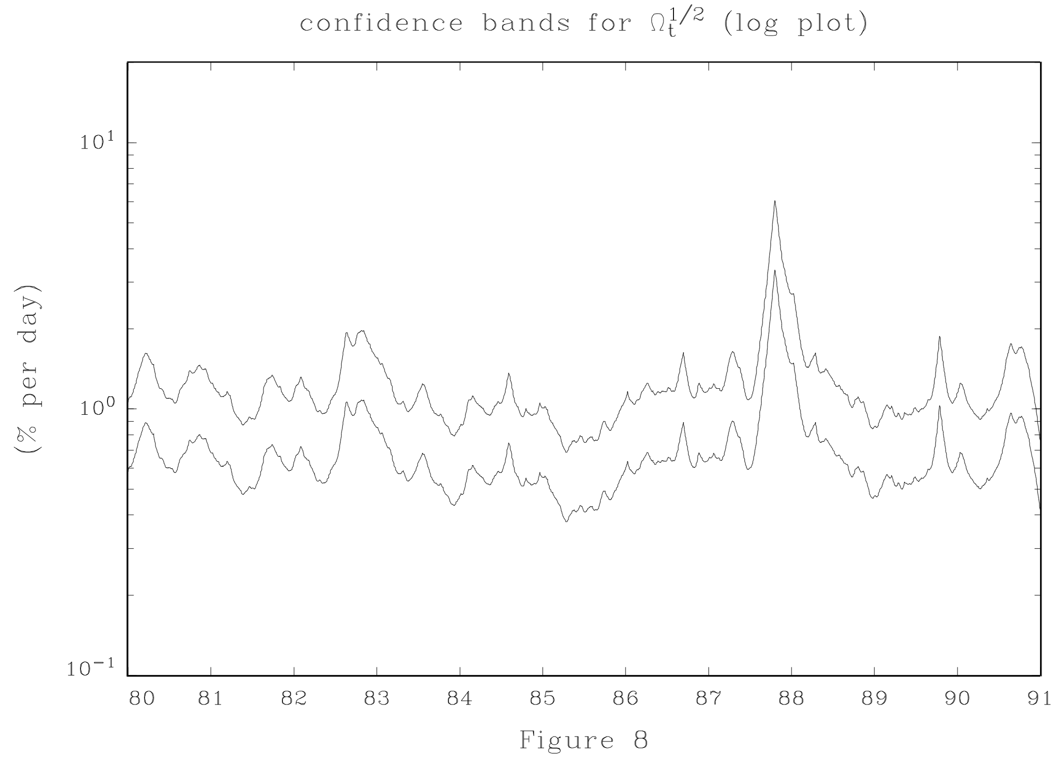

Figure 5: **95% confidence bands:** For each fixed time figure 4 showed the estimate of Ω. This graph shows a 95% confidence interval around that estimate. ($\widehat{\Omega}_t \pm 2$ S. E.) The 95% holds pointwise, not uniformly over the entire interval so we would expect that 5% of the time that the confidence interval does not cover the truth. Notice that this is a log scale plot of the $\Omega_t^{1/2}$.

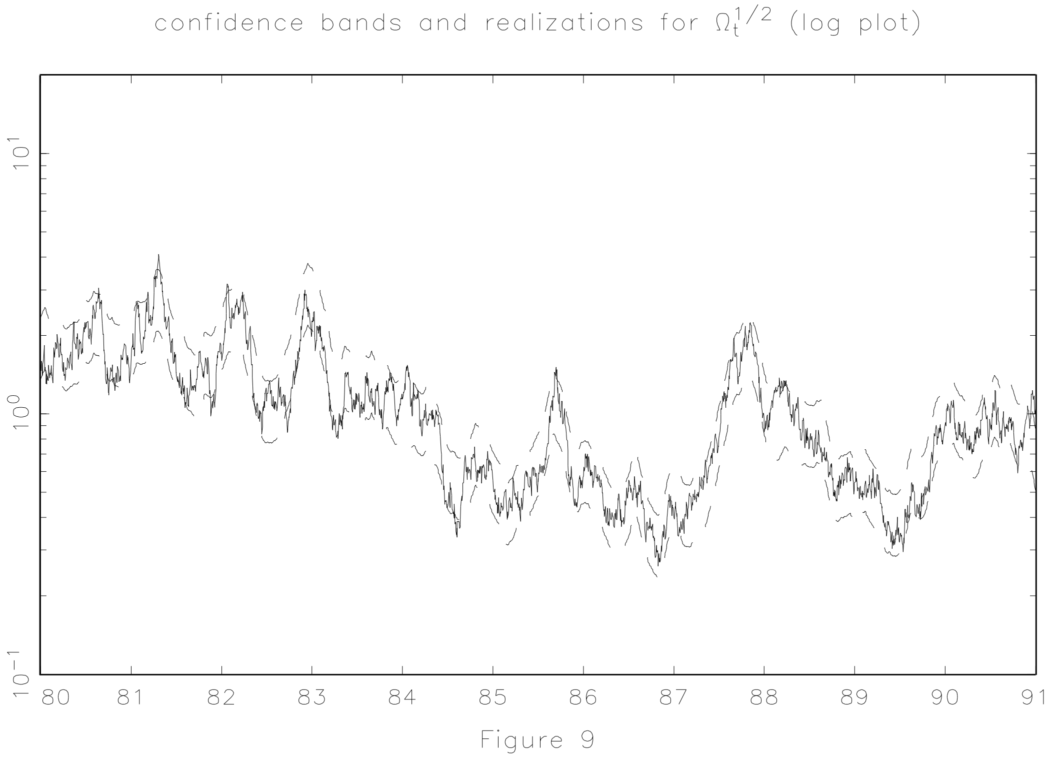

Figure 6: **Simulated data with 95% confidence bands:** The "truth" is seen to bounce around mostly between the confidence bounds–95% of the time lieing between the bounds and 5% of the time bouncing outside them. Since these bands hold pointwise, the "truth" *should* be outside them 5% of the time.