The Local Scale Model 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 model by assigning equal variance to the innovations to the volatility. The implied volatility gain at first declines sharply as in the classical Local Level Model, rather than being constant throughout as in traditional IGARCH (McCulloch 1985a; Engle and Bollerslev 1986).

The improved model is fit to monthly CRSP Value-weighted stock returns. The Maximum Likelihood estimate of the variance of the volatility shocks implies an upper bound of 7.76 on the 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.

Although the Student t densities generated by the Gaussian-based LSM account for much of the conditional leptokurtosis in the data, further research is required to adequately model the pronounced negative skewness in the returns.

The author is indebted to Martin Weitzman for crucial inspiration.

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

Keywords: Volatility clustering; State-space model; GARCH; Stock returns
1. Introduction and summary

The Local Scale Model (LSM) of Shephard (1994) is a state-space model of volatility clustering that models the unobserved volatility as evolving stochastically with beta-distributed shocks that are realistically independent of the observed errors themselves, rather than being 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 Student t densities, without the tedious numerical integrations that are ordinarily required by non-Gaussian state-space models.

The present paper improves upon the original Shephard 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).

The paper also goes beyond Shephard by fitting the model to empirical data, namely 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 nimbly to the observed return shocks. 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 the pronounced negative skewness in the stock returns. As a result, the Maximum Likelihood (ML) estimate of the mean excess return far exceeds the Weighted Least Squares (WLS) estimate using LSM-generated weights. Suggestions are offered for further research along these lines.

2. The Local Scale Model

Shephard (1994) models a time series $y_t$ as being Gaussian, conditional on a known mean $\mu$ and an unobserved time-varying precision, or reciprocal variance, $\theta$: 
$$ y_t \mid \theta_t \sim N(\mu, 1/\theta_t). $$
(1)
Conditional on last period’s experience $Y_{t-1} = \{y_1, ..., y_{t-1}\}$, last period’s precision is assumed to have a gamma distribution, with count parameter $a_{t-1}$ and intensity $b_{t-1}$ (see appendix):
The precision is assumed to evolve over time with beta-distributed multiplicative shocks \( \eta_t \). Generalizing Shephard’s notation somewhat so as to permit modifications,\(^1\) I set

\[
\theta_t = \theta_{t-1} \eta_t,
\]

\[
\eta_t \sim k_t B(a'_{t-1}, a_{t-1} - a'_{t-1}).
\]

for some \( a'_{t-1} < a_{t-1} \) 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 \( Y_t \), \( \theta_t \) also has a gamma distribution:

\[
\theta_t \mid Y_t = \theta_t \mid y_t, Y_{t-1}, \varepsilon_t \sim \text{Gamma}(a_t, b_t),
\]

with parameters

\[
a_t = a'_{t} + .5, \\
b_t = b'_t + \varepsilon_t^2 / 2.
\]

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

\[
v_t = 1 / E_{t-1}, \theta_t = b'_{t} / a'_{t}
\]

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:

\[
y_t \mid Y_{t-1} \sim \sqrt{v_t} T(d_t) + \mu .
\]

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

\[
v_t = \lambda_t v_{t-1} + \gamma_t \varepsilon_{t-1}^2,
\]

where

\[
\lambda_t = \frac{d_{t-1}}{k_t d_t}, \quad \gamma_t = \frac{1}{k_t d_t}.
\]

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

\(^1\) Note that whereas Shephard’s “\( \eta_t \)” 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 \( a_{\beta,1} \).
Shephard (1994) notes that in order to prevent the precision $\theta_t$ from converging in
probability to either 0 or $+\infty$, it is necessary to set
$$E \log(\theta_t / \theta_{t-1}) = E \log \eta_t = 0.$$ \hfill (6)

This in turn requires (see appendix)
$$\log k_t = \Psi(a_{t-1}) - \Psi(a'_t),$$
or equivalently,
$$k_t = \exp(\Psi((a_{t-1} + 1)/2) - \Psi(d_t / 2)),$$ \hfill (7)

where $\Psi(a)$ is the digamma function, defined by
$$\Psi(a) = \frac{d}{da} \ln \Gamma(a).$$ \hfill (8)

