Filtering with heavy tails

Andrew HARVEY and Alessandra LUATI *

Abstract

An unobserved components model in which the signal is buried in noise that is non-Gaussian may throw up observations that, when judged by the Gaussian yardstick, are outliers. We describe an observation driven model, based on a conditional Student t-distribution, that is tractable and retains some of the desirable features of the linear Gaussian model. Letting the dynamics be driven by the score of the conditional distribution leads to a specification that is not only easy to implement, but which also facilitates the development of a comprehensive and relatively straightforward theory for the asymptotic distribution of the maximum likelihood estimator. The methods are illustrated with an application to rail travel in the UK. The final part of the article shows how the model may be extended to include explanatory variables.

Keywords: outlier; robustness; score; seasonal; t-distribution; trend.

*Andrew Harvey is Professor of Econometrics at the Faculty of Economics, University of Cambridge, UK (E-mail: ach34@econ.cam.ac.uk). Alessandra Luati is Associate Professor of Statistics at the Department of Statistics, University of Bologna, Italy (E-mail: alessandra.luati@unibo.it). We would like to thank two Referees for their helpful comments.

1 Introduction

Linear Gaussian unobserved components models play an important role in time series modeling. The Kalman filter and associated smoother provide the basis for a comprehensive statistical treatment. The filtered and smoothed estimators of the signal are optimal, in the sense of minimizing the mean square error (MSE), the likelihood is given as a by-product of one-step prediction errors produced by the Kalman filter and the full multi-step predictive distribution has a known Gaussian distribution.

A model in which the signal is buried in noise that is non-Gaussian may throw up observations that, when judged by the Gaussian yardstick, are outliers. The purpose of this article is to investigate the practical value of an observation driven model that is tractable and retains some of the desirable features of the linear Gaussian model. The principal feature of the model is that the dynamics are driven by the score of the conditional distribution of the observations. As a result it is not only easy to implement, but its form also facilitates the development of a comprehensive and relatively straightforward theory for the asymptotic distribution of the maximum likelihood estimator. Models of this kind are called dynamic conditional score (DCS) models and they have already proved useful for modeling volatility; see Creal, Koopman and Lucas (2011) and Harvey (2013, ch. 4).

Modeling the additive noise with a Student t-distribution is effective and theoretically straightforward. Indeed the attractions of using the t-distribution to guard against outliers in static models are well-documented; see, for example, Lange, Little and Taylor (1989) and Delaigle, Hall and Jin (2011). The approach based on specifying a heavy tail distribution for the underlying process may be contrasted with the methods adopted in the robustness literature; see, for example, Muler, Peña and Yohai (2009).

The plan of the article is as follows. Section 2 sets out a simple unobserved components model and discusses the rationale for letting the dynamics depend on the conditional score. The first-order conditional score model for a Student t-distribution is described in Section 3. The asymptotic distribution of the maximum likelihood estimator is given in Section 4 and

complemented by a Monte Carlo study on small sample properties. Section 5 then extends DCS models using the state space form and Section 6 discusses how to model trend and seasonality. The viability of a DCS model with trend and seasonal components is demonstrated with real data in Section 6. Explanatory variables are introduced into the model in Section 7 and asymptotic results are presented. Section 8 concludes.

2 Unobserved components and filters

A simple Gaussian signal plus noise model is

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

for $t = 1, \ldots, T$ and where the irregular and level disturbances, ε_t and η_t respectively, are mutually independent and the notation $NID\left(0, \sigma^2\right)$ denotes normally and independently distributed with mean zero and variance σ^2. The autoregressive 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) of (1) is an ARMA(1,1) process

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

but with restrictions on θ. For example, when $\phi = 1$, $0 \leq \theta \leq 1$. The latter are obtained by equating the autocorrelation function (ACF) of y_t in (1), with the ACF expressed in terms of the parameters of the ARMA reduced form, see Harvey (1989, section 2.5.3). The forecasts from the unobserved components (UC) model and RF 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). The parameters ϕ and q may be estimated by maximum likelihood (ML), with the likelihood function constructed from the one-step ahead prediction errors. The KF can be expressed as a single equation which combines $\mu_{t|t-1}$, the optimal estimator of μ_t based on information at time $t-1$, with y_t 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 of the KF:

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

The Kalman gain, k_t, depends on ϕ and q. In the steady-state, k_t is constant. Setting it equal to κ 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 (1) 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. The RF is still an ARMA(1,1) process, 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 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 these issues in the robustness literature; see, for example the book by Maronna, Martin and Yohai (2006, ch. 8).

Simulation methods, such as Markov chain Monte Carlo (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). 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 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 filter in which the prediction error, v_t, in the KF equation is replaced by a variable, u_t, that is proportional

to the score of the conditional distribution. Thus the second equation in (3) becomes

$$\mu_{t+1|t} = \phi\mu_{t|t-1} + \kappa u_t \tag{4}$$

where κ is treated as an unknown parameter. The attraction of this observation driven model is that it becomes possible to derive the asymptotic distribution of the maximum likelihood estimator and generalize in various directions.

3 Dynamic Student-t location model

When the location changes over time, it may be captured by a model in which, conditional on past observations, y_t has a t_ν-distribution

$$f_t(y_t|Y_{t-1},\mu_1) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)(\nu e^{2\lambda})^{\frac{1}{2}}}\left(1+\frac{(y_t-\mu_{t|t-1})^2}{\nu e^{2\lambda}}\right)^{-\frac{\nu+1}{2}},$$

where $Y_{t-1} = \{y_{t-1,}y_{t-2},..\}, \exp(\lambda)$ is the scale and the location, $\mu_{t|t-1}$, is generated by a linear function of

$$u_t = \left(1+\nu^{-1}e^{-2\lambda}(y_t-\mu_{t|t-1})^2\right)^{-1} v_t, \quad t=1,...,T, \tag{5}$$

where $v_t = y_t - \mu_{t|t-1}$ is the prediction error. Differentiating the log-density shows that u_t is proportional to the conditional score, $\partial \ln f_t/\partial\mu_{t|t-1} = (\nu+1)\nu^{-1}\exp(-2\lambda)u_t$. No restriction is put on the degrees of freedom, ν, apart from requiring that it be positive: hence the reference to location rather than the mean. The scaling factor, $\exp(2\lambda)$, cancels out if the score is divided by the information quantity for the location.

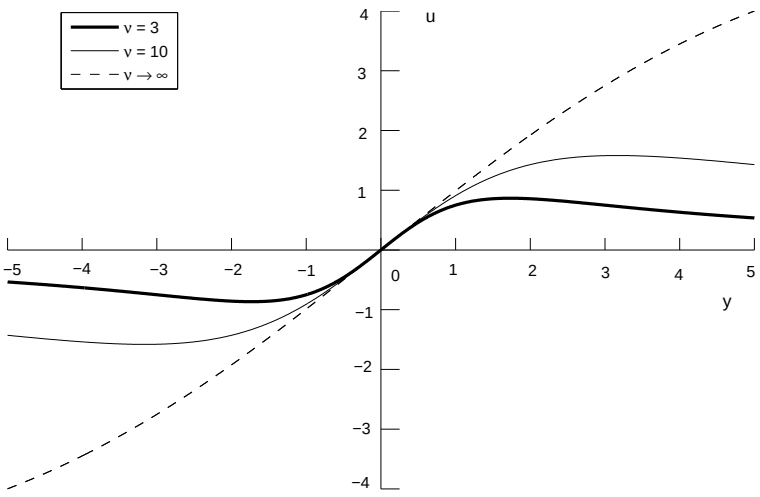
The first-order model corresponds to the Gaussian innovations form, (3), and is

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t = \mu_{t|t-1} + \exp(\lambda)\varepsilon_t, \qquad t=1,...,T \\ \mu_{t+1|t} &= \delta + \phi\mu_{t|t-1} + \kappa u_t, \end{aligned} \tag{6}$$

where ε_t is serially independent, standard t-variate. More generally, a model of order (p,r) is

$$\mu_{t+1|t} = \delta + \phi_1\mu_{t|t-1} + ... + \phi_p\mu_{t-p+1|t-p} + \kappa_0 u_t + \kappa_1 u_{t-1} + ... + \kappa_r u_{t-r}. \tag{7}$$

Figure 1: Plot of u_t against observations y_t from a zero mean and unit scale t_ν distribution, for $\nu = 3$ (thick line), $\nu = 10$ (thin line) and $\nu \to \infty$ (dashed line).



