Kernel density estimation for time series data

Andrew Harvey*, Vitaliy Oryshchenko

Faculty of Economics, University of Cambridge, United Kingdom

Abstract

A time-varying probability density function, or the corresponding cumulative distribution function, may be estimated nonparametrically by using a kernel and weighting the observations using schemes derived from time series modelling. The parameters, including the bandwidth, may be estimated by maximum likelihood or cross-validation. Diagnostic checks may be carried out directly on residuals given by the predictive cumulative distribution function. Since tracking the distribution is only viable if it changes relatively slowly, the technique may need to be combined with a filter for scale and/or location. The methods are applied to data on the NASDAQ index and the Hong Kong and Korean stock market indices.

Keywords: exponential smoothing, probability integral transform, time-varying quantiles, signal extraction, stock returns.

1. Introduction

A probability density function (PDF), or the corresponding cumulative distribution function (CDF), may be estimated nonparametrically by using a kernel. If the density is thought to change over time, observations may be weighted by introducing ideas derived from time series modelling. Although it has long been known that updating can be carried out recursively—see the discussion in Markovich (2007, pp.73-74)—there has been little or no exploration of the kind of weighting typically used in filtering for the mean or variance. For example, Hall & Patil (1994), suggest analysing evolving densities by moving blocks of data (which are then combined with suitable weighting).

*Corresponding author. Address for correspondence: Faculty of Economics, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. Tel.: +44 (0) 1223 335228; fax: +44 (0) 1223 335200.

Email addresses: ach34@cam.ac.uk (Andrew Harvey), vo213@cam.ac.uk (Vitaliy Oryshchenko)

One of the simplest time series weighting schemes takes the form of an exponentially weighted moving average (EWMA). This is widely used to estimate the level of a series and hence future observations. A similar scheme may be used to estimate the conditional variance. Such a scheme is widely used under the heading 'Riskmetrics' but a firmer theoretical underpinning is given by the integrated generalized autoregressive heteroskedasticity (GARCH) model. Other models imply other weighting schemes and hence other recursions for updating the estimates of parameters that are evolving over time. For example, changing growth rates and seasonal patterns can easily be accommodated. The recursions are usually combined with an assumption about the form of the one-step ahead predictive distribution and as a result a likelihood function can be constructed and then maximized with respect to the unknown parameters in the model. Once a model has been fitted, the one-step ahead predictions may be subjected to diagnostic checking by reference to the predictive distribution. Most commonly the predictive distribution is Gaussian and tests are carried out on the standardized residuals.

It is shown here that similar ideas carry over to the nonparametric estimation of a time-varying density or distribution function. Not only can updating be carried out recursively, but a likelihood function can be constructed from the predictive distributions. Hence dynamic parameters, such as the discount parameter in the EWMA, may be estimated by maximum likelihood. Furthermore the dynamic specification may be checked by using the residuals given by the predictive cumulative distribution function. The methods are those appropriate for the probability integral transform, as described in Diebold, Gunther, & Tay (1998).

Time varying quantiles may be extracted from the cumulative distribution function. In the time-invariant case there are efficiency gains for estimating quantiles this way as compared with simply using the sample quantiles calculated from the order statistics, but the gains are small; see Sheather & Marron (1990). There has been considerable interest in the last few years in estimating changing quantiles. The conditional autoregressive value at risk (CAViaR) approach of Engle & Manganelli (2004) models quantiles in terms of functions of past observations. De Rossi & Harvey (2009) adopt a different method, based on ideas from signal extraction and using only indicator variables. One drawback to the CAViaR approach is that, as pointed out by Gourieroux & Jasiak (2008), the quantiles may cross. This cannot happen with the cumulative distribution function.

Section 2 discusses linear filters and in section 3 filters for estimating time-varying densities are developed. Attention is focussed on the EWMA and a stable filter with an extra parameter. We also explain how to estimate the densities using a two-sided filter that is the equivalent of smoothing, or signal extraction, in time series and how to construct algorithms for weighting schemes associated with more general time series models. The ways in which bandwidth selection methods designed for time-invariant distributions may be adapted to deal with changing distributions are explored and estimation by maximum likelihood and cross-validation is discussed. Section 4 describes diagnostic checking with the probability integral transforms of the predictions. Section 5 discusses time-varying quantiles. Section 6 applies the methods to the NASDAQ index, while the link with tracking changes in the copula is illustrated in section 7.

2. Filters

A linear filter is a scheme for weighting current and past observations in order to estimate an unobserved component or a future value of the series. Thus the estimator of the level at time t could be written as

$$m_t = \sum_{i=0}^{t-1} w_{t,i} y_{t-i}, \qquad t = 1, \ldots, T,$$

where $w_{t,i}$ are weights. One way of putting more weight on the most recent observations is to let the weights decline exponentially. If t is large then $w_{t,i} = (1-\omega)\omega^i$, $i = 0, 1, 2, \ldots$, where ω is a discount parameter in the range $0 \leq \omega < 1$. (The weights sum to unity in the limit as $t \to \infty$). The attraction of exponential weighting is that estimates can be updated by a simple recursion, that is

$$m_t = \omega m_{t-1} + (1-\omega) y_t, \qquad t = 1, \ldots, T$$

with $m_0 = 0$ or $m_1 = y_1$. The filter can also be expressed in terms of the one step ahead prediction, so m_t is replaced by $m_{t+1|t}$. These are also the predictions of the series, that is $\hat{y}_{t+1|t} = m_{t+1|t}$. Thus the recursion can be written

$$m_{t+1|t} = m_{t|t-1} + (1-\omega)\nu_t, \qquad t = 1, \ldots, T, \tag{1}$$

where $\nu_t = y_t - \hat{y}_{t|t-1}$ is the one-step ahead prediction error or innovation.

The exponential weighting filter may be rationalized as the steady-state solution of an unobserved components model consisting of a random walk plus noise. The model, known as the local level model, is

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \quad t = 1, \ldots, T, \tag{2}$$
$$\mu_t = \mu_{t-1} + \eta_t, \qquad \eta_t \sim NID(0, \sigma_\eta^2).$$

where the disturbances ε_t and η_t 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 Kalman filter for the optimal estimator of μ_t based on information at time t is

$$m_{t+1|t} = (1 - k_t)\, m_{t|t-1} + k_t y_t, \quad t = 1, \ldots, T, \tag{3}$$

where $k_t = p_{t|t-1} / \left(p_{t|t-1} + 1\right)$ is the gain, and

$$p_{t+1|t} = p_{t|t-1} - \left[p_{t|t-1}^2 / \left(1 + p_{t|t-1}\right)\right] + q, \quad t = 1, \ldots, T,$$

where $q = \sigma_\eta^2 / \sigma_\varepsilon^2$ is the signal-noise ratio; see Harvey (2006, 1989, p.175). The MSE of $m_{t+1|t}$ is $\sigma_\varepsilon^2 p_{t+1|t}$. With a diffuse prior, $m_{1|0} = 0$ and as $p_{1|0} \to \infty$, $k_1 \to 1$. Hence $m_{2|1} = y_1$ and $p_{2|1} = 1 + q$. The steady-state solution for k_t is $1 - \omega$, where the parameter ω is a monotonic function of $q = \sigma_\eta^2 / \sigma_\varepsilon^2$. The likelihood function may be constructed from the one-step ahead prediction errors and maximized with respect to ω. Diagnostics may be performed on the residuals.

A backward smoothing filter is described in the appendix. The weights implicitly used in the smoother, that is the weights in

$$m_{t|T} = \sum_{t=1}^{T} w_{t,i} y_i, \quad t = 1, \ldots, T,$$

may be computed using the algorithm of Koopman & Harvey (2003). In a large sample, it follows from Whittle (1984) that

$$w_{t,i} \approx \{(1 - \omega)/(1 + \omega)\} \left[\omega^{|t-i|} + \omega^{2T-t-i+1} + \omega^{t+i-1}\right], \qquad t, i = 1, \ldots, T, \tag{4}$$

while in the middle of such a sample

$$w_{t,i} \approx \frac{1 - \omega}{1 + \omega} \omega^{|t-i|}, \qquad i = 1, \ldots, T. \tag{5}$$

These formulae are not used in our computations, but they are useful in showing the nature of the weighting patterns.

The random walk in (2) may be replaced by a stationary first-order autoregressive process. More complex models, perhaps with slopes and seasonals, may be set up and the appropriate filters derived by putting the model in state space form. Again the likelihood function may be constructed from the one-step ahead prediction errors given by the Kalman filter and the implicit weights for filtering and smoothing obtained from the algorithm of Koopman & Harvey (2003).

