Dynamic Models for Volatility and Heavy Tails

1. Introduction and Asymptotic Theory

Andrew Harvey

Cambridge University

November 2012

.

Econometric Society *Monograph, No. 52*
Cambridge University Press
to appear April 2013.
http://www.cambridge.org/gb/knowledge
/isbn/item7091594/?site_locale=en_GB

For Table of contents and Chapter 1, see
http://www.econ.cam.ac.uk/faculty/harvey/Pages-from-AHbook.pdf

Course outline

These lectures will introduce a new class of time series models that are particularly important for modelling volatility.
The main applications are in **financial econometrics**, but some of the techniques are also relevant to **macroeconometrics**.

Lecture 1: a) Introduction. Models for changing scale and location. GARCH and EGARCH. Dynamic conditional score (DCS) models.

b) Gamma and beta distributions and their relation to Student's t and the general error distribution.

Maximum likelihood estimation of DCS models.

Course outline

Lecture 2: a) Dynamic location. Outliers and robustness. Trend and seasonality. Smoothing.
b) Dynamic volatility. Beta-t-EGARCH models. Leverage effects.
Long memory and components. Application to stock returns.

Lecture 3: a) Location/scale models for non-negative variables. Application to intra-day data.
b) Multivariate models. Changing location, scale and correlation. Dynamic copulas.

Introduction: Dynamic Conditional Score (DCS) Models

1) A unified and comprehensive theory for a class of nonlinear time series models in which the conditional distribution of an observation may be heavy-tailed and the location and/or scale changes over time.
2) **The defining feature of these models is that the dynamics are driven by the score of the conditional distribution.** *(Score = first derivative of log-density wrt changing parameter)*
3) 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.
4) Furthermore a full *asymptotic distribution theory for maximum likelihood estimators* can be obtained, including analytic expressions for the asymptotic covariance matrix of the estimators.

Introduction: Dynamic Conditional Score (DCS) Models

The class of DCS models includes:
(i) standard linear time series models observed with an error which may be subject to outliers,
(ii) models which capture changing conditional variance, and
models for non-negative variables.
The last two of these are of considerable importance in financial econometrics.
(a) Volatility - Exponential GARCH (EGARCH)
(b) Duration (time between trades) and volatility as measured by range and realised volatility - Gamma, Weibull, logistic and F-distributions with changing scale and exponential link functions,

Introduction: Dynamic Conditional Score (DCS) Models

A guiding principle is **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 (UC) 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 (SV) models, where the variance is treated as an unobserved component. The close ties with UC and SV 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.

Unobserved component models

A simple Gaussian signal plus noise model is

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right), \quad t = 1, \dots, T$$

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

where the irregular and level disturbances, ε_t and η_t, are mutually independent. The AR parameter is ϕ, 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) is an ARMA(1,1) process

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

but with restrictions on θ. For example, when $\phi = 1$, $0 \leq \theta \leq 1$. The forecasts from the UC model and RF are the same.

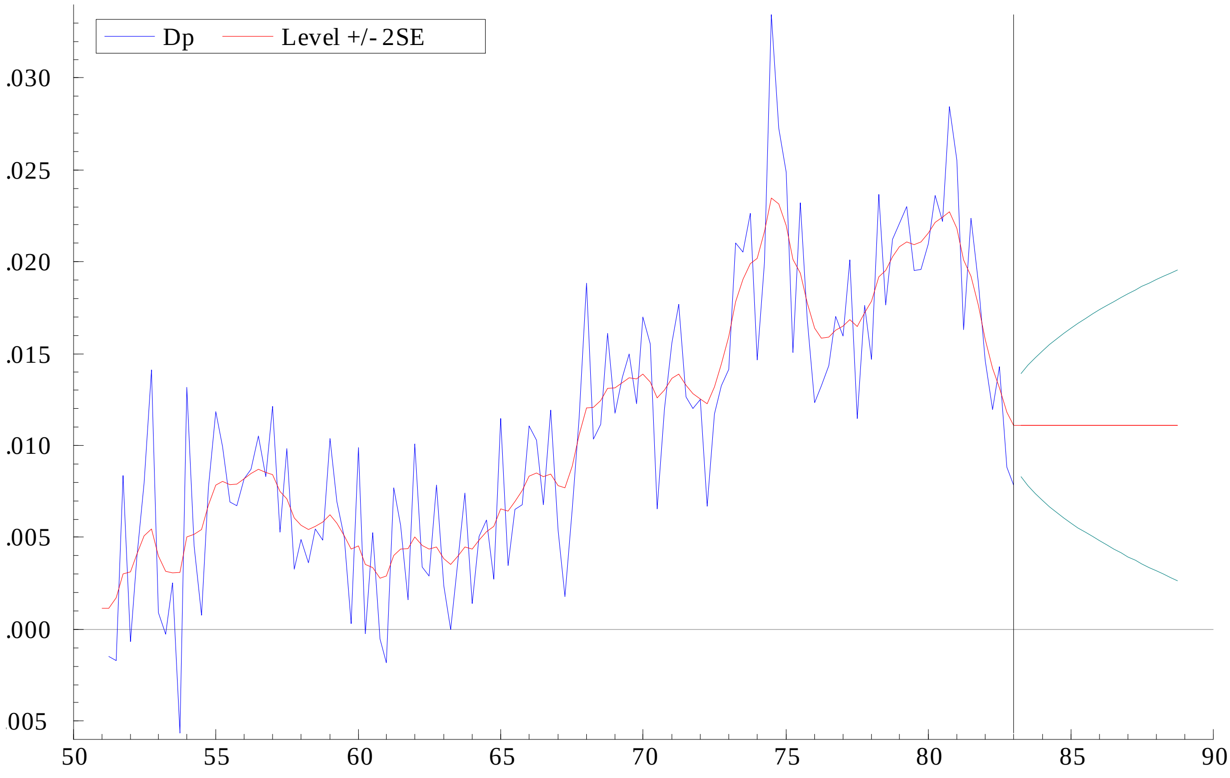


Figure: Forecasts of US inflation with smoothed level

Unobserved component models

The UC model is effectively in state space form (SSF) and, as such, it may be handled by the Kalman filter (KF). The parameters ϕ and q can be estimated by ML, with the likelihood function constructed from the one-step ahead prediction errors.

The KF can be expressed as a single equation. Writing this equation together with an equation for the one-step ahead prediction error, v_t, gives the innovations form (IF) of the KF:

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + k_t v_t \end{aligned}$$

The Kalman gain, k_t, depends on ϕ and q.

In the steady-state, k_t is constant. Setting it equal to κ and re-arranging gives the **ARMA(1,1)** model with $\xi_t = v_t$ and $\phi - \kappa = \theta$.

Outliers

Suppose noise is from a heavy tailed distribution, such as Student's t. Outliers.

The RF is still an ARMA(1,1), but 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.

An ARMA models with a heavy-tailed distribution is designed to handle *innovations outliers*, as opposed to *additive outliers*. See the **robustness** literature.

But a *model-based approach* is not only simpler than the usual robust methods, but is also more amenable to diagnostic checking and generalization.

Unobserved component models for non-Gaussian noise

Simulation methods, such as MCMC, 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).

But simulation-based estimation can be time-consuming and subject to a degree of uncertainty.

Also the statistical properties of the estimators are not easy to establish.

Observation driven model based on the score

The DCS approach 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 *observation driven,* as opposed to a UC model which is *parameter driven.* (Cox's terminology). In a *linear Gaussian UC* model, the KF is driven by the one step-ahead prediction error, v_t. The DCS filter replaces v_t in the KF equation by a variable, u_t, that is proportional to the score of the conditional distribution.

The IF becomes

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t, \qquad t = 1, ..., T \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + \kappa u_t \end{aligned}$$

where κ is an unknown parameter.

Why the score ?

If the signal in AR(1)+noise model 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 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 normal, ML estimator is the same as the sample mean, the moment estimator.

For a non-Gaussian distribution, the moment estimator and the ML estimator differ. Once the signal in a Gaussian model becomes dynamic, its estimate can be updated using the KF. 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.

Why the score ?

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), '..insufficiently justified and remains controversial.' Nevertheless, the procedure has been found to perform well both in simulation studies and with real data.

Why the score ?

(1) The attraction of treating the score-driven filter 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.
(2) 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.
(3) The justification for the class of DCS models is not that they approximate corresponding UC models, but rather that their statistical properties are both comprehensive and straightforward.
(4) An immediate practical advantage is seen from the response of the score to an outlier.

Dynamic location model

$$\begin{aligned} y_t &= \omega + \mu_{t|t-1} + v_t = \omega + \mu_{t|t-1} + \exp(\lambda)\varepsilon_t, \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + \kappa u_t, \end{aligned}$$

where ε_t is serially independent, standard t-variate and

$$u_t = \left(1 + \frac{(y_t - \mu_{t|t-1})^2}{\nu e^{2\lambda}}\right)^{-1} v_t,$$

where $v_t = y_t - \mu_{t|t-1}$ is the prediction error and $\varphi = \exp(\lambda)$ is the (time-invariant) scale.
Further details in Harvey and Luati (2012).

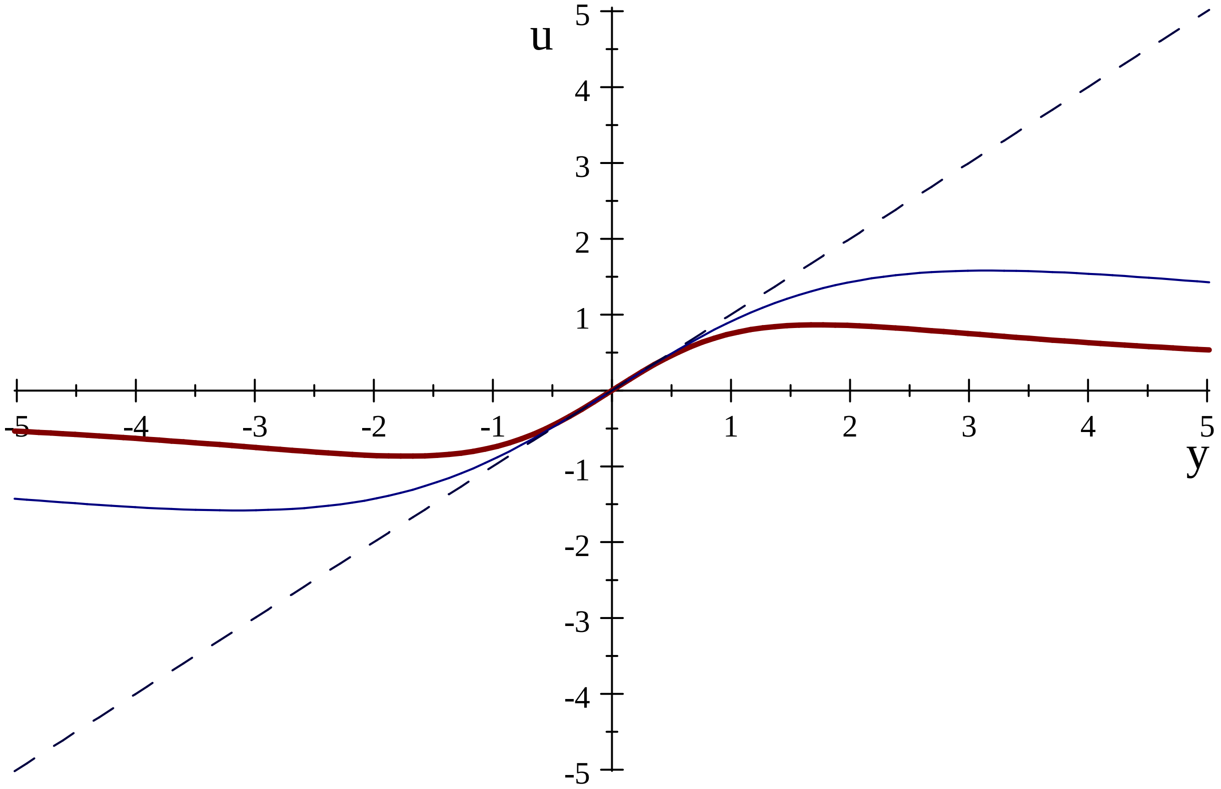
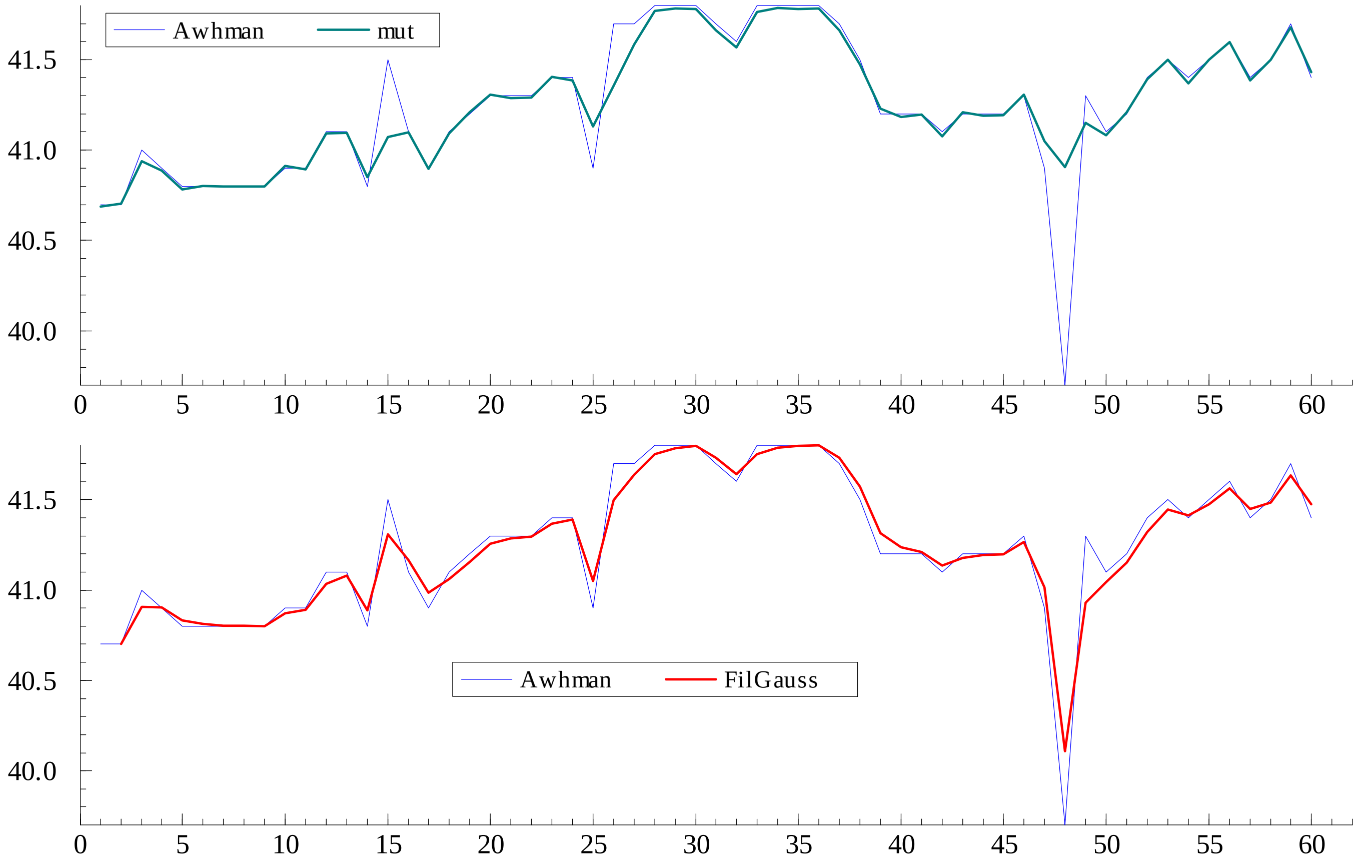