In the Gaussian case $u_t = v_t$. If q is defined as $max(p, r+1)$, y_t is an $ARMA(p,q)$ with MA coefficients $\theta_i = \phi_i - \kappa_{i-1}$, $i = 1, .., q$.

Re-parameterization in terms of the unconditional mean, ω, gives

$$\mu_{t|t-1} = \omega + \mu^{\dagger}_{t|t-1}, \quad t = 1, ..., T, \tag{8}$$

where $\mu^{\dagger}_{t|t-1}$ is as in (7), but without δ, and $\omega = \delta/(1 - \phi_1 - ... - \phi_p)$.

Figure 1 shows the impact of u_t against observations y_t from a zero mean and unit scale t distribution with various degrees of freedom. The Gaussian response is the 45 degree line. For low degrees of freedom, observations that would be seen as outliers for a Gaussian distribution are far less influential. As $|y| \to \infty$, the response tends to zero. Redescending M-estimators, which feature in the robustness literature, have the same property. On the other hand, the Huber M-estimator has a Gaussian response until a certain point, whereupon it is constant; see Maronna *et al* (2006, p 25-31). The implementation of M-estimates usually requires a (robust) estimate of scale to be pre-computed.

The variable u_t can be written

$$u_t = (1 - b_t)(y_t - \mu_{t|t-1}), \tag{9}$$

where

$$b_t = \frac{(y_t - \mu_{t|t-1})^2/\nu \exp(2\lambda)}{1 + (y_t - \mu_{t|t-1})^2/\nu \exp(2\lambda)}, \qquad 0 \le b_t \le 1, \quad 0 < \nu < \infty, \tag{10}$$

is distributed as $beta(1/2, \nu/2)$; see Harvey (2013, Chapter 3). The $u_t's$ are $IID(0, \sigma_u^2)$ and symmetrically distributed. Even moments of all orders exist. In particular

$$\mathrm{Var}(u_t) = \sigma_u^2 = \nu \exp(2\lambda)\mathrm{E}(b_t(1 - b_t)) = \nu^2(\nu + 3)^{-1}(\nu + 1)^{-1} \exp(2\lambda), \tag{11}$$

and the kurtosis is less than three for $\nu < \infty$. Since the $u_t's$ are $IID(0, \sigma_u^2)$, $\mu_{t|t-1}$ is weakly and strictly stationary so long as $|\phi| < 1$. Although determining the statistical properties of $\mu_{t|t-1}$ requires assuming that it started in the infinite past, the filter needs to be initialized in practice and this may be done by setting $\mu_{1|0} = \omega$ or $\mu_{1|0}^{\dagger} = 0$ in (8).

The existence of moments of y_t is not affected by the dynamics. The autocorrelations can be found from the infinite moving average representation; the patterns are as they would be for a Gaussian model.

The minimum mean square error (MMSE) predictor of $\mu_{T+\ell|T+\ell-1}$ can be computed recursively as in Gaussian model. Thus in the stationary first-order model

$$\mu_{T+\ell|T} = \omega(1 - \phi^{\ell-1}) + \phi^{\ell-1}\mu_{T+1|T}, \quad \ell = 2, 3, \ldots \tag{12}$$

the prediction error associated with $\mu_{T+\ell|T}$ is $\sum_{j=1}^{\ell-1} \psi_j u_{T+\ell-j}$, where $\psi_j = \kappa\phi^j$ for $j = 1, 2, \ldots$.and so $MSE(\mu_{T+\ell|T}) = \sigma_u^2 \sum_{j=1}^{\ell-1} \psi_j^2$, $\ell = 2, 3, \ldots$, where σ_u^2 is given by (11). The predictor of the observation at time $T + \ell$, that is $y_{T+\ell} = \mu_{T+\ell|T+\ell-1} + v_{T+\ell}$, is $y_{T+\ell|T} = \mu_{T+\ell|T}$, $\quad \ell = 2, 3, \ldots$, and, when $\nu > 2$, $y_{T+\ell|t}$ is the MMSE ℓ-step ahead prediction of $y_{T+\ell}$. A formula for the multi-step predictive distribution cannot be found unless the model is Gaussian. However, simulation is a viable option. The prediction error associated with $y_{T+\ell|t}$ is $\sum_{j=1}^{\ell-1} \psi_j u_{T+\ell-j} + v_{T+\ell}, \ell = 2, 3, \ldots$ and so u_{T+j}, $j = 1, .., \ell - 1$, and $v_{T+\ell}$ can be generated from independent t variates.

4 Maximum likelihood estimation

The asymptotic distribution of the maximum likelihood estimator is derived in Harvey (2013, p. 65) and outlined in the appendix. Let $y_t \mid Y_{t-1}$ have a t_ν-distribution with location, $\mu_{t|t-1}$, generated by (8) where $\mu^{\dagger}_{t|t-1}$ is a stationary first-order model, as in (4), and $|\phi| < 1$. Define

$$a = \phi - \kappa\frac{\nu}{\nu+3}, \qquad b = \phi^2 - 2\phi\kappa\frac{\nu}{\nu+3} + \kappa^2\frac{\nu\left(\nu^3+10\nu^2+35\nu+38\right)}{(\nu+1)(\nu+3)(\nu+5)(\nu+7)}$$

and let $\boldsymbol{\psi} = (\kappa,\ \phi, \omega)'$. Assuming that $b < 1$ and $\kappa \neq 0$, $(\widetilde{\boldsymbol{\psi}}', \widetilde{\lambda}, \widetilde{\nu})'$, the ML estimator of $(\boldsymbol{\psi}', \lambda, \nu)'$, is consistent and the limiting distribution of $\sqrt{T}(\widetilde{\boldsymbol{\psi}}' - \boldsymbol{\psi}', \widetilde{\lambda}-\lambda, \widetilde{\nu}-\nu)'$ is multivariate normal with mean vector zero and covariance matrix given by inverse of the information matrix

$$\mathbf{I}(\boldsymbol{\psi}, \lambda, \nu) = \begin{bmatrix} \frac{\nu+1}{\nu+3}\exp(-2\lambda)\mathbf{D}(\boldsymbol{\psi}) & 0 & 0 \\ 0 & \frac{2\nu}{\nu+3} & \frac{-2}{(\nu+3)(\nu+1)} \\ 0 & \frac{-2}{(\nu+3)(\nu+1)} & h(\nu)/2 \end{bmatrix} \tag{13}$$

with

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

where $\psi'(.)$ is the trigamma function, and

$$\mathbf{D}(\boldsymbol{\psi}) = \mathbf{D}\begin{pmatrix} \kappa \\ \phi \\ \omega \end{pmatrix} = \frac{1}{1-b}\begin{bmatrix} \sigma_u^2 & \frac{\sigma_u^2 a\kappa}{1-a\phi} & 0 \\ \frac{\sigma_u^2 a\kappa}{1-a\phi} & \frac{\sigma_u^2\kappa^2(1+a\phi)}{(1-\phi^2)(1-a\phi)} & 0 \\ 0 & 0 & \frac{(1-\phi)^2(1+a)}{1-a} \end{bmatrix} \tag{14}$$

A series of Monte Carlo experiments were carried out to investigate small sample properties. Table 1 reports the sample means and root mean square errors (RMSEs) from 1000 replications[1] for $T = 500$ and 1000 observations from first-order models with $\nu = 6$ and a range of (realistic) values of κ and ϕ. The expression for the information matrix shows that the asymptotic standard errors (ASEs) are independent of ω and that λ only appears as a scaling factor. Hence setting $\omega = \lambda = 0$ implies no loss in generality.

[1]We carried out some simulations using 5000 and 1000 replications, but since the results were the same up to the third decimal, we concentrated on 1000 replications (Matlab codes are available upon request).

Table 1: Simulation results for ML estimation of first-order DCS model.