The variance of the volatility shocks $\log \eta_t$ is in general given (see appendix) by
$$\operatorname{var} \log \eta_t = \Psi_1(a'_t) - \Psi_1(a_{t-1}),$$ \hfill (9)

where $\Psi_1(a)$ is the trigamma function, defined by
$$\Psi_1(a) = \frac{d^2}{da^2} \ln \Gamma(a).$$ \hfill (10)

Shephard (1994) assumes that $a'_t$ is some fixed constant $\omega < 1$ times $a_{t-1}$. However, this
specification implies that var $\log \eta_t$ 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
McCulloch 2005), that these shocks are homoskedastic, with constant variance $v_{\log \eta}$.

This in turn implies
$$a'_t = \Psi_1^{-1}(\Psi_1(a_{t-1}) + v_{\log \eta}),$$
or terms of the predictive DOF $d_t$,
$$d_t = 2\Psi_1^{-1}(\Psi_1((d_{t-1} + 1)/2) + v_{\log \eta}).$$ \hfill (11)

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

The familiar Local Level Model 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 $a_0$ 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
$$a_1 = 1/2, \quad b_1 = e^2_1 / 2.$$  

The first condition is equivalent to setting $a'_1 = 0$, which in turn implies $b'_1 = 0$ and
therefore $b_1 = e^2_1 / 2$ for any choice of $v_1$. Equivalently, we may simply set

---

2 The DIGAMMA function is supported by GAUSS, as is the TRIGAMMA function employed below.
3 On the spelling of heteroskedasticity, see McCulloch (1985b).
4 E.g. GAUSS.
\[ d_i = 0, \quad v_i = 0 \]  \hspace{1cm} (12)
in the recursions (4), (7), (11). Since the gamma count parameter, or equivalently the \( \chi^2 \) DOF, measures the precision of the estimate of the volatility, initializing either of these to zero is equivalent to starting with no information at all.\(^5\)

If \( v \log \eta \) is 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 \( v \log \eta > 0 \). 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 \( d_\infty \) determined by the unique fixed point of (11):
\[
\Psi_1 (d_\infty / 2) = \Psi_1 ((d_\infty + 1) / 2) + v \log \eta.
\]  \hspace{1cm} (13)

It can be shown graphically, if not analytically, that for small values of \( v \log \eta \),
\[ d_\infty \approx \sqrt{2 / v \log \eta} + 1. \]

The solid line in Figure 1 below shows the first 20 values of \( d_t \), using \( v \log \eta = .03744 \), 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.

\(^5\) Taking the limit of the gamma density for \( \theta | Y_0 \) as \( a' \) falls to 0 while holding \( v_1 = a_1' / b_1' \) constant yields an uninformative improper prior density for \( \theta \) that is proportional to \( 1 / \theta \).
Under Shephard’s specification \( \alpha_t = \omega \alpha_{t-1} \) for some constant \( \omega < 1 \), the predictive DOF obey \( d_t = \omega(d_{t-1}) + 1 \) and will eventually approach the asymptotic value \( d_\infty = 1/(1-\omega) \). My \( \nu_{\log \eta} \) thus replaces Shephard’s \( \omega \) 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 \( d_\infty \). 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.

Figure 2 below shows the gain \( \gamma_t \) and attrition \( \lambda_t \) from the GARCH-like recursion for \( \nu_t \) in (4) and (5), using the uninformative prior (12) and the same \( \nu_{\log \eta} \) 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.
For any values of the two hyperparameters $\mu$ and $\nu_{\log \eta}$, the log-likelihood implied by the modified LSM model (4), (5), (7), (11), (12) becomes

$$L(\mu, \nu_{\log \eta}) = \sum_{t=2}^{n} \log \left( t_d \left( \frac{\varepsilon_t}{\sqrt{\nu_t}} \right) / \sqrt{\nu_t} \right)$$

where

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

is the standard Student t density with $d$ DOF.\(^6\) This may be maximized numerically to obtain Maximum Likelihood (ML) estimates of $\mu$ and $\nu_{\log \eta}$ when, as is ordinarily the case, these are not really known, as has been assumed to this point. Since the Student t density 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

$$\text{se}(\hat{\mu}_{ML}) = \left( \frac{\partial^2}{\partial \mu^2} L(\mu, \nu_{\log \eta}) \right)^{-1/2}.$$