Figure: Impact of u_t for t_ν (with a scale of one) for $\nu = 3$ (thick), $\nu = 10$ (thin) and $\nu = \infty$ (dashed).



Volatility models: GARCH

Standard approach is GARCH(1,1).

$$y_t = \sigma_{t|t-1} z_t, \quad z_t \backsim NID\left(0, 1\right)$$

with conditional variance

$$\sigma^2_{t|t-1} = \gamma + \beta \sigma^2_{t-1|t-2} + \alpha y^2_{t-1}, \qquad \gamma > 0, \beta \geq 0, \alpha \geq 0$$

or

$$\sigma^2_{t|t-1} = \gamma + \phi \sigma^2_{t-1|t-2} + \alpha \sigma^2_{t-1|t-2} u_{t-1}, \qquad \phi = \alpha + \beta,$$

where

$$u_{t-1} = y^2_{t-1} / \sigma^2_{t-1|t-2} - 1$$

is a martingale difference (MD). Weakly stationary if $\phi < 1$.

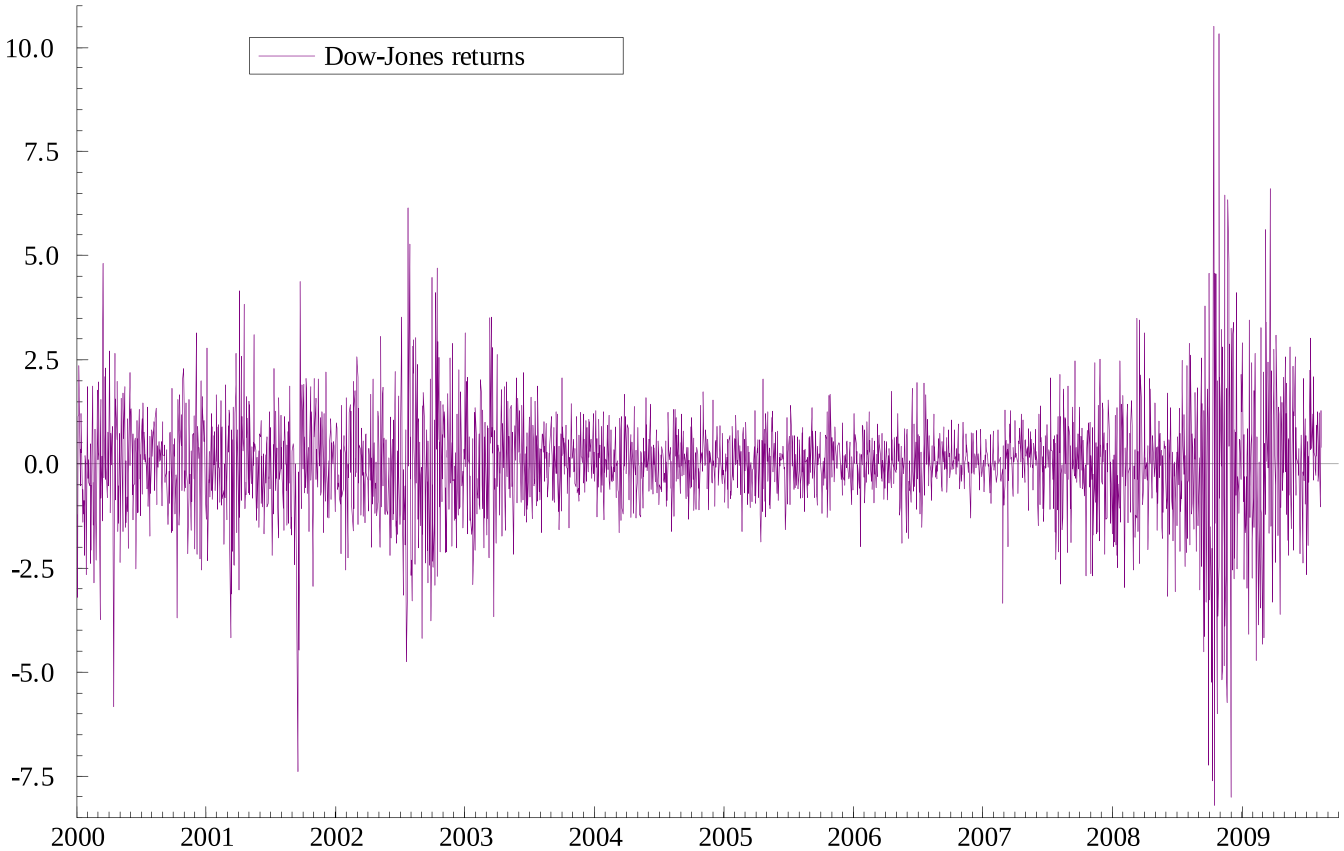
Volatility models: SV

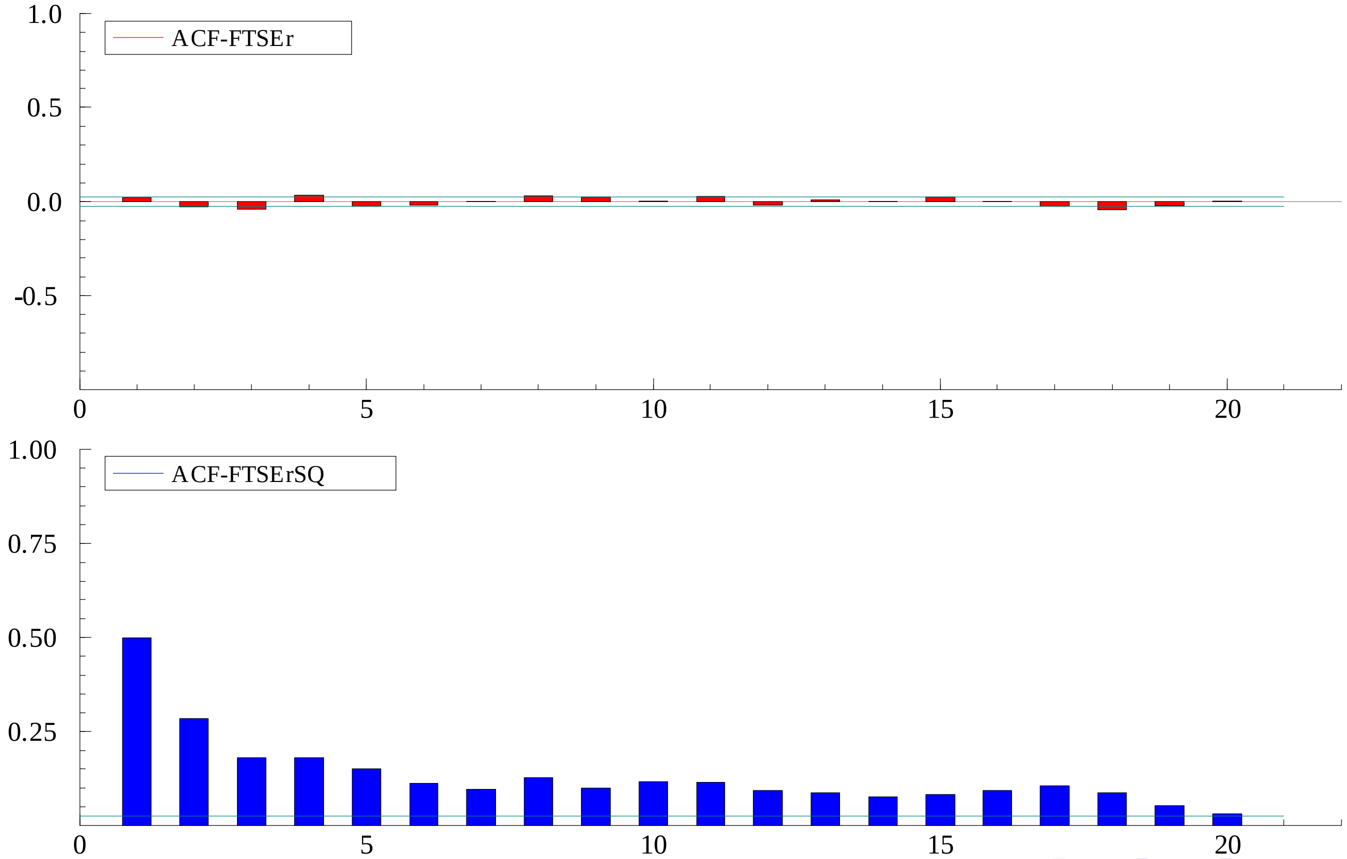
Observation driven models - parameter(s) of conditional distribution are functions of past observations. Contrast with parameter driven, eg stochastic volatility (SV) models
The variance in SV models is driven by an unobserved process. The first-order model is

$$y_t = \sigma_t \varepsilon_t, \qquad \sigma^2_t = \exp\left(\lambda_t\right), \qquad \varepsilon_t \sim IID\left(0, 1\right)$$

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

with ε_t and η_t mutually independent.


Dow-Jones returns
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009


ACF-FTSEr
1.0
0.5
0.0
-0.5
0
5
10
15
20
ACF-FTSErSQ
1.00
0.75
0.50
0.25
0
5
10
15
20

Independence, uncorrelatedness and martingale differences

ACFs of observations for some series, such as stock returns, and their squares show only serial correlation in the squares. The observations are uncorrelated but not independent. When this is the case, future observations may or may not be predictable.