	$T = 500$					$T = 1000$				
	$\phi = 0.8$	$\kappa = 0.5$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.8$	$\kappa = 0.5$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.784	0.501	0.002	-0.007	6.358	0.793	0.499	0.000	-0.005	6.164
RMSE	0.055	0.076	0.128	0.050	1.853	0.037	0.053	0.093	0.035	1.161
NSE	0.053	0.075	0.129	0.050	1.674	0.036	0.052	0.093	0.035	1.066
ASE	0.050	0.061	0.133	0.053	1.545	0.037	0.043	0.094	0.038	1.092
	$\phi = 0.8$	$\kappa = 1$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.8$	$\kappa = 1$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.791	1.004	0.003	-0.009	6.153	0.796	1.001	-0.000	-0.005	6.088
RMSE	0.036	0.092	0.196	0.045	1.305	0.025	0.067	0.144	0.031	0.920
NSE	0.035	0.088	0.200	0.044	1.263	0.025	0.063	0.144	0.031	0.855
ASE	0.034	0.063	0.208	0.053	1.545	0.024	0.045	0.147	0.038	1.092
	$\phi = 0.8$	$\kappa = 1.3$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.8$	$\kappa = 1.3$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.792	1.301	0.012	-0.007	6.149	0.796	1.302	-0.000	-0.005	6.054
RMSE	0.031	0.103	0.228	0.041	1.208	0.022	0.071	0.167	0.029	0.781
NSE	0.031	0.089	0.234	0.041	1.090	0.021	0.069	0.169	0.029	0.730
ASE	0.030	0.061	0.245	0.053	1.545	0.021	0.043	0.174	0.038	1.092
	$T = 500$					$T = 1000$				
	$\phi = 0.95$	$\kappa = 0.5$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.95$	$\kappa = 0.5$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.939	0.500	0.002	-0.007	6.337	0.945	0.499	0.013	-0.004	6.140
RMSE	0.023	0.070	0.325	0.049	1.797	0.015	0.048	0.244	0.035	1.100
NSE	0.020	0.068	0.319	0.049	1.616	0.013	0.048	0.245	0.035	1.030
ASE	0.017	0.053	0.381	0.053	1.545	0.012	0.038	0.269	0.038	1.092
	$\phi = 0.95$	$\kappa = 1$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.95$	$\kappa = 1$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.942	0.994	0.019	-0.007	6.243	0.946	1.001	0.023	-0.005	6.066
RMSE	0.019	0.093	0.486	0.045	1.352	0.0121	0.064	0.387	0.031	0.882
NSE	0.016	0.091	0.531	0.044	1.280	0.011	0.064	0.416	0.031	0.836
ASE	0.014	0.061	0.684	0.053	1.545	0.010	0.043	0.484	0.038	1.092
	$\phi = 0.95$	$\kappa = 1.3$	$\omega = 0$	$\lambda = 0$	$\nu = 6$	$\phi = 0.95$	$\kappa = 1.3$	$\omega = 0$	$\lambda = 0$	$\nu = 6$
Mean	0.943	1.307	0.014	-0.009	6.080	0.947	1.303	-0.005	-0.004	6.048
RMSE	0.017	0.107	0.561	0.042	1.081	0.011	0.071	0.445	0.029	0.740
NSE	0.015	0.092	0.609	0.041	1.061	0.010	0.069	0.495	0.029	0.728
ASE	0.014	0.061	0.843	0.053	1.545	0.010	0.043	0.596	0.038	1.092

In most cases convergence was rapid and few computational problems were encountered. The estimates were stable with respect to the initial values for the parameters. Problems only emerged for a significant number of replications when κ was assigned a value close to zero, the reason being that the model is not identifiable when $\kappa = 0$. The sample means give little indication of any significant bias. The ASEs, which were obtained from the square roots of the diagonal elements of the inverse of (13) divided by the sample size, are generally not far from the empirical RMSEs. The numerical standard errors (NSEs) were computed from the Hessian matrix of the log-likelihood function and averaged over all replications. On the whole they are very close to the corresponding RMSEs.

A set of experiments was also conducted to see what might be lost by using our Student-t model when the observations are Gaussian. The answer appears to be very little because in 5000 replications we encountered no estimate of ν less than 200.

Estimation of the unknown parameters adds another element of uncertainty to the predictions. However, because the parameters are estimated consistently, the contribution to the MSE of $\mu_{T+\ell|T}$, and hence $y_{T+\ell|T}$, is of $O(1/T)$. Nevertheless it may be of some significance in small samples. When the parameters are estimated by ML, (12) is replaced by

$$\widetilde{\mu}_{T+\ell|T} = \widetilde{\omega}(1 - \widetilde{\phi}^{\ell-1}) + \widetilde{\phi}^{\ell-1}\widetilde{\mu}_{T+1|T}, \quad \ell = 1, 2, 3, \ldots,$$

and so the term $\widetilde{\mu}_{T+\ell|T} - \mu_{T+\ell|T}$ is added to the estimation error associated with $\mu_{T+\ell|T}$. To illustrate the effect of estimating the unknown parameters, we simulated 1000 replications of the model and for each replication computed the difference between $\mu_{T+\ell|T}$ and $\widetilde{\mu}_{T+\ell|T}$ and hence constructed an estimate of the additional contribution to the MSE of the estimator of $\mu_{T+\ell|T+\ell-1}$. Table 2 shows the results for one of the parameter configurations in Table 1, namely $\phi = 0.8,\ \kappa = 0.5,\ \omega = 0,\ \lambda = 0$ and $\nu = 6$, with $T = 500$ and $T = 1000$. As can be seen, the bias is insignificant and the increase in the MSE is small in relation to $MSE(\mu_{T+\ell|T})$.

Table 2: Estimation error, simulation results, $M = 1000$

	Steps ahead for $T = 500$			Steps ahead for $T = 1000$		
	$\ell = 2$	$\ell = 3$	$\ell = 10$	$\ell = 2$	$\ell = 3$	$\ell = 10$
Mean of $\widetilde{\mu}_{T+\ell\|T} - \mu_{T+\ell\|T}$	-0.0005	0.0007	0.0045	0.0021	-0.0009	-0.0027
MSE of $\widetilde{\mu}_{T+\ell\|T} - \mu_{T+\ell\|T}$	0.0014	0.0012	0.0002	0.0005	0.0004	0.0001
$MSE(\mu_{T+\ell\|T})$	0.0914	0.1499	0.2494	0.0914	0.1499	0.2494

5 Higher-order models and the state space form

The general statistical treatment of unobserved components models is based on the state space form. The corresponding innovations form facilitates the handling of higher-order DCS models.

5.1 Linear Gaussian models and the Kalman filter

For simplicity let us assume a time-invariant univariate time series model and exclude any deterministic components. The general case is set out in Harvey (1989, Chapter 3). The observation in the Gaussian state space model is related to an $m \times 1$ state vector, $\boldsymbol{\alpha}_t$, through a measurement equation, $y_t = \omega + \mathbf{z}'\boldsymbol{\alpha}_t + \varepsilon_t$, $t = 1, ..., T$, where ω is a constant, $\mathbf{z}$ is an $m \times 1$ vector and $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$. The elements of $\boldsymbol{\alpha}_t$ are usually unobservable but are known to be generated by a transition equation $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\delta} + \mathbf{T}\boldsymbol{\alpha}_t + \boldsymbol{\eta}_t$, $t = 1, ..., T$,where $\boldsymbol{\delta}$ is a vector of constants and $\boldsymbol{\eta}_t \sim NID(\mathbf{0}, \mathbf{Q})$. The specification is completed by assuming that $\mathrm{E}(\boldsymbol{\alpha}_1) = \boldsymbol{\alpha}_{1|0}$ and $\mathrm{Var}(\boldsymbol{\alpha}_1) = \mathbf{P}_{1|0}$, where $\mathbf{P}_{1|0}$ is positive a semi-definite matrix, and that $\mathrm{E}(\varepsilon_t \boldsymbol{\alpha}_0') = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{\eta}_t \boldsymbol{\alpha}_0') = \mathbf{0}$ for $t = 1, , ..., T$. It is usually assumed that the disturbances are uncorrelated with each other in all time periods, that is $\mathrm{E}(\varepsilon_t \boldsymbol{\eta}_s') = \mathbf{0}$ for all $s, t = 1, ..., T$, though this assumption may be relaxed.

When the disturbances and initial state are normally distributed, the minimum mean square error estimates of the state and observation at time t, based on information at time $t - 1$, are their conditional expectations. The Kalman filter is a recursive procedure for

computing these estimates, given $\mathbf{z}, \sigma_\varepsilon^2, \mathbf{T}$ and $\mathbf{Q}$ together with the initial conditions, $\boldsymbol{\alpha}_{1|0}$ and $\mathbf{P}_{1|0}$. When the initial conditions are unknown, the filter may be started off as discussed in Durbin and Koopman (2012).