A nonlinear class of models may be constructed by applying the linear filters obtained from unobserved component models to transformations of the observations that reflect quantities of interest. For example, if the mean is fixed at zero, but the variance changes we might consider the filter

$$\sigma^2_{t+1|t} = (1-\omega)y_t^2 + \omega\sigma^2_{t|t-1} = \sigma^2_{t|t-1} + (1-\omega)(y_t^2 - \sigma^2_{t|t-1}), \quad t = 1, \ldots, T,$$

where the notation $\sigma^2_{t+1|t}$ accords with that used by Andersen, Bollerslev, Christoffersen, & Diebold (2006) for the variance in a GARCH model. This scheme is an EWMA in squares, with $y_t^2 - \sigma^2_{t|t-1}$ playing a similar role to the innovation in (1). It corresponds to integrated GARCH, where the predictive distribution in the Gaussian case is $y_t \mid Y_{t-1} \sim N(0, \sigma^2_{t|t-1})$. The more general filter

$$\sigma^2_{t+1|t} = (1-\omega^* - \omega)\sigma^2 + \omega^* y_t^2 + \omega\sigma^2_{t|t-1}, \quad t = 1, \ldots, T,$$

is stable when $\omega^* + \omega < 1$ and hence is able to generate a stationary series. Estimation may be simplified by setting σ^2 equal to the (unconditional) variance in the sample; this is known as 'variance targeting', as in Laurent (2007, p. 25).

If the above filtering schemes are viewed as approximations to an unobserved variance, the smoother that would correspond to the filter in a linear unobserved components model may be useful as a descriptive device.

The next section shows how filters may be applied to the whole distribution, rather than to selected moments.

3. Dynamic kernel density estimation

Using a sample of T observations drawn from a distribution $F(y)$ with a corresponding probability density function $f(y)$, a kernel estimator of $f(y)$ at point y is given by

$$\hat{f}_T(y) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{y - y_i}{h}\right), \tag{6}$$

where $K(\cdot)$ is the kernel and h is the bandwidth. The kernel, $K(\cdot)$, is a bounded PDF which is symmetric about the origin. The quadratic (Epanechnikov) kernel is 'optimal' if the choice is restricted to nonnegative kernels and the criterion is taken to be the asymptotic minimum integrated squared error for a fixed density; cf. Tsybakov (2009, Ch. 1). However, as the efficiency loss from using suboptimal kernels is typically small, the Gaussian kernel, whose relative efficiency is 0.95, is often used in practice; see e.g. Wand & Jones (1995, p. 31).

The choice of bandwidth is more important than the choice of kernel. One possible approach is cross-validation, but rule-of-thumb methods are common in applied work and they usually deliver satisfactory results. Examples include a normal reference rule and rule-of-thumb bandwidths as in Silverman (1986, p. 47).

The kernel estimator of the cumulative distribution function is given by

$$\widehat{F}_T(y) = \frac{1}{T}\sum_{i=1}^{T} H\left(\frac{y-y_i}{h}\right),$$

where $H(\cdot)$ is a kernel which now takes the form of a CDF. A kernel of this form may be obtained by integrating the kernel in (6).

The properties of kernel density estimators have been studied for dependent data; see Wand & Jones (1995). However the target PDF is an unconditional distribution, whereas here the aim is to estimate a conditional distribution.

3.1. Filtering and smoothing

In order to estimate a time varying density, a weighting scheme may be introduced into the kernel estimator so that (6) becomes

$$\hat{f}_t(y) = \frac{1}{h}\sum_{i=1}^{t} K\left(\frac{y-y_i}{h}\right) w_{t,i}, \quad t = 1,\ldots,T, \tag{7}$$

while, for the distribution function,

$$\widehat{F}_t(y) = \sum_{i=1}^{t} H\left(\frac{y-y_i}{h}\right) w_{t,i}. \tag{8}$$

In both cases, $\sum_{i=1}^{t} w_{t,i} = 1$, $t = 1,\ldots,T$. The weights, $w_{t,i}$, $i = 1,\ldots,t$, $t = 1,\ldots,T$, change over time, although in the steady-state, $w_{t,i} = w_{t-i}$.

Similarly for smoothing

$$\hat{f}_{t|T}(y) = \frac{1}{h}\sum_{i=1}^{T} K\left(\frac{y-y_i}{h}\right) w_{t,i}, \quad t = 1,\ldots,T,$$

and

$$\widehat{F}_{t|T}(y) = \sum_{i=1}^{T} H\left(\frac{y - y_i}{h}\right) w_{t,i}, \tag{9}$$

with $\sum_{i=1}^{T} w_{t,i} = 1$, $t = 1, \ldots, T$.

3.2. Recursions

Simple exponential weighting gives recursions similar to those of section 2. Thus for the CDF

$$\widehat{F}_t(y) = \omega \widehat{F}_{t-1}(y) + (1-\omega) H\left(\frac{y - y_t}{h}\right), \quad t = 1, \ldots, T.$$

Schemes of this kind are not new; see, for example, Wegman & Davies (1979).

The above recursion can be re-written with $\widehat{F}_{t+1|t}(y)$ replacing $\widehat{F}_t(y)$. A simple re-arrangement then gives

$$\widehat{F}_{t+1|t}(y) = \widehat{F}_{t|t-1}(y) + (1-\omega) V_t(y), \quad 0 \leq \omega < 1, \quad t = 1, \ldots, T,$$

where

$$V_t(y) = H\left(\frac{y - y_t}{h}\right) - \widehat{F}_{t|t-1}(y) \tag{10}$$

plays a similar role to the innovation[1] in (1). However, $V_t(y) < 0$ when $y_t > y$. Note also that $-\widehat{F}_{t|t-1}(y) \leq V_t(y) \leq 1 - \widehat{F}_{t|t-1}(y)$.

An analogous recursion can be written down for the PDF. To be specific

$$\widehat{f}_{t+1|t}(y) = \widehat{f}_{t|t-1}(y) + (1-\omega) \nu_t(y), \quad 0 \leq \omega < 1, \quad t = 1, \ldots, T,$$

where the innovation is

$$\nu_t(y) = \frac{1}{h} K\left(\frac{y - y_t}{h}\right) - \widehat{f}_{t|t-1}(y), \tag{11}$$

with $-\widehat{f}_{t|t-1}(y) \leq \nu_t(y) \leq h^{-1} K(0)$.

The filter can be initialized with $\hat{f}_{1|0}(y) = 0$ and, in order to ensure that the weights discounting past observations sum to unity, ω may be set to $1 - k_t$, where k_t is defined in (3), until such time, $t = m$, as the filter is deemed to have converged. Alternatively $\hat{f}_{m+1|m}(y)$ may be computed directly from (7). The CDF recursion for $\widehat{F}_{t+1|t}(y)$ may be similarly initialized from the first m observations.

[1]In a Gaussian model, $H(y_t) = y_t$ and $\widehat{F}_{t|t-1}(y) = \widehat{y}_{t|t-1}$. The only impact is on location and ν_t is a scalar.

The stable filter is

$$\widehat{F}_{t+1|t}(y) = (1 - \omega^* - \omega)\overline{F}(y) + \omega^* H\left(\frac{y - y_t}{h}\right) + \omega \widehat{F}_{t|t-1}(y), \quad t = 1, \ldots, T, \tag{12}$$

where $\overline{F}(y)$ is the unconditional kernel density for the whole sample ('distribution targeting'). Setting the initial condition as $\widehat{F}_{1|0}(y) = \overline{F}(y)$ means that the weight attached to $\overline{F}(y)$ at time t is $(1 - \omega^*)$, but that it gradually goes to $(1 - \omega^* - \omega)$. We can also write

$$\widehat{F}_{t+1|t}(y) = (1 - \omega^* - \omega)\overline{F}(y) + (\omega^* + \omega)\widehat{F}_{t|t-1}(y) + \omega^* \nu_t, \quad t = 1, \ldots, T.$$

More complex weighting schemes, derived from unobserved components models, may also be adopted. For example an integrated random walk trend yields a cubic spline and the Kalman filter may be reduced to a single equation recursion which for the CDF is

$$\widehat{F}_{t+1|t}(y) = 2\widehat{F}_{t|t-1}(y) - \widehat{F}_{t-1|t-2}(y) + k_1 \omega^* H\left(\frac{y - y_t}{h}\right) + k_2 \omega^* H\left(\frac{y - y_{t-1}}{h}\right),$$

where k_1 and k_2 are parameters that depend on a signal-noise ratio in the original unobserved components model.