Example

The observations in the model

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2}, \quad \varepsilon_0 = \varepsilon_{-1} = 0, \quad t = 1, ..., T$$

are serially uncorrelated, but the one-step ahead prediction is not zero.

When is White Noise not White Noise ?

(a) *WN - uncorrelated* i.e. $E(y_t y_s) = 0$ if $t \neq s$ (and mean is 0). Constant variance

(b) *Strict WN - independent*

N.B. Strict *WN* implies *WN* (provided that variance is finite)

$WN + Gaussianity$ is strict *WN*

For *WN* there may be a non-trivial (nonlinear) predictor.

When is White Noise not White Noise ?

(c) *Martingale Difference (MD)*

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

It is also necessary that $E|y_t| < \infty$.
By law of iterated expectations (LIE)

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

y_t is uncorrelated with any function of past observations as

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

and so unconditional expectation of $y_t f(Y_{t-1})$ is 0

1. All MD's are serially uncorrelated, but not converse (*WN* if variance constant)
2. All zero mean independent sequences are MDs but not converse.

Volatility: GARCH-t

Stock returns are known to be **non-normal**
(i) Assume that z_t has a Student t_ν-distribution, where ν denotes degrees of freedom - GARCH-t model.
(ii) The *t*-distribution is employed in the predictive distribution of returns and used as the basis for maximum likelihood (ML) estimation of the parameters, but it is not acknowledged in the design of the equation for the conditional variance.
(iii) *The specification of the* $\sigma^2_{t|t-1}$ *as a linear combination of squared observations is taken for granted*, but the consequences are that $\sigma^2_{t|t-1}$ responds too much to extreme observations and the effect is slow to dissipate.
(iv) Note that QML estimation procedures do not question this linearity assumption.

Volatility: Exponential GARCH (EGARCH)

In the EGARCH model (Nelson, 1991)

$$y_t = \sigma_{t|t-1} z_t, \qquad z_t \text{ is } IID(0,1),$$

with first-order dynamics

$$\ln \sigma^2_{t|t-1} = \delta + \phi \ln \sigma^2_{t-1|t-2} + \theta(|z_{t-1}| - E\,|z_{t-1}|) + \theta^* z_{t-1}$$

The role of z_t is to capture leverage effects.

Volatility: EGARCH

Weak and covariance stationary if $|\phi| < 1$. More general infinite MA representation. Moments of $\sigma^2_{t|t-1}$ and y_t exist for the $GED(\upsilon)$ distribution with $\upsilon > 1$. The normal distribution is $GED(2)$.
Two major problems:
(a) If z_t is t_ν distributed, the conditions needed for the **existence of the moments** of $\sigma^2_{t|t-1}$ and y_t are rarely (if ever) satisfied in practice.
(b) **No asymptotic theory for ML**. See reviews by Linton (2008) and Zivot (2009). For GARCH there is no comprehensive theory.

DCS Volatility Models

What does the assumption of a t_ν-distribution imply about the specification of an equation for the conditional variance?
The possible inappropriateness of letting $\sigma^2_{t|t-1}$ be a linear function of past squared observations when ν is finite becomes apparent on noting that, if the variance were constant, the sample variance would be an inefficient estimator of it.
Therefore replace u_t in the conditional variance equation

$$\sigma^2_{t+1|t} = \gamma + \phi\sigma^2_{t|t-1} + \alpha\sigma^2_{t|t-1}u_t,$$

by another MD

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

which is proportional to the **score** of the conditional distribution.

Exponential DCS Volatility Models

$$y_t = \varepsilon_t \exp(\lambda_{t|t-1}/2), \quad t = 1, ..., T,$$

where the serially independent, zero mean variable ε_t has a $t_\nu-$distribution with degrees of freedom, $\nu > 0$, and the dynamic equation for the log of scale is

$$\lambda_{t|t-1} = \delta + \phi\lambda_{t-1|t-2} + \kappa u_{t-1}.$$

The conditional score is

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

NB The variance is equal to the square of the **scale**, that is $(\nu-2)\sigma^2_{t|t-1}/\nu$ for $\nu > 2$.

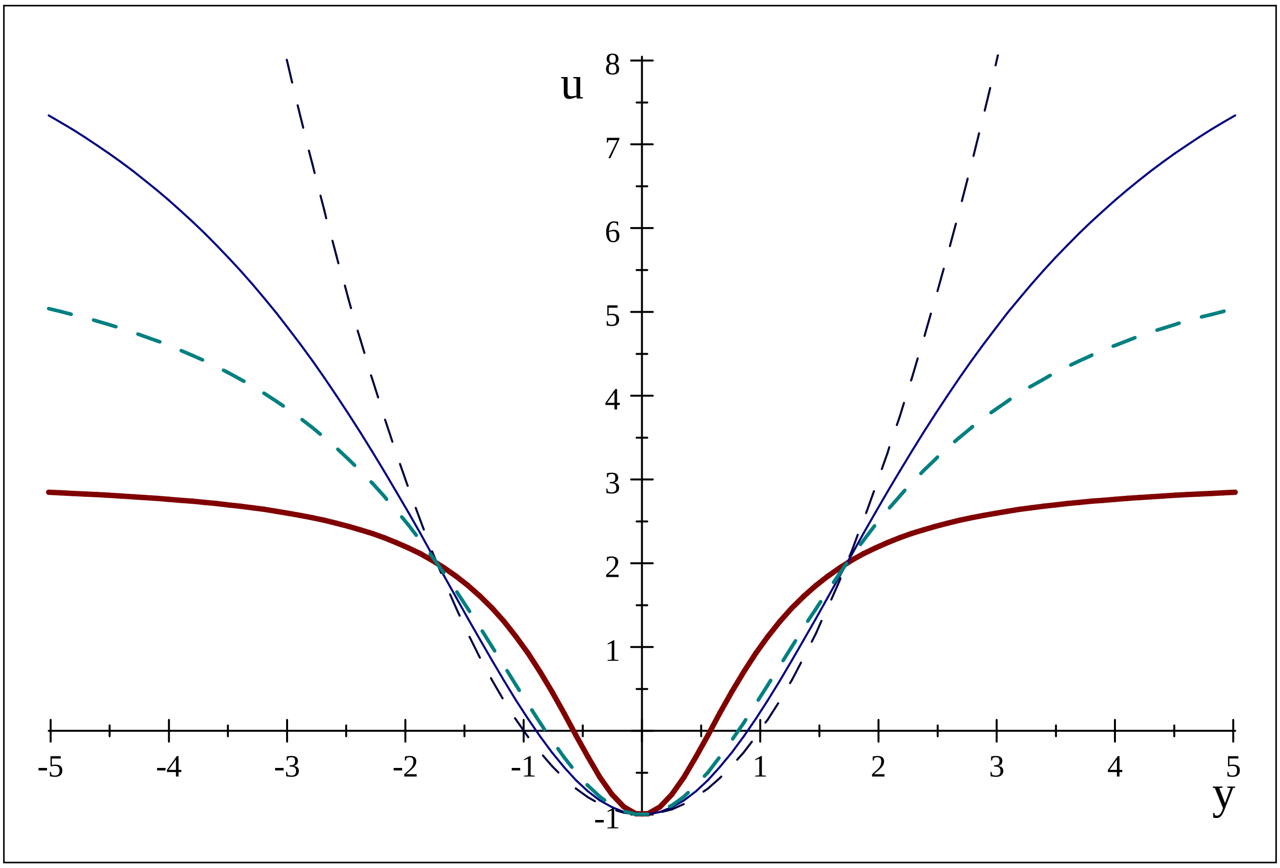
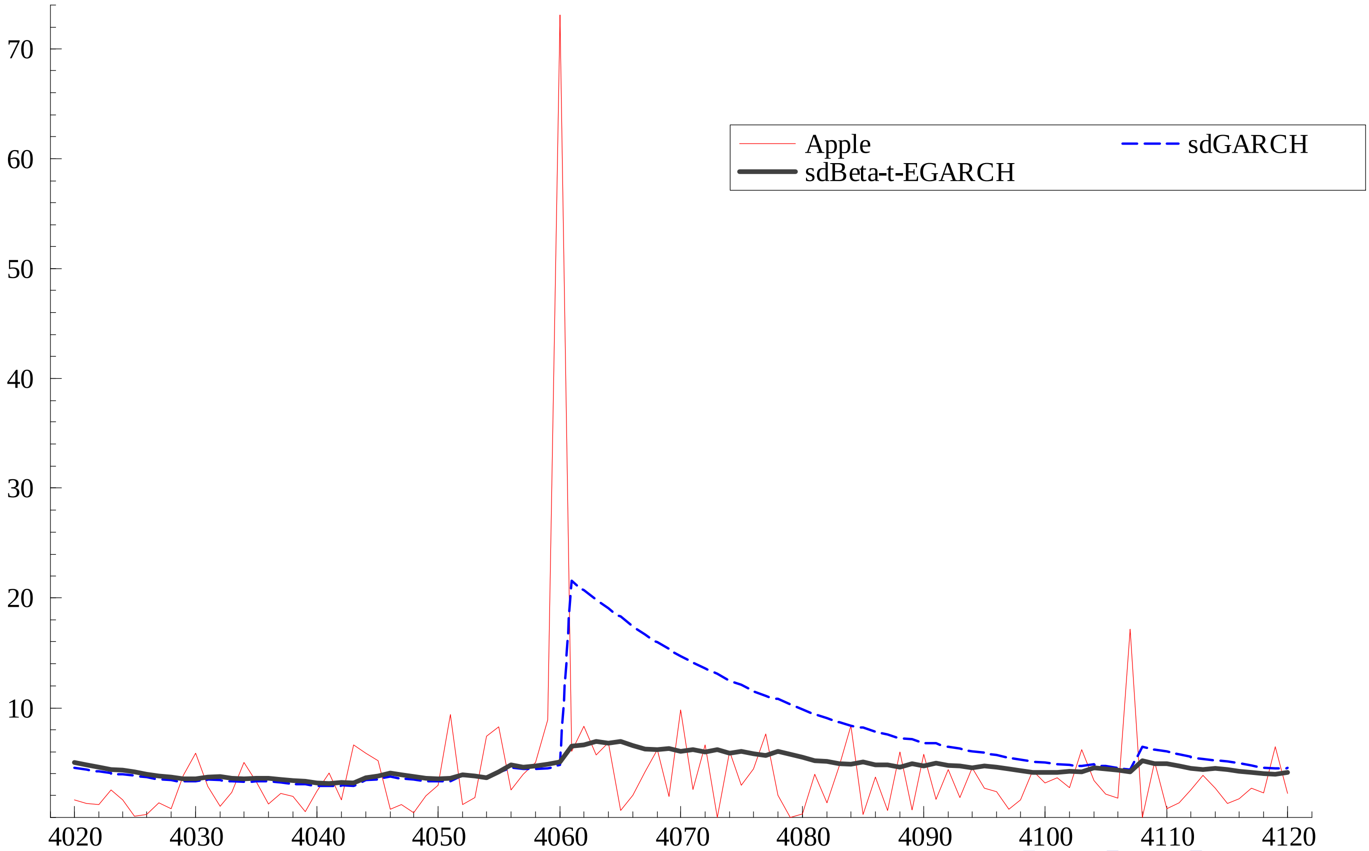


Figure: Impact of u_t for t_ν with $\nu = 3$ (thick), $\nu = 6$ (medium dashed) $\nu = 10$ (thin) and $\nu = \infty$ (dashed).



Beta-t-EGARCH

The variable u_t may be expressed as

$$u_t = (\nu + 1)b_t - 1,$$

where

$$b_t = \frac{y_t^2/\nu \exp(\lambda_{t|t-1})}{1 + y_t^2/\nu \exp(\lambda_{t|t-1})}, \qquad 0 \leq b_t \leq 1, \quad 0 < \nu < \infty,$$

is distributed as $Beta(1/2, \nu/2)$, a **Beta distribution**. Thus the $u_t's$ are IID.
Since $E(b_t) = 1/(\nu+1)$ and $Var(b_t) = 2\nu/\{(\nu+3)(\nu+1)^2\}$, u_t has zero mean and variance $2\nu/(\nu+3)$.

