# Dynamic Models for Volatility and Heavy Tails

with applications to financial time series
Chapter 1.

Andrew Harvey,
Faculty of Economics, Cambridge University
and Fellow of Corpus Christi College.

April 13, 2012

## 1 Introduction

The aim of this monograph is to set out a unified and comprehensive theory for a class of nonlinear time series models that can deal with distributions that change over time. The emphasis is models in which the conditional distribution of an observation may be heavy-tailed and the location and/or scale changes over time. The defining feature of these models is that the dynamics are driven by the score of the conditional distribution. When a suitable link function is employed for the dynamic parameter, analytic expressions may be derived for (unconditional) moments, autocorrelations and moments of multi-step forecasts. Furthermore a full asymptotic distributional theory for maximum likelihood estimators can be obtained, including analytic expressions for the asymptotic covariance matrix of the estimators.

The class of what we call *dynamic conditional score* (DCS) models includes standard linear time series models observed with an error which may be subject to outliers, models which capture changing conditional variance and models for non-negative variables. The last two of these are of considerable importance in financial econometrics where they are used for forecasting volatility. A guiding principle underlying the proposed class of models is that of signal extraction. When combined with basic ideas of maximum likelihood estimation, the signal extraction approach leads to models which, in contrast to many in the literature, are relatively simple in their form and yield analytic expressions for their principal features.

For estimating location, DCS models are closely related to the unobserved components models described in Harvey (1989). Such models can be handled using state space methods and they are easily accessible using the STAMP package of Koopman et al (2008). For estimating scale, the models are close to stochastic volatility models, where the variance is treated as an unobserved component. The close ties with unobserved component and stochastic volatility models provides insight into the structure of the DCS models, particularly with respect to modeling trend and seasonality, and into possible restrictions on the parameters.

# 2 Unobserved components and filters

Autoregressive and autoregressive integrated moving average (ARIMA) models focus on forecasting future values of a series. A more general framework is given by the signal plus noise paradigm. Signal extraction is of interest in itself and once the problem has been solved, the forecasting solution follows.

A simple Gaussian signal plus noise model is

$$\begin{gathered} y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right), \quad t = 1, ..., T \qquad (1) \\ \mu_{t+1} = \phi\mu_t + \eta_t, \quad \eta_t \sim NID(0, \sigma_\eta^2), \end{gathered}$$

where the irregular and signal disturbances, $\varepsilon_t$ and $\eta_t$ respectively, are mutually independent and the notation $NID\left(0, \sigma^2\right)$ denotes normally and independently distributed with mean zero and variance $\sigma^2$. The autoregressive parameter is $\phi$, while the signal-noise ratio, $q = \sigma_\eta^2/\sigma_\varepsilon^2$, plays the key role in determining how observations should be weighted for prediction and signal extraction. The reduced form (RF) of (1) is an ARMA(1,1) process

$$y_t = \phi y_{t-1} + \xi_t - \theta\xi_{t-1}, \quad \xi_t \sim NID\left(0, \sigma^2\right), \quad t = 1, ..., T \qquad (2)$$

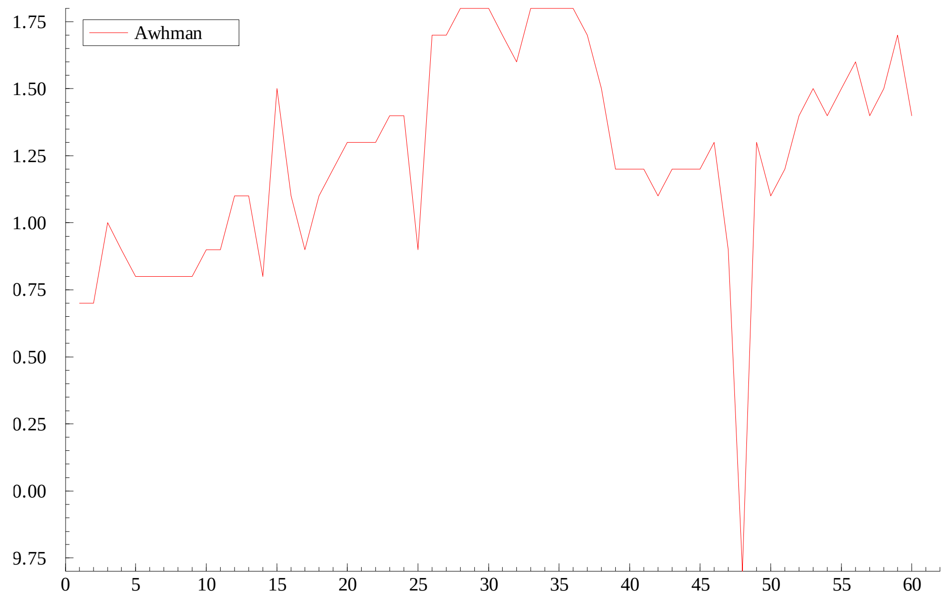
but with restrictions on $\theta$. For example, when $\phi = 1$, $0 \leq \theta \leq 1$. The forecasts from the unobserved components (UC) model and reduced form are the same.

The UC model in (1) is effectively in state space form (SSF) and, as such, it may be handled by the Kalman filter (KF); see Harvey (1989). The parameters $\phi$ and $q$ can be estimated by maximum likelihood, with the likelihood function constructed from the one-step ahead prediction errors. The KF can be expressed as a single equation which combines the estimator of $\mu_t$ based on information at time $t-1$ with the $t-th$ observation in order to produce the best estimator of $\mu_{t+1}$. Writing this equation together with an equation that defines the one-step ahead prediction error, $v_t$, gives the innovations form (IF) of the KF:

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t, \quad t = 1, ..., T, \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + k_t v_t. \end{aligned} \tag{3}$$

The Kalman gain, $k_t$, depends on $\phi$ and $q$. In the steady-state, $k_t$ is constant. Setting it equal to $\kappa$ in (3) and re-arranging gives the ARMA model,(2), with $\xi_t = v_t$ and $\phi - \kappa = \theta$. A pure autoregressive model is a special case in which $\kappa = \phi$, so that $\mu_{t|t-1} = \phi y_{t-1}$.