The above filters for $\widehat{F}_{t+1|t}(y)$ and $\widehat{f}_{t+1|t}(y)$ may be run by defining a grid of N points in the range $[y_{\min}, y_{\max}]$. To implement the smoother recursively, as described in the appendix for the random walk plus noise, it is necessary to store the $N \times (T - m)$ matrix of innovations. It is also necessary to store r_t or $m_{t|t-1}$, depending on which algorithm is used. Alternatively we could just compute the weights for a given t, $t = m + 1, \ldots, T$, with the algorithm in Koopman & Harvey (2003), and so construct filtered and smoothed estimates of the PDF or CDF directly from the formulae in the previous sub-section. When the aim is to compute estimation criteria, residuals and a limited number of quantiles, algorithms based on the direct approach seem to be more computationally efficient. A full set of filtering and smoothing recursions for a grid is not necessary unless an estimate of the density is required for each time period.

3.3. Estimation

The recursive nature of the filter leads naturally to maximum likelihood (ML) estimation of the bandwidth, h, and any parameters governing the dynamics, such as the discount factor, ω, in exponential weighting. The log-likelihood function, normalized by the sample size, is

$$\ell(\omega, h) = \frac{1}{T - m} \sum_{t=m}^{T-1} \ln \hat{f}_{t|t-1}(y_{t+1}) = \frac{1}{T - m} \sum_{t=m}^{T-1} \ln \left[\frac{1}{h} \sum_{i=1}^{t} K\left(\frac{y_{t+1} - y_i}{h}\right) w_{t,i}(\omega)\right], \tag{13}$$

where $w_{t,i}(\omega)$ are the weights, which may be obtained as described in section 2, and m is some preset number of observations used to initialise the procedure. The value of m will depend on the sample at hand, but it may not be unreasonable to suggest setting $m = 50$ or 100 if the sample size is big. The main consideration is that the predictions are meaningful.

The log-likelihood (13) can be maximized subject to $\omega \in (0, 1]$ and $h > 0$ using constrained maximization with numerical derivatives obtained via finite differencing. Using a non-negative kernel with unbounded support, such as a Gaussian kernel, theoretically guarantees that $\widehat{f}_{t|t-1}(y_{t+1}) > 0$ for all $t = m, \ldots, T-1$. The problem arises when the density is evaluated at outlier points in that the estimate is numerically zero. In those cases $\widehat{f}_{t|t-1}(\cdot)$ can be set equal to a very small positive number.

From the theoretical point of view, it is interesting to note that as in a linear Gaussian model, such as (2), the likelihood can be written in terms of the innovations since, from (11), $\widehat{f}_{t|t-1}(y_t) = h^{-1}K(0) - \nu_t(y_t)$ for $t = m+1, \ldots, T$. Thus, instead of re-computing the density estimate at each t using the data up to $t-1$ inclusive, the recursive formulae given in section 3 can, in principle, be used. However, in order to evaluate the log-likelihood (13), the grid for the recursion will need to include all the sample values of y_t.

For smoothing, the parameters can be estimated by maximizing the likelihood cross-validation (CV) criterion

$$CV(\omega, h) = \frac{1}{T}\sum_{t=1}^{T} \ln \widehat{f}_{(-t)|T}(y_t) = \frac{1}{T}\sum_{t=1}^{T} \ln\left[\frac{1}{h}\sum_{\substack{i=1\\ i\neq t}}^{T} K\left(\frac{y_t - y_i}{h}\right) w_{t,T,i}(\omega)\right], \tag{14}$$

where $w_{t,T,i}(\omega)$ is given by a two-sided smoothing filter such as (4).

Alternatively, one can simply choose the same parameters as for filtering.

The number of parameters to be estimated could be reduced by setting the bandwidth according to a rule of thumb, $h = cT^{-1/5}$, where the constant c depends on the spread of the data[2] and $T = T(\omega)$ is set equal to the effective sample size. In this case the likelihood and the CV criterion are maximized only with respect to ω. In the steady-state of the local level model, the mean square error (MSE) of the contemporaneous filtered estimator, m_t, of the level is $\sigma_\varepsilon^2(1-\omega)$. If the level were

[2]For instance, if the kernel is the Gaussian density, and the underlying distribution is normal with variance σ^2, the constant in the asymptotically optimal bandwidth is $c = 1.06\sigma$. Another popular choice is $c = 1.06\min\left(\hat{\sigma}, \widehat{IQR}/1.34\right)$, where $\widehat{IQR}$ is the sample interquartile range; see Silverman (1986).

fixed, the MSE of the sample mean would be σ_ε^2/T. This suggests an effective sample size for the filtering of $T(\omega) = 1/(1-\omega)$. For smoothing the suggestion is $T(\omega) = (1+\omega)/(1-\omega) \approx 2/(1-\omega)$, provided that t is not too close to the beginning or end of the sample. Thus when the bandwidth selection criterion is proportional to $T^{-1/5}$, the bandwidth for filtering will be bigger by a factor of approximately $2^{1/5} = 1.15$.

Estimation procedure thus involves first maximizing the likelihood function (13) or the CV criterion (14) whereby obtaining estimates of the smoothing parameter, ω, and the bandwidth h. These estimates are then used to compute the estimates of the PDF, CDF and quantiles. CDF (filtered or smoothed) can be computed by applying formulae (8) and (9) directly. Quantile functions can be obtained by inverting estimated CDFs as described in section 5.1 below.

3.4. Correcting for changing mean and variance

If the series displays trending movements there is clearly a problem in implementing the preceding algorithms for estimating time-varying distributions. A possible solution is to model the level separately, for example by a random walk plus noise, and then to adjust the observations so that the dynamic kernel estimation is applied to the innovations. Thus $H(\cdot)$ in (10), or $K(\cdot)$, is re-defined by replacing y_t by $y_t - m_{t|t-1}$. Serial correlation may be similarly handled by fitting an autoregressive–moving average (ARMA) model.

The most straightforward option for dealing with short-term movements in variance is to fit a GARCH model for the conditional variance. Then $H(\cdot)$ becomes

$$H\left(\frac{y-(y_t-m_{t|t-1})}{h\sigma_{t|t-1}}\right) = H\left(\frac{y-y_t+m_{t|t-1}}{h\sigma_{t|t-1}}\right).$$

The disadvantage of pre-filtering is that the treatment of the scale and mean becomes decoupled from the estimation of the distribution as a whole.

4. Specification and diagnostic checking

The probability integral transform (PIT) of an observation from a given distribution has a uniform distribution in the range $[0, 1]$. Hence the hypothesis that a set of observations come from a particular parametric distribution can be tested. One possibility is to use the Kolmogorov-Smirnov test.

The PITs are often used to assess forecasting schemes; see Dawid (1984) or Diebold et al. (1998). Here the PIT is given directly by the predictive kernel CDF, that is the PIT of the t-th observation is $\widehat{F}_{t|t-1}(y_t)$, $t = m+1, \ldots, T$. As with the evaluation of $\widehat{f}_{t|t-1}(y_t)$ in the likelihood function, the calculation at each point in time need only be done for $y = y_t$.

The PITs may be expressed in terms of innovations. Specifically,

$$\widehat{F}_{t|t-1}(y_t) = H(0) - V_{t|t-1}(y_t) = 0.5 - V_{t|t-1}(y_t).$$

Hence $E(V_{t|t-1}(y_t)) = 0$ as $E(F_t(y_t)) = 0.5$.

If the PITs are not uniformly distributed, their shape can be informative. For example, a humped distribution indicates that the forecasts are too narrow and that the tails are not adequately accounted for; see Laurent (2007, p. 98). Plots of the autocorrelation functions (ACFs) of the PITs, and of absolute values[3] and powers of the demeaned PITs, may indicate the source of serial dependence. Tests statistics for detecting serial correlation, such as Box-Ljung, and stationarity test statistics may be used, but it should be noted that their asymptotic distribution is unknown. There may sometimes be advantages in transforming to normality as in Berkowitz (2001).

5. Time-varying quantiles

A plot showing how the quantiles have evolved over time provides a good visual impression of the changing distribution. The first sub-section below explains how quantiles can be computed from the kernel estimates.

Rather than estimating a time-varying distribution, time-varying quantiles may be computed directly, either by formulating a model for a particular quantile or using a nonparametric procedure. The second sub-section reviews some of these procedures and contrasts them with the kernel approach.

5.1. Kernel-based estimation

When the distribution is constant, the τ-quantile, $\xi(\tau)$, $0 < \tau < 1$, can be estimated from the distribution function by solving $\widehat{F}(y) = \tau$, ie $\widehat{F}^{-1}(\tau) = \widehat{\xi}(\tau)$. Nadaraya (1964) shows that $\widehat{\xi}(\tau)$ is consistent and asymptotically normal with the same asymptotic distribution as the sample quantile. Azzalini (1981) proposes the use of a Newton-Raphson procedures for finding $\widehat{\xi}(\tau)$.