Beta-t-EGARCH

1) Moments exist and ACF of $|y_t|^c$, $c \geq 0$, can be derived.
2) Closed form expressions for moments of multi-step forecasts of volatility can be derived and full distribution easily simulated.
3) Asymptotic distribution of ML estimators with analytic expressions for standard errors.
4) Can handle time-varying trends (eg splines) and seasonals (eg time of day or day of week).

Gamma–GED-EGARCH

When the conditional distribution of y_t has a $GED(\upsilon)$ distribution, u_t is a linear function of $|y_t|^{\upsilon}$. These variables can be transformed so as to have a **gamma** distribution and the properties of the model are again derived. The normal distribution is a special case of the GED, as is the double exponential, or Laplace, distribution. The conditional variance equation for the Laplace model has the same form as the conditional variance equation in the EGARCH model of Nelson (1991).

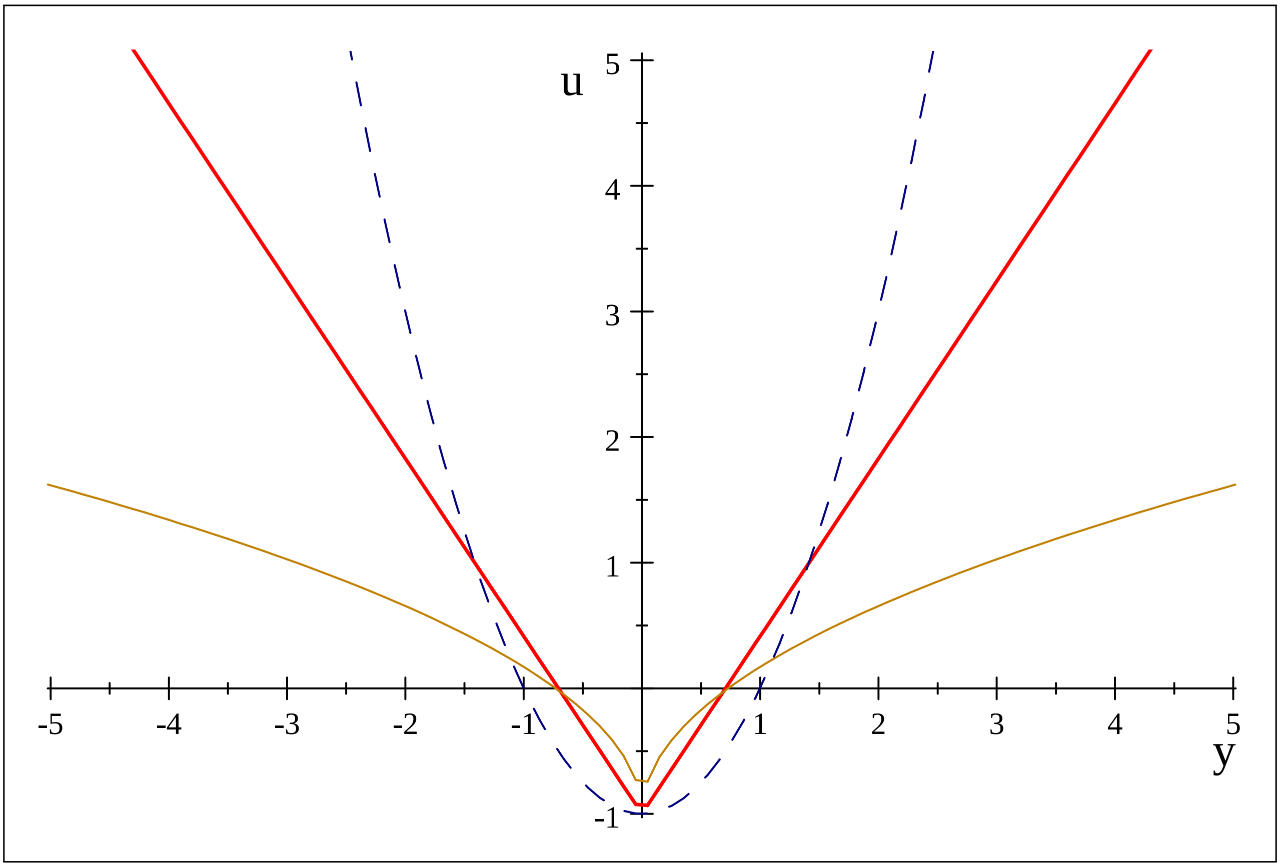


Figure: Impact of u_t for GED with $\upsilon = 1$ (thick), $\upsilon = 0.5$ (thin) and $\upsilon = 2$ (dashed).

Non-negative variables: duration, realized volatility and range

Many variables, particularly those associated with intra-day financial data, are intrinsically non-negative. Examples include the time between trades, the range of a price over a day and realized volatility; see Brownlees and Gallo (2010). Distributions appropriate for non-negative variables include the gamma, Weibull, Burr and F.

As a rule, the location and scale for such distributions are closely connected, usually depending on the same parameter. If the location/scale is to change over time, the use of an exponential link function ensures that it remains positive. The unobserved components model is then

$$y_t = \varepsilon_t \exp(\lambda_t), \qquad 0 \leq y_t < \infty, \quad t = 1, ..., T,$$

where $\lambda_t = \ln \mu_t$ depends on a disturbance term, η_t, which may or may not be correlated with the IID variable, ε_t.

In the first-order model

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

Non-negative variables: unobserved component model

Taking logarithms, that is

$$\ln y_t = \lambda_t + \ln \varepsilon_t, \quad t = 1, ..., T,$$

gives a linear state space form.

For some variables, like the logarithm of range, QML estimation using the Kalman filter may reasonably good because $\ln \varepsilon_t$ is often close to a normal distribution.

Nevertheless efficient estimation usually requires the use of simulation-based methods.

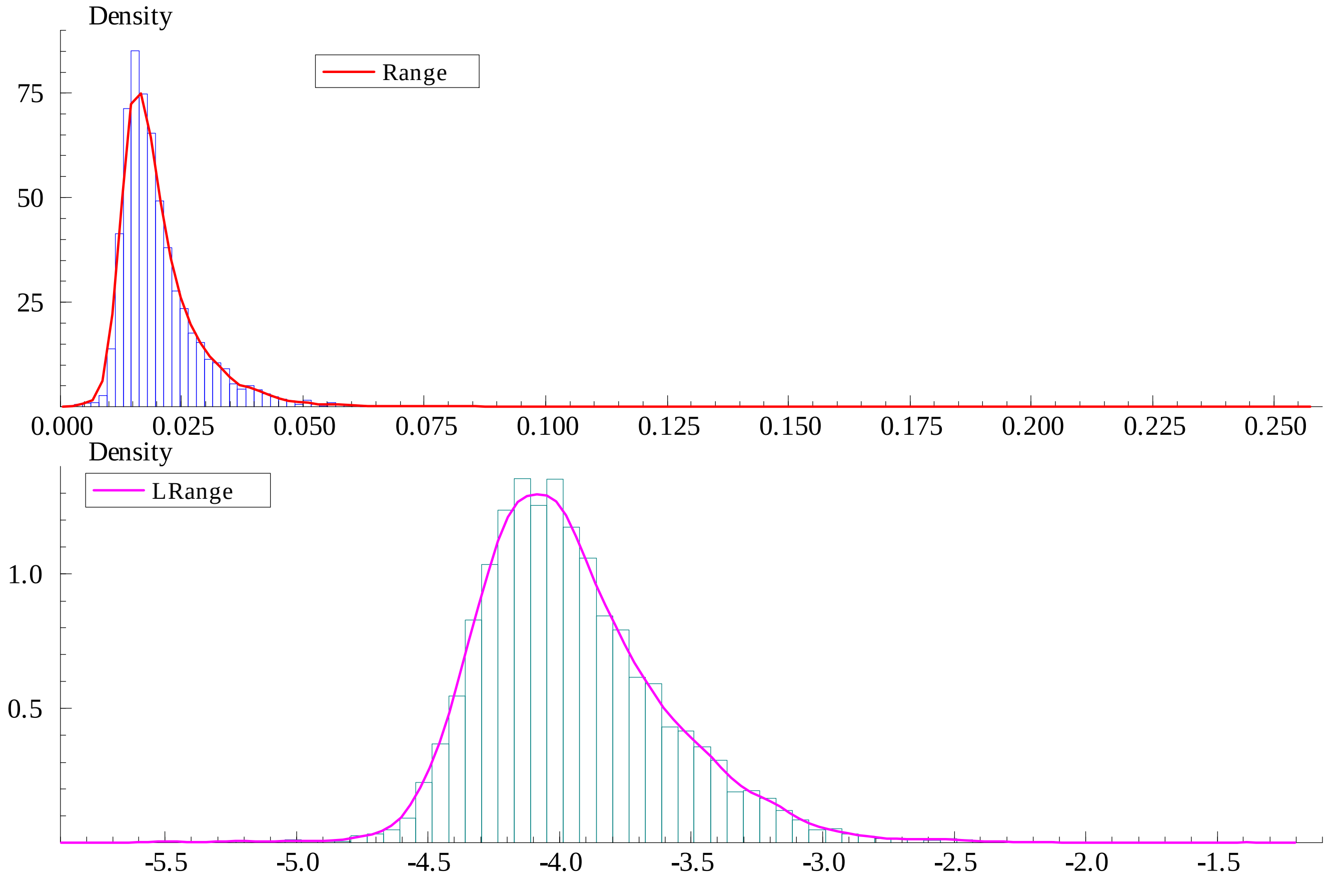


Figure: Distribution of range and its logarithm for Dow-Jones

Non-negative variables: duration, realized volatility and range

Multiplicative error models (MEMs) provide an observation-driven approach for dynamic non-negative variables; see Russell and Engle (2010) for a recent survey. In these models, the conditional mean, $\mu_{t|t-1}$, and hence the conditional scale, is a linear function of past observations. The model can be written

$$\begin{array}{rcl} y_t & = & \varepsilon_t \mu_{t|t-1}, \qquad 0 \leq y_t < \infty, \quad t = 1, ..., T, \\ \mu_{t+1|t} & = & \delta + \beta \mu_{t|t-1} + \alpha y_t, \qquad \delta, \alpha, \beta > 0 \end{array}$$

where ε_t has a distribution with mean one. The emphasis in early work was on the gamma and Weibull distributions, both of which include the exponential distribution as a special case.
An exponential link function, $\mu_{t|t-1} = \exp(\lambda_{t|t-1})$ ensures that $\mu_{t|t-1}$ is positive. Exponential link functions have been studied and applied by Brandt and Jones (2006) and others.

Non-negative variables: DCS models

It is the combination of the exponential link function with the conditional score that facilitates the development of an asymptotic distribution theory and enables comprehensive expressions for the moments, autocorrelations and forecasts to be derived.

The practical implication is that the conditional score for a heavy-tailed distribution will give extreme observations less weight than they would receive in the standard MEM framework.

It is not always convenient to define ε_t so that its mean is one. For many purposes it is better to work with a measure of scale and to set its logarithm equal to $\lambda_{t|t-1}$. Since scale and location only differ by a factor of proportionality, the statistical properties of parameters estimated with an exponential link function are essentially unchanged.

Non-negative variables: DCS models

The model can be written

$$y_t = \varepsilon_t \exp(\lambda_{t|t-1}), \quad t = 1, ..., T,$$

$$\lambda_{t+1|t} = (1-\phi)\omega + \phi\lambda_{t|t-1} + \kappa u_t, \qquad |\phi| < 1,$$

where ω is the unconditional mean of $\lambda_{t|t-1}$ and $\exp(\lambda_{t|t-1})$ is equal to a measure of scale, with the distribution of ε_t standardized accordingly.

The dynamics are driven by the (standardized) score, u_t.

The score is linear for a gamma distribution, but exteme observations are discounted when the conditional distribution has a heavy tail.

