Learning about Volatility:
The Local Scale Model with Homoskedastic Innovations, with Application to Stock Returns

J. Huston McCulloch *
Ohio State University

Oct. 27, 2006

SHORT ABSTRACT

The Local Scale Model (LSM) of Shephard (1994) is similar in effect to IGARCH, but has an unobserved volatility that realistically evolves independently of the observations. It has one fewer parameter to estimate than IGARCH, and a closed form likelihood, despite the unobservability of the volatility. Although the errors are Gaussian conditional on the unobserved stochastic variance, they are Student t conditional on experience.

The present paper improves on the Shephard model by assigning equal variance to the innovations to the volatility. The improved model is fit to monthly stock returns by Maximum Likelihood, implying 7.76 steady-state degrees of freedom. (*100 wds.*)

This paper was presented at the 12th International Conference on Computing in Economics and Finance, Limassol, Cyprus, June 2006. The author is indebted to Martin Weitzman for crucial inspiration.

JEL codes:
C22 – Time Series Models
G10 – General Financial Markets

Keywords: Volatility clustering; adaptive learning; stock returns, Student t distribution, Neyman smooth test

Updates online via <www.econ.ohio-state.edu/jhm/jhm.html>

* Economics Department
Ohio State University
1945 N. High St.
Columbus, OH 43210
mcculloch.2@osu.edu

ABSTRACT

The Local Scale Model (LSM) of Shephard (1994) is a state-space model of volatility clustering similar in effect to IGARCH, but with an unobserved volatility that realistically evolves independently of the observed errors, instead of being mechanically determined by them. It has one fewer parameter to estimate than IGARCH, and a closed form likelihood, despite the unobservability of the volatility. Although the errors are assumed to be Gaussian conditional on the unobserved stochastic variance, they are Student t when conditioned on experience, with degrees of freedom that grow to a finite bound.

The present paper improves on the Shephard (1994) model by assigning equal variance to the innovations to the volatility. The implied volatility gain at first declines sharply as in the Local Level Model, rather than being constant throughout as in traditional IGARCH.

The improved model is fit to monthly stock returns by Maximum Likelihood. The parameter estimates imply 7.76 steady-state degrees of freedom. A short-lived "Great Moderation" is evident during the mid-1990's, but expires by 1998. Otherwise the period since 1970 was generally more volatile than the 1950s and 60s, though less so than the 1930s and 40s. The LSM volatility responds more nimbly to the data than does an IGARCH model.

Although the Student t densities generated by the Gaussian-based LSM account for much of the conditional leptokurtosis in the data, further refinements will be required to adequately model the pronounced negative skewness and/or residual leptokurtosis in stock returns.

1. Introduction and summary

The Local Scale Model (LSM) of Shephard (1994) is a model of volatility clustering that treats the volatility as an unobserved state variable that evolves stochastically with shocks that are realistically independent of the observed errors themselves, rather than being observed and mechanically determined by them as in ARCH (Engle 1982) or GARCH (Bollerslev 1986). The result is an IGARCH-like recursion for the parameters governing the volatility, but with one fewer parameter than is required by IGARCH. The LSM permits the likelihood to be computed exactly in terms of standard densities, without the tedious numerical integrations that are ordinarily required by non-Gaussian state-space models. The LSM does for stochastic volatility what Adaptive Learning (Evans and Honkapohja 2001, McCulloch 2005) does for stochastic regression coefficients.

The present paper improves upon the Shephard (1994) model by assigning equal variance to the innovations to the unobserved volatility. This results in a volatility gain that at first declines sharply with the number of observations as in the classical Local Level Model and in Adaptive Least Squares (McCulloch 2005), rather than being constant throughout as in traditional IGARCH models (McCulloch 1985a, Engle and Bollerslev 1986), or declining more slowly as in the Shephard (1994) model.

The paper also goes beyond Shephard (1994), by fitting the model to empirical data, specifically monthly excess stock returns for 1926-2003. Although the excess returns are assumed to be Gaussian conditional on the unobserved stochastic variance, they are in fact Student t when conditioned on investor experience. The estimated variance of the volatility innovations implies degrees of freedom (DOF) that are bounded above by 7.76.

The estimated volatility reacts more nimbly to the observed return shocks than does a conventional IGARCH(1,1) model. A short-lived "Great Moderation" in stock market volatility is apparent during the mid-1990's, but expires by 1998. Otherwise the period since 1970 was generally more volatile than the 1950s or 1970s, but far less so than the 1930s.

Unlike a conventional IGARCH model, the LSM accounts for considerable leptokurtosis after conditioning on experience. However, it does nothing to account for pronounced negative skewness in the conditional stock returns. The symmetric Gaussian-based model can be formally rejected using the Neyman Smooth Test for goodness-of-fit. Suggestions are offered for further refinements of the model to account for skewness and/or residual leptokurtosis.

Section 2 below reviews the Shephard LSM and modifies it to make the innovations to the volatility homoskedastic. Section 3 compares the improved LSM to related models of volatility clustering. Section 4 provides an application to monthly stock returns. Section 5 tests the transformed residuals for uniformity using the Neyman

smooth test, and finds that this can be easily rejected. Section 6 considers a WLS estimator of the mean. Section 7 concludes. The Appendix reviews key properties of the gamma and beta probability distributions.

2. The Local Scale Model

Shephard (1994) models a time series y_t as being Gaussian, conditional on a known mean μ and an unobserved time-varying precision, or reciprocal variance, θ_t:

$$y_t \mid \theta_t \sim N(\mu, 1/\theta_t).^{1} \tag{1}$$

Conditional on last period's experience $Y_{t-1} = \{y_1, \ldots y_{t-1}\}$, last period's precision is assumed to have a gamma distribution, with count parameter a_{t-1} and intensity b_{t-1} (see appendix):

$$\theta_{t-1} \mid Y_{t-1} \sim G(a_{t-1}, b_{t-1}).$$

The precision is assumed to evolve over time with beta-distributed multiplicative shocks η_t. Generalizing Shephard's notation somewhat so as to permit an important modification,[2] I set

$$\theta_t = \theta_{t-1}\eta_t,$$
$$\eta_t \sim k_t B(a'_t, a_{t-1} - a'_t),$$

for some $a'_t < a_{t-1}$ and $k_t > 1$ to be determined from a_{t-1}. It follows (see appendix) that

$$\theta_t \mid Y_{t-1} \sim G(a'_t, b'_t),$$

where

$$b'_t = b_{t-1} / k_t.$$

Furthermore, using Bayes' Rule and setting $\varepsilon_t = y_t - \mu$, it can easily be shown that conditional on the new information set Y_t, θ_t also has a gamma distribution:

$$\begin{aligned}
\theta_t \mid Y_t &= \theta_t \mid y_t, Y_{t-1} \\
&\propto (y_t \mid \theta_t, Y_{t-1})(\theta_t \mid Y_{t-1}) \\
&= (y_t \mid \theta_t)(\theta_t \mid Y_{t-1}) \\
&\propto \theta^{.5} \exp(-\theta_t \varepsilon_t^2 / 2)\theta_t^{a'_t - 1} \exp(-b'_t \theta_t) \\
&= \theta^{a'_t - .5} \exp(-b_t \theta_t) \\
&\sim G(a_t, b_t),
\end{aligned}$$

with parameters

$$\begin{aligned}
a_t &= a'_t + .5, \\
b_t &= b'_t + \varepsilon_t^2 / 2.
\end{aligned} \tag{2}$$

Exploiting the equivalence (see appendix) of a gamma RV with count parameter a to a scaled χ^2 RV with $d = 2a$ degrees of freedom (DOF), and setting

---

[1] As noted below, the model can easily be generalized to replace the fixed mean μ with a linear combination of exogenous regressors.