The Kalman filter can be written as a single set of recursions going directly from $\boldsymbol{\alpha}_{t|t-1}$ to $\boldsymbol{\alpha}_{t+1|t}$. The innovations form, generalizing (3), is

$$\begin{aligned} y_t &= \omega + \mathbf{z}'\boldsymbol{\alpha}_{t|t-1} + v_t, \quad t = 1, ..., T, \\ \boldsymbol{\alpha}_{t+1|t} &= \boldsymbol{\delta} + \mathbf{T}\boldsymbol{\alpha}_{t|t-1} + \mathbf{k}_t v_t, \end{aligned} \tag{15}$$

where $v_t = y_t - \omega - \mathbf{z}_t'\boldsymbol{\alpha}_{t|t-1}$ is the innovation and $f_t = \mathbf{z}'\mathbf{P}_{t|t-1}\mathbf{z} + \sigma_\varepsilon^2$ is its variance. The gain vector is $\mathbf{k}_t = (1/f_t)\mathbf{T}\mathbf{P}_{t|t-1}\mathbf{z}$ and $\mathbf{P}_{t|t-1}$ is calculated by a matrix recursion. Since (15) contains only one disturbance term, it may be regarded as a reduced form model with $\mathbf{k}_t$ subject to restrictions coming from the original structural form. In the steady-state, $\mathbf{k}_t$ and f_t are time-invariant.

5.2 The DCS model

A general location DCS model may be set up in the same way as the innovations form of a Gaussian state space model. The model corresponding to the steady-state of (15) is

$$\begin{aligned} y_t &= \omega + \mathbf{z}'\boldsymbol{\alpha}_{t|t-1} + v_t, \quad t = 1, ..., T, \\ \boldsymbol{\alpha}_{t+1|t} &= \boldsymbol{\delta} + \mathbf{T}\boldsymbol{\alpha}_{t|t-1} + \boldsymbol{\kappa} u_t. \end{aligned} \tag{16}$$

The $\mathbf{z}$ vector and $\mathbf{T}$ matrix may be specified in the same way as for the Gaussian UC models. The transition equation in (16) is stationary provided that the roots of the transition matrix $\mathbf{T}$ have modulus less than one. When this is the case, $\boldsymbol{\delta}$ is superfluous and initialization is achieved by setting $\boldsymbol{\alpha}_{1|0} = \mathbf{0}$. If $\boldsymbol{\alpha}_{t|t-1}$ contains nonstationary elements, the best option seems to be to treat their initial values as unknown parameters.

There remains the question of how to specify the parameters in the vector κ. More specifically, what restrictions should be imposed? The issues are explored for trend and seasonal components below.

Remark The general model, (7), of order (p, r) may be put in the state space form of (16) in a similar way to an $ARMA(p, r)$ plus noise unobserved components models.

6 Trend and seasonality

Stochastic trend and seasonal components may be introduced into UC models for location. These models, called structural time series models, are described in Harvey (1989) and implemented in the STAMP package of Koopman *et al* (2009). The way in which the innovations forms of structural time series models lead to corresponding DCS-t models is explored below.

6.1 Local level model

The Gaussian random walk plus noise or *local level* model is

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_t = \mu_{t-1} + \eta_t, \tag{17}$$

where $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$, $\eta_t \sim NID(0, \sigma_\eta^2)$ and $\mathrm{E}(\varepsilon_t \eta_s) = 0$ for all t and s. The signal noise ratio is $q = \sigma_\eta^2/\sigma_\varepsilon^2$ and the parameter, θ in the $ARIMA(0, 1, 1)$ reduced form representation, (2), lies in the range $0 \leq \theta < 1$ when $\sigma_\varepsilon^2 > 0$. Since $\theta = 1 - \kappa$, the range of κ in the steady-state innovations form is $0 < \kappa \leq 1$. In this case $\mu_{t+1|t}$ is an exponentially weighted moving average in which the weights on current and past observations are non-negative.

The local level DCS-t model is

$$y_t = \mu_{t|t-1} + v_t, \qquad \mu_{t+1|t} = \mu_{t|t-1} + \kappa u_t. \tag{18}$$

The initialization of the KF in (17) is best done using a diffuse prior; see Harvey (1989, pp 107-8). This is not an option for the DCS model. One possibility is to set $\mu_{2|1} = y_1$, but the filter could be adversely affected if the first observation is an outlier. An alternative approach is to treat the initial value, $\mu_{1|0}$, as an unknown parameter that must be estimated along with κ and ν. This is the technique used by Ord, Koehler and Snyder (1997) to initialize nonlinear single source of error models (see also Hyndman et al, 2008).

Because $u_t = (1-b_t)(y_t - \mu_{t|t-1})$, re-arranging the dynamic equation in (18) gives

$$\mu_{t+1|t} = (1-\kappa(1-b_t))\mu_{t|t-1} + \kappa(1-b_t)y_t. \tag{19}$$

A sufficient condition for the weights on current and past observations to be non-negative is that $\kappa(1-b_t) < 1$ and, because $0 \leq b_t \leq 1$, this is guaranteed by $0 < \kappa \leq 1$. However, the restriction $\kappa \leq 1$ is neither necessary nor desirable. Estimates of κ greater than one are not unusual and are entirely appropriate when the signal is strong relative to the noise.

As regards asymptotic properties, the result in Section 4 can be modified to deal with the nonstationary case. to be specific, when $b < 1$ and $\mu_{1|0}$ is fixed and known or $\mu_{2|1} = y_1$, where y_1 is fixed, the ML estimator of κ in (18) is consistent and $\sqrt{T}(\widetilde{\kappa} - \kappa)$ has a limiting normal distribution with mean zero and variance

$$\text{Var}(\widetilde{\kappa}) = \left(2\kappa\frac{\nu}{\nu+3} - \kappa^2\frac{\nu\left(\nu^3+10\nu^2+35\nu+38\right)}{(\nu+1)(\nu+3)(\nu+5)(\nu+7)}\right)\left(\frac{\nu+3}{\nu}\right)^2.$$

It can be seen that $\kappa > 0$ is a necessary condition for $b < 1$ and hence $\text{Var}(\widetilde{\kappa}) > 0$. When the initial value, $\mu_{1|0}$, is treated as a parameter to be estimated, it appears from some limited simulation evidence that the distribution of the ML estimator of κ is essentially unchanged.

The result extends to the random walk plus drift trend, that is

$$\mu_{t+1|t} = \beta + \mu_{t|t-1} + \kappa u_t, \tag{20}$$

where β is an unknown constant. The ML estimators of κ and β are asymptotically independent. Thus $\text{Var}(\widetilde{\kappa})$ is unchanged and adapting expression (2.44) in Harvey (2013, p 38) gives

$$\text{Var}(\widetilde{\beta}) = e^{2\lambda}\left(2\kappa\frac{\nu}{\nu+3} - \kappa^2\frac{\nu\left(\nu^3+10\nu^2+35\nu+38\right)}{(\nu+1)(\nu+3)(\nu+5)(\nu+7)}\right)\frac{\nu+3}{\nu+1}\frac{\nu\kappa}{(2-\kappa)\nu+6}.$$

6.2 Local linear trend

The DCS filter corresponding to the UC local linear trend model is

$$\begin{aligned} y_t &= \mu_{t|t-1} + v_t, \\ \mu_{t+1|t} &= \mu_{t|t-1} + \beta_{t|t-1} + \kappa_1 u_t, \qquad \beta_{t+1|t} = \beta_{t|t-1} + \kappa_2 u_t. \end{aligned} \tag{21}$$

The initialization $\beta_{3|2} = y_2 - y_1$ and $\mu_{3|2} = y_2$ can be used, but, as in the local level model, initializing in this way is vulnerable to outliers at the beginning. Estimating the fixed starting values, $\mu_{1|0}$ and $\beta_{1|0}$, may be a better option.

An integrated random walk trend in the UC local linear trend model implies the contraint $\kappa_2 = \kappa_1^2/(2-\kappa_1), \quad 0 < \kappa_1 < 1$, which may be found using formulae in Harvey (1989, p. 177). The restriction can be imposed on the DCS-t model by treating $\kappa_1 = \kappa$ as the unknown parameter, but without unity imposed as an upper bound.