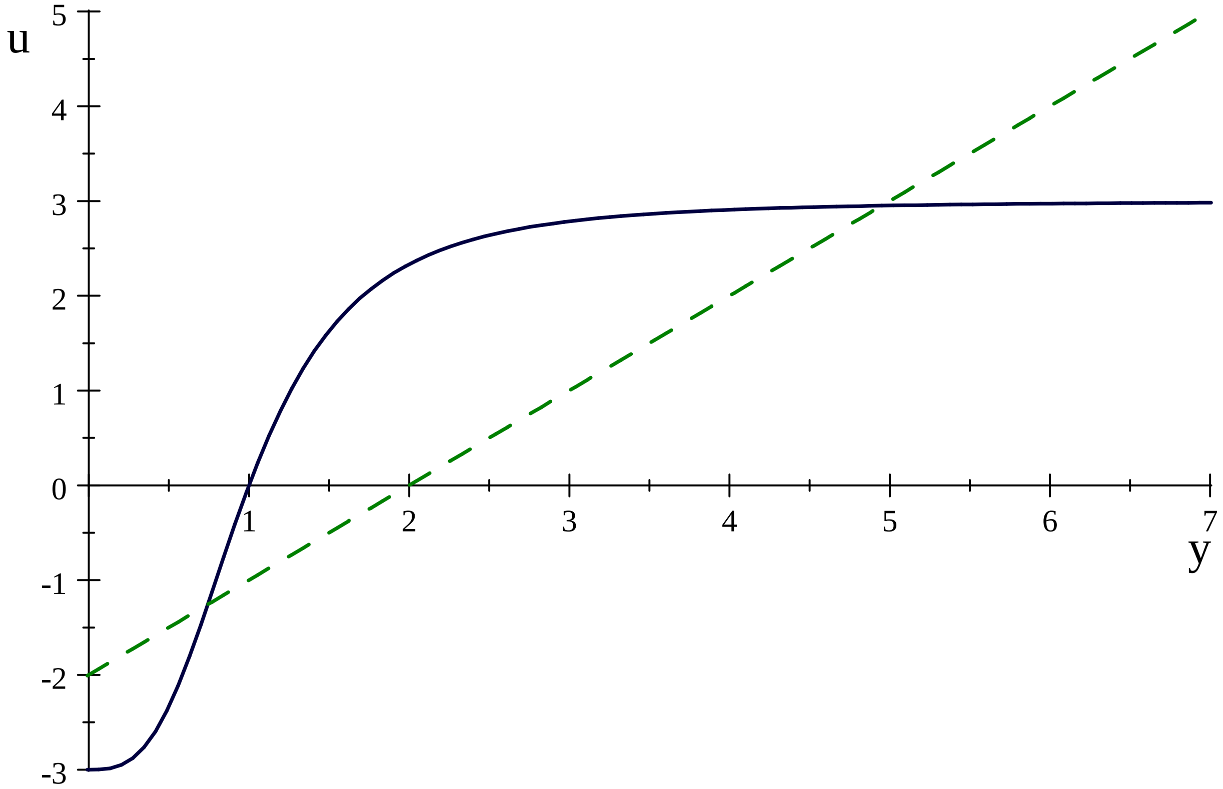


Figure: Impact of u for a log-logistic distribution and a gamma (dashed), with shape parameters $\nu = 3$ and $\gamma = 2$ respectively.

Generalized gamma and beta distributions

The statistical theory of DCS models for non-negative variables is simplified by the fact that for the gamma and Weibull distributions the score and its derivatives are dependent on a gamma variate, while for the Burr, log-logistic and F-distributions the dependence is on a beta variate.
(1) Gamma and Weibull distributions are special cases of the *generalized gamma distribution*.
(2) Burr and log-logistic distributions are special cases of the **generalized beta distribution**.
The $F-$distribution is related to the generalized beta distribution
Members of the generalized beta class are particularly useful in situations where there is evidence of heavy tails.

Multivariate models

The DCS location model is

$$\mathbf{y}_t = \boldsymbol{\omega} + \boldsymbol{\mu}_{t|t-1} + \boldsymbol{\nu}_\mathbf{t}, \quad \boldsymbol{\nu}_t \sim t_\nu\left(\mathbf{0},\Omega\right), \quad t = 1, ..., T$$

$$\boldsymbol{\mu}_{t+1|t} = \Phi\boldsymbol{\mu}_{t|t-1} + \mathbf{K}\mathbf{u}_t.$$

Asymptotic theory relatively straightforward. Constraints for common trends and other factors similar to UC models - Harvey (1989, ch 8).

Multivariate models

$$\mathbf{y}_t = \boldsymbol{\omega} + \boldsymbol{\nu}_\mathbf{t}, \quad \boldsymbol{\nu}_t \sim t_\nu\left(\mathbf{0},\, \Omega_{t|t-1}\right)$$

A direct extension of Beta-t-EGARCH to model changing scale, $\Omega_{t|t-1}$, is difficult. Matrix exponential is $\Omega_{t|t-1} = \exp \Lambda_{t|t-1}$. As a result, $\Omega_{t|t-1}$ is always p.d. and if $\Lambda_{t|t-1}$ is symmetric then so is $^{\cdot}{}_{t|t-1}$; see Kawakatsu (2006, JE). Unfortunately, the relationship between the elements of $\Omega_{t|t-1}$ and those of $\Lambda_{t|t-1}$ is hard to disentangle. Can't separate scale from association.

Issues of interpretation aside, differentiation of the matrix exponential is needed to obtain the score and this is not straightforward.

Multivariate models for changing scale

A better way forward is to follow the approach in Creal, Koopman and Lucas (2011, JBES) and let

$$\Omega_{t|t-1} = \mathbf{D}_{t|t-1}\mathbf{R}_{t|t-1}\mathbf{D}_{t|t-1},$$

where $\mathbf{D}_{t|t-1}$ is *diagonal* and $\mathbf{R}_{t|t-1}$ is a pd correlation matrix with diagonal elements equal to unity. An exponential link function can be used for the volatilities in $\mathbf{D}_{t|t-1}$.
Changing correlations in $\mathbf{R}_{t|t-1}$.
The condititional score also provides guidance on dynamics for a *copula* - Creal et al (2012, JAE).

Distributions

Student's t distribution

The t_ν-distribution with a location of μ and scale of φ has probability density function (PDF)

$$f\left(y;\mu,\varphi,\nu\right) = \frac{\Gamma\left(\left(\nu+1\right)/2\right)}{\Gamma\left(\nu/2\right)\varphi\sqrt{\pi\nu}}\left(1+\frac{\left(y-\mu\right)^2}{\nu\varphi^2}\right)^{-(\nu+1)/2}, \qquad \varphi,\nu>0,$$

where ν is the degrees of freedom and $\Gamma\left(.\right)$ is the gamma function.
Moments exist only up to and including $\nu-1$.
For $\nu>2$, the variance is

$$\sigma^2 = \left\{\nu/\left(\nu-2\right)\right\}\varphi^2.$$

The excess kurtosis, that is the amount by which the normal distribution's kurtosis of three is exceeded, is $6/(\nu-4)$, provided that $\nu>4$.

Distributions

Student's t distribution

The Cauchy distribution is t_1 and famously has no moments. Its PDF is

$$f(y) = \frac{1}{\pi\varphi}\left(1 + \frac{(y-\mu)^2}{\varphi^2}\right)^{-1}.$$

The standardized PDF, that is $\mu = 0$ and $\varphi = 1$, is plotted together with that of the normal in Figure . The tails of the Cauchy distribution are much heavier than those of the normal.

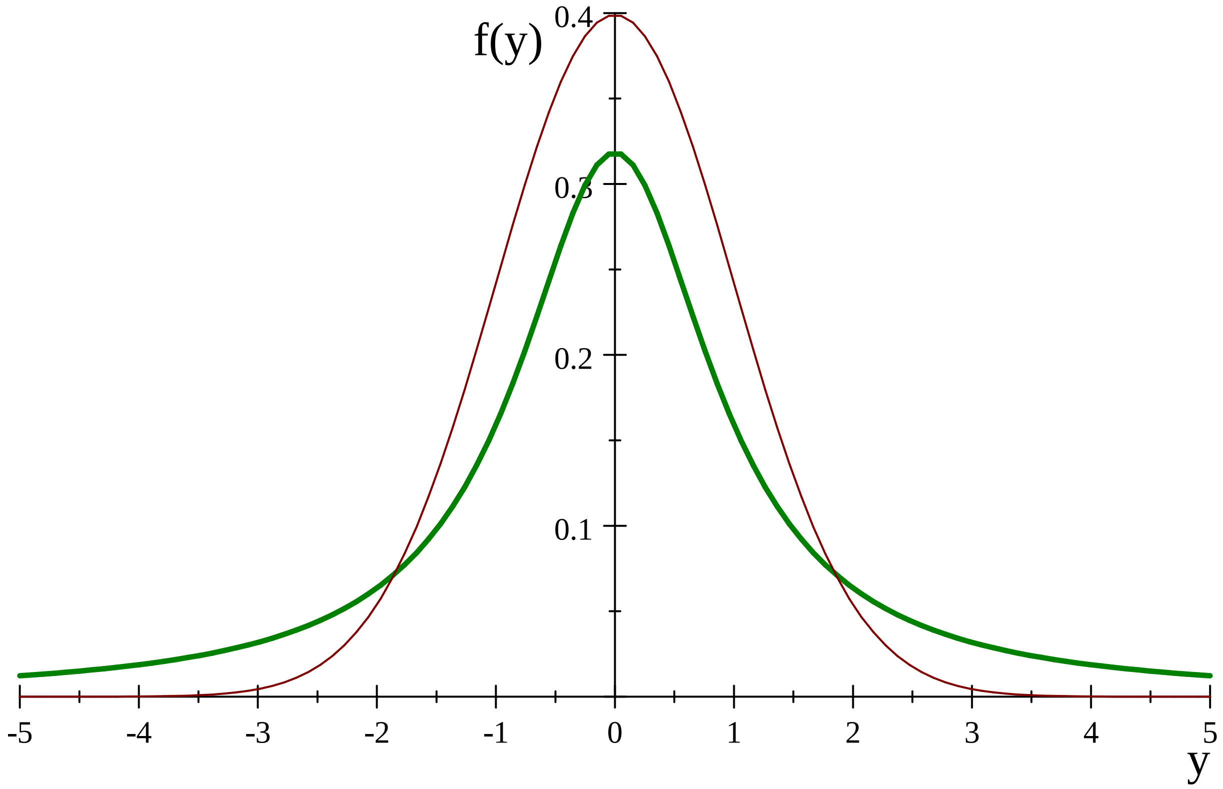


Figure: Cauchy (thick line) and Gaussian distributions

When observations are from a t_ν-distribution in which ν is small, the sample mean is a very inefficient estimator of μ. For example with $\nu = 3$, the efficiency is one-half. The sample variance is even more inefficient. Specifically

$$Eff\,(variance) = (\nu + 3)(\nu - 4)/\{\nu(\nu - 1)\}, \quad \nu > 4.$$

Distributions

General error distribution (GED)

The PDF of the general error distribution, denoted $GED(\upsilon)$, is

$$f\,(y;\mu,\varphi,\upsilon) = \left[2^{1+1/\upsilon}\varphi\Gamma(1+1/\upsilon)\right]^{-1}\exp(-\left|(y-\mu)/\varphi\right|^{\upsilon}/2), \qquad \varphi,\upsilon > 0,$$

where φ is a scale parameter, related to the standard deviation by the formula

$$\sigma = 2^{1/\upsilon}(\Gamma\left(3/\upsilon\right)/\Gamma\left(1/\upsilon\right))^{1/2}\varphi,$$

and υ is a tail-thickness parameter.
Figure shows the standardized $(\varphi = 1)$ PDF for $\upsilon = 1.5$ and $\upsilon = 1$, the second these being the double exponential or Laplace distribution. The normal distribution is obtained when $\upsilon = 2$, in which case $\sigma = \varphi$.