Now suppose that the noise in a UC model comes from a heavy tailed distribution, such as Student's t. Such a distribution can give rise to observations which, when judged against the yardstick of a Gaussian distribution, are considered to be outliers. Figure 2 illustrates a situation of this kind. In the case of (1), the reduced form is still an ARMA(1,1) process, but the $\xi_t's$ in (2) are not independent and identically distributed (IID) and quasi-maximum likelihood (QML) estimation- that is estimation carried out under the assumption of normality - is inefficient. Approximating by a pure AR is even more problematic. (If, as may happen with a heavy-tailed distribution, the variance of $\varepsilon_t$ does not exist, the QML estimator may not even be consistent.)



Monthly observations on Canadian manufacturing.

Allowing the $\xi_t's$ to have a heavy-tailed distribution does not deal with the problem as a large observation becomes incorporated into the level and takes time to work through the system. To be specific, the AR representation of (2) is

$$y_t = (\phi - \theta) \sum_{j=1}^{\infty} \phi^{j-1} y_{t-j} + \xi_t = \mu_{t|t-1} + \xi_t.$$

If the $t-th$ observation is contaminated by adding an arbitrary amount, $C$, then after $\tau$ periods, the prediction of the next observation is still contaminated by $C$ since it contains the component $(\phi - \theta)\phi^{\tau} C$.

An ARMA or AR model in which the disturbances are allowed to have a heavy-tailed distribution is designed to handle *innovations outliers*, as opposed to *additive outliers*. There is a good deal of discussion of outliers, and how to handle them, in the robustness literature; see, for example the book by Maronna, Martin and Yohai (2006, ch 8) and the recent paper by Muler, Pena and Yohai (2009) on robust estimation for ARMA models. The argument in this monograph is that a model-based approach is not only simpler, both conceptually and computationally, than the usual robust methods, but is also more amenable to diagnostic checking and generalization.

Simulation methods, such as Markov chain Monte Carlo (MCMC) and particle filtering, provide the basis for a direct attack on models that are nonlinear and/or non-Gaussian. The aim is to extend the Kalman filtering and smoothing algorithms that have proved so effective in handling linear Gaussian models. Considerable progress has been made in recent years; see Durbin and Koopman (2012). However, the fact remains that simulation-based estimation can be time-consuming and subject to a degree of uncertainty. In addition the statistical properties of the estimators are not easy to establish.

The approach here begins by writing down the distribution of the $t-th$ observation, conditional on past observations. Time-varying parameters are then updated by a suitably defined filter. Such a model is what Cox (1981) called *observation driven.* In a linear Gaussian UC model, which is called *parameter driven* in Cox's terminology, the KF is driven by the one step-ahead prediction error, as in (3). The main ingredient in the filter developed here for non-Gaussian distributions is the replacement of $v_t$ in the KF equation by a variable, $u_t$, that is proportional to the score of the conditional distribution, that is the logarithm of the probability density function at time $t$ differentiated with respect to $\mu_{t|t-1}$. Thus the second equation in (3) becomes

$$\mu_{t+1|t} = \phi\mu_{t|t-1} + \kappa u_t$$

where $\kappa$ is treated as an unknown parameter.

Why the score ? If the signal in (1) were fixed, that is $\phi = 1$ and $\sigma_\eta^2 = 0$, $\mu_{t+1} = \mu$, the sample mean, $\widehat{\mu}$, would satisfy the condition

$$\sum_{t=1}^{T}(y_t - \widehat{\mu}) = 0.$$

The maximum likelihood (ML) estimator is obtained by differentiating the log-likelihood function with respect to $\mu$ and setting the resulting derivative, the score, equal to zero. When the observations are normally distributed, the ML estimator is the same as the sample mean, the moment estimator. However, for a non-Gaussian distribution, the moment estimator and the ML estimator differ. Once the signal in a Gaussian model becomes dynamic, as in (1), its estimate can be updated with each new observation using the Kalman filter. With a non-normal distribution exact updating is no longer possible, but the fact that ML estimation in the static case sets the score to zero provides the rationale for replacing the prediction error, which has mean zero, by the score, which for each individual observation, also has mean zero. The resulting filter might therefore be regarded as an approximation to the computer intensive solution for the UC model and the evidence presented later lends support to this notion.

The attraction of treating the filter driven by the score of the conditional distribution as a model in its own right is that it becomes possible to derive the asymptotic distribution of the ML estimator and to generalize in various directions. Thus the same approach can then be used to model scale, using an exponential link function, and to model location and scale for non-negative variables. The first equation in (3) is then nonlinear. The justification for the class of dynamic conditional score models is not that they approximate corresponding UC models, but rather that their statistical properties are both comprehensive and straightforward.

The use of the score of the conditional distribution to robustify the KF was originally proposed by Masreliez (1975). However, it has often been argued that a crucial assumption made by Masreliez (concerning the approximate normality of the prior at each time step) is, to quote Schick and Mitter (1994, p 1054), '..insufficiently justified and remains controversial.' Nevertheless, the procedure has been found to perform well both in simulation studies and with real data.' Schick and Mitter (1994) suggest a generalization of the Masreliez filter based on somewhat stronger theoretical foundations. The observation noise is assumed to come from a contaminated normal distribution and the resulting estimator employs banks of Kalman filters and optimal smoothers weighted by posterior probabilities. As a result it is considerably more complicated than the Masreliez filter. Once the realm of computationally intensive techniques has been entered it seems better to follow adopt the simulation based methods alluded to earlier.

The situations tackled by Masreliez are more complicated than those considered here because the system matrices in the state space model may be time-varying. The models in this monograph are simpler in structure and as a result the use of the score to drive the dynamics can be put on much firmer statistical foundations.

# 3 Independence, white noise and martingale differences

The study of models that are not linear and Gaussian requires a careful distinction to be made between the concepts of independence, uncorrelatedness and martingale differences. The definitions are as follows.

(a) *White noise* (WN) variables are serially uncorrelated with constant mean and variance.

(b) *A martingale difference* has a zero (or constant) conditional expectation, that is

$$\underset{t-1}{E}\left(y_t\right) = E\left(y_t \mid Y_{t-1}\right) = 0.$$

It is also necessary for the unconditional expectation of the absolute value to be finite, that is $E\left|y_t\right| < \infty$; see Davidson (2000, p 121-2).

(c) *Strict white noise* variables are independent and identically distributed (IID).