[2] Note that whereas Shephard's "η_t" is the beta-distributed shock itself, I have incorporated the scale factor k_t (Shephard's $\exp(r_t)$) into it. My a'_t is equivalent to Shephard's $a_{t|t-1}$.

$$v_t = 1/E_{t-1}\theta_t = b'_t / a'_t$$

and

$$d_t = 2a'_t,$$

it follows that conditional on past experience Y_{t-1}, the new observation y_t has a scaled and shifted Student t distribution with d_t DOF:

$$y_t \mid Y_{t-1} \sim \sqrt{v_t}\,T(d_t) + \mu\,. \tag{3}$$

Equations (2) then imply the following GARCH-like recursion for v_t:

$$v_t = \lambda_t v_{t-1} + \gamma_t \varepsilon^2_{t-1}, \tag{4}$$

where

$$\lambda_t = \frac{d_{t-1}}{k_t d_t}, \quad \gamma_t = \frac{1}{k_t d_t}. \tag{5}$$

Note that there is no constant term in (4), and that $\lambda_t + \gamma_t$ is not necessarily unity.

Shephard (1994) notes that in order to prevent the precision θ_t from converging in probability to either 0 or +∞, it is necessary to set

$$E\log(\theta_t / \theta_{t-1}) = E\log\eta_t = 0\,. \tag{6}$$

He observes that this in turn requires (see appendix)

$$\log k_t = \Psi(a_{t-1}) - \Psi(a'_t),$$

or equivalently,

$$k_t = \exp\big(\Psi\big((d_{t-1}+1)/2\big) - \Psi\big(d_t/2\big)\big), \tag{7}$$

where Ψ(a) is the *digamma function*, defined by

$$\Psi(a) = \frac{d}{da}\ln\Gamma(a).^{3} \tag{8}$$

The variance of the volatility shocks log η_t is in general given (see appendix) by

$$\operatorname{var}\log\eta_t = \Psi_1(a'_t) - \Psi_1(a_{t-1}), \tag{9}$$

where Ψ_1(a) is the *trigamma function*, defined by

$$\Psi_1(a) = \frac{d}{da}\Psi(a) = \frac{d^2}{da^2}\ln\Gamma(a)\,. \tag{10}$$

Shephard (1994) assumes that a'_t is some fixed constant $\omega < 1$ times a_{t-1}. However, this specification implies that the variance of log η_t is not constant, but rather declines sharply initially under the uninformative prior specified below. The present paper instead adopts the more appropriate assumption, in the spirit of the Local Level Model for the mean (see e.g. McCulloch 2005), that these shocks are homoskedastic,[4] with some constant variance φ. This in turn requires

$$a'_t = \Psi_1^{-1}\big(\Psi_1(a_{t-1}) + \varphi\big),$$

or terms of the predictive DOF d_t,

[3] The DIGAMMA function is supported by GAUSS, as is the TRIGAMMA function employed below.

[4] On the spelling of *heteroskedasticity*, see McCulloch (1985b).

$$d_t = 2\Psi_1^{-1}\left(\Psi_1((d_{t-1}+1)/2)+\varphi\right).^5 \qquad (11)$$

The familiar Local Level Model (see, e.g., McCulloch 2003) in fact goes one step further than homoskedastic shocks to the unobserved mean of the process, by assuming that they are actually *identically* distributed. However, in the present model this is not feasible since a_t, and therefore the parameters of the requisite beta distribution, change over time. Making the shocks homoskedastic as in (11) is the next best thing to making them identical.

Shephard (1994, p. 187) appropriately suggests that the Local Scale Model be initialized by specifying that $a_1 = 1/2$. This is equivalent to setting $a_1' = 0$, which in turn implies $b_1' = 0$ and therefore $b_1 = \varepsilon_1^2/2$ for any choice of v_1. Equivalently, we may simply set

$$d_1 = 0, \quad v_1 = 0 \qquad (12)$$

in the recursions (4), (7), (11). Since the gamma count parameter, or equivalently the χ^2 DOF, measures the precision of the estimate of the volatility, initializing it to zero is equivalent to starting with no information at all.[6]

If $\varphi = 0$, $d_t = t$-1 under the uninformative prior (12) , and indeed d_t behaves much like t-1 for small values of t even when $\varphi > 0$. For sample size n, d_n can thus be thought of as the effective average size of individual variance regimes. As t becomes large, d_t will approach a constant value d_∞ determined by the unique fixed point of (11):

$$\Psi_1(d_\infty/2) = \Psi_1((d_\infty+1)/2)+\varphi. \qquad (13)$$

It can be shown graphically, if not analytically, that for small values of φ,

$$d_\infty \approx \sqrt{2/\varphi}+1.$$

The solid line in Figure 1 below shows the first 20 values of d_t, using $\varphi = .03744$, the value estimated for monthly stock returns in section 4 below, in conjunction with (11) and (12). Initially, d(t) behaves much like t-1, shown as the dot-dash line, but then levels off as it quickly approaches its asymptotic value $d_\infty = 7.759$, represented by long dashes.

---

[5] If the inverse trigamma function required by (11) is not supported by the software at hand (e.g. GAUSS), it can easily be evaluated by means of a binary search.

[6] Taking the limit of the gamma density for $\theta_1|Y_0$ as a_1' falls to 0 while holding $v_1 = a_1'/b_1'$ constant yields an uninformative improper prior density for θ_1 that is proportional to $1/\theta_1$.

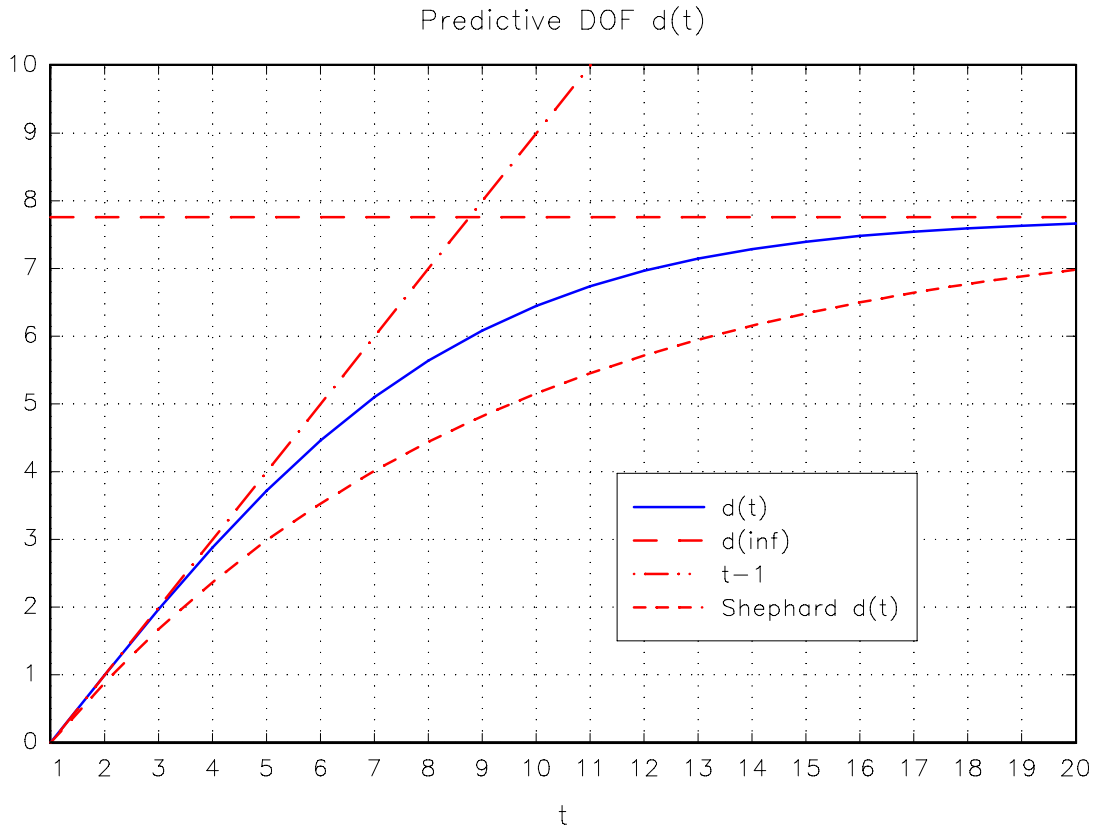

Figure 1

Under Shephard's specification that $a_t' = \omega a_{t-1}$ for some constant $\omega < 1$, the predictive DOF obey $d_t = \omega(d_{t-1} + 1)$ and will eventually approach the asymptotic value $d_\infty = \omega/(1-\omega)$. The present φ thus replaces Shephard's ω as the key parameter determining the asymptotic learning rate of the process. The short dashed line in Figure 1 depicts how d_t behaves under Shephard's specification, using $\omega = 0.8858$ so as to obtain the same asymptotic d_∞. It may be seen that under Shephard's heteroskedastic specification for the innovations to the log variance, d_t grows more slowly than t-1 initially, and approaches its asymptotic value much more slowly. The Shephard model thus takes longer to obtain a good fix on the volatility than does the improved model.