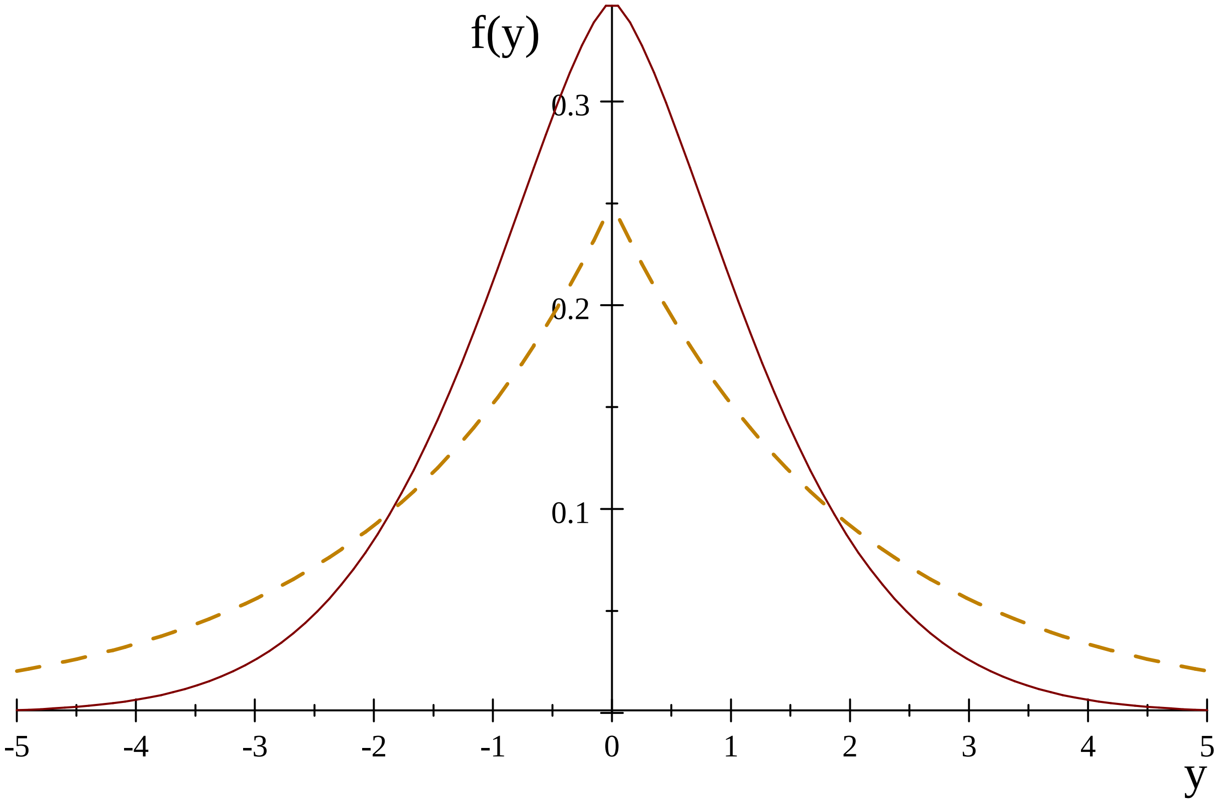


Figure: GED distributions with $\upsilon = 1.5$ and $\upsilon = 1$ (dashed).

Distributions

Beta distribution

When b has a $beta(\alpha, \beta)$ distribution, the PDF is

$$f(b) = \frac{1}{B(\alpha, \beta)} b^{\alpha-1}(1-b)^{\beta-1}, \quad 0 \leq b \leq 1, \qquad \alpha, \beta > 0,$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function.
Lemma When b has a $beta(\alpha, \beta)$ distribution and $w(b)$ is a function of b with finite expectation,

$$E[b^h(1-b)^k w(b)] = \frac{B(\alpha+h, \beta+k)}{B(\alpha, \beta)} E[w(b)], \qquad h > -\alpha, \ k > -\beta,$$

where the expectation on the right hand side is now understood to be with respect to a $beta(\alpha+h, \beta+k)$ distribution.
Note that

$$E(b^h(1-b)^k) = \frac{B(\alpha+h, \beta+k)}{B(\alpha, \beta)}, \quad h > -\alpha, k > -\beta$$

and the moments are obtained by setting $k = 0$.

Distributions
Beta distribution

The moment generating function (MGF) of a beta distribution plays a key role in deriving results in later chapters. Suppose b has a $beta(\alpha, \beta)$ distribution and c is a finite number. Then

$$M_\beta(c;\alpha,\beta) = E(e^{cb}) = 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{c^k}{k!}, \quad \alpha, \beta > 0.$$

The above expression is Kummer's (confluent hypergeometric) function, ${}_1F_1(\alpha;\beta;c)$.
Finally note that when b has a $beta(\alpha,\beta)$ distribution, the distribution of $1-b$ is $beta(\beta,\alpha)$.

Distributions
Gamma distribution

The *PDF* of a $gamma(\alpha,\gamma)$ variable is

$$f(g) = \alpha^{-\gamma}g^{\gamma-1}e^{-g/\alpha}/\Gamma(\gamma), \quad 0 \leq g < \infty, \quad \alpha, \gamma > 0,$$

where α is the scale parameter and γ is the shape parameter. The chi-square distribution with ν degrees of freedom is $gamma(2, \nu/2)$.
Setting $\gamma = 1$ gives the exponential distribution.
The moments of all orders exist, with the raw moments given by

$$E(g^h) = \alpha^h\Gamma(h+\gamma)/\Gamma(\gamma), \quad h > 0.$$

Hence the mean is $\gamma\alpha$, while the variance is $\gamma\alpha^2$. The MGF is

$$M(c;\alpha,\gamma) = E(e^{cg}) = (1-\alpha c)^{-\gamma}, \quad -\infty < c < 1/\alpha, \quad \gamma > 0.$$

There are a number of important connections between the gamma and other distributions.

Distributions

Gamma distribution

Lemma The expected value of the absolute value of a standardized t_ν-variate, ε_t, raised to a power c is

$$E(|\varepsilon_t|^c) = \nu^{c/2}\Gamma(c/2+1/2)\Gamma(-c/2+\nu/2)/(\Gamma(1/2)\Gamma(\nu/2)), \quad -1 < c < \nu.$$

Lemma If z is $gamma(\theta, \alpha)$ and w is $gamma(\theta, \beta)$, then $x = z/(w+z)$ is beta(α, β).
Lemma The variable $(t^2/\nu)/(1+t^2/\nu)$ has a $beta(1/2, \nu/2)$ distribution, while $1/(1+t^2/\nu)$ has a $beta(\nu/2, 1/2)$ distribution.

Distributions

Gamma distribution

Lemma For the GED, $|(y-\mu)/\varphi|^\upsilon$ has a $gamma(2,\ 1/\upsilon)$ distribution
The above result can be easily proved by change of variable. The next result is immediate.
Corollary The expected value of the absolute value of a standardized $GED(\upsilon)$ variate raised to a power c is

$$E(|(y-\mu)/\varphi|^c) = 2^{c/\upsilon}\Gamma((c+1)/\upsilon)/\Gamma(1/\upsilon), \quad c > -1, \quad \upsilon > 0.$$

The above expression gives the even (central) moments of a GED variate. The odd central moments are zero.

Maximum likelihood (ML)

Let y_t, $t = 1, ..., T$, be a set of time series observations, each of which is drawn from a distribution with PDF $f(y_t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. When the observations are IID, the joint density function is just the product of the individual density functions. The likelihood function has the same form as the joint density function. It is more convenient to work with its logarithm

$$\ln L(\boldsymbol{\theta}; y_1, .., y_T) = \sum_{t=1}^{T} \ln f(y_t; \boldsymbol{\theta}).$$

The likelihood function differs from the joint density function in that the observations are taken as given, while $\boldsymbol{\theta}$ is variable. The aim is to find the value of $\boldsymbol{\theta}$ that makes the sample 'most likely'. The global ML estimator, $\widetilde{\boldsymbol{\theta}}$, maximizes $\ln L(\boldsymbol{\theta})$ over the full parameter space. Provided that $\ln L(\boldsymbol{\theta})$ is differentiable at the true parameter value, $\widetilde{\boldsymbol{\theta}}$ will be given by solving the likelihood equations

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1}$$

Maximum likelihood

The information matrix for a single observation is

$$\mathbf{I}(\boldsymbol{\theta}_0) = E_0\left(\frac{\partial \ln f}{\partial \boldsymbol{\theta}} \frac{\partial \ln f}{\partial \boldsymbol{\theta}'}\right) = -E_0\left(\frac{\partial^2 \ln f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right), \quad t = 1, ..., T, \tag{2}$$

where the expectation is taken at the true value of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}_0$. The full information matrix is $T.\mathbf{I}(\boldsymbol{\theta}_0)$. Provided that $f(y_t; \boldsymbol{\theta})$ satisfies certain regularity conditions, $\widetilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ and it is asymptotically normal in that $\sqrt{T}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate normal with mean vector zero and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\theta}_0)$. The information matrix is positive definite provided that the model is identifiable.

The vector in (1) is called the **score** of the likelihood function and, as such, is a function of $\boldsymbol{\theta}$ for a given set of observations. The statistical properties of DCS models are derived with respect to the score vector, $\partial \ln f(y_t; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, where the observations are random variables, just as they are in the formulae for the information matrix, (2).

Maximum likelihood

When observations are from a t_ν-distribution, the logarithm of the PDF, the log-density, is

$$\begin{aligned}\ln f(y_t;\mu,\varphi,\nu) &= \ln\Gamma\left((\nu+1)/2\right)-\frac{1}{2}\ln\pi-\ln\Gamma\left(\nu/2\right)\\ &\quad -\frac{1}{2}\ln\nu-\ln\varphi-\frac{(\nu+1)}{2}\ln\left(1+\frac{(y_t-\mu)^2}{\nu\varphi^2}\right).\end{aligned}$$

The information matrix is

$$\mathbf{I}\begin{pmatrix}\mu\\ \varphi^2\\ \nu\end{pmatrix}=\begin{bmatrix}\frac{\nu+1}{\nu+3}\varphi^{-2} & 0 & 0\\ 0 & \frac{\nu}{2\varphi^4(\nu+3)} & \frac{1}{2\varphi^2(\nu+3)(\nu+1)}\\ 0 & \frac{1}{2\varphi^2(\nu+3)(\nu+1)} & h(\nu)/2\end{bmatrix},$$

where

$$h(\nu)=\frac{1}{2}\psi'\left(\nu/2\right)-\frac{1}{2}\psi'\left((\nu+1)/2\right)-\frac{\nu+5}{\nu\left(\nu+3\right)\left(\nu+1\right)},$$

with $\psi'(.)$ being the trigamma function.

Maximum likelihood

The parameter λ, where $\varphi=\exp(\lambda)$, will normally be employed here when modeling dynamic scale. The information matrix is then

$$\mathbf{I}\begin{pmatrix}\mu\\ \lambda\\ \nu\end{pmatrix}=\begin{bmatrix}\frac{\nu+1}{\nu+3}\exp(-2\lambda) & 0 & 0\\ 0 & \frac{2\nu}{\nu+3} & \frac{1}{(\nu+3)(\nu+1)}\\ 0 & \frac{1}{(\nu+3)(\nu+1)} & h(\nu)/2\end{bmatrix},$$

and a scale parameter no longer appears in the lower block. This feature of the link function turns out to be of crucial importance.

Maximum likelihood estimation of dynamic conditional score models

In DCS models, some or all of the parameters are time-varying and the distribution of y_t is defined conditional on these parameters. For a single time-varying parameter, the dynamics are driven by the conditional score vector, $\partial \ln \ell(\theta_{t|t-1}; y_t)/\partial\theta_{t|t-1}$. A crucial requirement - though not the only one - for establishing results on asymptotic distributions of ML estimators of the parameters governing the movements in $\theta_{t|t-1}$ is that it does not appear in the expression for its information quantity. The fulfillment of this condition may require a careful choice of link function.
1)Basic lemma on the information matrix
2) Information matrix for the first-order model.
3) Asymptotic distribution of the ML estimator
4) Nonstationarity

An information matrix lemma

Suppose initially that there is just one parameter, θ, in a static model. Define

$$u_t = k.\partial \ln f(y_t;\theta)/\partial\theta, \qquad t = 1, ..., T,$$

where k is a finite constant which may be the information quantity.
The derivative $\partial \ln f(y_t;\theta)\partial\theta$ is a random variable which has zero mean at the true parameter value, θ_0. Hence u_t has zero mean at $\theta = \theta_0$ and its variance, σ_u^2, is finite under standard regularity conditions.
The *information quantity* for one observation is

$$I(\theta_0) = -E(\partial^2 \ln f/\partial\theta^2) = E[(\partial \ln f/\partial\theta)^2] = E(u_t^2)/k^2 = \sigma_u^2/k^2 < \infty. \tag{3}$$

An information matrix lemma

Now let $\theta = \theta_{t|t-1}$ evolve over time as a function of past observations and past values of the score of the conditional distribution. Since the conditional score depends on past observatons through $\theta_{t|t-1}$, it can be broken down into two parts:

$$\frac{\partial \ln f_t(y_t \mid Y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \frac{\partial \ln f_t(y_t; \theta_{t|t-1})}{\partial \theta_{t|t-1}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}, \qquad (4)$$

where the notation $f_t(y_t; \theta_{t|t-1})$ indicates that the distribution of y_t depends on the time-varying parameter, $\theta_{t|t-1}$, and $\boldsymbol{\psi}$ denotes the vector of parameters governing the dynamics. Since $\theta_{t|t-1}$ and its derivatives depend only on past information, the distribution of the score conditional on information at time $t-1$ is the same as its unconditional distribution and so is time invariant.

The above decomposition of the conditional score leads to the following result.