The relationship between the two types of white noise and martingale differences is as follows;

i) All zero mean independent sequences are MDs but not the converse.

ii) All MDs are WN, provided that the variance is finite. The converse is not true.

Finally note that when a variable is normally distributed, the distinction between WN, strict WN and MDs disappears, the reason being that a normal distribution is fully described by its first two moments. Thus Gaussian white noise is strict white noise.

A crucial element in the understanding of the above concepts, and in the statistical derivations that follow, is the *Law of Iterated Expectations* (LIE) which states that, if $g\left(y_t\right)$ is a function of $y_t$, then an expected value several steps ahead can be found from the sequence of one-step ahead expectations. Thus

$$\underset{t-J}{E}\left[g\left(y_t\right)\right] = \underset{t-J}{E} \cdots \underset{t-1}{E}\left[g\left(y_t\right)\right], \quad J = 2, 3, ...$$

The unconditional expectation is found by letting $J \to \infty$. For predicting a function of the observation at time $T + \ell$ at $T$, set $t = T + \ell$ and $J = \ell$ so

$$\underset{T}{E}\left[g\left(y_{T+\ell}\right)\right] = \underset{T}{E} \cdots \underset{T+\ell-1}{E}\left[g(y_{T+\ell})\right]$$

The proof can be found in many introductory time series and econometrics texts.

The LIE enables us to show that all MD's have zero unconditional mean and are serially uncorrelated. Specifically

$$E\left(y_t\right) = E\left[E\left(y_t \mid Y_{t-1}\right)\right] = 0$$

and $y_t$ is uncorrelated with any function of past observations because

$$E\left[y_t f\left(Y_{t-1}\right) \mid Y_{t-1}\right] = f\left(Y_{t-1}\right) E\left(y_t \mid Y_{t-1}\right) = 0.$$

Hence the unconditional expectation of $y_t f\left(Y_{t-1}\right)$ is zero.

A WN sequence is not necessarily a MD because there may be a nontrivial nonlinear predictor. For example, the model

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2}, \quad \varepsilon_t \sim IID(0, \sigma^2)$$

where $\varepsilon_0$ and $\varepsilon_{-1}$ are fixed and known, is white noise, but not a MD as $E\left(y_{T+1} \mid Y_T\right) = \beta\varepsilon_T\varepsilon_{T-1}$.

A linear process is usually defined as one which can be written as an infinite moving average in IID disturbances, with zero mean and constant variance, $\sigma^2$, with the sum of the squares of the coefficients being finite, that is

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty, \quad \varepsilon_t \sim IID(0, \sigma^2) \tag{4}$$

More generally a linear process may be defined as a linear combination of past observations and/or strict white noise disturbances, with appropriately defined initial conditions. (Though even this is not straightforward - see see Terasvirta et al (2010, p 1-2). For a stationary process, the representation in (4) means that all the information about the dynamics is in the autocorrelation function (ACF). Furthermore the minimum mean square error (MMSE) predictor of $y_{T+\ell}$ is linear and its MSE is $\sigma^2 \sum_{j=0}^{\ell-1} \psi_j^2$. However, unless the disturbances are Gaussian, a linear model is of limited value since it is not usually possible to derive the unconditional distribution or the multi-step predictive distribution. On the other hand the optimal forecasts in a model which is a linear function of current and past MDs is the same as if the MDs were strict WN and, if the conditional variances are constant, the MSE is the same.

# 4 Volatility

If markets are working efficiently, financial returns are MDs. In other words they should not be predictable on the basis of past information. However, returns are not usually independent and so features of the conditional distribution apart from the mean may be predictable. In particular non-trivial predictions can be made for the variance or, more generally, the scale.

## 4.1 Stochastic Volatility

The variance in Stochastic Volatility (SV) models is driven by an unobserved process. The first-order model for $y_t$, $t = 1, .., T$, is

$$\begin{gathered} y_t = \mu_t + \sigma_t \varepsilon_t, \qquad \sigma_t^2 = \exp\left(2\lambda_t\right), \qquad \varepsilon_t \sim IID\left(0,1\right) \\ \lambda_{t+1} = \delta + \phi\lambda_t + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma_\eta^2\right) \end{gathered} \tag{5}$$

with $\varepsilon_t$ and $\eta_t$ mutually independent. Leverage effects, which enable $\sigma_t^2$ to respond asymmetrically to positive and negative values of $y_t$, can be introduced by allowing these disturbances to be correlated, as in Harvey and Shephard (1996). Shephard and Andersen (2009) discuss the relationship between SV models and continous time models in the finance literature.

The *exponential link function* ensures that the variance remains positive and the restrictions needed for $\lambda_t$ and $y_t$ to be stationary are straightforward, that is $|\phi| < 1$. Furthermore, analytic expressions for moments and ACFs of the absolute values of the observations raised to any power can be derived. Instead of using an exponential link function, we could have

$$\sigma_{t+1}^2 = \delta + \phi\sigma_t^2 + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma_\eta^2\right)$$

but with $\sigma_t$ taken to be the positive square root of $\sigma_t^2$. However, this model, which corresponds to the square root process in continuous time, is less satisfactory.

Unfortunately, direct maximum likelihood (ML) estimation of the SV model is not possible. A quasi-maximum likelihood (QML) procedure can be based on the linear state space form obtained by taking logarithms of the absolute values of the observations, corrected for the mean, to give the following measurement equation:

$$\ln|y_t| = \lambda_t + \ln|\varepsilon_t|, \quad t = 1, .., T. \tag{6}$$

The parameters in the model are then estimated by using the Kalman filter, as in Harvey, Ruiz and Shephard (1994). However, there is a loss in efficiency because the distribution of $\ln|\varepsilon_t|$ is far from Gaussian. Efficient estimation can be achived by computer intensive methods, as described in Durbin and Koopman (2001) and elsewhere.