Figure 2 below shows the gain γ_t and attrition λ_t from the GARCH-like recursion for v_t in (4) and (5), using the uninformative prior (12) and the same φ as Figure 1. It may be seen that the gain behaves approximately like 1/(t-1) initially, and that the sum of the coefficients is slightly less than unity.

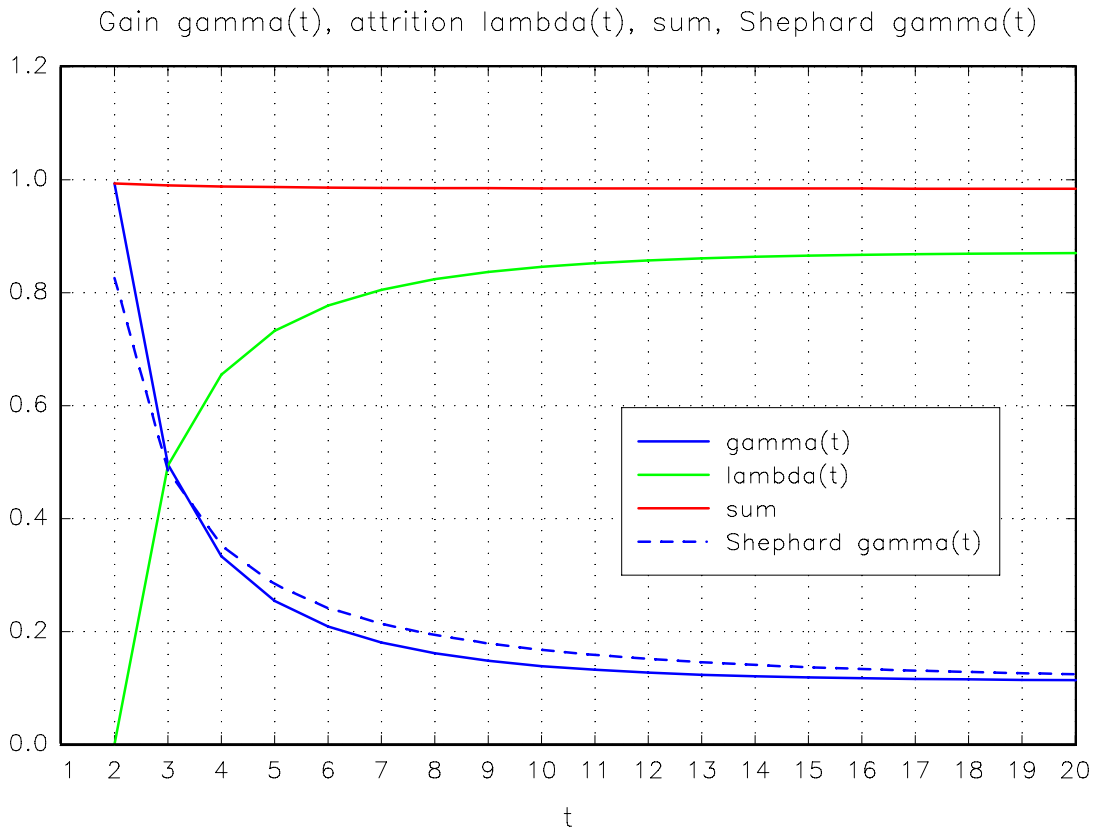

Figure 2

The dashed line in Figure 2 shows the gain under the Shephard specification. Because this specification attributes a very high variance to the initial volatility shocks, the y_2 predictive DOF d_2 is already substantially less than unity. The gain then rises above that in the improved specification, since the higher initial volatility shock variance implies that earlier experience becomes obsolete faster.

For any values of the two hyperparameters μ and φ, the log-likelihood implied by the modified LSM model (4), (5), (7), (11), (12) becomes

$$\mathcal{L}(\mu, \varphi) = \sum_{t=2}^{n} \log\left(t_{d_t}\left(\varepsilon_t / v_t^{1/2}\right) / v_t^{1/2}\right)$$

where

$$t_d(x) = \frac{\Gamma\left((d+1)/2\right)}{\Gamma(d/2)\sqrt{d\pi}\left(1 + x^2/d\right)^{(d+1)/2}}$$

is the standard Student t density with d DOF.[7] This may be maximized numerically to obtain Maximum Likelihood (ML) estimates of μ and φ when, as is ordinarily the case, these are not really known, as has been assumed to this point. Since the Student t density

[7] $\log\Gamma(a)$ may be computed to adequate precision in GAUSS as LN(GAMMA(a)) for $a < 7$, and as LNFACT(a-1) for $a \geq 7$. Neither function is adequate over the entire required range, however.

is symmetric, the two parameter estimates are asymptotically orthogonal, and hence the variance of the ML estimate of the mean may be estimated simply as

$$se(\hat{\mu}_{ML}) = (\frac{\partial^2}{\partial \mu^2} \mathcal{L}(\mu, \varphi))^{-1/2}.$$

3. Related Models

In the pioneering ARCH(p) model of Engle (1982), it was assumed that

$$y_t \mid Y_{t-1} \sim N(\mu, \sigma_t^2),$$

where

$$\sigma_t^2 = \delta + \sum_{j=1}^{p} \gamma_j \varepsilon_{t-j}^2 .$$

In this model, the shocks that drive the variance are the squares of the observation errors themselves, rather than being independent as in the LSM. This assumption is contrived, but is very conveniently computationally. Because it implies that the variance is actually observed for $t > p$, the joint probability of $y_{p+1}, \ldots y_n$ can be written down in closed form as a product of normal densities without the tedious numerical integrals that would ordinarily be necessary if the variance were more realistically treated as an unobserved state variable (see, e.g., Harvey 1989, pp. 162-4; Bidarkota and McCulloch 1998). This computational convenience accounts for the overwhelming success of ARCH and ARCH-like models.

It was quickly recognized (McCulloch 1985a, Bollerslev 1986) that a high degree of persistence can be obtained with far fewer parameters than ARCH, simply by adding one or more lags of the variance itself to a small number of ARCH terms. In the popular GARCH(1,1) model,

$$\sigma_t^2 = \delta + \lambda \sigma_{t-1}^2 + \gamma \varepsilon_{t-1}^2,^{8} \qquad (14)$$

the unconditional variance

$$\mathrm{E}\, \varepsilon_t^2 = \delta / (1 - \lambda - \gamma)$$

is finite when $\lambda + \gamma < 1$.[9]

---

[8] Bollerslev (1986) in fact considered a GARCH(p,q) model with p lags of the squared errors and q of the variance, but $p = q = 1$ is ordinarily adequate.

[9] In GARCH (1,1), the joint probability of $y_1, \ldots y_n$ unfortunately depends on the unobserved initial variance σ_1^2. In principle, if the process is strictly stationary, as it is even for $\lambda + \gamma = 1$, the unconditional likelihood could be found by first finding the unconditional distribution of σ_1^2 by iterating numerically on (14), and then taking the expectation of the conditional likelihood under this distribution. However, the effect of the unobserved initial variance quickly dies out, so that practitioners invariably resort instead to simpler expedients such as using pre-sample values (McCulloch 1985; Bollerslev 1986, p. 315, n. 4), using the full-sample variance (Engle and Bollerslev 1986), treating σ_1^2 as an additional parameter to be estimated by ML (Hamilton and Susmel 1994, Bidarkota and McCulloch 1998), backcasting using an arbitrary geometric decay factor (EViews 4.0 2000, p. 385), or backcasting from the end of the sample using the GARCH coefficients themselves (McCulloch 2005). The LSM does not require such expedients.