\(^6\) $\log \Gamma(a)$ may be computed to adequate precision in GAUSS as $\text{LN} (\text{GAMMA}(a))$ for $a < 7$, and as $\text{LNFACT}(a-1)$ for $a \geq 7$. Neither function is adequate over the entire required range, however.
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
\[ \sigma_t^2 = \delta + \sum_{j=1}^{\rho} \gamma_j \varepsilon_{t-j}^2. \]
In this model, the shocks that are driving 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,
\[ \sigma_t^2 = \delta + \lambda \sigma_{t-1}^2 + \gamma \varepsilon_{t-1}^2, \]
the unconditional variance
\[ \mathbb{E}\varepsilon_t^2 = \delta/(1-\lambda-\gamma) \]
is finite when \(\lambda + \gamma < 1.\)

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 originally thought that the further restriction \(\delta = 0\) would then be required as in the AR\((1)\) case, to prevent convergence in probability to infinity. The original IGARCH model\(^9\) was thus

\(\text{Bollerslev (1986) in fact considered a GARCH}\((p,q)\) model with } p \text{ lags of the squared errors and } q \text{ of the variance, but } p = q = 1 \text{ is ordinarily adequate.} \)

\(\text{In GARCH\((1,1)\), the joint probability of } y_1, \ldots, y_n \text{ unfortunately depends on the unobserved initial variance } \sigma_1^2. \text{ In principle, if the process is strictly stationary, as it is even for } \lambda + \gamma = 1, \text{ the unconditional likelihood could be found by first finding the unconditional distribution of } \sigma_1^2 \text{ by iterating numerically on} (14), \text{ 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 } \sigma_1^2 \text{ 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.} \)

\(\text{McCulloch (1985) called essentially this model “Adaptive Conditional Heteroskedasticity” (ACH, to be pronounced as } \text{ach!} \text{ in German), but Bollerslev’s (1986) “IGARCH” caught on instead. Since McCulloch}\)
\[ \sigma_i^2 = (1 - \gamma)\sigma_{i-1}^2 + \gamma e_{i-1}^2 \] (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 \( \sigma_i^2 \) increases without bound, yet \( \sigma_i^2 \) is bounded below by zero, this requires that virtually all the density must eventually converge to near 0. In order to prevent \( \sigma_i^2 \) from invariably collapsing on 0 or exploding to infinity, its log must be a martingale. This in turn requires, by Jensen’s inequality, that \( \sigma_i^2 \) 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):

\[ \sigma_i^2 = \delta + (1 - \gamma)\sigma_{i-1}^2 + \gamma e_{i-1}^2 \] (16)

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

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 bounded by \( d_\alpha \), rather than a Gaussian distribution as in ARCH or GARCH. As Weitzman (2006) points out, this has very dramatic implications for the equity premium, when the model is used to fit continuously compounded returns.

In the modified LSM of Section 2, \( v_t = 1/E \theta_t \) follows the GARCH-like recursion (4) with no intercept term, and with time-varying coefficients \( \lambda_t \) and \( \gamma_t \) 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,

(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 \( \alpha \) 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.
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 \( d_t \leq 2 \), and

\[
h_t = E(e_t^2 | Y_{t-1}) = E(1/\theta_t | Y_{t-1}) = b_t'(a_t' - 1) = v_t d_t / (d_t - 2)
\]

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

\[
h_t = \lambda_t h_{t-1} + \gamma^* t e_{t-1}^2,
\]

where

\[
\gamma^* = \gamma_t d_t / (d_t - 2).
\]