## 4.2 Generalized Autoregressive Conditional Heteroscedasticity (GARCH)

The Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, introduced, as ARCH, by Engle (1982) and generalized by Bollerslev (1986) and Taylor (1986), is the classic way of modeling changes in the volatility of returns. It does so by letting the variance be a linear function of past squared observations. In the first-order case, the GARCH $(1,1)$ model, with the mean of the observations, $y_t$, $t = 1, .., T$, assumed to be zero, is

$$y_t = \mu_{t|t-1} + \sigma_{t|t-1}\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \tag{7}$$

and

$$\sigma^2_{t|t-1} = \delta + \beta\sigma^2_{t-1|t-2} + \alpha y^2_{t-1}, \quad \delta > 0, \beta \geq 0, \alpha \geq 0. \tag{8}$$

The conditions on $\alpha$ and $\beta$ ensure that the variance remains positive. The sum of $\alpha$ and $\beta$ is typically close to one and the Integrated GARCH (IGARCH) model is obtained when the sum is equal to one. The variance in IGARCH is an exponentially weighted moving average (EWMA) of past squared observations and, as such, is often used by practitioners.

The principal advantage of GARCH models over SV models is that, because they are observation driven, the likelihood function is immediately available. As noted in the previous sub-section, this is not the case with the parameter driven SV model.

The model may be extended by adding lags of the variance and the squared observations. Heavy tails are accomodated by letting the conditional distribution be t-distributed, as proposed by Bollerslev (1987). The $GARCH(1,1) - t$ model has become something of an industry standard.

Leverage effects, which enable $\sigma^2_{t|t-1}$ to respond asymmetrically to positive and negative values of $y_t$, are typically incorporated into GARCH models by including a variable in which the squared observations are multiplied by an indicator that takes a value of unity when an observation is negative and is zero otherwise; see Taylor (2005, p 220-1). The techique is often known as GJR, after the originators, Glosten, Jagannanthan and Runckle (1993).

The autocorrelation function (ACF) of squared observations may be obtained relatively easily as they obey an ARMA process. For example, for GARCH(1,1) with zero mean

$$y_t^2 = \gamma + \phi y_{t-1}^2 + v_t + \theta^* v_{t-1} \tag{9}$$

where $v_t$ is white noise and $\phi = \alpha + \beta$ and $\theta^* = -\beta$. The drawback to working with squared observations is that outlying observations seriously weaken the serial correlation and it is a well-established stylized fact that the autocorrelations of absolute values tend to be larger and hence provide a better vehicle for detecting dynamic volatility and assessing its nature; see Taylor (2005).

## 4.3 Exponential GARCH

Nelson (1991) introduced the exponential GARCH (EGARCH) model in which the dynamic equation for volatility is formulated in terms of the logarithm of the conditional variance in (7). The leading case is

$$\ln \sigma_{t|t-1}^2 = \delta + \phi \ln \sigma_{t-1|t-2}^2 + \alpha \left[ |\varepsilon_{t-1}| - E\,|\varepsilon_{t-1}| \right] + \alpha^* \varepsilon_{t-1}, \tag{10}$$

where $\alpha$ and $\alpha^*$ are parameters and, for a Gaussian model, $E\,|\varepsilon_t| = \sqrt{2/\pi}$. The role of $\varepsilon_t$ is to capture leverage effects. As in the SV model, the exponential link function ensures that the variance is always positive. Indeed the model has a structure similar to SV since $\alpha \left[ |\varepsilon_{t-1}| - E\,|\varepsilon_{t-1}| \right] + \alpha^* \varepsilon_{t-1}$ is a MD. Stationarity restrictions are similar to those in the SV model; for example, in the equation above $|\phi| < 1$. The exponential link permits models that would be problematic with GARCH because of the need to ensure a positive variance. In particular cycles and seasonal effects are possible.

Nelson (1991) notes that if the conditional distribution of the observations is Student's t, with finite degrees of freedom, the conditions needed for the existence of the moments of $\sigma_{t|t-1}^2$ and $y_t$ are rarely satisfied in practice. Hence the model is of little practical value since, without a first moment, even the sample mean is inconsistent. The lack of moments for Student's t and the fact that there is no asymptotic theory for ML has limited the application of EGARCH.

## 4.4 Variance, scale and outliers

Substituting in (8) gives an infinite autoregression in squared observations. In an ARCH(p) model, forecasts are made directly from a finite number of past squared observations - hence the name ARCH. From our perspective, the reason that GARCH is more plausible than ARCH(p) is that estimating variance is an exercise in signal extraction and as such the conditional variance cannot normally be a finite autoregression. The ARCH(1) model is particularly problematic as it is based on a single squared observation which is bound to be a poor estimator of variance.

A linear combination of past squares (even if infinite) may not be a good choice for modeling dynamics when the conditional distribution is non-Gaussian observations. This stems from the fact that the sample variance in a static model can be very inefficient. Indeed, for some heavy-tailed distributions, the variance may not exist. This difficulty may be avoided by modeling scale instead. Since scale is necessarily positive (as is variance), an exponential link function is appropriate. Furthermore a model for the logarithm of volatility may be regarded as an approximation to an SV model. This reasoning lead to Nelson proposing EGARCH. The only flaw was to use absolute values in the dynamic equation. Replacing the absolute value by the score resolves the difficulties.

Outliers present a practical problem for GARCH models, even if the conditional distribution is allowed to have heavy-tails, as in GARCH-t. The reason is that a large value becomes embedded in the conditional variance and typically takes a long time to work through. This is the same difficulty that was 'noted earlier in connection with additive outliers.

## 4.5 Location/scale models

Many variables are intrinsically non-negative. Examples in finance include duration, realized volatility and spreads; see, for example, Russell and Engle (2010) and Brownlees and Gallo (2010). Other situations in economics where distributions for non-negative variables are appropriate are in the study of incomes and the size of firms; the book by Kleiber and Kotz (2003) describes many case studies.

Engle (2002) introduced a class of *multiplicative error models* (MEMs) for time series modeling of non-negative variables. In these models, the conditional mean, $\mu_{t|t-1}$, is driven by a GARCH-type equation and so

$$y_t = \varepsilon_t \mu_{t|t-1}, \qquad 0 \leq y_t < \infty, \quad t = 1, ...., T,$$

where $\varepsilon_t$ has a distribution with mean one and, analogous to GARCH(1,1),

$$\mu_{t|t-1} = \delta + \beta \mu_{t-1|t-2} + \alpha y_{t-1}, \quad \delta > 0, \beta \geq 0, \alpha \geq 0. \tag{11}$$

The gamma distribution is often used for $\varepsilon_t$, with the exponential distribution being an important special case. The gamma distribution does not have a particularly heavy tail. However, other distributions, such as the Weibull and Burr, can have a heavy tail and observations that are outliers for a gamma distribution can become embedded in the predictions. Thus the linearity of (11) must be questioned, just as the use of a linear combination of squares was questioned for GARCH.