6.3 Stochastic seasonal

A fixed seasonal pattern may be modeled as $\gamma_t = \sum_{j=1}^{s} \gamma_j z_{jt}$, where s is the number of seasons and the dummy variable z_{jt} is one in season j and zero otherwise. In order not to confound trend with seasonality, the coefficients, γ_j, $j = 1, ..., s$, are constrained to sum to zero. The seasonal pattern may be allowed to change over time by letting the coefficients evolve as random walks. If γ_{jt} denotes the effect of season j at time t, then

$$\gamma_{jt} = \gamma_{j,t-1} + \omega_{jt}, \quad \omega_t \sim NID(0, \sigma_\omega^2), \quad j = 1, ..., s. \tag{22}$$

Although all s seasonal components are continually evolving, only one affects the observations at any particular point in time, that is $\gamma_t = \gamma_{jt}$ when season j is prevailing at time t. The requirement that the seasonal components evolve in such a way that they always sum to zero, that is $\sum_{j=1}^{s} \gamma_{jt} = 0$, is enforced by the restriction that the disturbances sum to zero at each point in time. This restriction is implemented by the correlation structure in $\mathrm{Var}(\boldsymbol{\omega}_t) = \sigma_\omega^2 \left(\mathbf{I} - s^{-1}\mathbf{i}\mathbf{i}'\right)$, where $\boldsymbol{\omega}_t = (\omega_{1t}, ..., \omega_{st})'$, coupled with initial conditions requiring that the seasonals sum to zero at $t = 0$. It can be seen that $\mathrm{Var}(\mathbf{i}'\boldsymbol{\omega}_t) = 0$.

In the state space form, the transition matrix is just the identity matrix, but the $\mathbf{z}$ vector must change over time to accommodate the current season. Apart from replacing $\mathbf{z}$ by $\mathbf{z}_t$, the form of the KF remains unchanged. Adapting the innovations form to the DCS observation driven framework, (16), gives

$$y_t = \mathbf{z}_t'\boldsymbol{\alpha}_{t|t-1} + v_t, \qquad \boldsymbol{\alpha}_{t+1|t} = \boldsymbol{\alpha}_{t|t-1} + \boldsymbol{\kappa}_t u_t, \tag{23}$$

where $\mathbf{z}_t$ picks out the current season, $\gamma_{t|t-1}$, that is $\gamma_{t|t-1} = \mathbf{z}_t'\boldsymbol{\alpha}_{t|t-1}$. The only question is how to parameterize $\boldsymbol{\kappa}_t$.

The seasonal components in the UC model are constrained to sum to zero and the same is true of their filtered estimates. Thus $\mathbf{i}'\boldsymbol{\kappa}_t = 0$ in the Kalman filter and this property should carry across to the DCS filter. If κ_{jt}, $j = 1, .., s$, denotes the $j-th$ element of $\boldsymbol{\kappa}_t$ in (23), then in season j we set $\kappa_{jt} = \kappa_s$, where κ_s is a non-negative unknown parameter, while $\kappa_{it} = -\kappa_s/(s-1)$ for $i \neq j$. The amounts by which the seasonal effects change therefore sum to zero.

The seasonal recursions can be combined with the trend filtering equations of (21) in order to give a structure similar in form to that of the Kalman filter for the stochastic trend plus seasonal plus noise UC model, sometimes known as the 'basic structural model'. Thus

$$y_t = \mu_{t|t-1} + \gamma_{t|t-1} + v_t, \tag{24}$$

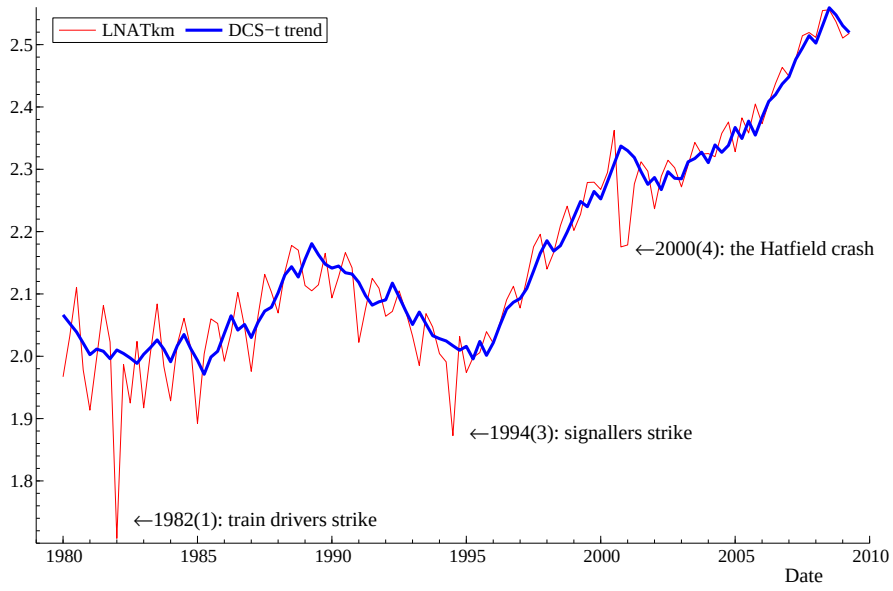
where $\mu_{t|t-1}$ is as defined in (21). The initial conditions at time $t = 0$ are estimated by treating them as parameters; there are $s-1$ seasonal parameters because the remaining initial seasonal state is minus the sum of the others.

6.4 Application to rail travel

In a project carried out for the UK Department for Transport by one of the authors, the STAMP 8 package of Koopman et al (2009) was used to fit an unobserved components model to the logarithm of National Rail Travel, defined as the number of kilometres traveled by UK passengers. (Source: National Rail Trends). The observations started in the first quarter of 1980 and finished in the second quarter of 2009. Trend, seasonal and irregular components were included but the model was augmented with intervention variables to take out the effects of observations that were known to be unrepresentative. The intervention dummies were: (i) the train drivers strikes in 1982(1,3); (ii) the Hatfield crash and its aftermath, 2000(4) and 2001(1); and (iii) the signallers strike in 1994(3).

Fitting a DCS model with trend and seasonal, that is (24), avoids the need to deal explicitly

Figure 2: Trend from a DCS-t model fitted to UK National Rail Travel.



with the outliers. The ML estimates for the parameters in a model with a random walk plus drift trend, (20), are

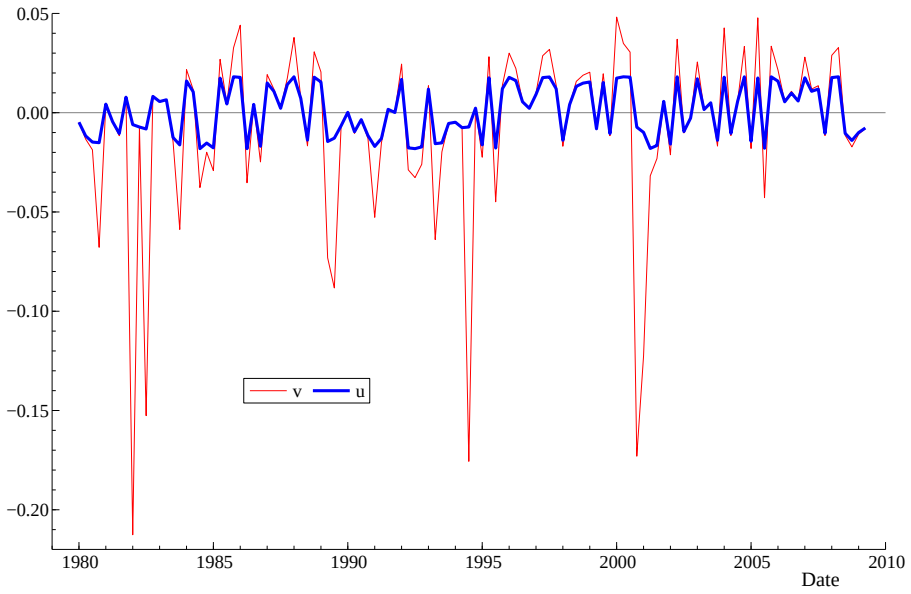
$$\begin{aligned} \widetilde{\kappa} &= 1.421(0.161) \qquad \widetilde{\kappa}_s = 0.539\ (0.070) \qquad \widetilde{\beta} = 0.003\ (0.001) \\ \widetilde{\nu} &= 2.564\ (0.553) \quad \widetilde{\lambda} = -3.787\ (0.107) \end{aligned}$$

with initial values $\widetilde{\mu} = 2.066\ (0.009)$, $\widetilde{\gamma}_1 = -0.094\ (0.007)$, $\widetilde{\gamma}_2 = -0.010\ (0.006)$ and $\widetilde{\gamma}_3 = 0.086\ (0.006)$. The figures in parentheses are numerical standard errors. The last seasonal is $\widetilde{\gamma}_4 = 0.018$; it has no SE as it was constructed from the others.