In the spirit of Adaptive Expectations and the Local Level Model, McCulloch (1985a) and Engle and Bollerslev (1986) imposed the further restriction $\lambda + \gamma = 1$, in what came to be known as the Integrated GARCH, or IGARCH model. It was at first believed that the additional restriction $\delta = 0$ would then be required, in order to prevent convergence in probability to infinity. The original, unaugmented IGARCH(1,1) model was thus

$$\sigma_t^2 = (1-\gamma)\sigma_{t-1}^2 + \gamma\varepsilon_{t-1}^2 .^{10} \qquad (15)$$

However, Nelson (1990) soon pointed out that with no intercept, (15) implies that variance and therefore the errors themselves must in fact converge in probability to 0. This occurs because under this specification, the variance is a martingale. Since the variance of σ_t^2 increases without bound, yet σ_t^2 is bounded below by zero, this requires that virtually all the density must eventually converge to near 0. In order to prevent σ_t^2 from invariably collapsing on 0 or exploding to infinity, its *log* must be a martingale. This in turn requires, by Jensen's inequality, that σ_t^2 itself must be a supermartingale, i.e. $\lambda + \gamma$ must exceed unity by some small amount. Since it is difficult to compute the boundary, most practitioners since 1990 have instead simply added a positive constant to (15):

$$\sigma_t^2 = \delta + (1-\gamma)\sigma_{t-1}^2 + \gamma\varepsilon_{t-1}^2 \qquad (16)$$

This augmented IGARCH(1,1) process is strictly stationary for positive δ, despite the infinite expectation of σ_t^2, and is bounded below by δ.

The Local Scale Model eliminates the artificial assumption of ARCH and GARCH, that the disturbances to the time-changing variance are just the squares of the observation errors themselves and therefore that the variance is actually observed (in the ARCH case) or virtually observed (in the GARCH case), and replaces it with the much more natural assumption that the variance is unobserved, and evolves with shocks that are independent of the observation errors. This does not generate tedious numerical integrals, because Shephard's (1994) special assumption that the precision shocks are beta of a certain form implies that conditional on experience, the precision always has a closed form gamma distribution. The LSM is also more parsimonious than the augmented IGARCH model (16), in that it has only one parameter to estimate, rather than two. Since the underlying process is not strictly stationary, the variance is permitted to wander and remain arbitrarily high or low, and is not bounded below by $\delta > 0$.

As a consequence of its realistically modeling the variance as an unobserved state variable to be inferred by signal extraction techniques, the LSM makes it clear that conditional on experience, the errors in fact have a Student t distribution with DOF

---

[10] McCulloch (1985) called essentially this model "Adaptive Conditional Heteroskedasticity" (ACH, to be pronounced as *ach*! in German), but Bollerslev's (1986) "IGARCH" caught on instead. Since McCulloch (1985) generalized the conditional errors to be symmetric stable, which have infinite variance except in the Gaussian special case, the variance in (15) was replaced by the stable scale parameter c_t and the squared error by the absolute error, which has a finite mean when the stable characteristic exponent α exceeds unity. In retrospect, the squared scale and squared errors could just as easily have been retained from ARCH despite the infinite expectation of the latter in the non-Gaussian cases.

bounded by d_∞, rather than a Gaussian distribution as in ARCH or GARCH. As Weitzman (2006) points out, this has very dramatic implications for the arithmetic equity premium, when the model is applied to logarithmic returns. The power tails of the conditionally Student t distribution of the log returns make the arithmetic equity premium infinite, requiring a re-evaluation of what we mean by an equity premium.[11]

In the modified LSM of Section 2, $v_t = 1/\mathrm{E}\theta_t$ follows the GARCH-like recursion (4) with no intercept term, and with time-varying coefficients λ_t and γ_t that sum to slightly less than unity, as illustrated in Figure 2 above. However, this process does not collapse on zero, as does (15), since the LSM errors are in fact conditionally Student t, whereas the IGARCH errors are conditionally Gaussian. The heavy tails of the Student t errors provide the extra kick to keep the process alive. The conditional variance of the LSM errors is not v_t itself, but rather is infinite for $d_t \le 2$, and

$$h_t = \mathrm{E}(\varepsilon_t^2 \mid Y_{t-1}) = \mathrm{E}(1/\theta_t \mid Y_{t-1}) = b_t'/(a_t' - 1) = v_t d_t/(d_t - 2) \qquad (17)$$

for $d_t > 2$ (see appendix). This conditional variance obeys the recursion

$$h_t = \lambda_t h_{t-1} + \gamma_t^* \varepsilon_{t-1}^2,$$

where

$$\gamma_t^* = \gamma_t d_t/(d_t - 2).$$

The coefficients λ_t and γ_t^* (not plotted) in fact sum to more than unity.

Uhlig (1997) discusses how a Shephard-like LSM could be applied to the estimation of vector autoregressions, using the multivariate Wishart distribution to generalize the gamma. However, Uhlig considers only the long-run case when the predictive count parameter a_t' and therefore k_t have reached their limiting values a_∞' (Uhlig's ν) and k_∞ ($1/\lambda$ in Uhlig's notation). He notes (p. 61), following Shephard (1994), that the process will tend a.s. to 0 or ∞ unless k_∞ is governed by (6), but then, without explanation, sets his $\lambda = \nu/(\nu + 1)$ instead, which, as he notes, is not even the condition for a martingale in θ_t. Rather than estimating his ν from the data, he instructs the reader (p. 71) to set it to 20 for quarterly data and to 60 for monthly data. He provides no empirical application of his method.

Hamilton and Susmel (HS 1994, p. 310) reject continuous-state GARCH-like models of stock returns in favor of discrete-state Markov-switching models of volatility clustering, and even in favor of a naive constant-variance model, on the erroneous grounds that if the variance has been correctly modeled, the model should minimize the mean squared deviation of the *squared* errors *from the modeled variance*.

For a Gaussian distribution, the log likelihood is affine and decreasing in the squared deviation about the mean, and hence the average squared deviation of the errors about the mean themselves can be taken as an equivalent loss function for evaluating the

[11] McCulloch (2003) demonstrates than in the case of log-stable returns, which can have a similar upper power tail and infinite arithmetic equity premium, the risk-neutral measure is not a simple location shift of the frequency measure as in the Gaussian case, but actually has a different shape, and finite expectation. A similar change of shape may occur in the log-Student case.

mean estimate. However, even if the errors are Gaussian, the *squared* errors have a scaled χ_1^2 distribution, which is far from Gaussian. Likewise, if (as in HS's best GARCH-like model) the scaled errors are Student t with ν DOF, the squared errors have a scaled $F(1,\nu)$ distribution, which again is far from Gaussian. The sum of squared deviations of the squared errors from the modeled variance is therefore an entirely inappropriate loss function.

In fact, the correctly computed forecasting loss function under each model postulated by HS is simply the negative of the log likelihood. As can be seen from their Table 1, the GARCH and t-GARCH-L models greatly outperform the constant variance model by this correct criterion, even using the Schwartz penalty for number of parameters. In fact, by SIC, the t-GARCH-L model is the best one tabulated. There is therefore no reason to reject an elegantly continuous-state GARCH-type model in favor of HS's cumbersome discrete-state switching models on the basis of this criterion.

Shephard (1994) provides no empirical application of the LSM. The following section fills this void by applying the improved LSM to U.S. stock returns.

4. Application to Stock Returns

Figure 3 plots monthly continuously compounded CRSP Value-Weighed stock returns, including distributions, in excess of the Fama 1-month Treasury bill rate, as obtained from Wharton Research Data Services, for Jan. 1926 – Dec. 2003 (936 observations). For this purpose, the arithmetic CRSP returns were converted to log returns. The T-bill rates, which have already been converted to continuous compounding with a 365-day year, were divided by 1200 to give monthly log returns. They were then lagged one month relative to the stock returns, since the Fama T-bill rate for e.g. Jan. 1926 is the rate on a bill purchased at the end of January, whose payoff at the end of February is already known, whereas the CRSP stock return for the same date is the return on stocks purchased at the end of Dec. 1925, whose payoff is not known until the end of January. Bid and asked yields were averaged, with a few missing asked yields constructed from the average of the spreads for adjacent months. Missing asked yields for 1/35 – 3/36 were set to 0, the actual asked quote for 3/34 – 12/34 and 4/36 – 11/36. A few negative average yields were left as quoted.[12]

[12] The only important negative yield was Nov. 1930, the month of the failure of the Bank of United States, when T-bills, which were evidently regarded as safer than interbank deposits, yielded -1.074 % per annum (bid) and -1.188% (asked). Slightly negative average yields in the late 30's were all within transactions costs of 0.