An exponential link function is sometimes used so as to ensure that $\mu_{t|t-1}$ remains positive; see, for example, Brandt and Jones (2006). However, an exponential link does not, in itself, deal with the problem noted at the end of the previous paragraph.

# 5 Dynamic conditional score models

An observation driven model is set up in terms of a conditional distribution for the $t - th$ observation. Thus

$$p(y_t | \lambda_{t|t-1}, Y_{t-1}), \qquad t = 1, ...., T \tag{12}$$

$$\lambda_{t+1|t} = g(\lambda_{t|t-1}, \lambda_{t-1|t-2}, ..., Y_t)$$

where $Y_t$ denotes observations up to, and including $y_t$, and $\lambda_{t|t-1}$ is a parameter that changes over time. The second equation in (12) may be regarded as a data generating process or as a way of writing a filter that approximates a nonlinear UC model. In both cases the notation $\lambda_{t+1|t}$ stresses its status as a parameter of the conditional distribution and as a filter, that is a function of past observations. The likelihood function for an observation driven model is immediately available since the joint density of a set of $T$ observations is

$$L(\boldsymbol{\psi}) = \prod_{t=1}^{T} p(y_t | \lambda_{t|t-1}, Y_{t-1}; \boldsymbol{\psi}),$$

where $\boldsymbol{\psi}$ denotes a vector of unknown parameters.

The first-order Gaussian GARCH model, (7) and (8), is an observation driven model in which $\lambda_{t|t-1} = \sigma^2_{t|t-1}$. (Andersen *et al* (2006) use the same notation for the conditional variance.) As such it may be written

$$y_t \mid Y_{t-1} \sim NID\left(0, \sigma^2_{t|t-1}\right)$$

$$\sigma^2_{t+1|t} = \delta + \phi\sigma^2_{t|t-1} + \alpha v_t, \quad \delta > 0,\ \phi \geq \alpha,\ \alpha \geq 0, \tag{13}$$

where $\phi = \alpha + \beta$ and $v_t = y_t^2 - \sigma^2_{t|t-1}$ is a martingale difference. Writing the dynamic equation with $\sigma^2_{t+1|t}$, as opposed to $\sigma^2_{t|t-1}$, on the left hand side stresses the link with signal extraction.

Once the assumption of Gaussianity is dropped, the case for weighting the squared observations is much weaker. In a DCS model, $\sigma^2_{t+1|t}$ depends on current and past values of a variable, $u_t$, that is defined as being proportional to the (standardized) score of the conditional distribution at time $t$. This variable is a MD by construction. When $y_t$ has a conditional $t$-distribution with $\nu$ degrees of freedom, the DCS modification replaces $v_t$ in the conditional variance equation, (13), by another MD, $v_t = \sigma^2_{t|t-1}u_t$, where

$$u_t = \frac{(\nu+1)y_t^2}{(\nu-2)\sigma^2_{t|t-1} + y_t^2} - 1, \quad -1 \leq u_t \leq \nu, \quad \nu > 2. \tag{14}$$

This model is called Beta-t-GARCH because $u_t$ is a linear function of a variable with a beta distribution. Note that $u_t$ is the score standardized by dividing by the information quantity, $I(\sigma^2_{t|t-1}) = \sigma^{-4}_{t|t-1}$, and then multiplying by two. When $\nu = \infty$, $u_t = y_t^2/\sigma^2_{t|t-1} - 1$ and the standard GARCH model, (13), is obtained by setting $v_t = \sigma^2_{t|t-1}u_t$.

Figure 1 plots the conditional score function, $u$, against $y/\sigma$ for $t-$distributions with $\nu = 3$ and 10 and for the normal distribution ($\nu = \infty$). When $\nu = 3$ an extreme observation has only a moderate impact as it is treated as coming from a $t_\nu-$ distribution rather than from a normal distribution with an abnormally high variance. As $|y_t| \to \infty$, $u_t \to \nu$ so $u_t$ is bounded for finite $\nu$, as is the robust conditional variance equation proposed by Muler and Yohai (2008, p 2922).

The DCS volatility models have particularly attractive properties when an exponential link function is used. In the Gaussian case, this implies that the dynamic equation applies to $\ln \sigma^2_{t+1|t}$, as in EGARCH. For the $t$-distribution it is better to work with the scale, which, for $\nu > 2$, is related to the standard

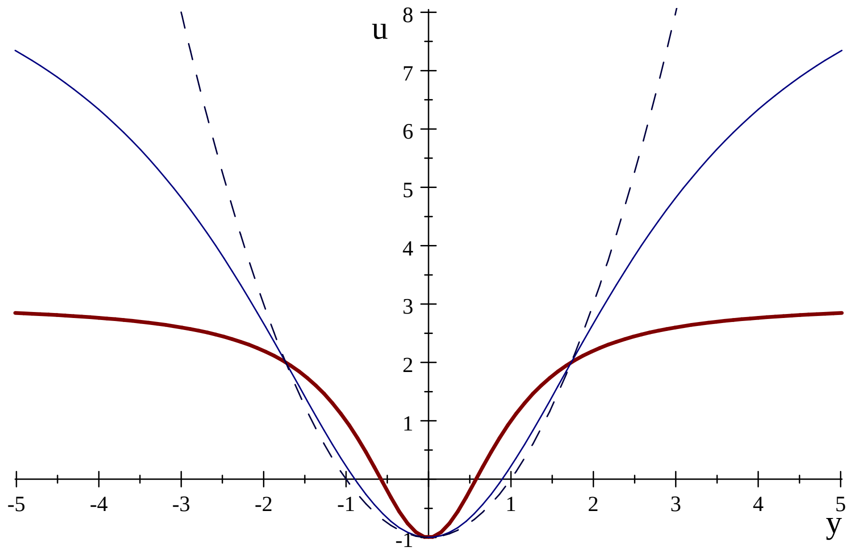


Figure 1: Impact of $u$ for $t_\nu$ with $\nu = 3$ (thick), $\nu = 10$ (thin) and $\nu = \infty$ (dashed).