The filtered DCS-t trend shown in Figure 2 appears not to be affected by the outliers. We also found that it is very close to the filtered trend obtained from the UC model with interventions. The same is true of the filtered seasonal.

Figure 3 shows the residuals, that is the one-step ahead prediction errors, for the DCS model, together with the scores. The outliers, which were removed by dummies in the UC model, show up clearly in the residuals. In the score series the outliers are downweighted and the autocorrelations are slightly bigger than those of the residuals, presumably because they

Figure 3: Residuals, v_t and (scaled) scores, u_t, from DCS-t model.



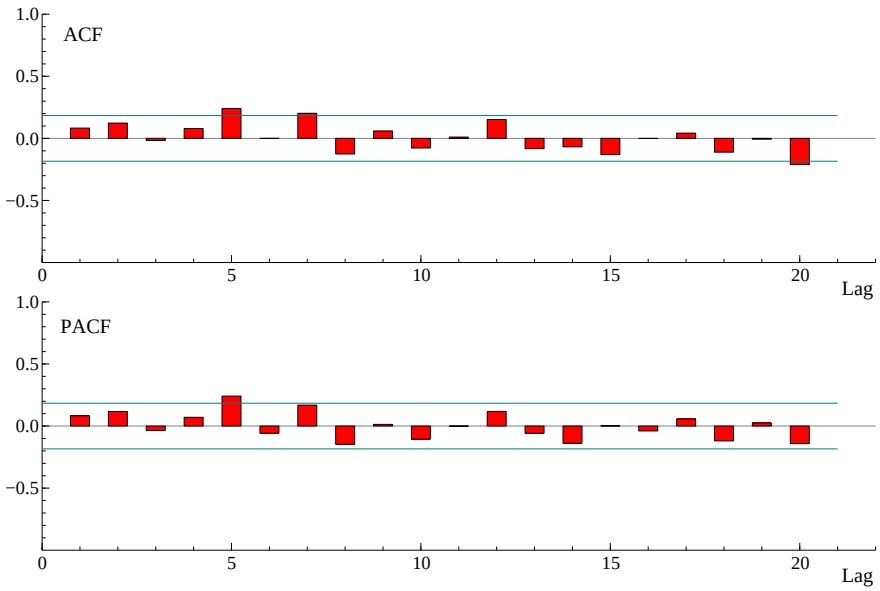
are not weakened by aberrant values. The Box-Ljung $Q(12)$ statistic is 19.78 for the scores and 12.40 for the residuals. If it can be assumed that only the number of fitted dynamic parameters affects the distribution of the Box-Ljung statistic, its distribution under the null hypothesis of correct model specification is χ^2_{10}, which had a 5% critical value of 18.3. Thus the scores reject the null hypothesis, albeit only marginally, while the residuals do not. Having said that, the score autocorrelations do not exhibit any clear pattern and the ACF shown in Figure 4 is almost indistinguishable from the corresponding sample partial autocorrelation function (PACF). Hence it is difficult to see how the dynamic specification could be improved.

7 Explanatory variables

The location 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

$$y_t = \mu^{\dagger}_{t|t-1} + \mathbf{w}_t'\boldsymbol{\gamma} + \varepsilon_t \exp(\lambda), \qquad t = 1, ..., T, \tag{25}$$

Figure 4: ACF and PACF of the scores from DCS-t model.



where $\mu^{\dagger}_{t|t-1}$ could be a stationary process, as in (8), or a stochastic trend such as (21). The model may be augmented by a seasonal component as in sub-section 6.4.

If it is possible to make a sensible guess of initial values of the explanatory variable coefficients, the degrees of freedom parameter, ν, and the dynamic parameters, ϕ and κ for a stationary first-order model or β and κ for a random walk with drift, can be estimated by fitting a univariate model to the residuals, $y_t - \mathbf{w}_t'\widehat{\boldsymbol{\gamma}}$, $t = 1, .., T$. These values are then used to start off numerical optimization with respect to all the parameters in the model.

7.1 Asymptotic distribution

The following result is obtained by specializing Corollary 10 in Harvey (2013, Section 2.6). Consider model (25) with a stationary first-order component. Assume that the explanatory variables are weakly stationary with mean $\boldsymbol{\mu}_w$ and second moment $\boldsymbol{\Lambda}_w$ and are strictly exogenous in the sense that they are independent of the $\varepsilon_t's$ and therefore of the $u_t's$. Provided that $b < 1$ and $\kappa \neq 0$, the ML estimator of $(\kappa, \phi, \boldsymbol{\gamma}', \lambda, \nu)'$, is consistent and the limiting distribution of $\sqrt{T}(\widetilde{\kappa} - \kappa, \widetilde{\phi} - \phi, \widetilde{\boldsymbol{\gamma}}' - \boldsymbol{\gamma}', \widetilde{\lambda} - \lambda, \widetilde{\nu} - \nu)'$ is multivariate normal with mean vector zero and

covariance matrix given by the inverse of the information matrix in (13) but with $\boldsymbol{\psi}$ replaced by $(\kappa, \phi)'$ and $\mathbf{D}(\boldsymbol{\psi})$ replaced by

$$\mathbf{D}\begin{pmatrix}\kappa\\ \phi\\ \boldsymbol{\gamma}\end{pmatrix} = \frac{1}{1-b}\begin{bmatrix} \sigma_u^2 & \frac{\sigma_u^2 a\kappa}{1-a\phi} & \mathbf{0}' \\ \frac{\sigma_u^2 a\kappa}{1-a\phi} & \frac{\sigma_u^2\kappa^2(1+a\phi)}{(1-\phi^2)(1-a\phi)} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_w \end{bmatrix},$$

with

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

with $\boldsymbol{\Lambda}_w(1) = \mathrm{E}(\mathbf{w}_t\mathbf{w}_{t-1}') = \mathrm{E}(\mathbf{w}_{t-1}\mathbf{w}_t')$.

An estimator of the asymptotic covariance matrix can be obtained by replacing $\boldsymbol{\Lambda}_w$ and $\boldsymbol{\Lambda}_w(1)$ by $T^{-1}\sum \mathbf{w}_t\mathbf{w}_t'$ and $T^{-1}\sum \mathbf{w}_t\mathbf{w}_{t-1}'$ respectively. The constant term, ω, will normally appear as an explanatory variable in which case the corresponding element in $\mathbf{w}_t$ will be unity.

When $\mu_{t|t-1}^{\dagger}$ is known to be a random walk with drift, β, as in (20), and $\mu_{1|0}^{\dagger}$ is fixed and known, the information matrix is as in (13) but with

$$\mathbf{D}\begin{pmatrix}\kappa\\ \boldsymbol{\gamma}\\ \beta\end{pmatrix} = \frac{1}{1-b}\begin{bmatrix} \sigma_u^2 & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0} & \mathbf{C}_{\Delta w} & \boldsymbol{\mu}_{\Delta w} \\ \mathbf{0} & \boldsymbol{\mu}_{\Delta w}' & 1 \end{bmatrix}, \quad b<1,$$

where $\boldsymbol{\mu}_{\Delta w} = \mathrm{E}(\Delta\mathbf{w}_t)$ and $\mathbf{C}_{\Delta w} = \mathrm{E}(\Delta\mathbf{w}_t\Delta\mathbf{w}_t')$. The first differences of the explanatory variables must be weakly stationary but their levels may be nonstationary. It follows that the covariance matrix of the limiting distribution of $\sqrt{T}\widetilde{\boldsymbol{\gamma}}$ is

$$\mathrm{Var}(\widetilde{\boldsymbol{\gamma}}) = \left(2\kappa\frac{\nu}{\nu+1} - \kappa^2\frac{\nu\left(\nu^3+10\nu^2+35\nu+38\right)}{(\nu+1)^2(\nu+5)(\nu+7)}\right)e^{2\lambda}(\mathbf{C}_{\Delta w} - \boldsymbol{\mu}_{\Delta w}\boldsymbol{\mu}_{\Delta w}')^{-1}. \tag{26}$$

7.2 Application to rail travel

Potential explanatory variables for the rail travel series of Sub-section 6.5 are: (i) Real GDP (in £2003 prices), (ii) Real Fares, obtained by dividing total revenue by the number of kilometres travelled and the retail price index (RPI), and (iii) Petrol and Oil index (POI), divided by RPI. The fares series was smoothed by fitting a univariate UC model.