Figure 3

The mean excess return is .005105/month (s.e. = .001777), i.e. 6.126%/year (s.e. = 2.132). However, this OLS estimate is inefficient and its standard error invalid, given the obvious presence of conditional heteroskedasticity.

The ML estimates of the improved LSM parameters and implied values are given in Table 1 below. Since the null hypothesis of homoskedasticity ($\varphi = 0$) is on the boundary of the parameter space, the Likelihood Ratio (LR) statistic does not necessarily have its customary χ_1^2 distribution. Nevertheless, this large statistic presumably provides strong evidence against the null.

Table 1

μ (se)	0.008662/mo. = 10.39%/yr. (0.001320) (1.58)
φ	0.03744
$\varphi^{1/2}$	0.1935
d_n	7.756
d_∞	7.759
LR (φ=0)	300.033

Figure 4 shows the demeaned excess returns, along with the local scale $v_t^{1/2}$. It may be seen that the LSM scale adapts quickly to lulls in market volatility such as occurred in the mid-1930s, -1960s, and -1990s. Although there was a brief "Great Moderation" during the mid-1990s, it expires by 1998. Otherwise, the period since 1970 has generally been more volatile than the 1950s and 60s, though less volatile than the 1930s and early 40s.

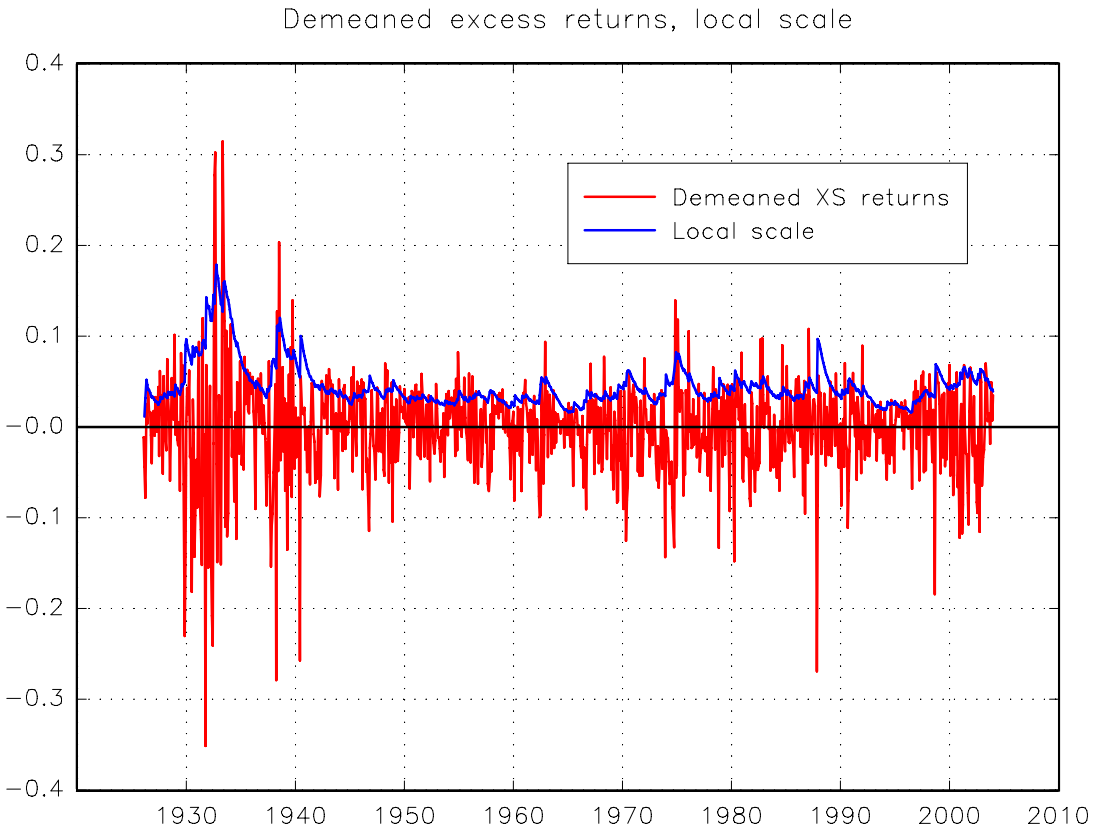

Figure 4

Hamilton and Susmel (1994, pp. 314-6) criticize GARCH-type models for overpredicting the squared errors during the weeks immediately following the 10/87 crash. Figure 4 above exhibits a similar phenomenon in the monthly returns. Again their criticism is irrelevant, however, since no forecasting model can be expected to predict every episode correctly. Models should be judged by their overall fit, and not on the basis of single episodes. Obviously there are many instances in which the LSM comparably conversely underpredicts the squared errors.[13]

[13] This is not to say that the fit could not be improved by adequately modeling the conditional skewness and/or residual leptokurtosis in the data, as discussed below. Such a refinement goes beyond the scope of the present paper.

For comparison, the data was also fit to the GARCH(1,1) model (14) and augmented IGARCH(1,1) model (15), as indicated in Table 2 below. Since the null of homoskedasticity ($\lambda = \gamma = 0$) is again on the boundary of the parameter space, the LR for this hypothesis once again does not necessarily have a χ^2 distribution. Nevertheless, the large value almost surely justifies rejecting the null. Even though the GARCH parameters fall short of the IGARCH boundary $\lambda + \gamma = 1$, they are very close to it and this restriction cannot be rejected by the LR statistic in the last line.

Table 2

	GARCH(1,1)	IGARCH(1,1)
μ	0.006622	0.006568
δ	7.01e-5	4.58e-5
λ	0.8679	0.8694
γ	0.1121	0.1306
$\lambda + \gamma$	0.9800	1.0000
LR ($\lambda = \gamma = 0$)	258.364	
LR ($\lambda + \gamma = 1$)		2.390

Figure 5 below compares the LSM local standard deviation $h_t^{1/2}$ as computed from (17) to the IGARCH local standard deviation. Although they have the same general pattern after the initial startup, it may be seen that the LSM adjusts more sensitively to periods of high and low variance than does IGARCH.