Filtered and smoothed estimators of changing quantiles can be similarly computed from time-varying CDF's. Thus, for filtering, $\widehat{\xi}_{t|t-1}(\tau) = \widehat{F}^{-1}_{t|t-1}(\tau)$, for $t = m, \ldots, T$. The iterative procedure

[3] The absolute value of a demeaned PIT is also uniformly distributed, unlike its square.

to calculate $\widehat{\xi}_{t|t-1}(\tau)$ is based on the direct evaluation of $\widehat{F}_{t|t-1}(y)$ in the vicinity of the quantile. To reduce computational time, a good starting value can be obtained from a preliminary estimate of a CDF by (linear) interpolation[4]. Alternatively, for $t = m+1, \ldots, T$, the estimate in the previous time period may be used as a starting value.

The estimates of bandwidth obtained by ML or CV suffer from the drawback that the asymptotically optimal choice of bandwidth for a kernel estimator of a CDF is proportional to $T^{-\frac{1}{3}}$, whilst the optimal bandwidth for a PDF is proportional to $T^{-\frac{1}{5}}$; see, for example, Azzalini (1981). A bandwidth for a kernel estimator of a CDF can be found by cross-validation, as in Bowman, Hall, & Prvan (1998), or by a rule of thumb approach, as in Altman & Léger (1995). It may be worth experimenting with these bandwidth selection criteria for quantile estimation. Similar considerations might apply to computation of the PITs.

5.2. Direct estimation of individual quantiles

Yu & Jones (1998) adopt a nonparametric approach. Their (smoothed) estimate, $\widehat{\xi}_t(\tau)$, of the τ-quantile is obtained by (iteratively) solving

$$\sum_{j=-h}^{h} K(\frac{j}{h}) IQ(y_{t+j} - \widehat{\xi}_t) = 0,$$

where $\widehat{\xi}_t = \widehat{\xi}_t(\tau)$, $K(\cdot)$ is a weighting kernel (applied over time), h is a bandwidth and $IQ(\cdot)$ is the quantile indicator function

$$IQ(y_t - \xi_t) = \begin{cases} \tau - 1, & \text{if} \quad y_t < \xi_t, \\ \tau, & \text{if} \quad y_t > \xi_t, \end{cases} \qquad t = 1, \ldots, T.$$

$IQ(0)$ is not determined, but in the present context we can set $IQ(0) = 0$. Adding and subtracting $\widehat{\xi}_t$ to each of the $IQ(y_{t+j} - \widehat{\xi}_t)$ terms in the sum leads to an alternative expression

$$\widehat{\xi}_t = \frac{1}{\sum_{j=-h}^{h} K(j/h)} \sum_{j=-h}^{h} K(\frac{j}{h})[\widehat{\xi}_t + IQ(y_{t+j} - \widehat{\xi}_t)]. \tag{15}$$

[4] To be precise, in our code, the CDF is first estimated on a grid of K points $\xi_1, \ldots, \xi_K$, and the initial estimate of ξ_t is obtained by finding $\xi_{lo} = \max_j \left(\xi_j : \widehat{F}_t(\xi_j) \leq \tau\right)$ and $\xi_{up} = \min_j \left(\xi_j : \widehat{F}_t(\xi_j) \geq \tau\right)$, and linearly interpolating between them. This is then used as a starting value in solving $\widehat{F}_t(\xi_t) = \tau$ for ξ_t. The final solution can usually be found in just a few iterations (we used Matlab routine `fzero`). In fact, with large K, the precision of the initial estimate of ξ_t will be sufficient for all practical purposes.

De Rossi & Harvey (2006, 2009) estimate time-varying quantiles by smoothing with weighting patterns derived from linear models for signal extraction. These quantiles have no more than $T\tau$ observations below and no more than $T(1-\tau)$ above. The weighting scheme derived from the local level model gives

$$\widetilde{\xi}_t = \frac{1-\omega}{1+\omega} \sum_{j=-\infty}^{\infty} \omega^{|j|}[\widetilde{\xi}_t + IQ(y_{t+j} - \widetilde{\xi}_{t+j})],$$

in a doubly infinite sample; cf. (5). The nonparametric kernel $K(j/h)$ in (15) is replaced by $\omega^{|j|}$ so giving an exponential decay. Note that the smoothed estimate, $\widehat{\xi}_{t+j}$, is used instead of $\widehat{\xi}_t$ when j is not zero. The time series model determines the shape of the kernel while the signal-noise ratio plays a role similar to that of the bandwidth.

The smoothed estimate of a quantile at the end of the sample is the filtered estimate. The model-based approach automatically determines a weighting pattern at the end of the sample. For the EWMA scheme derived from the local level model, the filtered estimator must satisfy

$$\widetilde{\xi}_{t|t} = (1-\omega) \sum_{j=0}^{\infty} \omega^j [\widetilde{\xi}_{t-j|t} + IQ(y_{t-j} - \widetilde{\xi}_{t-j|t})].$$

Thus $\widetilde{\xi}_{t|t}$ is an EWMA of the synthetic observations, $\widetilde{\xi}_{t-j|t} + IQ(y_{t-j} - \widetilde{\xi}_{t-j|t})$. As new observations become available, the smoothed estimates need to be revised. However, filtered estimates could be used instead, so

$$\widehat{\xi}_{t+1|t}(\tau) = \widehat{\xi}_{t|t-1}(\tau) + (1-\omega)\nu_t(\tau), \tag{16}$$

where $\nu_t(\tau) = IQ(y_t - \widehat{\xi}_{t|t-1}(\tau))$ is an indicator that plays an analogous role to that of the innovation in the Kalman filter. Such a scheme would belong to the class of CAViaR models proposed by Engle & Manganelli (2004) in the context of tracking value at risk. In CAViaR, the conditional quantile is

$$\widehat{\xi}_{t+1|t}(\tau) = \alpha_0 + \sum_{i=1}^{q} \beta_i \widehat{\xi}_{t+1-i|t-i}(\tau) + \sum_{j=1}^{r} \alpha_j f(y_{t-j}),$$

where $f(y_t)$ is a function of y_t. Suggested forms include an *adaptive* model

$$\xi_t(\tau) = \xi_{t-1}(\tau) + \gamma\{[1 + \exp(\delta[y_{t-1} - \xi_{t-1}(\tau)])]^{-1} - \tau\}, \tag{17}$$

where δ is a positive parameter. The recursion in (16) has the same form as the limiting case $(\delta \to \infty)$ of (17). Other CAViaR specifications, which are based on actual values, rather than indicators, may suffer from a lack of robustness to additive outliers. That this is the case is clear from an examination of Fig. 1 in Engle & Manganelli (2004, p. 373). More generally, recent evidence on predictive performance in Kuester, Mittnik, & Paolella (2006, pp. 80–81) indicates a preference for the adaptive specification.

The advantage of fitting individual quantiles is that different parameters may be estimated for different quantiles. The disadvantage of having different parameters is that the quantiles may cross; see Gourieroux & Jasiak (2008). If the parameters across quantiles have to be the same to prevent quantiles crossing, the ability to have different models for different quantiles loses much of its appeal.

6. Empirical application: NASDAQ index

Data on the NASDAQ index was obtained from Yahoo-Finance (`http://uk.finance.yahoo.com`). The sample starts on 5th February 1971 and ends on 20th February 2009, thus covering 13,896 days. Once weekends and holidays are excluded, there are 9,597 observations. As is usually the case with financial series, there is clear volatility clustering and the correlograms of the absolute values and squares of demeaned returns are large and slowly decaying; see Fig. 1. Some of the sample autocorrelations for the actual returns and their cubes also lie outside the lines drawn at ± 2 standard deviations from the horizontal axis. The distribution of returns is heavy-tailed and asymmetric.

6.1. Time-varying kernel

Fig. 2 shows filtered (upper panel) and smoothed (lower panel) time-varying quantiles of NASDAQ returns for $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$. Exponential weights and an Epanechnikov kernel were used throughout. The discount parameters for filtering and smoothing were estimated by maximizing the log-likelihood and likelihood cross-validation criterion respectively[5]. The ML estimates of the discount parameter and bandwidth are, respectively, $\widetilde{\omega} = 0.9928$ and $\widetilde{h} = 0.4286$. The CV estimates (for smoothing) are $\widehat{\omega} = 0.9928$ and $\widehat{h} = 0.2555$.

The quantiles, which are plotted in Fig. 2, seem to track the changing distribution well. However, as Fig. 3 shows, there is still some residual serial correlation in absolute values and squares of the PITs. With raw data, changing volatility tends to show up more in absolute values

[5]Computations were performed in Matlab (`www.mathworks.com`); code is available on request from the second named author.