Fitting an unobserved components time series model using STAMP gave the following estimates for the coefficients of the logarithms of the explanatory variables: GDP was 0.716 (0.267), fares was -0.416 (0.245) and POI was 0.050 (0.065). Because the explanatory variables enter the model in logarithms, their coefficients are elasticities. All the estimates are plausible. The coefficient of the petrol index is not statistically significant at any conventional level, but at least it has the right sign.

Failure to deal with outliers in a time series regression can lead to serious distortions and this is well-illustrated by the rail series when the intervention variables are not included. In particular the fare estimate is *plus* 0.28.

When rail travel was seasonally adjusted by removing the seasonal component obtained from the univariate DCS-t model fitted in sub-section 6.5 and LPOI was also seasonally adjusted, estimating the DCS-t model without a seasonal component gave

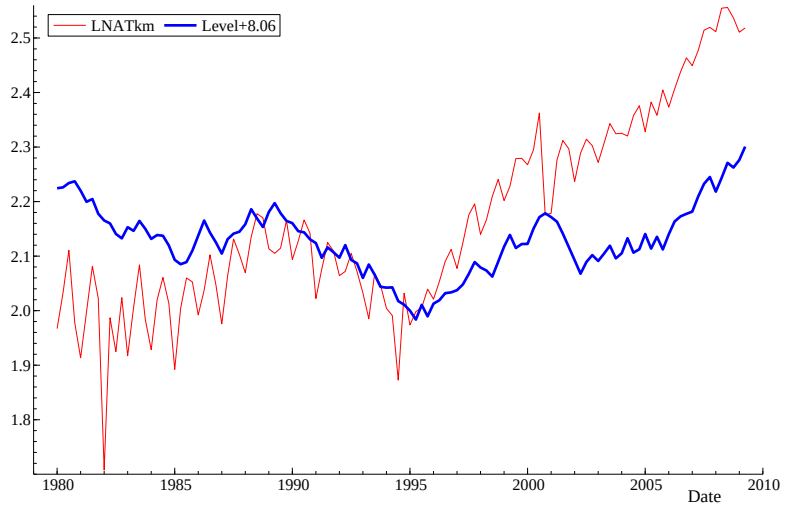
$$\widetilde{\kappa} = 1.346(0.151) \qquad \widetilde{\lambda} = -3.879\ (0.123) \qquad \widetilde{\nu} = 2.436\ (0.648) \qquad \widetilde{\beta} = 0.001\ (0.002),$$

where the figures in parentheses are asymptotic standard errors. The ASEs calculated for the coefficients of LGDP, Lfare (level) and LPOI (seasonally adjusted) using Var($\widetilde{\boldsymbol{\gamma}}$) in (26) were 0.251, 0.246 and 0.050 respectively. These figures are close to the standard errors for the UC model (with seasonal component) reported in the first paragraph of this sub-section. (The estimated SEs obtained from a UC model fitted to seasonal adjusted data were similar).

Fitting the full DCS-t model with the seasonal gave $\widetilde{\kappa} = 2.212$, $\widetilde{\kappa}_s = 0.771$, $\widetilde{\lambda} = -4.059$, $\widetilde{\nu} = 2.070$ and $\widetilde{\beta} = 0.0004$, with initial values $\widetilde{\mu} = -6.162$, $\widetilde{\gamma}_1 = -0.084$, $\widetilde{\gamma}_2 = -0.007$ and $\widetilde{\gamma}_3 = 0.070$. The coefficients of the explanatory variables were: $LGDP = 0.734$, $Lfare = -0.427$ and $LPOI = 0.056$. The Box-Ljung $Q(12)$ statistic is 5.30 for the score and 16.12 for the residuals. This result is a little surprising because in the univariate model the Q-statistic for the score was bigger than that of the residuals.

A good deal, but by no means all, of the growth in rail travel from the mid-nineties is due to the increase in GDP. The continued fall after the economy had moved out of the recession of the early nineties is partly explained by the fact that fares increased sharply in

Figure 5: Trend in rail travel after explanatory variables have been taken into account. (A constant has been added to the trend so that it is at a level comparable with that of the series.)



1993 in anticipation of rail privatisation and continued to increase till 1995. Nevertheless, as is apparent from Figure 5, there remain long-term movements in rail travel that cannot be accounted for by the exogenous variables.

8 Conclusions

In this paper we develop, analyse and apply a robust time series model based on a conditional $t-$distribution. Our Monte Carlo results show that maximum likelihood estimation works well in moderate size samples, with the asymptotic standard errors giving a good indication of empirical RMSEs. Furthermore, the theoretical MSEs of the predictions appear not to be significantly affected when parameters are estimated.

The model is extended to include trend and seasonal components and its viability is illustrated with real data containing outliers. Finally, explanatory variables are introduced into the model and the asymptotic distribution of the estimated coefficients is presented. The

application shows that the model deals effectively with the outliers.

References

[1] Creal, D., Koopman, S.J. and Lucas, A. (2011), A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations, *Journal of Business and Economics Statistics*, 29, 4, 552–563.

[2] Delaigle, A., Hall, P. and Jin, J. (2011), Robustness and Accuracy of Methods for High Dimensional Data Analysis Based on Student's t-Statistic *Journal of the Royal Statistical Society, Series B*, 73, 283-301.

[3] Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, 2nd edition, Oxford Statistical Science Series, Oxford.

[4] Harvey A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[5] Harvey, A.C. (2013). *Dynamic Models for Volatility and Heavy Tails: with applications to financial and economic time series*, Econometric Society Monograph, Cambridge University Press, New York.

[6] Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. (2008).*Forecasting with Exponential Smoothing*, Springer Series in Statistics, Springer, Berlin.

[7] Koopman, S. J. Harvey, A. C., Doornik, J. A. and Shephard, N. (2009), *STAMP 8.2 Structural Time Series Analysis Modeller and Predictor*, Timberlake Consultants Ltd, London.

[8] Lange, K. L., Little, R. J. A. and Taylor, M.G. (1989), Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association* **84**, 881-896.

[9] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006), *Robust Statistics*, Wiley, Chichester.

[10] Muler, N., Peña, D. and Yohai, V.J. (2009), Robust Estimation for ARMA Models, *Annals of Statistics*, 37, 2, 816–840.

[11] Ord, J. K., Koehler, A.B. and Snyder, R.D. (1997), Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models, *Journal of the American Statistical Association*, 92, 1621-1629.

APPENDIX

A Consistency and asymptotic normality of the ML estimator

This appendix explains how to derive the information matrix of the ML estimator for the first-order model and outlines a proof for consistency and asymptotic normality. As noted in the text, if the model is to be identified, κ must not be zero and or such that the constraint $b < 1$ is violated. A more formal statement is that the parameters should be interior points of the compact parameter space which will be taken to be $|\phi| < 1$, $|\omega| < \infty$ and $0 < \kappa < \kappa_u$, $\kappa_L < \kappa < 0$ where κ_u and κ_L are values determined by the condition $b < 1$.

The first step is to decompose the derivatives of the log density wrt $\boldsymbol{\psi}$ into derivatives wrt $\mu_{t|t-1}$ and derivatives of $\mu_{t|t-1}$ wrt $\boldsymbol{\psi}$, that is

$$\frac{\partial \ln f_t}{\partial \boldsymbol{\psi}} = \frac{\partial \ln f_t}{\partial \mu_{t|t-1}} \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}}.$$

Since the scores $\partial \ln f_t / \partial \mu_{t|t-1}$ are $IID(0, \sigma_u^2)$ and so do not depend on $\mu_{t|t-1}$,

$$\mathrm{E}_{t-1}\left[\left(\frac{\partial \ln f_t}{\partial \mu_{t|t-1}} \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \ln f_t}{\partial \mu_{t|t-1}} \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}}\right)'\right] = \left[\mathrm{E}\left(\frac{\partial \ln f_t}{\partial \mu}\right)^2\right] \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}} \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}'} = \sigma_u^2 \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}} \frac{\partial \mu_{t|t-1}}{\partial \boldsymbol{\psi}'}.$$

Thus the unconditional expectation requires evaluating the last term.