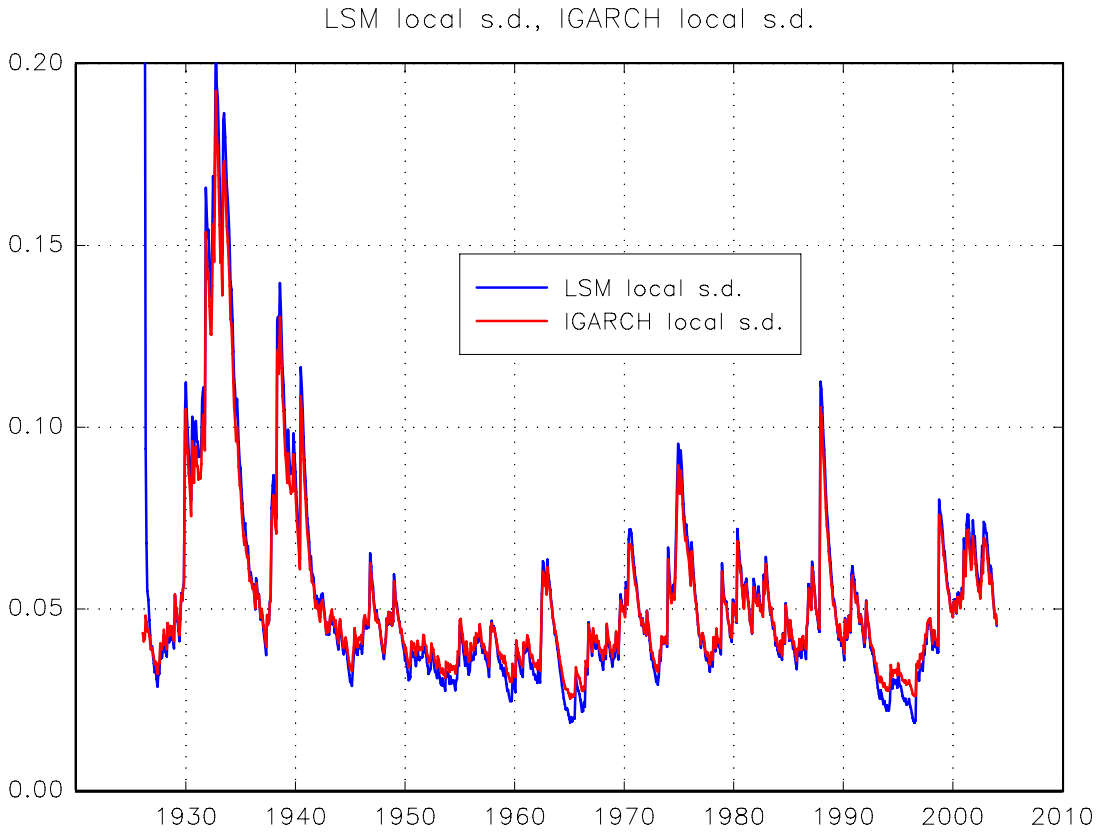

Figure 5

5. Analysis of residuals

Figure 6 depicts the scale-adjusted errors $\hat{\xi}_t = \hat{\varepsilon}_t / v_t^{1/2}$. Even though the errors are assumed to be Gaussian conditional on the unobserved precision θ_t, the LSM implies that they are Student t with d_t DOF when conditioned on experience to date. Figure 1 above shows the first 20 values of the predictive DOF d_t, along with (t-1) itself. For the first 3 or 4 observations, d_t follows (t-1) closely, but then it quickly approaches its limiting value of 7.759.

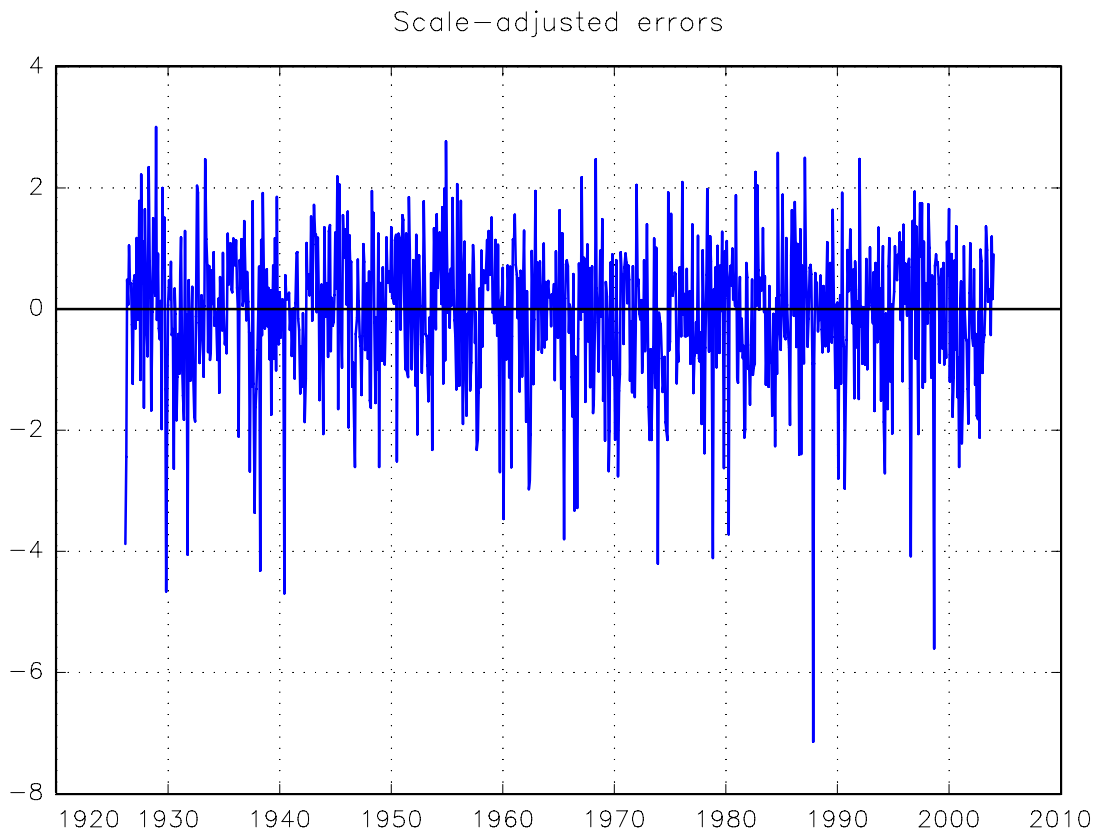

Figure 6

It is obvious from Figure 6 that there is considerable downward skewness to the scale-adjusted errors that is inconsistent with the intrinsically symmetrical Student t distribution (3) implied by the underlying Gaussian model (1). This downward skewness is almost as pronounced in the arithmetic returns, and hence is not simply due to the present study's focus on logarithmic returns.

Unfortunately, the standard test for symmetry based on the skewness statistic is inappropriate here, since the null hypothesis of that test is that the errors are iid Gaussian (and therefore symmetrical), while in fact the adjusted forecast errors should be Student t under the assumed model. Furthermore the errors are not even iid Student, since their DOF are not constant, at least not for the first several periods.

Nevertheless, if the LSM correctly characterizes the data, the transformed errors $\hat{u}_t = T_{d_t}\left(\hat{\xi}_t\right)$ should be iid U(0,1), where $T_d(\ )$ represents the Student t cumulative distribution function with d DOF.[14] Figure 7 gives the histogram of the transformed errors, using 10 equally spaced bins, along with a horizontal line at 93.7, the expected

[14] Strictly speaking, this is only true if the two hyperparameters μ and φ are known. If they have been estimated, the transformed residuals will tend to look even more uniform than they should. See Percy (2006) and discussion below.

frequency per bin. It may be seen that too many errors accumulate in the first bin, and too little in the last bin, with compensating deviations from uniformity in the 3rd, 4th, 8th, and 9th bins.