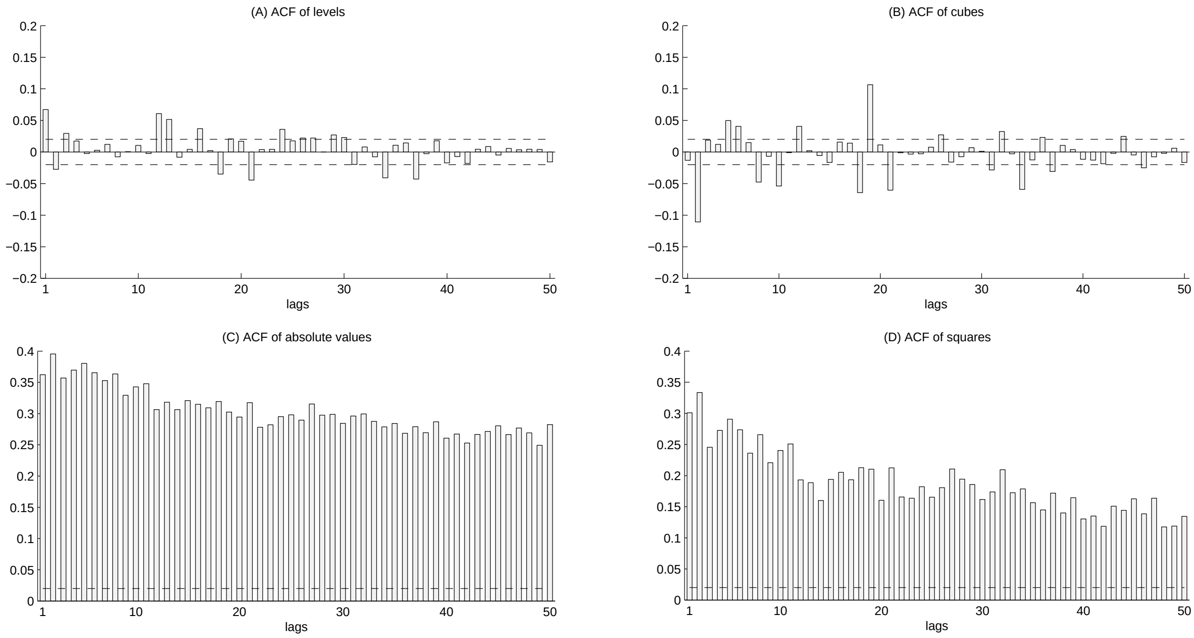

Figure 1: ACFs of NASDAQ returns.

[a]Panel A shows ACF of returns, y_t, panels B, C and D show ACFs of $(y_t - \bar{y})^3$, $|y_t - \bar{y}|$ and $(y_t - \bar{y})^2$ respectively. Lines parallel to the horizontal axis show ± 2 standard deviations (ie $2/\sqrt{T}$).

than in squares, as in Fig. 1. One reason for this happening is that sample autocorrelations are less sensitive to outliers when constructed from absolute values rather squares. However, the PITs do not have heavy tails, and the absolute value sample autocorrelations are, in most cases, slightly less than the corresponding sample autocorrelations computed from squares.

The first-order sample autocorrelation in the raw returns is rather high. It is even higher in the PITs. This may be partly a consequence of the transformation, though the higher order autocorrelations are, if anything, smaller than the corresponding autocorrelations for the raw returns.

The sample autocorrelations of the third and fourth powers of the demeaned PITs (not shown here) are, like those of the absolute values, small but persistent.

The histogram of PITs, shown in Fig. 3, is too high in the middle and too low at the ends, showing departures from uniformity and hence imperfections in the forecasting scheme. The hump-shaped distribution of the PITs indicates that the tail behaviour is not adequately captured. The problem could be caused by the bandwidth being too wide, resulting in a degree of oversmoothing. Forecasting performance might be improved by using different bandwidths for the tails and middle

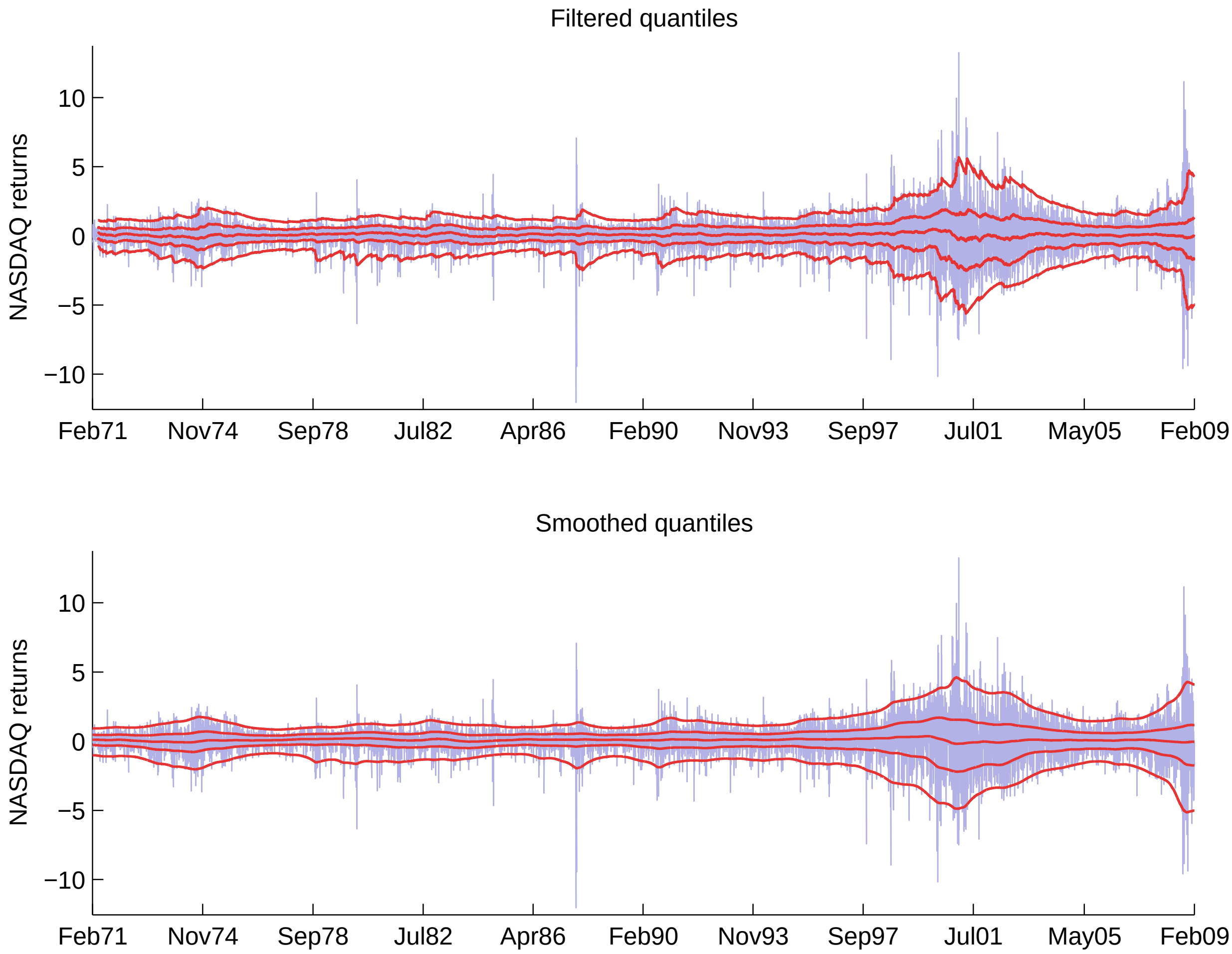

Figure 2: Filtered (upper panel) and smoothed (lower panel) time-varying quantiles of NASDAQ returns.