deviation by the formula, $\varphi_{t+1|t} = (\nu - 2)^{1/2}\sigma_{t+1|t}$. The dynamic equation is then set up for the logarithm of scale, $\lambda_{t+1|t} = \ln \varphi_{t+1|t}$, and so the first-order model is

$$\lambda_{t+1|t} = \delta + \phi\lambda_{t|t-1} + \kappa u_t, \qquad t = 1, ...., T \tag{15}$$

where

$$u_t = \frac{(\nu+1)y_t^2}{\nu \exp(2\lambda_{t|t-1}) + y_t^2} - 1, \quad -1 \leq u_t \leq \nu, \quad \nu > 0,$$

is just the conditional score; compare (14). The class of models obtained by combining the conditional score with an exponential link function is called Beta-t-EGARCH. A complementary class is based on the general error distribution (GED) distribution. The conditional score then has a gamma distribution, leading to the name Gamma-GED-EGARCH.

**Example 1** *An announcement made by the electronics firm Apple illustrates very clearly advantage of Beta-t-EGARCH over the standard GARCH(1,1) model. On Thursday 28 September 2000 a profit warning was issued, which* *led the value of the stock to plunge from an end-of-trading value of \$26.75 to \$12.88 on the subsequent day. This change corresponds to a drop of about 73% in term of the log-difference. In terms of volatility this fall was a one-off event, since it apparently had no effect on the variability of the price changes on the following days. Figure* **??***, contains a snapshot of the event and the surrounding period. The figure plots absolute returns, the fitted conditional standard deviations of a GARCH(1,1)-t specification with leverage, and the fitted conditional standard deviations of the comparable Beta-t-EGARCH model; a full set of estimation results are given later in table* **??***) and in Harvey and Sucarrat (2012). As is clear from the figure, the GARCH forecasts of one-step standard deviations exceed absolute returns for almost two months after the event, a clear-cut example of forecast failure. By contrast, the Beta-t-EGARCH forecasts remain in the same range of variation as the absolute returns.*

Similar considerations arise when dealing with location/scale models. Again $u_t$ is chosen so as to be proportional to the conditional score. Figure 3 shows a plot of $u_t$ against $y/\mu$ for a Weibull distribution, with a shape parameter of 0.5, and contrasts it with the response for a gamma distribution, which is linear. While the DCS approach for a gamma distribution is consistent with the conditional mean dynamic equation of (11), it suggests a dampening down of the impact of a large observation from a Weibull.

**Remark 2** *There is a considerable literature on QML estimation of GARCH models. In this context, QML estimates the parameters under the assumption of Gaussianity. Similarly QML can be used to estimate the parameters in an ARMA model. QML is then essentially just least squares. For a location/scale model, QML is based on the exponential distribution. The estimators can be shown to be consistent for certain distributions other than the assumed distribution and asymptotically correct standard errors may be computed. Unfortunately standard QML asymptotic theory breaks down for GARCH models with heavy-tailed distributions (specifically those without fourth moments) and modified bootstrap procedures have to be used, as in Hall and Yao (2003). However, even though QML can be adapted to give consistent estimators, even when the correct distribution is not specified, it is of little use if the dynamic equation is incorrect.*

In this monograph attention is directed towards score-driven models for which an asymptotic distribution for the ML estimator can be derived. The

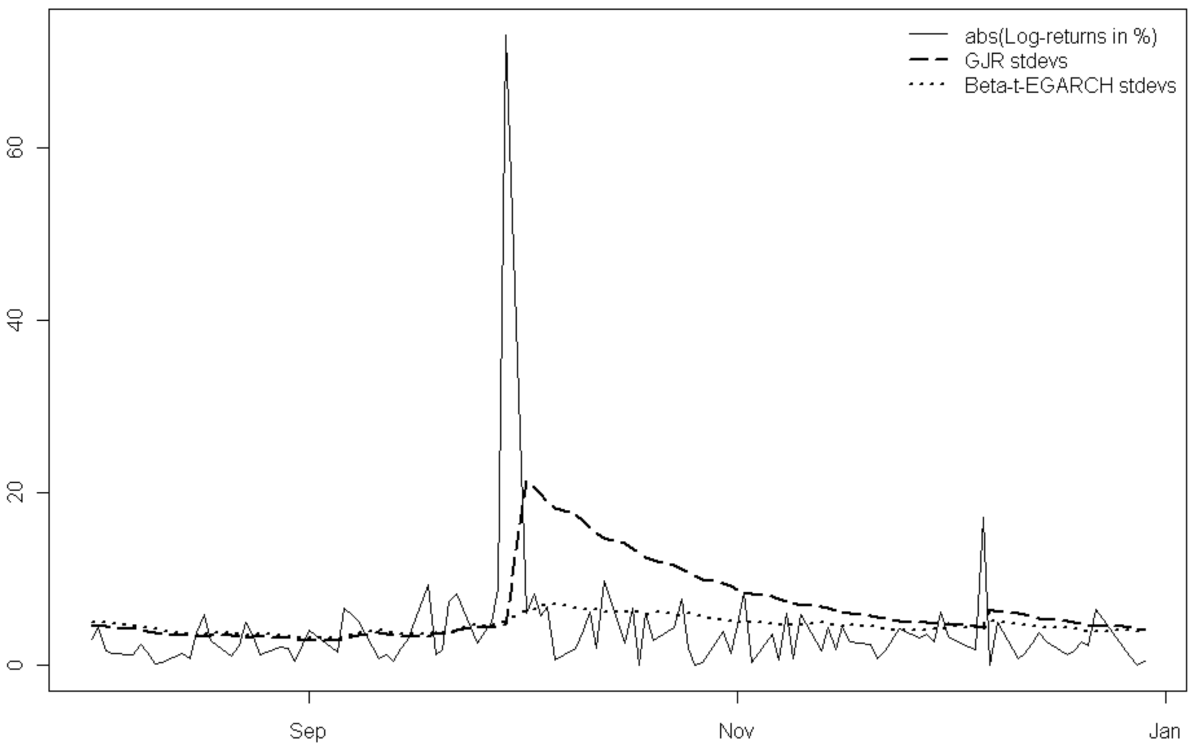


Figure 2: Absolute values of de-meaned Apple returns with filtered GARCH and Beta-t-EGARCH