The derivative of $\mu_{t|t-1}$ wrt κ is

$$\frac{\partial \mu_{t|t-1}}{\partial \kappa} = \phi \frac{\partial \mu_{t-1|t-2}}{\partial \kappa} + \kappa \frac{\partial u_{t-1}}{\partial \kappa} + u_{t-1}, \qquad t = 2, ..., T.$$

However,

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

Therefore

$$\frac{\partial \mu_{t|t-1}}{\partial \kappa} = x_{t-1} \frac{\partial \mu_{t-1|t-2}}{\partial \kappa} + u_{t-1} \tag{27}$$

where

$$x_t = \phi + \kappa \frac{\partial u_t}{\partial \mu_{t|t-1}}, \quad t = 1, ..., T. \tag{28}$$

Define

$$a = \mathrm{E}_{t-1}(x_t) = \phi + \kappa \mathrm{E}_{t-1}\left(\frac{\partial u_t}{\partial \mu_{t|t-1}}\right) = \phi + \kappa \mathrm{E}\left(\frac{\partial u_t}{\partial \mu}\right)$$

Since $\partial u_t / \partial \mu_{t|t-1}$ is IID, unconditional expectations can replace conditional ones. When the process for $\mu_{t|t-1}$ starts in the infinite past and $|a| < 1$, taking conditional expectations of the derivatives at time $t-2$, followed by unconditional expectations gives

$$\mathrm{E}\left(\frac{\partial \mu_{t|t-1}}{\partial \kappa}\right) = \mathrm{E}\left(\frac{\partial \mu_{t|t-1}}{\partial \phi}\right) = 0 \quad and \quad \mathrm{E}\left(\frac{\partial \mu_{t|t-1}}{\partial \omega}\right) = \frac{1-\phi}{1-a}.$$

To derive the information matrix, square both sides of (27) and take conditional expectations to give

$$\begin{aligned} \mathrm{E}_{t-2}\left(\frac{\partial \mu_{t|t-1}}{\partial \kappa}\right)^2 &= \mathrm{E}_{t-2}\left(x_{t-1}\frac{\partial \mu_{t-1|t-2}}{\partial \kappa} + u_{t-1}\right)^2 \\ &= b\left(\frac{\partial \mu_{t-1|t-2}}{\partial \kappa}\right)^2 + 2c\frac{\partial \mu_{t-1|t-2}}{\partial \kappa} + \sigma_u^2, \end{aligned} \tag{29}$$

where

$$\begin{aligned} b &= \mathrm{E}_{t-1}(x_t^2) = \phi^2 + 2\phi\kappa \mathrm{E}\left(\frac{\partial u_t}{\partial \mu}\right) + \kappa^2 \mathrm{E}\left(\frac{\partial u_t}{\partial \mu}\right)^2 \geq 0, \qquad and \\ c &= \mathrm{E}_{t-1}(u_t x_t) = \kappa \mathrm{E}\left(u_t \frac{\partial u_t}{\partial \mu}\right) \end{aligned}$$

Taking unconditional expectations gives

$$\mathrm{E}\left(\frac{\partial \mu_{t|t-1}}{\partial \kappa}\right)^2 = bE\left(\frac{\partial \mu_{t-1|t-2}}{\partial \kappa}\right)^2 + 2cE\left(\frac{\partial \mu_{t-1|t-2}}{\partial \kappa}\right) + \sigma_u^2$$

and so, provided that $b < 1$,

$$\mathrm{E}\left(\frac{\partial \mu_{t|t-1}}{\partial \kappa}\right)^2 = \frac{\sigma_u^2}{1-b}.$$

Expressions for other elements in the information matrix may be similarly derived; see Harvey (2013, Appendix A). Fulfillment of the condition $b < 1$ implies $|a| < 1$. That this is the case follows directly from the Cauchy-Schwarz inequality $\mathrm{E}(x_t^2) \geq [\mathrm{E}(x_t)]^2$.

The information matrix, (13), is given by noting that

$$\frac{\partial u_t}{\partial \mu} = 2(1-b_t)b_t - (1-b_t). \tag{30}$$

The distribution of (30) does not depend on μ and $\mathrm{E}(\partial u_t/\partial \mu) = -\nu/(\nu+3)$. Similarly

$$\mathrm{E}\left(u_t \frac{\partial u_t}{\partial \mu}\right) = \mathrm{E}(2(1-b_t)b_t - (1-b_t))(y_t - \mu_{t|t-1})(1-b_t) = 0$$

because $\mathrm{E}((y_t - \mu_{t|t-1}) \mid b_t) = 0$, and

$$\mathrm{E}\left(\frac{\partial u_t}{\partial \mu}\right)^2 = \mathrm{E}(2(1-b_t)b_t - (1-b_t))^2 = \frac{\nu\left(\nu^3 + 10\nu^2 + 35\nu + 38\right)}{(\nu+1)(\nu+3)(\nu+5)(\nu+7)} \leq 1.$$

Consistency and asymptotic normality can be proved by showing that the conditions for Lemma 1 in Jensen and Rahbek (2004, p 1206) hold. The main point to note is that the first three derivatives of $\mu_{t|t-1}$ wrt κ, ϕ and ω are stochastic recurrence equations (SREs); see Brandt (1986) and Straumann and Mikosch (2006, p 2450-1). The condition $b < 1$ is sufficient[2] to ensure that they are strictly stationarity and ergodic at the true parameter value. Similarly $b < 1$ is sufficient to ensure that the squares of the first derivatives are strictly stationary and ergodic.

Let $\boldsymbol{\psi}_0$ denote the true value of $\boldsymbol{\psi}$. Since the score and its derivatives wrt μ in the static model possess the required moments, it is straightforward to show that (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. and (ii) as $T \to \infty$, $(-1/T)\partial^2 \ln L(\boldsymbol{\psi}_0)/\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}' \xrightarrow{P} \mathbf{I}(\boldsymbol{\psi}_0)$. The final condition in Jensen and Rahbek (2004) is concerned

[2]The necessary condition for strict stationarity is $\mathrm{E}(\ln|x_t|) < 0$. This condition is satisfied at the true parameter value when $|a| < 1$ since, from Jensen's inequality, $\mathrm{E}(\ln|x_t|) \leq \ln \mathrm{E}(|x_t|) < 0$ and as already noted $b < 1$ implies $|a| < 1$.

with boundedness of the third derivative of the log-likelihood function in the neighbourhood of $\boldsymbol{\psi}_0$. The first derivative of u_t, (30) is a linear function of terms of the form $b_t^* = b_t^h(1-b_t)^k$, where h and k are non-negative integers, as is the second derivative. As regards u_t itself, since $u_t = (1-b_t)(y_t - \mu_{t|t-1})$, it can be seen that that $u_t = 0$ when $y_t = 0$ and $u_t \to 0$ as $|y_t| \to \infty$. Thus u_t, like its derivatives, is bounded for any admissible $\boldsymbol{\psi}$. Since

$$b_t = h(y_t;\boldsymbol{\psi})/(1+h(y_t;\boldsymbol{\psi})), \qquad 0 \leq h(y_t;\boldsymbol{\psi}) \leq \infty,$$

where $h(y_t;\boldsymbol{\psi})$ depends on y_t and $\boldsymbol{\psi}$, it is clear that for any admissible $\boldsymbol{\psi}$, $0 \leq b_t \leq 1$ and so $0 \leq b_t^* \leq 1$. Furthermore the derivatives of $\mu_{t|t-1}$ must be bounded at $\boldsymbol{\psi}_0$ since they are stable SREs which are ultimately dependent on u_t and its derivatives. They must also be bounded in the neighbourhood of $\boldsymbol{\psi}_0$ since the condition $b < 1$ is more than enough to guarantee the stability condition $\mathrm{E}(\ln|x_t|) < 0$.

Unknown shape parameters, including degrees of freedom, pose no problem as the third derivatives (including cross-derivatives) associated with them are almost invariably non-stochastic.

Additional references

Brandt, A. (1986). The Stochastic Equation $Y_{n+1} = A_nY_n + B_n$ with Stationary Coefficients. *Advances in Applied Probability* **18**, 211–220.

Jensen, S.T. and A. Rahbek (2004). Asymptotic Inference for Nonstationary GARCH. *Econometric Theory,* 20, 1203-26.

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.