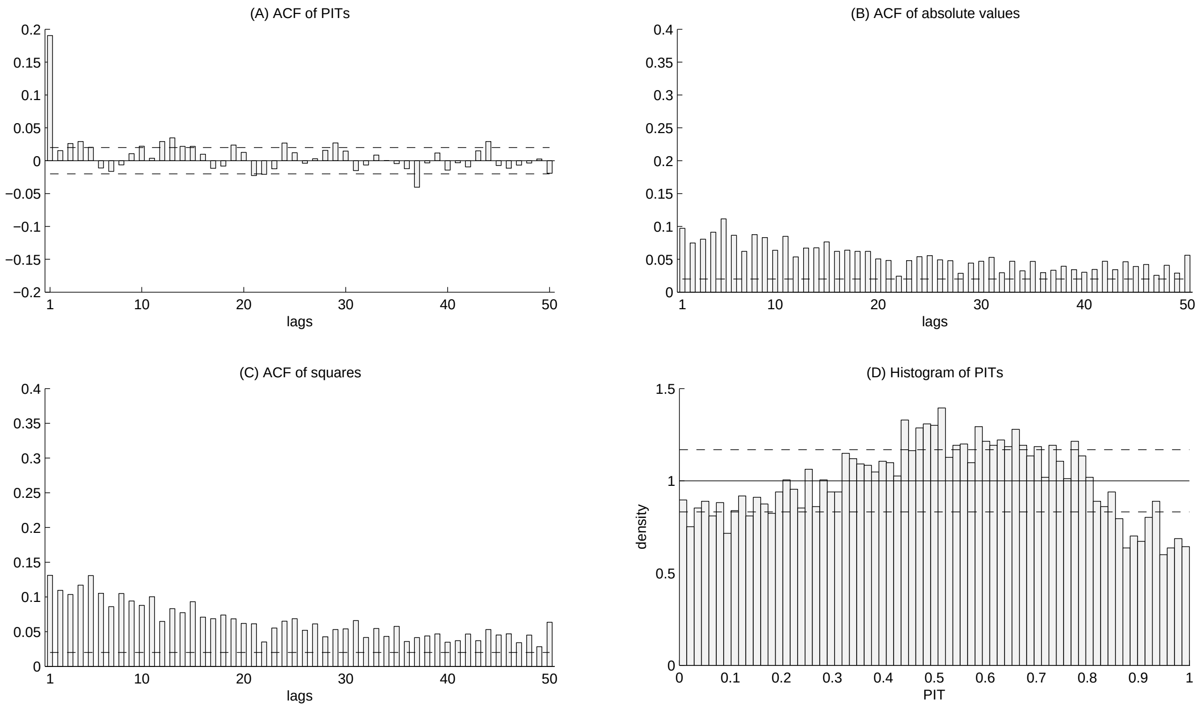

Figure 3: ACFs and histogram of PITs.

[a]Panels A, B and C show ACFs of PITs, z_t, absolute values, $|z_t - \bar{z}|$, and squares of the demeaned PITs, $(z_t - \bar{z})^2$, respectively. Lines parallel to the horizontal axis show ± 2 standard deviations (ie $2/\sqrt{T}$).
Panel D shows the histogram of PITs. Dashed lines show ± 2 standard deviations (ie $2\sqrt{(k-1)/T}$, where k is the number of bins).

of the distribution.

Changing the basis for bandwidth selection is unlikely to correct the failure to pick up short term serial correlation (at lag one) or to remove all the movements in volatility. The reason is that a time-varying kernel can really only pick up long-term changes. Hence there may be a case for pre-filtering.

6.2. ARMA-GARCH residuals

In order to pre-filter the NASDAQ data an MA(1) model with a GARCH(1,1)-t conditional variance equation was fitted using the G@RCH 5 program of Laurent (2007). The GARCH parameters were estimated to be 0.0979 (the coefficient of the lagged squared observation) and 0.9010, so the sum is close to the IGARCH boundary. The estimated MA(1) parameter was 0.2102, while the degrees of freedom of the t-distribution was estimated to be 7.04.

Fitting a time-varying kernel to the GARCH residuals gave ML estimates of $\widetilde{\omega} = 0.9996$ and $\widetilde{h} = 0.3595$, and CV estimates $\widehat{\omega} = 0.9991$ and $\widehat{h} = 0.3339$. The discount parameters are bigger than those estimated for the raw data and since they are closer to one there is less scope for picking up time variation, as can be seen from the quantiles in Fig. 4 (quantiles are shown for $\tau = 0.01$, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99). As might be anticipated, the pre-filtering effectively renders the median and inter-quartile range constant. Any remaining time variation is to be found in the high and low quantiles.

Some notion of the way in which tail dispersion changes can be obtained by plotting the ratio of the τ to $1-\tau$ range, for small τ, to the interquartile range, that is

$$\widetilde{\alpha}_t(\tau) = \frac{\widetilde{\xi}_t(1-\tau) - \widetilde{\xi}_t(\tau)}{\widetilde{\xi}_t(0.75) - \widetilde{\xi}_t(0.25)}, \quad \tau < 0.25,$$

where $\widetilde{\xi}_t(\tau)$ is an estimator that might be obtained by filtering or smoothing. Fig. 5 plots $\widetilde{\alpha}_t(\tau)$ for $\tau = 0.01$ and 0.05 computed using smoothed quantiles. Note that $\alpha(0.05)$ is 2.44 for a normal distribution and 2.66 for t_7; the corresponding figures for $\alpha(0.01)$ are 3.45 and 4.22 respectively.

For a symmetric distribution $\xi(\tau) + \xi(1-\tau) - 2\xi(0.5)$ is zero for all $t = 1, \ldots, T$. Hence a plot of the skewness measure

$$\widetilde{\beta}_t(\tau) = \frac{\widetilde{\xi}_t(1-\tau) + \widetilde{\xi}_t(\tau) - 2\widetilde{\xi}_t(0.5)}{\widetilde{\xi}_t(1-\tau) - \widetilde{\xi}_t(\tau)}, \quad \tau < 0.5,$$

shows how the asymmetry captured by the complementary quantiles, $\xi_t(\tau)$ and $\xi_t(1-\tau)$, changes over time. The statistic $\beta(0.25)$ was originally proposed by Bowley in 1920; see Groeneveld &

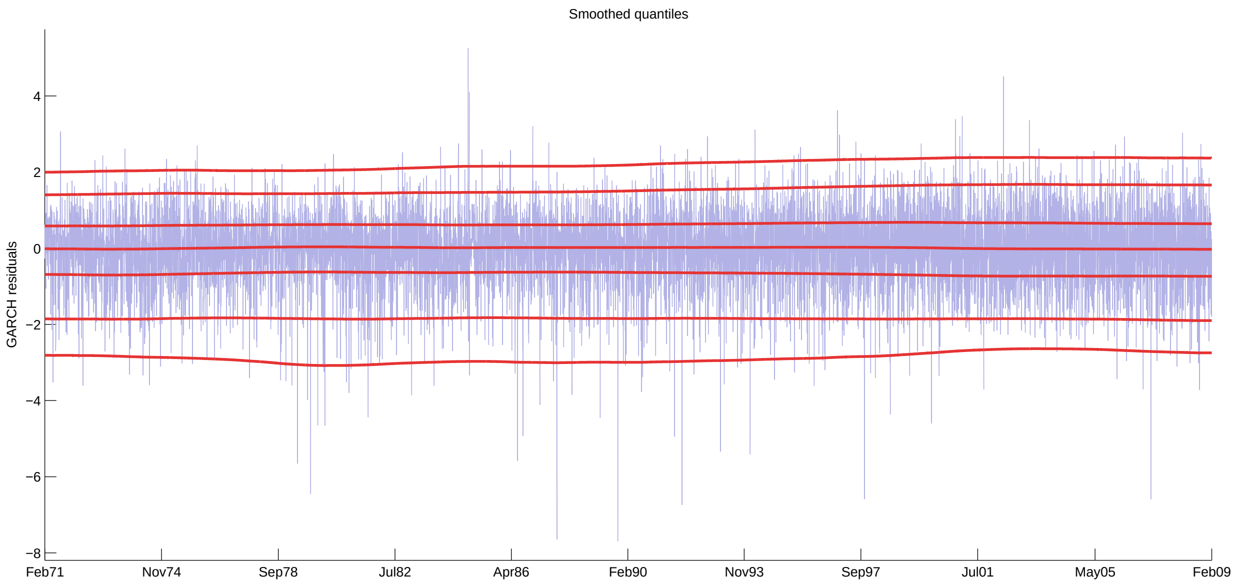

Figure 4: Smoothed time-varying quantiles of GARCH residuals.