The coefficients \( \lambda_t \) and \( \gamma_t \) (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 \( a_{\infty} \) (Uhlig’s \( \nu \)) and \( k_{\infty} / \lambda_{\infty} \) in Uhlig’s notation. He notes (p. 61), following Shephard (1994), that the process will tend a.s. to 0 or \( \infty \) unless \( k_{\infty} \) 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 \( E(\theta_t) = \theta_{t-1} \). Rather than estimating his \( \nu \) 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 about the mean can be taken as an equivalent loss function for evaluating the mean estimate. However, even if the errors are Gaussian, the squared errors are scaled \( \chi^2_t \), which is far from Gaussian. Likewise, if (as in HS’s best GARCH-like model) the scaled errors are Student t with \( \nu \) 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.
4. Application to Stock Returns

Figure 3 plots monthly continuously compounded CRSP Value-Weighted 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 are already continuously compounded and based on 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 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 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.10

---

10 The most notable of these 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.
The mean excess return is .005105/month (se = .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

\[
\hat{\mu}_{ML} = 0.008662/\text{mo.} = 10.39%/\text{yr.}
\]

\[
(\text{se}) \quad (0.001320) \quad (1.58)
\]

\[
\hat{\eta}_{\log v} = 0.03744
\]

\[
\sqrt{\hat{\eta}_{\log v}^2} = 0.1935
\]

\[
d_n = 7.756
\]

\[
d_{\infty} = 7.759.
\]

Figure 4 shows the demeaned excess returns, along with the local scale \( \sqrt{v_t} \). 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.

![Demeaned excess returns, local scale](image)

**Figure 4**

Hamilton and Susmel (1994, pp. 314-6) criticize GARCH-type models for overestimating the squared errors during the weeks immediately following the 10/87
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.

5. Analysis of residuals

Figure 5 depicts the scale-adjusted errors $\tilde{\epsilon}_i = \hat{\epsilon}_i / \sqrt{v_i}$. Even though the errors are assumed to be Gaussian conditional on the unobserved precision $\theta_t$, 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.

![Scale-adjusted errors](image)

Figure 5

It is obvious from Figure 5 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 use of 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 \( \hat{u}_i = T_{d_i}(\hat{\xi}) \) should be iid U(0,1), where \( T_d(\cdot) \) represents the Student t cumulative distribution function with \( d \) DOF.\(^{11}\) Figure 6 depicts these transformed errors. The issue then is whether too many of the transformed errors are clustered near the bottom of this diagram, and too few near the top, for them to be uniformly distributed.

![Figure 6](Image)

There are two ways to compare the distribution of the transformed errors in Figure 6 to a uniform distribution: in terms of their CDF or in terms of a histogram of their PDF. Figure 7 shows the PP plot of the transformed errors \( \hat{u}_i \), i.e. their empirical

\(^{11}\) Strictly speaking, this is only true if the two hyperparameters \( \mu \) and \( \nu_{\log} \) are known. If they have been estimated, the transformed residuals will tend to look even more uniform than they should. See Percy (2006) and footnote 12 below.
CDF versus the unit interval. It may be seen that indeed too many errors accumulate in the lower portion of the distribution, while there are too few in the upper portion.

Figure 7

Figure 8 gives the histogram of the transformed errors, using 10 equally spaced bins, along with a horizontal line at 93.7, the expected frequency per bin. Again, there is too much density in the first two bins, and too little in the last bin, with compensating deviations from uniformity in the 3rd, 4th, 8th, and 9th bins.
The Pearson chi-square test uses frequency counts such as those in Figure 8 to test that the deviation from uniformity is not just sampling error. However, the Pearson test implicitly assumes as its alternative hypothesis that the density is constant in each bin, and then changes abruptly to a completely dissimilar 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 indeed found the Neyman Smooth Test to have much more power for discriminating among heavy-tailed probability distributions than the Pearson test.

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

$$LM = s'Y^{-1}s,$$
where \( s = (s_1, \ldots, s_k)' \) is the score vector (the gradient of the log-likelihood), and \( I = (I_{jj}) \) is the information matrix, both evaluated under the null of uniformity. For a sample of size \( n \), these are

\[
s_j = \sum_{i=1}^{n} \hat{u}_i^j - n/(j+1),
\]

\[
I_{jj} = n((j + j' + 1)^{-1} - (j + 1)^{-1}(j' + 1)^{-1}).
\]

(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.) Under the null of uniformity, the LM statistic is asymptotically \( \chi^2 \) when the parameters are known.