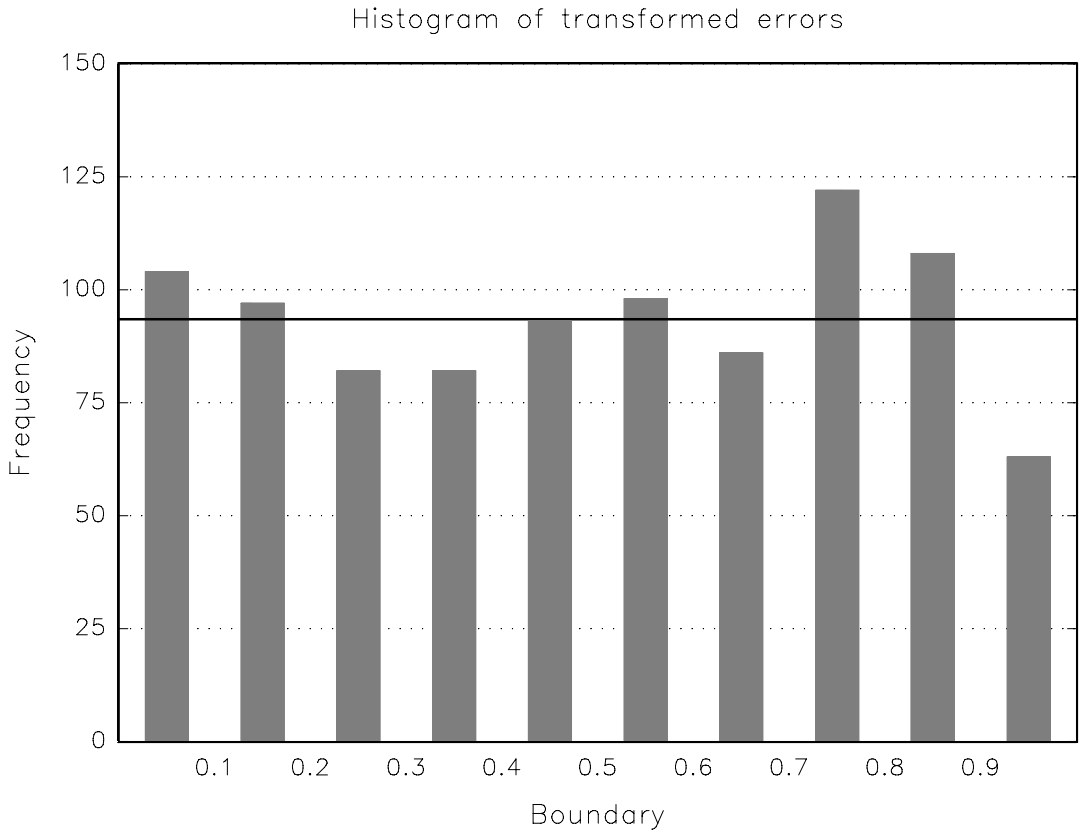

Figure 7

The classic Pearson χ^2 test uses frequency counts such as those in Figure 7 to test that the deviation from uniformity is not just sampling error. However, the Pearson test's alternative hypothesis is that the density is constant in each bin, and then changes abruptly to a completely unrelated value (aside from an adding-up constraint) in the adjacent bins.

The Neyman (1936) Smooth Test for uniformity instead poses as its alternative that the density is a polynomial of degree *k.* This alternative hypothesis allows the density to change continually, without the arbitrary discontinuities of the Pearson alternative hypothesis. Percy (2006) has found that the Neyman Smooth Test does indeed have much more power than the Pearson test, for discriminating among heavy-tailed probability distributions.

When the hyperparameters of the model are known, the Lagrange Multiplier (LM) version of the Neyman test statistic is

$$LM = \mathbf{s}'\mathbf{I}^{-1}\mathbf{s},$$

where $\mathbf{s} = (s_1, \ldots, s_k)'$ is the score vector (the gradient of the log-likelihood), and $\mathbf{I} = \left(I_{jj'}\right)$ is the information matrix (the expected Hessian of the log-likelihood), both evaluated under the null of uniformity.[15] For a typical sample of size *n*, these are

$$s_j = \sum_{t=1}^{n} \hat{u}_t^j - n/(j+1),$$
$$I_{jj'} = n\left((j+j'+1)^{-1} - (j+1)^{-1}(j'+1)^{-1}\right).$$

Under the null of uniformity, the LM statistic is asymptotically $\chi^2_{(k)}$ when the parameters are known. (In the present application, the first predictive density is missing, so that the sum is taken from t = 2 to n, and n replaced by n-1 in the above.)

Although it is not obvious what particular value of *k* would be ideal, we are concerned with potential skewness in the underlying errors, yet have already removed the mean and scale from the data, and therefore want to consider $k \geq 3$. The Neyman LM statistic is given in Table 3 for $k = 3, \ldots 10$, along with the corresponding $\chi^2_{(k)}$ *p*-values. In every case, the null of uniformity (and therefore the underlying symmetric model) can be rejected at well under the .001 level. The p-values are in fact surprisingly insensitive to the tabulated values of k.[16]

[15] Neyman in fact used a Likelihood Ratio (LR) form of the test, which required him to use an exponentiated polynomial perturbation to uniformity as his alternative to preclude negative density estimates, and then to perform a potentially ill-conditioned estimate of his model under the alternative. The LM form investigated by Percy (2006) is much easier to use, since it only requires estimation under the null, and can employ a simple polynomial perturbation.

[16] The LM statistics for k = 1 and 2 were 0.423 and 0.463, resp., with p-values of 0.515 and 0.793, resp. However, since the mean and scale have been estimated from the data and removed from the errors, it would be very surprising if either of these registered significance, no matter how bad the fit.

Table 3
Neyman Smooth Test for Goodness of Fit

k	LM	p
3	21.91	.00007
4	22.47	.00016
5	25.83	.00010
6	25.92	.00023
7	26.77	.00037
8	27.36	.00061
9	30.92	.00031
10	31.07	.00057

Because the two hyperparameters μ and φ are in fact not known, but have been estimated from the data, χ^2 critical values will tend to underreject the null, as noted by Percy (2006). The true rejection of the model is therefore even stronger than suggested by Table 1. The present paper makes no attempt to apply the correction for estimated parameters, as implemented by Percy (2006).

Correctly modeling the highly skewed stock market returns would require replacing the intrinsically Gaussian assumption (1) with a distribution such as the skew-stable class (McCulloch 1998) that generalizes the Central Limit considerations that motivate the typical Gaussian assumption, yet permits skewness and/or intrinsic leptokurtosis. Such a generalization would be desirable, but goes far beyond the scope of the present paper.

A far simpler approximate solution would be to replace (3) with the ad hoc assumption that conditional on experience, returns are themselves either of the Pearson Type IV class (Heinrich 2004), or else of the Bauwens and Laurent (2002) type, with scale and DOF determined as if the Gaussian model were somehow valid. These distributions generalize the Student t class to include a skewness parameter that is effective even with infinite DOF. Such an approach could give empirically useful results without any deep reworking of the model, but again would go far beyond the scope of the present study.

6. Robust WLS Estimation of the Mean

The Best Linear Unbiased Estimate (BLUE) of the mean, if not the best global estimate, is given by Weighted Least Squares (WLS),

$$\hat{\mu}_{WLS} = \frac{\sum_{t=1}^{n} y_t / h_t}{\sum_{t=1}^{n} 1/h_t}, \tag{18}$$

using weights determined by (17).[17] Since $h_t = \infty$ for $d_t \leq 2$, the first 3 or more values of y_t are completely ignored by WLS. Although ML is asymptotically efficient under the assumed model, WLS may be more robust to deviations from the posited model, such as skewness.

The WLS estimate of the mean and its standard error, using the weights (17) implied by the LSM, are given in Table 4 below. Althugh the ML and WLS standard errors are quite similar, the WLS estimate of the mean is dramatically smaller than the ML estimate, by 3.78%/yr, a difference of more than 2 standard errors.

Table 4
WLS Estimate of mean excess return

$\hat{\mu}_{WLS}$	0.005506/mo. = 6.607 %/yr.
(se)	(0.001351) (1.621)

If the model were true, or even approximately true, we would ordinarily prefer the ML estimate of the mean to the WLS estimate. However, in the present instance, since we know that there is substantial downward skewness to the returns not captured by the model, the WLS estimate may actually be preferable. WLS may therefore be giving us a more robust estimate of the true mean than the mis-specified ML in the present instance.

7. Conclusion

The Local Scale Model (LSM) of Shephard (1994) provides a computationally simple model of volatility clustering that is at once more realistic and more parsimonious than a conventional augmented IGARCH(1,1) model. Even though the observations are assumed to be Gaussian conditional on the unobserved volatility, they are Student t with finite degrees of freedom conditional on experience. The LSM thus accounts for much of the leptokurtosis in the data, without assuming a leptokurtic underlying distribution.

The present paper improves upon Shephard's original formulation by making the innovations to the log of the volatility homoskedastic, as in the classic Local Level Model. This modification makes more realistic and efficient use of the initial observations, and thereby facilitates ML estimation of the hyperparameters.

The present paper also goes beyond Shephard (1994) by providing an empirical application of the model, using CRSP monthly excess stock returns. The estimated LSM implies that conditional on steady-state experience, returns are Student t with 7.76 DOF. The LSM volatility adjusts to changing the data with greater sensitivity than does a conventional IGARCH(1,1) model.

[17] The WLS weights may be computed iteratively from the residuals about WLS estimates of the mean, starting with the OLS estimate, though in the illustration below they are computed about the full ML estimate.

Even though the log returns are assumed to be Gaussian conditional on the unobserved true volatility, the conventional Black/Scholes option model cannot be used to price options, since returns are Student t conditional on investor experience. Furthermore, expected arithmetic returns, and therefore the conventionally measured equity premium, are both infinite. A similar problem arises with log-stable asset returns, but may be solved by considering that the Risk-Neutral Measure for these distributions is not a simple location shift as in the log-Gaussian case (McCulloch 2003). Future research should study the nature of the Risk-Neutral Measure for the Student t returns implied by the LSM.