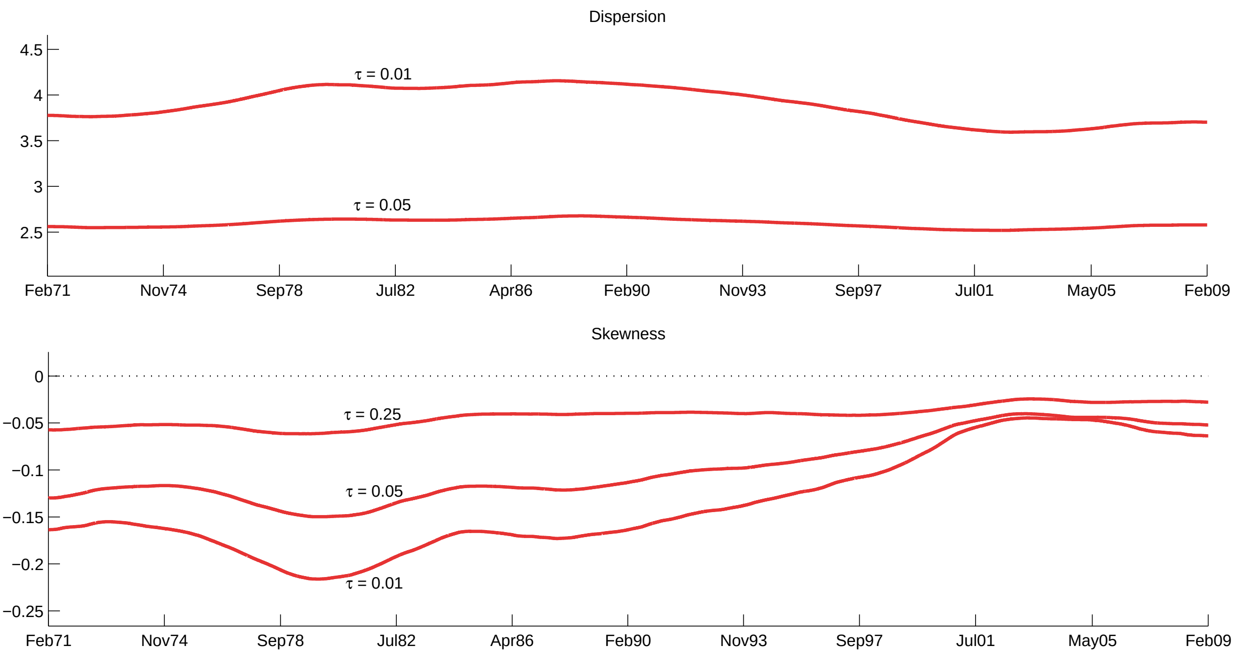

Figure 5: Changing tail dispersion and skewness for GARCH residuals.

Meeden (1984) for a detailed discussion. The maximum value of $\widetilde{\beta}_t(\tau)$ is one, representing extreme right (positive) skewness and the minimum value is minus one, representing extreme left skewness. Fig. 5 plots $\widetilde{\beta}_t(\tau)$ for $\tau = 0.01$, 0.05 and 0.25 using the smoothed quantiles. There is substantial time variation in skewness: it is high in the late 70s, whereas around 2002–2005, the distribution is almost symmetric.

The ACFs of the PITs, their squares and absolute values are shown in Fig. 6. There is far less serial correlation than in the corresponding correlograms in Fig. 3 The histogram of PITs from a time-varying kernel fitted to the ARMA-GARCH residuals, shown in Fig. 6, displays the same hump-shaped pattern as was evident in the PITs from the raw data, but arguably to a lesser extent.

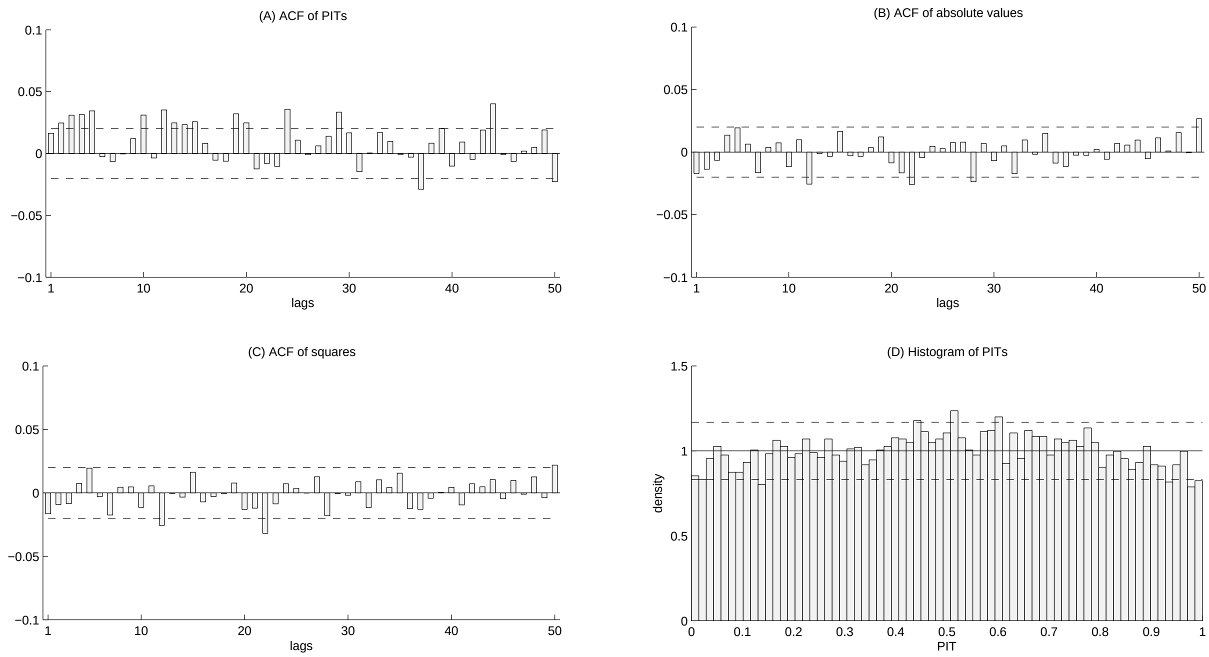

Figure 6: ACFs and histogram of PITs of GARCH residuals.

Panels A, B and C show ACFs of PITs, z_t, absolute values, $|z_t - \bar{z}|$, and squares of the demeaned PITs, $(z_t - \bar{z})^2$, respectively. Lines parallel to the horizontal axis show ± 2 standard deviations (ie $2/\sqrt{T}$).
Panel D shows the histogram of PITs. Dashed lines show ± 2 standard deviations (ie $2\sqrt{(k-1)/T}$, where k is the number of bins).

7. Quantiles and copulas

The quantiles can be used as a first step in tracking probabilities associated with a copula; see Harvey (2010) for a detailed discussion on this topic. For example, we may be interested in the probability that observations in two series are both below a certain quantile. The application described in Harvey (2010) is for the Hong Kong (Hang Seng) and Korean (SET) stock market indices. The time-varying quantiles for the returns for the two indices are obtained by a method based on estimating time-varying histograms, rather than by the kernel approach adopted here.

The ML estimates for an exponentially weighted kernel density for Hong Kong returns are $\widehat{\omega} = 0.9947$ and $\widehat{h} = 0.0050$; for Korean returns they are $\widehat{\omega} = 0.9948$ and $\widehat{h} = 0.0036$. In both cases the Epanechnikov kernel was used. The filtered quantiles for $\tau = 0.05$, 0.10, 0.25 and 0.50 are plotted in Fig. 7.

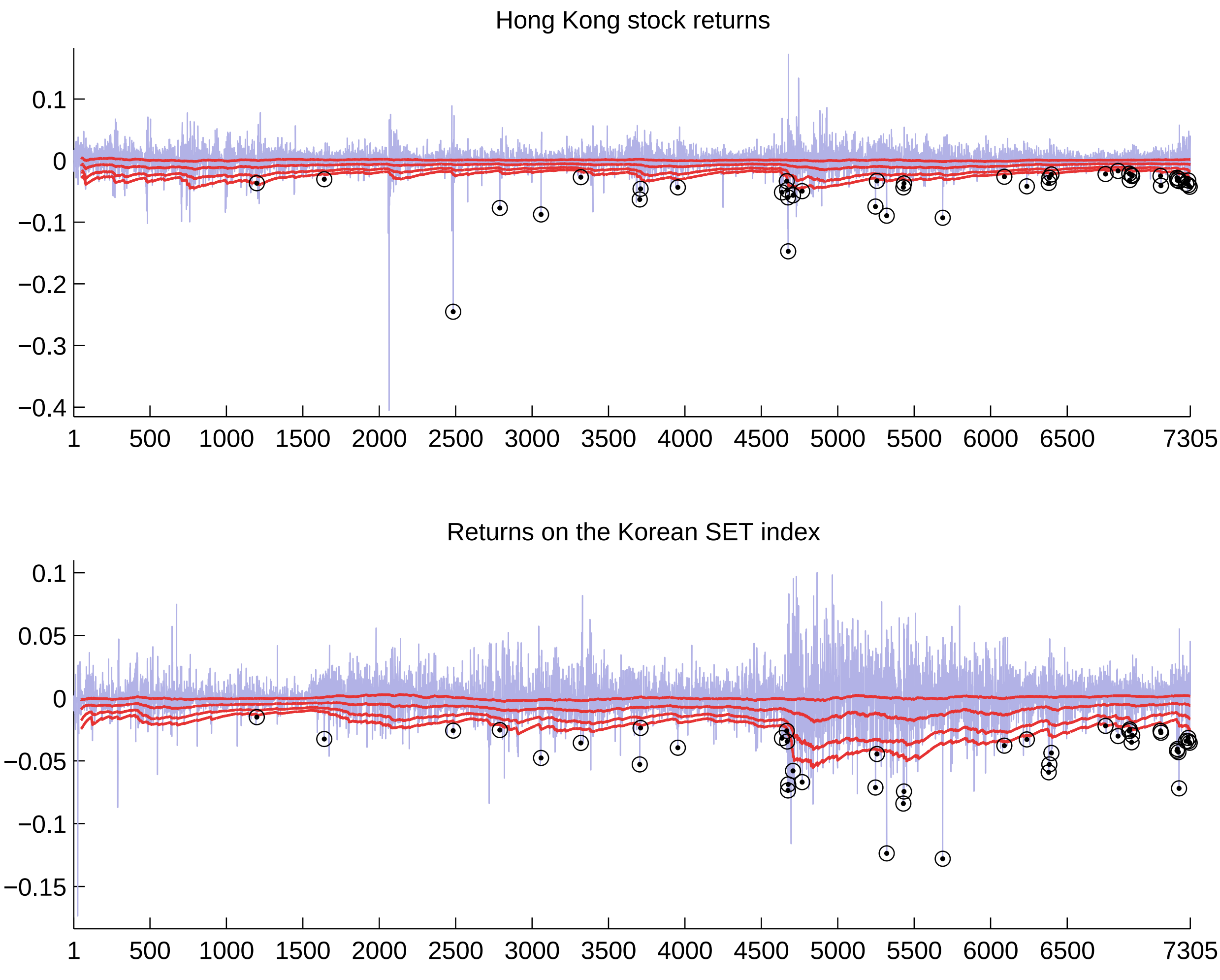

Figure 7: Returns and filtered 0.05, 0.10, 0.25, and 0.50-th quantiles.

An indicator variable which takes the value one whenever observations in both series fall below a certain quantile contains information on changes in the copula. For example, observations on

Hong Kong and Korea returns that both fall below their respective 0.05-th quantiles are highlighted with circles in Fig. 7. Filtering these indicator variables, perhaps also with exponential weighting, yields estimates of the probabilities that both series are below their respective τ-th quantiles at each point in time.

8. Conclusion

We have proposed a modification of the kernel density estimator that allows one to capture the changes in the density, and hence quantiles, by weighting the observations using schemes derived from time series models. The paper shows how the implied recursive procedures are of a similar form to those used for filtering time series observations to extract evolving means or variances. Associated smoothing schemes are obtained in the same way.

As is the case for many time series models, the likelihood function may be obtained from the predictive distribution. Hence the parameters governing the dynamics of the kernel can be estimated, together with the bandwidth, by maximum likelihood. Estimates for smoothing may be obtained by cross-validation. The innovations produced by the predictive CDF are probability integral transforms and can be used for diagnostic checking. If there is time variation in medians, asymmetry and the tails of distributions, tracking the changes in the whole distribution, or in a limited number of quantiles or quantile contrasts, may be informative.

Attention has been focussed on discounting past observations using exponential weighting. Exponential weighting is very simple to apply. However, generalizations to other weighting schemes are not difficult because the filters can be obtained from the state space forms of appropriate time series models. One scheme that certainly warrants future investigation is the stable filter corresponding to the standard stationary GARCH model.

The techniques were illustrated on NASDAQ, Hong Kong, and Korean stock market indices. The applications show the advantages of the proposed methods, but also expose their limitations. In particular the methods are only appropriate for monitoring distributions that change relatively slowly over time, since otherwise the effective sample size is too small. Short bursts of volatility may have to be accommodated by fitting a GARCH model. For tracking the copula, such prefiltering may not be necessary as the proposed technique is again only suitable for slow changes. A second limitation is that the bandwidth chosen by maximising the likelihood function or the likelihood cross-validation criterion appears to result in a degree of oversmoothing, which manifests itself in the hump-shaped histogram of the probability integral transforms. It may be possible to mitigate this effect by letting the bandwidth vary over the distribution, but the fundamental problem is that there is not enough information to provide an accurate description of tail behaviour. Modifications, such as combining kernel estimators with extreme value distributions for the tails, as in Markovich (2007, pp. 101–111), may be worth exploring.

Acknowledgements

We would like to thank two referees, Neil Shephard, Richard Smith, Ana Perez and participants at the Lisbon conference for helpful comments. An earlier version of the paper was presented at the **cemmap**/ ESRC Econometric Study Group Workshop on Quantile Regression (London, June 2009), ESRC Econometric Study Group Annual Conference (Bristol, July 2009), 7th OxMetrics User Conference (London, September 2009), 'Stats in the Château' summer school (Jouy-en-Josas, France, September 2009) and Econometrics Seminar (University of Cambridge, October 2009). Any errors are our responsibility. We would also like to thank Mardi Dungey for supplying the data on Hong Kong and Korea. Financial support from the ESRC under the grant Time-Varying Quantiles, RES-062-23-0129, is gratefully acknowledged.

Appendix. Smoothing in the local level model

The smoothed estimates for the Gaussian local level model (2) can be computed by saving the innovations and Kalman gains from the filter (3) and using them in the backward recursions

$$r_{t-1} = (1-k_t)r_t + (1-k_t)\nu_t, \qquad t = T, \ldots, 2,$$

where $\nu_t = y_t - m_{t|t-1}$ and $r_T = 0$, and

$$\begin{aligned} m_{t|T} &= m_{t|t-1} + p_{t|t-1}r_{t-1}, \qquad t = 1, \ldots, T, \\ &= m_{t|t-1} + k_t(r_t + \nu_t) \end{aligned}$$

Since $r_0 = (1-k_1)r_1 + (1-k_1)\nu_1$, initializing with a diffuse prior will give $m_{1|T} = (p_{1|0}/(p_{1|0} + 1))(r_1 + y_1)$ which goes to $r_1 + y_1$ as $p_{1|0}$ goes to infinity. The following forward recursion can also be used

$$m_{t+1|T} = m_{t|T} + qr_t, \qquad t = 1, \ldots, T-1,$$

with $m_{1|T} = r_1 + y_1$; see Koopman (1993).

References

Altman, N. & Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, *46*(2), 195–214.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. In: Elliott, G., Granger, C., & Timmermann, A. (eds.), *Handbook of Economic Forecasting*, chap. 15, pp. 777–878, Amsterdam: North Holland.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, *68*(1), 326–328.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4), 465–474.

Bowman, A., Hall, P., & Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, *85*(4), 799–808.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, *147*(2), 278–292.

De Rossi, G. & Harvey, A. C. (2006). Time-varying quantiles. CWPE 0649, University of Cambridge.

De Rossi, G. & Harvey, A. C. (2009). Quantiles, expectiles and splines. *Journal of Econometrics*, *152*(2), 179–185.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts, with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Engle, R. F. & Manganelli, S. (2004). Caviar: conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistic*, *22*, 367–381.

Gourieroux, C. & Jasiak, J. (2008). Dynamic quantile models. *Journal of Econometrics*, *147*(1), 198–205.

Groeneveld, R. A. & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *33*(4), 391–399.

Hall, P. & Patil, P. (1994). On the efficiency of on-line density estimators. *IEEE Transactions on Information Theory*, *40*(5), 1504–1512.

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

Harvey, A. C. (2006). Forecasting with unobserved components time series models. In: Elliott, G., Granger, C., & Timmermann, A. (eds.), *Handbook of Economic Forecasting*, chap. 7, pp. 327–412, Amsterdam: North Holland.

Harvey, A. C. (2010). Dynamic distributions and changing copulas. Forthcoming in the *Journal of Empirical Finance*, DOI: 10.1016/j.jempfin.2009.10.004.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, *80*(1), 117–126.

Koopman, S. J. & Harvey, A. C. (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control*, *27*(7), 1317–1333.

Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, *4*(1), 53–89.

Laurent, S. (2007). *GARCH 5*. Timberlake Consultants Ltd., London.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*. Wiley series in probability and statistics, John Wiley & Sons.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability and its Applications*, *9*(3), 497–500.

Sheather, S. J. & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, *85*(410), 410–416.

Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. (Springer Series in Statistics), Springer.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

Wegman, E. J. & Davies, H. I. (1979). Remarks on some recursive estimators of a probability density. *The Annals of Statistics*, *7*(2), 316–327.

Whittle, P. (1984). *Prediction and regulation by linear least-square methods*. Oxford: Blackwell, 2nd edn.

Yu, K. & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, *93*(441), 228–237.