Although it is not obvious what particular value of \( k \) would be ideal, we are concerned with potential skewness in the underlying errors and therefore want to consider \( k \geq 3 \). The Neyman LM statistic is given in Table 1 for \( k = 3, \ldots, 10 \), along with the corresponding \( \chi^2 \) \( p \)-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 \( k \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>LM</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21.91</td>
<td>.00007</td>
</tr>
<tr>
<td>4</td>
<td>22.47</td>
<td>.00016</td>
</tr>
<tr>
<td>5</td>
<td>25.83</td>
<td>.00010</td>
</tr>
<tr>
<td>6</td>
<td>25.92</td>
<td>.00023</td>
</tr>
<tr>
<td>7</td>
<td>26.77</td>
<td>.00037</td>
</tr>
<tr>
<td>8</td>
<td>27.36</td>
<td>.00061</td>
</tr>
<tr>
<td>9</td>
<td>30.92</td>
<td>.00031</td>
</tr>
<tr>
<td>10</td>
<td>31.07</td>
<td>.00057</td>
</tr>
</tbody>
</table>

Because the two hyperparameters \( \mu \) and \( \nu \) are in fact not known, but have been estimated from the data, \( \chi^2 \) 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).

---

12 Pearson 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.

13 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.
The Best Linear Unbiased Estimate (BLUE) of the mean, if not the best global estimate, is given by Weighted Least Squares,

\[ \hat{\mu}_{WLS} = \frac{\sum_{t=1}^{n} y_t / h_t}{\sum_{t=1}^{n} 1 / h_t}, \]  

(18)

using weights determined by (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

\[ \hat{\mu}_{WLS} = 0.005506/\text{mo.} = 6.607 \%/\text{yr.} \]

(\( \text{se} \) \( 0.001351 \)) \( (1.621) \).

Although 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 2.33 standard errors.

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.

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

---

14 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.

15 It may also be convenient to use these weights when the mean and/or regression parameters themselves are estimated by a pseudo-Gaussian signal extraction process as in McCulloch (2005).
any deep rethinking of the model, but again would go far beyond the scope of the present study.
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 $\chi^2$ RV with $d = 2a$ DOF, and scaled by $1/2b$:

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

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

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

where the beta function $B(\alpha, \beta)$ is given by

$$B(\alpha, \beta) = \int_{0}^{1} x^{\alpha-1}(1-x)^{\beta-1} \, dx$$

$$= \Gamma(\alpha)\Gamma(\beta) / \Gamma(\alpha, \beta).$$

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 $X \sim G(a, b)$ and $Y \sim \text{Beta}(a', a-a')$, with $a' < a$, their product $Z$ is again gamma, but with reduced count parameter $a'$:

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

If $X \sim \text{Beta}(\alpha, \beta)$, the characteristic function of $Y = \log(X)$ is

$$c_fY(t) = E e^{itY}$$

$$= \frac{1}{B(\alpha, \beta)} \int_{0}^{1} e^{it\log(x)} x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$$= \frac{1}{B(\alpha, \beta)} \int_{0}^{1} x^{\alpha+it-1} (1-x)^{\beta-1} \, dx$$

$$= B(\alpha + it, \beta) / B(\alpha, \beta).$$

It follows that
\[ E Y = i^{-1} \frac{d}{dt} \left. \text{cf}_Y(t) \right|_{t=0} = \Psi(\alpha) - \Psi(\alpha + \beta), \]

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

\[ \text{var} Y = i^{-2} \left. \frac{d^2}{dt^2} \text{cf}_Y(t) \right|_{t=0} = \Psi_1(\alpha) - \Psi_1(\alpha + \beta), \]

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

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

\[
E(1/X) = \int_0^\infty x^{-1} x^{a-1} \exp(-x) / \Gamma(a) dx
= \frac{\Gamma(a-1)}{\Gamma(a)}
= \frac{1}{(a-1)}.
\]

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

\[
E(1/X) = b E(1/G(a,1))
= b/(a-1)
= a^{-1} E X.
\]

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

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

for \( d > 2 \), and infinity otherwise.
References:


