CHAPTER 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 dynamic distributions. The emphasis is on 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 changing parameter, analytic expressions may be derived for unconditional moments, autocorrelations and moments of multistep forecasts. Furthermore, a full asymptotic distribution theory for maximum likelihood estimators can be developed, including analytic expressions for asymptotic covariance matrices of the estimators.

The class of what we call *dynamic conditional score* (DCS) models includes standard linear time series models observed with an error that 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 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. (2009). For estimating scale, the models are close to stochastic volatility models, in which the variance is treated as an unobserved component. The close ties with unobserved component and stochastic volatility models provide insight into the structure of the DCS models, particularly with respect to modelling trend and seasonality, and into possible restrictions on the parameters.

The reference to location and scale rather than mean and variance is deliberate. Location and scale apply to all distributions, whereas mean and variance may not always exist, a point which is particularly relevant when dealing with heavy tails. Furthermore, although a knowledge of the mean and variance of a Gaussian distribution tells us all there is to know, this is not the case with many other distributions. Focussing too much attention on mean and variance is unwise, particularly in financial econometrics. By a similar token, correlation measures the strength of the relationship between two variables in a Gaussian world, but the more general concept of association is of wider relevance, as witnessed by the recent upsurge of interest in copulas.

Section 1.1 introduces a very basic, but important, unobserved components time series model. The idea of signal extraction for Gaussian models is explained, and the Kalman filter is written down in a form that leads to the development of a more general filter, based on the score of a conditional distribution for each observation. Some basic definitions are noted in Section 1.2, before moving on to a discussion of volatility models in Section 1.3. The relevance of dynamic conditional score models for volatility modelling is explained in Section 1.4, and the implications of outlying observations for conventional and DCS filters are explored. Section 1.5 stresses the importance of modelling the full conditional distribution of an observation, rather than just its first two moments. The last section outlines the contents of each chapter.

1.1 UNOBSERVED COMPONENTS AND FILTERS

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 for a sample of T observations, y_t, $t = 1, .., T$, is

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right), \quad t = 1, \ldots, T, \tag{1.1}$$

$$\mu_{t+1} = \phi\mu_t + \eta_t, \quad \eta_t \sim NID\left(0, \sigma_\eta^2\right),$$

where ϕ is the autoregressive parameter, the irregular and signal disturbances, ε_t and η_t respectively, are mutually and serially independent and the notation $NID\left(0, \sigma^2\right)$ denotes normally and independently distributed with mean zero and variance σ^2. The *signal-noise ratio*, $q = \sigma_\eta^2/\sigma_\varepsilon^2$, plays a key role in determining how observations should be weighted for prediction and signal extraction. The *reduced form* of (1.1) is an $ARMA(1, 1)$ process,

$$y_t = \phi y_{t-1} + \xi_t - \theta\xi_{t-1}, \quad \xi_t \sim NID\left(0, \sigma^2\right), \quad t = 1, \ldots, T, \tag{1.2}$$

but with restrictions on θ. For example, when $\phi = 1$, $0 \leq \theta \leq 1$. The forecasts from the unobserved components (UC) model and reduced form are the same. An autoregressive approximation to the reduced form is possible, but, if q is close to zero, a large number of lags may be needed for the approximation to yield acceptable forecasts.

Cambridge University Press & Assessment
978-1-107-63002-4 — Dynamic Models for Volatility and Heavy Tails
Andrew C. Harvey
Excerpt
More Information

The UC model in (1.1) is effectively in state space form (SSF), and as such, it may be handled by the Kalman filter (KF); see Harvey (1989). The parameters ϕ 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 μ_t based on information at time $t-1$ with the t-th observation in order to produce the best estimator of μ_{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 Kalman filter:

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t, \quad t = 1, \ldots, T, \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + k_t v_t. \end{aligned} \tag{1.3}$$

The Kalman gain, k_t, depends on ϕ and q. In the steady-state, k_t is constant. Setting it equal to a parameter, κ, and rearranging gives the ARMA model, (1.2), with $\xi_t = v_t$ and $\phi - \kappa = \theta$. A pure autoregressive (AR) 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. In the case of (1.1), the reduced form is still an $ARMA(1, 1)$ process, but with disturbances which, although they are serially uncorrelated, are not independently and identically distributed. Allowing the disturbances to have a heavy-tailed distribution does not deal with the problem. A large value of ε_t only affects the current observation, but in the reduced form, it is incorporated into the level and takes time to work through the system. To be specific, the AR representation of an $ARMA(1, 1)$ process 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 τ periods, the prediction of the next observation is still contaminated by C because it contains the quantity $(\phi - \theta)\phi^{\tau}C$.

An ARMA or AR model in which the disturbances are allowed to have a heavy-tailed distribution is designed to handle *innovation 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, Chapter 8) and the recent article by Muler, Pena and Yohai (2009) on robust estimation for ARMA models. The view taken here 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), importance sampling 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 Robert and Casella (2010), Durbin and Koopman (2012) and Koopman, Lucas and Schartha (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 *parameter-driven* in Cox's terminology, the KF is driven by the one-step-ahead prediction error, as in (1.3). 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 (1.3) becomes

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

where κ is treated as an unknown parameter.

Why the score? If the signal in (1.1) were fixed, that is $\phi = 1$ and $\sigma^2_\eta = 0$ so $\mu_t = \mu$, 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 μ 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 (1.1), 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 a 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. Further theoretical support comes from the conditional mode approach to smoothing for nonlinear models. Indeed the argument presented in Sub-section 3.7.3 is a more comprehensive one.

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

Cambridge University Press & Assessment
978-1-107-63002-4 — Dynamic Models for Volatility and Heavy Tails
Andrew C. Harvey
Excerpt
More Information

variables. The first equation in (1.3) 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 Kalman filter 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, they note that the procedure 'has been found to perform well both in simulation studies and with real data'. Schick and Mitter (1994) suggested 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 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 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.

1.2 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. But before proceeding, some basic statistical results need to be stated. The proofs can be found in many introductory time series and econometrics texts.

1.2.1 The Law of Iterated Expectations and Optimal Predictions

A key element in some of the statistical derivations that follow is the *law of iterated expectations* (LIE). Suppose that it is difficult to find the expected value of a random variable, y, but evaluating its expectation conditional on another random variable, x, is straightforward. Then $E(y)$ may be obtained as

$$E(y) = E_x[E(y \mid x)],$$

because

$$\begin{aligned} E_x[E(y|x)] &= \int \left[\int yf(y|x)\,dy \right] f(x)\,dx \\ &= \int\int yf(y,x)\,dydx = E(y). \end{aligned}$$

The above process may be generalized and repeated. Thus, if $g(y_t)$ is a function of y_t, an expected value several steps ahead can be found from the sequence of one-step-ahead conditional expectations because

$$\underset{t-j}{E}[g(y_t)] = \underset{t-j}{E} \cdots \underset{t-1}{E}[g(y_t)], \qquad j = 2, 3, \ldots .$$

The unconditional expectation is found by letting $j \to \infty$. The expectation of a function of the observation at time $T + \ell$ based on information available at time T is given by setting $t = T + \ell$ and $j = \ell$ so

$$\underset{T}{E}[g(y_{T+\ell})] = \underset{T}{E} \cdots \underset{T+\ell-1}{E}[g(y_{T+\ell})], \qquad \ell = 2, 3, \ldots . \tag{1.4}$$

When the objective is to predict a future observation based on current information, the conditional expectation, $E_T(y_{T+\ell})$, $\ell = 1, 2, 3, \ldots$, is optimal in the sense that it minimizes the mean square error (MSE) of the prediction error; see, for example, Harvey (1993, p. 33). As such, it is called the minimum mean square error (MMSE) predictor. For nonlinear models, expression (1.4) is of considerable practical importance for finding MMSE predictors.

1.2.2 Definitions and Properties

The following important definitions should be noted.

Definition 1 *White noise (WN) variables are serially uncorrelated with constant mean and variance.*

Definition 2 *A martingale difference (MD) has a zero (or constant) conditional expectation, that is,*

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

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

Definition 3 *Strict white noise variables are independent and identically distributed (IID).*

The relationship between a martingale difference and the two kinds of white noise is given by the following proposition.

Proposition 1 *(a) All zero mean independent sequences are martingale differences and (b) all martingale differences are white noise, provided that the variance is finite. In neither case is the converse true.*

Proof. Part (a) requires no proof. Part (b) follows because all MDs have zero unconditional mean and are serially uncorrelated. Specifically

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

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.

The heteroscedastic models introduced in the next section produce observations that are MDs but not IID. That a WN sequence is not necessarily an MD can be demonstrated by a simple example showing that there may be a nontrivial nonlinear predictor. To be specific, the observations in the model

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

where ε_0 and ε_{-1} are fixed and known, are white noise, but not an MD because $E\left(y_{T+1} \mid Y_T\right) = \beta\varepsilon_T\varepsilon_{T-1}$. ■

Remark 1 *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 *linear process* is usually defined as one that can be written as an infinite moving average in $IID(0, \sigma^2)$ disturbances, with the sum of the squares of the coefficients being finite, that is,

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty, \quad \varepsilon_t \sim IID(0, \sigma^2). \tag{1.5}$$

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.[1] For a stationary process, the representation in (1.5) means that all the information about the dynamics is in the autocorrelation function (ACF). Furthermore, the minimum mean square error 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, the linearity of (1.5) is of limited practical value since it is not usually possible to derive the multistep predictive distribution. On the other hand, the optimal forecasts in a model which is a linear function of current and past MDs are the same as in a model in which the MDs are replaced by strict WN, and if the conditional variances are constant, the MSEs are the same.

1.3 VOLATILITY

If dividends and other payments are ignored, financial returns can be defined as the first differences of the logarithm of the price; see Taylor (2005, Chapter 2 and pp. 100–2). When markets are working efficiently, returns are martingale differences. In other words, they should not be predictable on the basis of past information. However, returns are not usually independent, and so features

[1] For further discussion, see Terasvirta et al. (2010, pp. 1–2).

of the conditional distribution apart from the mean may be predictable. In particular, nontrivial predictions can be made for the variance or scale.

1.3.1 Stochastic Volatility

The variance in *stochastic volatility* (SV) models is driven by an unobserved process. The first-order model, with the mean of the observations, y_t, $t = 1, .., T$, assumed to be zero, is

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \exp(2\lambda_t), \quad \varepsilon_t \sim IID(0,1) \tag{1.6}$$

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

where the disturbances ε_t and η_t are mutually independent. Leverage effects, which enable σ_t^2 to respond asymmetrically to positive and negative values of y_t, can be introduced by allowing ε_t and η_t to be correlated, as in Harvey and Shephard (1996). Shephard and Andersen (2009) discuss the relationship between SV models and continuous time models in the finance literature.

The *exponential link function* ensures that the variance remains positive and the restrictions needed for λ_t and y_t to be stationary are straightforward; for (1.6), $|\phi| < 1$. Furthermore, analytic expressions for moments and ACFs of the absolute values of the observations raised to any power can be derived.

Unfortunately, direct maximum likelihood estimation of the SV model is not possible. A procedure can be based on the linear state space form obtained by taking logarithms of the absolute values of the demeaned observations to give the following measurement equation:

$$\ln|y_t| = \lambda_t + \ln|\varepsilon_t|, \quad t = 1, \ldots, T. \tag{1.7}$$

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 achieved by computer-intensive methods, as described in Creal (2012), Andrieu et al. (2011) and Durbin and Koopman (2012).

1.3.2 Generalized Autoregressive Conditional Heteroscedasticity

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 modelling changes in the volatility of returns. It does so by letting the variance be a linear function of past squared observations. The first-order model, $GARCH(1,1)$, is

$$y_t = \sigma_{t|t-1}\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \tag{1.8}$$

and

$$\sigma_{t|t-1}^2 = \delta + \beta\sigma_{t-1|t-2}^2 + \alpha y_{t-1}^2, \quad \delta > 0, \beta \geq 0, \alpha \geq 0. \tag{1.9}$$

The conditions on α and β ensure that the variance remains positive. The sum of α and β 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 of past squared observations and, as such, is often used by practitioners.

The model may be extended by adding lags of the variance and the squared observations. Heavy tails are accommodated by letting the conditional distribution be Student's t, 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, pp. 220–1). The technique is often known as GJR, after the originators, Glosten, Jagannanthan and Runckle (1993).

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

$$y_t^2 = \gamma + \phi y_{t-1}^2 + v_t + \theta^* v_{t-1}, \tag{1.10}$$

where v_t is white noise, $\phi = \alpha + \beta$ and $\theta^* = -\beta$. The drawback to working with squared observations is that outlying observations can seriously weaken the serial correlation. The autocorrelations of absolute values tend to be larger and so provide a better vehicle for detecting dynamic volatility and assessing its nature.

The principal advantage of GARCH models over SV models is that, because they are observation-driven, the likelihood function is immediately available.

1.3.3 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 (1.8). The leading case is

$$\ln \sigma^2_{t|t-1} = \delta + \phi \ln \sigma^2_{t-1|t-2} + \alpha\left[|\varepsilon_{t-1}| - E\,|\varepsilon_{t-1}|\right] + \alpha^* \varepsilon_{t-1}, \tag{1.11}$$

where α and α^* are parameters and, for a Gaussian model, $E\,|\varepsilon_t| = \sqrt{2/\pi}$. The role of ε_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 that of the SV model because $|\varepsilon_{t-1}| - E\,|\varepsilon_{t-1}|$, like ε_{t-1}, is an MD. Stationarity restrictions are similar to those in the SV model; for example, in the preceding equation, $|\phi| < 1$. 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) noted 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 $\sigma^2_{t|t-1}$ and y_t are rarely satisfied in practice. Hence the model is of little practical value because, 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.

1.3.4 Variance, Scale and Outliers

Substituting repeatedly for the conditional variance in (1.9) 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.

The great strength of the GARCH filter is its simple interpretation as an estimate of variance constructed by weighting the squared observations. This is also its weakness, because a linear combination of past squares (even if infinite) may not be a good choice for modelling dynamics when the conditional distribution is non-Gaussian. 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. The difficulties can be avoided by modelling 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 led 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 problem.

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.

1.3.5 Location/Scale Models

Many variables are intrinsically non-negative. Examples in finance include duration, range, realized volatility and spreads; see, for example, Brownlees and Gallo (2010) and Russell and Engle (2010). Other situations in economics in which 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.