An information matrix lemma

Consider a model with a single time-varying parameter, $\theta_{t|t-1}$, which satisfies an equation that depends on variables which are fixed at time $t-1$. The process is governed by a set of fixed parameters, $\boldsymbol{\psi}_0$. If the link function is such that $I(\theta_0)$ does not depend on θ_0, then the conditional score for the $t-$th observation, $\partial \ln f_t(y_t \mid Y_{t-1}; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$, is a MD at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, with conditional covariance matrix

$$E_{t-1}\left(\frac{\partial \ln f_t(y_t \mid Y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \ln f_t(y_t \mid Y_{t-1}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)' = I.\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right),$$

where the information quantitly, I, is constant over time and independent of $\boldsymbol{\psi}$.

Proof The fact that the score in (4) is a MD is confirmed by the fact that the derivative of the time-varying parameter, $\partial \theta_{t|t-1}/\partial \boldsymbol{\psi}$, is fixed at time $t-1$ and the expected value of the score in the static model is zero.

An information matrix lemma

Proof (contd) Write the outer product of (4) as

$$\left(\frac{\partial \ln f_t}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \ln f_t}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)' = \left(\frac{\partial \ln f_t}{\partial \theta_{t|t-1}}\right)^2\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right).$$

Now take expectations conditional on information at time $t-1$. If $E_{t-1}\left(\partial \ln f_t/\partial \theta_{t|t-1}\right)^2$ does not depend on $\theta_{t|t-1}$, it is fixed and equal to the unconditional expectation in the static model, that is (3). Therefore, since $\theta_{t|t-1}$ is fixed at time $t-1$,

$$E_{t-1}\left[\left(\frac{\partial \ln f_t}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \ln f_t}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)'\right] = \left[E\left(\frac{\partial \ln f_t}{\partial \theta}\right)^2\right]\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}.$$

*****

The information matrix is obtained by finding the unconditional expectation on the last matrix.

$$\mathbf{I}(\boldsymbol{\psi}) = I.\mathbf{D}(\boldsymbol{\psi}), \quad \textit{where} \quad \mathbf{D}(\boldsymbol{\psi}) = E\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right).$$

Information matrix for the first-order model

$$\theta_{t+1|t} = \delta + \phi\theta_{t|t-1} + \kappa u_t, \qquad |\phi| < 1, \qquad \kappa \neq 0, \tag{5}$$

is the conventional formulation of a first-order dynamic model, but it turns out that the information matrix takes a simpler form

$$\theta_{t+1|t} = \omega + \theta^{\dagger}_{t+1|t}, \qquad \theta^{\dagger}_{t+1|t} = \phi\theta^{\dagger}_{t|t-1} + \kappa u_t, \qquad t = 1, ..., T. \tag{6}$$

Re-writing the above model in a similar way to (5) gives

$$\theta_{t+1|t} = \omega(1-\phi) + \phi\theta_{t|t-1} + \kappa u_t. \tag{7}$$

where setting $\theta^{\dagger}_{1|0} = 0$ is the same as setting $\theta_{1|0} = \omega$.

Information matrix for the first-order model

The complications arise because u_t depends on $\theta_{t|t-1}$ and hence on the parameters in $\boldsymbol{\psi}$. Differentiating (6) and noting that $\partial\theta_{t|t-1}/\partial\kappa = \partial\theta^{\dagger}_{t|t-1}/\partial\kappa$, $\partial\theta_{t|t-1}/\partial\phi = \partial\theta^{\dagger}_{t|t-1}/\partial\phi$, but $\partial\theta_{t|t-1}/\partial\omega = 1 + \partial\theta^{\dagger}_{t|t-1}/\partial\omega$, the vector $\partial\theta_{t+1|t}/\partial\boldsymbol{\psi}$ becomes

$$
\begin{aligned}
\frac{\partial\theta_{t+1|t}}{\partial\kappa} &= \phi\frac{\partial\theta_{t|t-1}}{\partial\kappa} + \kappa\frac{\partial u_t}{\partial\kappa} + u_t \\
\frac{\partial\theta_{t+1|t}}{\partial\phi} &= \phi\frac{\partial\theta_{t|t-1}}{\partial\phi} + \kappa\frac{\partial u_t}{\partial\phi} + \theta_{t|t-1} - \omega \\
\frac{\partial\theta_{t+1|t}}{\partial\omega} &= \phi\frac{\partial\theta_{t|t-1}}{\partial\omega} + \kappa\frac{\partial u_t}{\partial\omega} + 1 - \phi.
\end{aligned}
$$

Information matrix for the first-order model

However,

$$\frac{\partial u_t}{\partial\kappa} = \frac{\partial u_t}{\partial\theta_{t|t-1}}\frac{\partial\theta_{t|t-1}}{\partial\kappa},$$

and similarly for the other two derivatives. Therefore

$$
\begin{aligned}
\frac{\partial\theta_{t+1|t}}{\partial\kappa} &= x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa} + u_t \\
\frac{\partial\theta_{t+1|t}}{\partial\phi} &= x_t\frac{\partial\theta_{t|t-1}}{\partial\phi} + \theta_{t|t-1} - \omega \\
\frac{\partial\theta_{t+1|t}}{\partial\omega} &= x_t\frac{\partial\theta_{t|t-1}}{\partial\omega} + 1 - \phi,
\end{aligned}
\qquad (8)
$$

where

$$x_t = \phi + \kappa\frac{\partial u_t}{\partial\theta_{t|t-1}}, \qquad t = 1, ..., T.$$

Information matrix for the first-order model

Assume that $E(u_t u_t') < \infty$ *and* $E(u_t'^2) < \infty$ *as well as* $E(u_t^2) < \infty$.
The following definitions are needed:

$$\begin{aligned}
a &= E_{t-1}(x_t) = \phi + \kappa E_{t-1}\left(\frac{\partial u_t}{\partial \theta_{t|t-1}}\right) = \phi + \kappa E\left(\frac{\partial u_t}{\partial \theta}\right) \qquad (9) \\
b &= E_{t-1}(x_t^2) = \phi^2 + 2\phi\kappa E\left(\frac{\partial u_t}{\partial \theta}\right) + \kappa^2 E\left(\frac{\partial u_t}{\partial \theta}\right)^2 \geq 0 \\
c &= E_{t-1}(u_t x_t) = \kappa E\left(u_t \frac{\partial u_t}{\partial \theta}\right)
\end{aligned}$$

Because the above expectations are time invariant the unconditional expectations can replace conditional ones.
The following lemma is a pre-requisite for Theorem 1.

Information matrix for the first-order model

Lemma When the process for $\theta_{t|t-1}$ starts in the infinite past and $|a| < 1$,

$$\begin{aligned}
E\left(\frac{\partial \theta_{t+1|t}}{\partial \kappa}\right) &= 0, \qquad t = ...0, 1, ..., T, \qquad (10) \\
E\left(\frac{\partial \theta_{t+1|t}}{\partial \phi}\right) &= 0, \\
E\left(\frac{\partial \theta_{t+1|t}}{\partial \omega}\right) &= \frac{1-\phi}{1-a}.
\end{aligned}$$

Information matrix for the first-order model

Proof Taking the conditional expectations of the first equation in (8) gives

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)=E_{t-1}\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa}+u_t\right)=a\frac{\partial\theta_{t|t-1}}{\partial\kappa}+0$$

and, from the law of iterated expectations, the expectation at time $t-2$ is given by

$$\begin{aligned}E_{t-2}E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right) &= aE_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right)\\ &= aE_{t-2}\left(x_{t-1}\frac{\partial\theta_{t-1|t-1}}{\partial\kappa}+u_{t-1}\right)=a^2\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}.\end{aligned}$$

Hence, if $|a|<1$,

$$\lim_{n\to\infty}E_{t-n}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)=0,\qquad t=...0,1,...,T.$$

Information matrix for the first-order model

Proof (cont'd) As regards ω,

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right)=a\frac{\partial\theta_{t|t-1}}{\partial\omega}+1-\phi. \tag{11}$$

We can continue to evaluate this expression by substituting for $\partial\theta_{t|t-1}/\partial\omega$, taking conditional expectations at time $t-2$, and then repeating this process. Once a solution has been shown to exist, the result can be confirmed by taking unconditional expectations in (11) to give

$$E\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right)=\frac{1-\phi}{1-a}.$$

Information matrix for the first-order model

Proof (cont'd) Similarly taking conditional expectations of $\partial\theta_{t+1|t}/\partial\phi$ at time $t-1$ gives

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\phi}\right)=a\frac{\partial\theta_{t|t-1}}{\partial\phi}+(\theta_{t|t-1}-\omega)$$

and then the unconditional expectations yields

$$E\left(\frac{\partial\theta_{t+1|t}}{\partial\phi}\right)=0,\quad t=...0,1,...,T.$$

The above Lemma requires that $|a|<1$. The result on the information matrix below requires $b<1$ and fulfillment of this condition implies $|a|<1$. That this is the case follows directly from the Cauchy-Schwarz inequality $E(x_t^2)\geq[E\,|x_t|)]^2$.

Information matrix for the first-order model: Theorem 1

Provided that $b<1$,

$$\mathbf{I}(\boldsymbol{\psi})=I.\mathbf{D}(\boldsymbol{\psi})=(\sigma_u^2/k^2)\mathbf{D}(\boldsymbol{\psi}),\tag{12}$$

where

$$\mathbf{D}(\boldsymbol{\psi})=\mathbf{D}\begin{pmatrix}\kappa\\ \phi\\ \omega\end{pmatrix}=\frac{1}{1-b}\begin{bmatrix}A & D & E\\ D & B & F\\ E & F & C\end{bmatrix}\tag{13}$$

with

$$\begin{aligned}A&=\sigma_u^2, & B&=\frac{\kappa^2\sigma_u^2(1+a\phi)}{(1-\phi^2)(1-a\phi)}, & C&=\frac{(1-\phi)^2(1+a)}{1-a},\\ D&=\frac{a\kappa\sigma_u^2}{1-a\phi}, & E&=\frac{c(1-\phi)}{1-a}\quad \text{and} & F&=\frac{ac\kappa(1-\phi)}{(1-a)(1-a\phi)}.\end{aligned}$$

Information matrix for the first-order model

Derivation of A is given to illustrate the method. This term is the unconditional expectation of the square of the first derivative in (‘). To evaluate it, first take conditional expectations at time $t-1$, to obtain

$$\begin{aligned} E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 &= E_{t-1}\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa}+u_t\right)^2 \\ &= b\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right)^2+2c\frac{\partial\theta_{t|t-1}}{\partial\kappa}+\sigma_u^2. \qquad (14) \end{aligned}$$

It follows from earlier Lemma that the unconditional expectation of the second term is zero. Eliminating this term, and taking expectations at $t-2$ gives

$$\begin{aligned} E_{t-2}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 &= bE_{t-2}\left(x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}+u_{t-1}\right)^2+\sigma_u^2 \\ &= b^2\left(\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}\right)^2+2cb\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}+b\sigma_u^2+\sigma_u^2. \end{aligned}$$

Information matrix for the first-order model

Again the second term can be eliminated and it is clear that

$$\lim_{n\to\infty} E_{t-n}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2=\frac{\sigma_u^2}{1-b}.$$

Taking unconditional expectations in (14) gives the same result.

Asymptotic theory for the first-order model

Theorem

The ML estimator of $\boldsymbol{\psi}$, denoted $\widetilde{\boldsymbol{\psi}}$, is consistent and the limiting distribution of $\sqrt{T}(\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)$ multivariate normal with mean vector zero and covariance matrix

$$Var(\widetilde{\boldsymbol{\psi}}) = \mathbf{I}^{-1}(\boldsymbol{\psi}_0) = I^{-1}\mathbf{D}^{-1}(\boldsymbol{\psi}_0). \qquad (15)$$

The interpretation is that $\widetilde{\boldsymbol{\psi}}$ is approximately normal with mean $\boldsymbol{\psi}$ and

$$Avar(\widetilde{\boldsymbol{\psi}}) = T^{-1}Var(\widetilde{\boldsymbol{\psi}}). \qquad (16)$$

The asymptotic standard error (ASE) of an element of $\widetilde{\boldsymbol{\psi}}$ is the square root of the corresponding diagonal element of $Avar(\widetilde{\boldsymbol{\psi}})$.
The proof follows along the lines of Jensen and Rahbek (ET, 2004), but the details are much simpler. Assume that $\ln L$ is three times continuously differentiable for $\boldsymbol{\psi}$, $\boldsymbol{\psi}_0$ is an interior point of the compact parameter space, and the following are true.