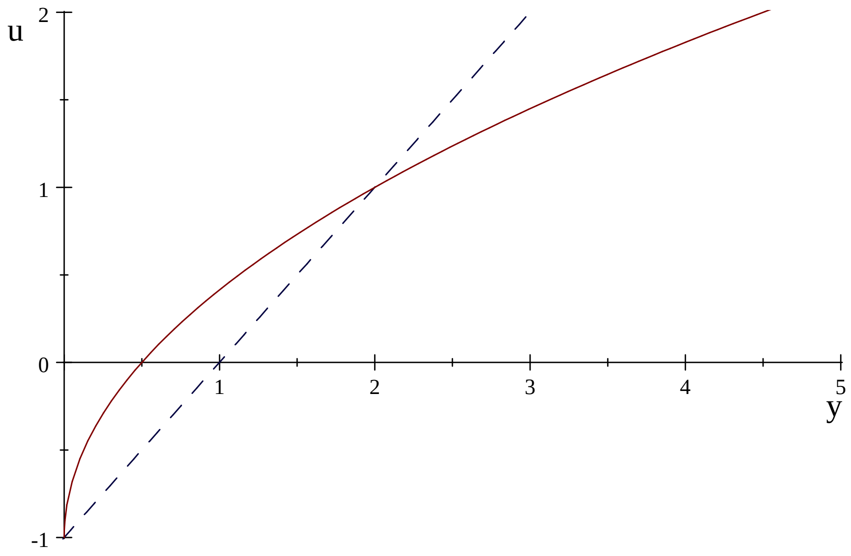


Figure 3: Impact of $u$ for a Weibull distribution and a Gamma (dashed).

reason the asymptotics work is that the score and its first derivative are distributed independently of the time-varying parameter(s) and have finite first and second moments. The main theorem is set out in chapter 2 and then applied in the next three chapters which deal respectively with location models ( that is for the mean or median), scale models ( primarily EGARCH), models and location/scale models. Other properties of the proposed models may also be found. In particular there are analytic expressions for : 1) for moments, 2) autocorrelation functions and 3) multi-step forecasts and their mean square errors. The properties, particularly for the volatility models, which employ an exponential link function, are more general than is usually the case. For example, expressions for unconditional moments, autocorrelations and the conditional moments of multi-step predictive distributions can be found for absolute values of the observations raised to any power.

Given $\lambda_{T+1|T}$, the distribution of $\lambda_{T+\ell|T+\ell-1}$, and in the case of volatility models, its exponent, depends only on future disturbances, as does the conditional distribution of $y_{T+\ell}$. Although these multi-step conditional distributions are difficult to find analytically, it is straightforward to simulate them.

For modeling volatility, the popularity of GARCH(1,1) suggests that the first-order model, (15), is likely to be the most widely used in practice. More generally, a linear dynamic model of order $(p,r)$ may be defined as

$$\lambda_{t+1|t} = \delta + \phi_1\lambda_{t|t-1} + ... + \phi_p\lambda_{t-p+1|t-p} + \kappa_0 u_t + \kappa_1 u_{t-1} + ... + \kappa_r u_{t-r}, \quad (16)$$

where $p \geq 0$ and $r \geq 0$ are finite integers and $\delta, \phi_1, .., \phi_p, \kappa_0, .., \kappa_r$ are (fixed) parameters. Stationarity (both strict and covariance) of $\lambda_{t|t-1}$ requires that the roots of the associated autoregressive polynomial equation[1] are less than one in absolute value, as in an autoregressive-moving average model. However, any conditions which need to be imposed on what, at first sight, look like moving average coefficients are not immediately apparent from an analogy with ARMA theory. Since (16) is a filter, rather than a conventional ARMA model, it will be referred to as a quasi-ARMA model and denoted $QARMA(p,r)$.

The terminology for the order of (16) follows that of Nelson (1991). Thus the first-order model, (15), is (1,0). This nomenclature is not consistent with GARCH, where the first-order model is labeled (1,1), but it is in keeping with the signal extraction interpretation because the filter reflects an underlying AR(1) dynamic process for volatility. ARCH(1) sets $\phi = \kappa$ which is a very special restriction when viewed in terms of (13). In the location case, the level in (1) is clearly an AR(1), while it is the reduced form, the ARMA model for $y_t$, that is of order (1,1). The series itself only becomes an AR(1) process when no noise is added. While such a model is fine for location, it is not really suitable for variance because variance cannot be observed directly. Further discussion on these matters can be found in Appendix E.

While equation (16) generalizes DCS models in the ARMA direction, another possibility is develop DCS models that mirror the unobserved component, or structural time series models, that are implemented in the STAMP package of Koopman *et al* (2008). Such models typically include trend, seasonal and cyclical components for capturing movements in location. The DCS approach leads to a filter that is suitable for a heavy-tailed irregular component. Furthermore, the use of an exponential link function allows the inclusion of trend, seasonal and cyclical components in dynamic volatility models, without the attendant dificulties experienced with GARCH because of the need to ensure a positive variance.

[1] The associated autoregressive polynomial equation is $x^p - \phi_1 x^{p-1} - ... - \phi_p = 0$. The roots may be complex conjugates. Hence the reference to absolute value ( or modulus).

# 6 Distributions and quantiles

Once the Gaussian assumption is dropped, the question arises as to why the focus should be on mean and variance. Admittedly, the mean is rather basic, but the attraction of the variance is limited since attention is typically on certain quantiles or indeed the whole distribution.

One of the reasons for the interest in quantiles is that they define, or help to define, certain measures of risk. In particular, *value at risk* (VaR) for a return, $y$, is

$$\Pr(y \leq VaR_\tau(y)) = \tau,$$

so $VaR_\tau(y)$ is just the $\tau-th$ quantile; see, for example, Tsay (2010). *Expected shortfall* (ES), defined as

$$ES_\tau(y) = E[y \mid y > VaR_\tau(y)),$$

is often preferred to VaR since it aggregates risks in a coherent manner.

For tabulated distributions, such as the normal and $t$, the quantiles can be read off directly and so VaR for the one-step ahead conditional distribution is readily available. As noted earlier, multi-step distributions for DCS models are easily simulated and hence VaR and ES can be calculated to a required degree of precision. In a similar way, expected loss can be computed by simulation for any loss function and the results employed in decision making; see Harvey (1989, pp222-26).