Unfortunately, pronounced skewness in the scale-adjusted errors from the stock return data permits us to reject the Gaussian-based LSM, using the Neyman Smooth Test of goodness-of-fit. This indicates that further refinements of the model itself are in order, at least for stock returns.

Although in the present paper, the observed data series was assumed to have a constant mean, the model could easily be generalized to incorporate a fixed linear combination of independent variables as the mean. Future research should focus on the more difficult problem of incorporating a time-varying mean and/or regression coefficients as in McCulloch (2005).

Appendix

This appendix states and/or develops certain key properties of the gamma and beta distributions used in the text.

A Gamma distributed random variable (RV) $G(a, b)$ with count parameter a and intensity b has density defined for $x \in (0,\infty)$ of

$$b^a x^{a-1} e^{-bx} / \Gamma(a)$$

and mean

$$EG(a,b) = a/b\,,$$

where the gamma function $\Gamma(a)$ is defined by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-bx} dx\,.$$

When a is an integer, the gamma distribution governs the waiting time until the a-th event in a Poisson-driven process with intensity (frequency per unit time) b, so that a counts the number of Poisson-driven arrivals. The intensity b is the reciprocal of the scale. As is well known (e.g. Casella and Berger 2002, p. 627), a Gamma RV is equivalent to a χ^2 RV with $d = 2a$ DOF, and scaled by $1/2b$:

$$\mathrm{G}(a,b) \sim \frac{1}{2b}\chi^2_{2a}\,.$$

A Beta distributed RV Beta(α, β) has density defined for $x \in [0,1]$ of

$$x^{\alpha-1}(1-x)^{\beta-1} / \mathrm{B}(\alpha,\beta)\,,$$

where the beta function B(α, β) is given by

$$\begin{aligned}\mathrm{B}(\alpha,\beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \\ &= \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha,\beta).\end{aligned}$$

Shephard (1994) exploits the well-known, if little-appreciated, fact (see e.g. Casella and Berger, 2002, p. 195, problem 4.24) that if X and Y are independent RVs with $X \sim \mathrm{G}(a, b)$ and $Y \sim \mathrm{Beta}(a', a-a')$, with $a' < a$, their product Z is again gamma, but with reduced count parameter a':

$$Z = XY \sim \mathrm{G}(a', b).$$

If X ~ Beta(α, β), the characteristic function of Y = log(X) is

$$\begin{aligned}\mathrm{cf}_Y(t) &= \mathrm{E}\, e^{iYt} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 e^{i\log(x)t} x^{\alpha-1}(1-x)^{\beta-1} dx \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 x^{\alpha+it-1}(1-x)^{\beta-1} dx \\ &= \mathrm{B}(\alpha+it,\beta)/\mathrm{B}(\alpha,\beta).\end{aligned}$$

It follows that

$$\mathrm{E}\,Y = i^{-1}\frac{d}{dt}\mathrm{cf}_Y(t)\,|_{t=0} = \Psi(\alpha) - \Psi(\alpha+\beta),$$

where $\Psi(a)$ is the *Euler Psi* or *digamma function*, defined in Equation (8) in the text. Furthermore,

$$\mathrm{var}\,Y = i^{-2}\frac{d^2}{dt^2}\mathrm{cf}_Y(t)\,|_{t=0} = \Psi_1(\alpha) - \Psi_1(\alpha+\beta),$$

where $\Psi_1(a)$ is the *trigamma function*, defined in Equation (10) in the text.

If X ~ G(a, 1) with a > 1, 1/X has mean

$$\begin{aligned}\mathrm{E}(1/X) &= \int_0^\infty x^{-1}x^{a-1}\exp(-x)/\Gamma(a)dx \\ &= \Gamma(a-1)/\Gamma(a) \\ &= 1/(a-1).\end{aligned}$$

Consequently, if X ~ G(a, b) ~ G(a,1)/b, 1/X has mean

$$\begin{aligned}\mathrm{E}(1/X) &= b\,\mathrm{E}(1/G(a,1)) \\ &= b/(a-1) \\ &= \frac{a}{a-1}\frac{1}{\mathrm{E}\,X}.\end{aligned}$$

This exactly quantifies Jensen's Inequality, by which E(1/X) > 1/EX for any nondegenerate distribution. It follows that if X has a scaled χ^2 distribution with d DOF,

$$E(1/X) = \frac{d}{d-2}\frac{1}{EX},$$

for d > 2, and infinity otherwise.

References:

Bauwens, Luc and Sebastien Laurent, "A New Class of Multivariate Skew Densities, with Application to GARCH Models." Society for Computational Economics Conference on Computing in Economics and Finance, Aix en Provence, France, 2002.

Bidarkota, Prasad V., and J. Huston McCulloch, "Optimal Univariate Inflation Forecasting with Symmetric Stable Shocks," *J. Applied Econometrics* **13** (1998): 659-670.

Bollerslev, Tim, "Generalized Autoregressive Conditional Heteroskedasticity," *J. Econometrics* **31** (1986): 1-50.

Casella, George, and Roger L. Berger, *Statistical Inference*, 2nd ed. Duxbury, 2002.

Durbin, James, and S.J. Koopman, *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

Engle, Robert F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflations,". *Econometrica* **50** (1982): 987-1007.

Engle, Robert F., and Tim Bollerslev, "Modelling the Persistence of Conditional Variances," *J. Econometrics* **31** (1986): 307-27.

Evans, George W., and Seppo Honkapohja. *Learning and Expectations in Macroeconomics*. Princeton University Press, 2001.

EViews 4 Command and Programming Reference. Quantitative Micro Software, Irvine CA, 2000.

Hamilton, James D., and Raul Susmel, "Autoregressive Conditional Heteroskedasticity and Changes in Regime," *J. Econometrics* **64** (1994): 307:33.

Harvey, Andrew C., *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge Univ. Press, 1989.

Heinrich, Joel, "A Guide to the Pearson Type IV Distribution," Collider Detector at Fermilab (CDF) publication # 6820, 2004. Online at <www.cdf.fnal.gov/publications/cdf6820_pearson4.pdf>.

McCulloch, J. Huston, "Interest-Risk Sensitive Deposit Insurance Premia: Stable ACH Estimates," *J. Banking and Finance* **9** (1985a): 137-156.

_____, "On Heteros*edasticity," *Econometrica* **53** (1985b): 483.

_____, "Financial Applications of Stable Distributions," in G.S. Maddala and C.R. Rao, eds., *Handbook of Statistics, Vol.* 14, Elsevier Science B.V., 1996, pp. 393-425.

_____, "The Risk-Neutral Measure and Option Pricing under Log-Stable Uncertainty," Ohio State University Economics Department Working Paper 03-07, June 2003, online at <http://econ.ohio-state.edu/jhm/papers/rnm.pdf>.

_____, "The Kalman Foundations of Adaptive Least Squares," Ohio State University Economics Dept. Working Paper 05-01, August 2005, online at <http://econ.ohio-state.edu/jhm/papers/KalmanAL.pdf>.

Nelson, Daniel B., "Stationarity and Persistence in the GARCH(1,1) Model," *Econometric Theory* **6** (1990): 318-34.

Neyman, Jerzy, "Smooth tests for goodness of fit," *Skand. Aktuar.* **20** (1937): 150-99.

Percy, E. Richard, Jr., "Corrected LM Goodness-of-Fit Tests with Application to Stock Returns," unpublished Ph.D. dissertation, Ohio State University, 2006.

Shephard, Neil, "Local Scale Models: State Space Alternative to Integrated GARCH Processes," *J. Econometrics* **60** (1994): 181-202.

Uhlig, Harald, "Bayesian Vector Autoregressions with Stochastic Volatility," *Econometrica* **65** (1997): 59-73.

Weitzman, Martin L., "Risk, Uncertainty, and Asset-Pricing 'Antipuzzles'," Harvard University, Feb. 15, 2006. Nov. 9, 2005 version of this working Paper online at <http://post.economics.harvard.edu/faculty/weitzman/papers/VersionZBayesLimit.pdf>.