Asymptotic theory for the first-order model

(i) As $T \to \infty$, $(1/\sqrt{T})\partial \ln L(\boldsymbol{\psi}_0)/\partial\boldsymbol{\psi} \to N(0, \mathbf{I}(\boldsymbol{\psi}_0))$, where $\mathbf{I}(\boldsymbol{\psi}_0)$ is p.d.
(ii) As $T \to \infty$, $(-1/T)\partial^2 \ln L(\boldsymbol{\psi}_0)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}' \xrightarrow{P} \mathbf{I}(\boldsymbol{\psi}_0)$
(iii) If $\psi_i, i = 1, .., n$ denote the parameters in $\boldsymbol{\psi}$,

$$\max_{h,i,j=1,\ldots,n} \sup_{\boldsymbol{\psi}\in N(\boldsymbol{\psi}_0)} \left| \frac{\partial^3 \ln L}{T.\partial\psi_h \partial\psi_i \partial\psi_j} \right| \leq c_T,$$

where $N(\boldsymbol{\psi}_0)$ is a neighbourhood of the true parameter value, $\boldsymbol{\psi}_0$, and $0 \leq c_T \xrightarrow{P} c$, $0 < c < \infty$.
It is not difficult to show that (iii) holds for IID observations from Student's t as the score and its first two derivatives are **bounded**. The argument is combined with the following result.
Proposition *The first three derivatives of $\theta_{t|t-1}$ wrt κ, ϕ and ω are stochastic recurrence equations (SREs) and the condition $b < 1$ is sufficient to ensure that they are strictly stationarity and ergodic at the true parameter value.*

Tests

Before fitting a DCS model it is advisable to test whether time-variation is, in fact, present. When a series is random, a first-order DCS model will not be identifiable because the information matrix is singular. Even if a numerical optimization procedure is coaxed into convergence, a Wald test in which the null hypothesis is $\kappa = 0$ should not be carried out.

The **Lagrange multiplier principle** may be used to construct suitable tests against serial correlation in the feature of interest by taking account of the form of the conditional distribution posited for the dynamic model. Such tests should be able to guard against low power as a consequence of outliers.

After a model has been fitted, diagnostics tests may be constructed based on similar principles. These tests may be complemented by procedures for assessing the goodness of fit of the assumed conditional distribution.

Serial correlation

The standard tests against serial correlation are based on the sample autocorrelations, $r(j) = c(j)/c(0)$, where

$$c(j) = T^{-1} \sum_{t=j+1}^{T} (y_t - \overline{y})(y_{t-j} - \overline{y}), \qquad j = 0, 1, 2, \ldots$$

The portmanteau test uses the statistic

$$Q(P) = T \sum_{j=1}^{P} r^2(j),$$

where P is a non-zero integer; the Ljung-Box modification is

$$Q_{LB}(P) = T(T+2) \sum_{\tau=1}^{P} (T-\tau)^{-1} r^2(\tau).$$

Both $Q(P)$ and $Q_{LB}(P)$ are asymptotically χ^2_P when the observations are independent. The weaker assumption that the observations are serially uncorrelated is not sufficient.

Serial correlation

The portmanteau test may be derived as a Lagrange multiplier (LM) test against a Gaussian moving average process of order P. Similarly a DCS test may be developed by considering the $(0, P-1)$ dynamic process

$$\theta_{t|t-1} = \omega + \kappa_0 u_{t-1} + ... + \kappa_{P-1} u_{t-P}, \quad t = 1, ..., T,$$

where u_t is the score for the $t - th$ observation.

Proposition

The LM test statistic for $\kappa_0 = \kappa_1 = ... = \kappa_{P-1} = 0$ *is the portmanteau statistic*

$$Q_u(P) = T \sum_{j=1}^{P} r_u^2(j), \tag{17}$$

where $r_u(j)$ *is the* $j - th$ *sample autocorrelation of* u_t*. The asymptotic distribution of* $Q_u(P)$ *under the null hypothesis is* χ_P^2.

Serial correlation

Proof For simplicity assume that $k = 1$; when $k \neq 1$, it cancels in the final expression. We have

$$\frac{\partial \theta_{t|t-1}}{\partial \kappa_j} = \sum_{i=1}^{P} \kappa_{i-1} \frac{\partial u_{t-i}}{\partial \kappa_j} + u_{t-j-1}, \quad j = 0, ..., P-1,$$

but under the null hypothesis, when $\boldsymbol{\kappa} = (\kappa_0, .., \kappa_{P-1})' = \mathbf{0}$,

$$\frac{\partial \theta_{t|t-1}}{\partial \kappa_j} = u_{t-j-1}, \quad j = 0, ..., P-1.$$

Hence

$$\mathbf{D}(\boldsymbol{\kappa}) = E\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\kappa}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\kappa}'}\right) = \sigma_u^2 \mathbf{I}_P$$

and so

$$\mathbf{I}(\boldsymbol{\kappa}) = \sigma_u^2 \mathbf{D}(\boldsymbol{\kappa}) = \sigma_u^4 \mathbf{I}_P.$$

Serial correlation

The result then follows from the standard LM formula, that is

$$LM = \frac{\partial \ln L}{\partial \boldsymbol{\kappa}'} \mathbf{I}^{-1}(\boldsymbol{\kappa}) \frac{\partial \ln L}{\partial \boldsymbol{\kappa}},$$

since

$$\frac{\partial \ln f_t}{\partial \kappa_j} = \frac{\partial \ln f_t}{\partial \theta_{t|t-1}} \frac{\partial \theta_{t|t-1}}{\partial \kappa_j} = u_t u_{t-j-1}, \quad j = 0, 1, ..., P-1$$

Estimation of fixed parameters, such as degrees of freedom or the unconditional mean, makes no difference to the form of the result. *Note that under the null hypothesis, the first-order condition,* $\partial \ln L / \partial \omega = 0$, *implies that* $\sum_{t=1}^{T} u_t = 0$ *and so* $\overline{u} = 0$.

Serial correlation

Remark

The Ljung-Box statistic

$$Q_u^*(P) = T(T+2) \sum_{j=1}^{P} (T-j)^{-1} r_u^2(j),$$

may also be used. The asymptotic distribution under the null hypothesis is also χ_P^2.

Estimation of dynamic parameters will affect the asymptotic distribution of the portmanteau statistic constructed from residuals. The solution, as in the Box-Pierce test, is to reduce the degrees of freedom by the number of dynamic parameters estimated, that is $p+q$ for when an $ARMA(p,q)$ model is fitted.

Another possibility is to develop LM tests.

Serial correlation

It is not unusual to be faced with a situation where the serial correlation in $\theta_{t|t-1}$ is very persistent. In the first-order model, persistence means that the parameter ϕ is close to one. In these circumstances a stationarity test, such as the one proposed by Nyblom and Mäkeläinen (1983) or the more general test of Kwiatkowski *et al* (1992) will tend to have higher power than the portmanteau test; see the evidence in Harvey and Streibel (1998). The test statistics could be constructed from conditional score variables.

Goodness of fit of distributions

Assessing how well a particular conditional distribution fits the data is an important part of the methodology of DCS models. The goodness of fit of a Gaussian distribution is routinely tested with the Bowman-Shenton (Jarque-Bera) statistic which is based on sample skewness and kurtosis. The fact that the test statistic consistes of two parts is helpful in determining whether a rejection is based primarily on skewness or kurtosis. More generally, goodness of fit is assessed by reference to the quantiles of the assumed theoretical distribution. There are two approaches. The first is based on the QQ plot, in which the T equally-spaced quantiles of the comparison distribution are plotted on the vertical axis and the order statistics of the sample, $y_{(1)} \leq y_{(2)} \leq ...... \leq y_{(T)}$, are plotted on the horizontal axis. If the sample comes from the comparison distribution, the plot will approximate a straight 45 degree line.

Goodness of fit of distributions

A second plot, which essentially presents the same information, but in a different way, uses the probability integral transform (PIT). The PIT of a variable is given by its CDF, that is $PIT(y) = F(y)$. By construction, the PIT has a standard uniformdistribution, that is the range is [0,1]. If the PITs are not uniformly distributed, the shape of the histogram can be informative. For example, a hump indicates that the forecasts are too narrow and that the tails are not adequately represented. The hypothesis that a set of T observations comes from a particular parametric distribution can be tested using the Kolmogorov-Smirnov statistic

$$KS_{PIT} = \max_j \left| PIT_{(j)} - j/T \right| ,$$

where $PIT_{(j)}$, $j = 1, ..., T$ denotes the ordered $PIT(y_t)'s$. When parameters have been estimated, the distribution of the Kolmogorov-Smirnov statistic will normally have to be found by simulation.

Goodness of fit of distributions

Example

The PIT of a variable that has an exponential distribution with mean θ is simply

$$PIT(y_t) = F(y_t) = \int_0^{y_t} \theta^{-1} \exp(-x/\theta)dx = 1 - \exp(-y_t/\theta).$$

Berkowitz (2001) suggests transforming the PITs to standard normal variables using an algorithm for computing the inverse of the Gaussian distribution function (the quantile function). This opens up a wide range of tests appropriate for the normal distribution, including the Bowman-Shenton test.

Goodness of fit of distributions

Diebold *et al* (1998) discuss various ways in which the PITs may be used to assess forecasting schemes. In the present context this means checking model specification using the PITs of the one-step ahead prediction errors. For example, plots of the autocorrelation functions of the PITs, and of their powers, may indicate the source of serial dependence.

Tests based on standardized residuals, their PITs and the normal variates created from PITs offer three alternatives for diagnostic checking. Since the variables are all transformations of each other, it is not clear which will be preferable. Indeed the answer almost certainly depends on the assumed distribution.

The fitting of DCS models also offers a fourth possibility, namely the use of the **conditional scores**. The conditional scores may be appropriate for testing against serial correlation and in cases where they have a known distribution comparing the empirical and theoretical distributions may be valuable.

Model fit

The overall fit of a model can be assessed in a number of ways. The various criteria which are discussed here are generally applicable and so can be used to compare DCS models with other models. It is assumed that estimation is by maximum likelihood.

The maximized log-likelihood is the basic measure of goodness of fit, but in order to compare different models an allowance is often made for the number of parameters, n, estimated. The *Akaike information criterion* (AIC) is defined as

$$AIC = -2\ln L(\widetilde{\boldsymbol{\psi}}) + 2n,$$

while the BIC has $2n$ replaced by $n\ln T$.

Post-sample predictive testing may also be employed. The predictive likelihood, sometimes called the log-score, is both simple and effective; see Mitchell and Wallis (2011). Looking at the post-sample PIts, residuals and scores may also be useful.

Explanatory variables

The changing parameter may depend on a set of observable explanatory variables, denoted by the $k \times 1$ vector $\mathbf{w}_t$, as well as on its own past values and the score. The model can be set up as

$$\theta_{t|t-1} = \mathbf{w}_t'\gamma + \theta^{\dagger}_{t|t-1}, \qquad t = 1, ..., T, \tag{18}$$

where, in the first-order case,

$$\theta^{\dagger}_{t+1|t} = \phi\theta^{\dagger}_{t|t-1} + \kappa u_t, \qquad \textit{with} \qquad \theta^{\dagger}_{1|0} = 0$$

The explanatory variables are strictly exogenous in the sense that they are independent of the other variables in the model in all time periods.
The following result is a generalization of Theorem 1. A constant mean, ω, is a special case, obtained when $\mathbf{w}_t$ is a scalar equal to unity.

Explanatory variables

Corollary Assume that the explanatory variables are weakly stationary with mean $\boldsymbol{\mu}_w$ and second moment Λ_w and are strictly exogenous. The information matrix for model (18) is

$$\mathbf{I}\begin{pmatrix} \kappa \\ \phi \\ \gamma \end{pmatrix} = \frac{\sigma_u^2}{k^2(1-b)}\begin{bmatrix} A & D & E\boldsymbol{\mu}_w' \\ D & B & F\boldsymbol{\mu}_w' \\ E\boldsymbol{\mu}_w & F\boldsymbol{\mu}_w & \mathbf{C}_w \end{bmatrix},$$

where A, B, C, D, E and F are as in (13) while

$$\mathbf{C}_w = (1+\phi^2)\Lambda_w - 2\phi\Lambda_w(1) + \frac{2a(1-\phi)^2}{1-a}\boldsymbol{\mu}_w\boldsymbol{\mu}_w',$$

with $\Lambda_w(1) = E(\mathbf{w}_t\mathbf{w}_{t-1}') = E(\mathbf{w}_{t-1}\mathbf{w}_t')$.