Sometimes analytic expressions are available for quantiles. The quantile function for a given distribution function, $F(y)$, is $F^{-1}(\tau)$, $0 \leq \tau \leq 1$.

What happens if we are not prepared to assume a distribution? The attraction of QML estimation for GARCH is that it is consistent, even if the distribution is not normal. As a result many researchers are more comfortable with QML than with an approach that assumes a specific distribution. However, setting aside the point that QML values consistency more than efficiency and a desire to explore different model specifications, the implied focus on variance is of limited value if what is required is knowledge of the quantiles. In any case, as was pointed out in Remark 2, the argument that QML is robust to misspecification misses the point because it assumes that the specification of the conditional variance as a linear combination of squares is correct.

A better approach to relaxing the dependence on distributional assumptions is to develop nonparametric methods for time series data. Rather than weighting squared observations, as in GARCH, weighting patterns implied by dynamics models can be applied to the kernels that are typically used for density estimation. Thus the whole distribution is tracked as it changes over time and, at the same time, features of the distribution, such as quantiles, can be extracted. Proceeding in this way raises various issues. For example, is it better to model the quantiles directly and how well is tail-behaviour captured?

# 7 Plan of book

The plan of the book is as follows. Chapter 2 provides some basic theory, beginning with a review of the Student-t and general error distributions. The principles of maximum likelihood estimation are discussed. The asymptotic theory for the properties of the maximum likelihood estimators of the parameters in the DCS class is then developed.

Chapters 3,4 and 5 develop the theory for location, scale and location/scale models. Attention is initially focussed on stationary time series, after which it is shown how trend and seasonal components may be handled by drawing on parallels with the unobserved component, or structural, time series models that have been sucessfully applied to modeling the level of Gaussian time series. The technical manipulations rest mainly on standard properties of the beta and gamma distributions. Once this is appreciated, most of the results and formulae follow in a straightforward and elegant fashion. Indeed the fact that the mathematics is so transparent is a strong indication that the statistical structure of the class of models is a sound one. However the appeal of the mathematics should not detract from the main purpose of the models which is to deal with heavy-tailed distributions in a manner that is efficient, both statistically and from the practical perspective.

Chapter 6 indicates how the ideas of the earlier chapters might be extended to nonparametric estimation of changing distributions, while chapter 7 provides an introduction to the challenges associated with modeling multivariate time series. The emphasis on correlation, like the focus on variance, stems from an implicit assumption of Gaussianity. Questioning this assumption for multivariate time series leads to an exploration of the opportunities afforded by copulas.

# References

[1] Andres, P. and A. C. Harvey (2012). The dynamic location/scale model. Mimeo.

[2] Bollerslev, T.: (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307-327.

[3] Bollerslev, T., (1987). A conditionally heteroskedastic time series model for security prices and rates of return data. *Review of Economics and Statistics* 59, 542-547.

[4] Brandt, M.W. and Jones, C.S. (2006). Volatility Forecasting with Range-Based EGARCH Models. *Journal of Business and Economic Statistics* 24, No. 4.

[5] Brownlees, Christian T., & Gallo, Giampiero M. 2010. Comparison of Volatility Measures: a Risk Management Perspective. Journal of Financial Econometrics, 8(1), 29-56.

[6] Creal, D., Koopman, S.J., Lucas, A., (2008). *A general framework for observation driven time-varying parameter models.* Tinbergen Institute Discussion Paper, TI 2008-108/4, Amsterdam.

[7] Creal, D., Koopman, S.J., Lucas, A., (2011). A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations, *Journal of Business and Economic Statistics,* 29, 552-63.

[8] Diebold F.X., Gunther, T.A., & Tay, A.S. (1998). Evaluating density forecasts. *International Economic Review, 39,* 863-83.

[9] Durbin, J., Koopman, S.J., (2012). *Time Series Analysis by State Space Methods.* 2nd Ed. Oxford University Press, Oxford

[10] Engle, R.F. and J.R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* 66, 1127-1162.

[11] Engle, R.F. and G.M. Gallo (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131, 3-27.

[12] Fiorentini, G., G. Calzolari and I. Panattoni (1996). Analytic derivatives and the computation of GARCH estimates. *Journal of Applied Econometrics* 11, 399-417.

[13] Hall, P. and Yao, Q. (2003). Inference in arch and garch models with heavy-tailed errors. *Econometrica*, 71, 285-317.

[14] Harvey A.C., (1989). *Forecasting, structural time series models and the Kalman filter* (Cambridge: Cambridge University Press).

[15] Harvey, A.C. and Chakravarty, T. (2009). *Beta-t-EGARCH.* Working paper. Earlier version appeared in 2008 as a Cambridge Working paper in Economics, CWPE 0840.

[16] Harvey, A. C. and G. Sucarrat (2012). EGARCH models with fat tails, skewness and leverage. Mimeo.

[17] Koopman, S-J., Harvey A.C., Doornik J.A., Shephard, N., (2007). *STAMP 8.0 Structural Time Series Analysis Modeller and Predictor.* Timberlake Consultants Ltd., London.

[18] Koopman, S.J., Lucas, A. and M. Schartha (2012) Predicting Time-Varying Parameters with Parameter-Driven and Observation-Driven Models. Tinbergen Institute Discussion Paper, TI 2012-020/4, Amsterdam.

[19] Maronna, R., Martin, D. and Yohai, V. (2006) *Robust Statistics: Theory and Methods.* Wiley

[20] Muler, N., Peña, D., and V.J. Yohai (2009) Robust estimation for ARMA models. *Annals of Statistics,* 37, 816–840.

[21] Muler, N., Yohai, V.J., 2008. Robust estimates for GARCH models. *Journal of Statistical Planning and Inference* 138, 2918-2940.

[22] Nelson, D.B., (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6, 318-24.

[23] Nelson, D.B., (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, 59, 347-370.

[24] Straumann, D. and T. Mikosch (2006) Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *Annals of Statistics*, 34, 2449-2495.

[25] Taylor, S. J, (2005). *Asset Price Dynamics, Volatility, and Prediction.* Princeton University Press, Princeton.