Generalised Density Forecast Combinations*

N. Fawcett
Bank of England

G. Kapetanios
Queen Mary, University of London

J. Mitchell
Warwick Business School

S. Price
Bank of England, City University London and CAMA

September 4, 2014

Abstract

Density forecast combinations are becoming increasingly popular as a means of improving forecast 'accuracy', as measured by a scoring rule. In this paper we generalise this literature by letting the combination weights follow more general schemes. Sieve estimation is used to optimise the score of the generalised density combination where the combination weights depend on the variable one is trying to forecast. Specific attention is paid to the use of piecewise linear weight functions that let the weights vary by region of the density. We analyse these schemes theoretically, in Monte Carlo experiments and in an empirical study. Our results show that the generalised combinations outperform their linear counterparts.

JEL Codes: C53

Keywords: Density Forecasting; Model Combination; Scoring Rules

1 Introduction

Density forecast combinations or weighted linear combinations, or pools, of prediction models are becoming increasingly popular in econometric applications as a means of improving forecast 'accuracy', as measured by a scoring rule (see Gneiting & Raftery (2007)), especially in the face of uncertain instabilities and uncertainty about the preferred model; e.g., see Jore et al. (2010), Geweke & Amisano (2012) and Rossi (2013). Geweke & Amisano (2011) contrast Bayesian model averaging with linear combinations of predictive densities, so-called "opinion pools", where the weights on the component density forecasts are optimised to maximise the score, typically the logarithmic score, of the density combination as suggested in Hall & Mitchell (2007).

*Address for correspondence: James Mitchell (James.Mitchell@wbs.ac.uk). We thank Gianni Amisano, Sheheryar Malik, Neil Shephard and participants at the UCL/Bank of England workshop in honour of Mark Watson (2013), ESEM 2013 and MMF 2013 for helpful comments and advice.

In this paper we generalise this literature by letting the combination weights follow more general schemes. Specifically, we let the combination weights depend on the variable one is trying to forecast. We let the weights in the density combination depend on, for example, where in the forecast density you are, which is often of interest to economists. This allows for the possibility that while one model may be particularly useful (and receive a high weight in the combination) when the economy or market is in recession or a bear market, for example, another model may be more informative when output growth is positive or there is a bull market. The idea of letting the weights on the component densities vary according to the (forecast) value of the variable of interest contrasts with two recent suggestions, to let the weights in the combination follow a Markov-switching structure (Waggoner & Zha (2012)), or to evolve over time according to a Bayesian learning mechanism (Billio et al. (2013)). Accommodating time-variation in the combination weights mimics our approach to the extent that over time one moves into different regions of the forecast density. Our approach is also distinct and not subsumed by the combination methods considered in Gneiting & Ranjan (2013) which take nonlinear transformations of a linear pool with fixed weights, rather than nonlinear or what we call generalised pools where the weights themselves induce the nonlinearities.

The plan of this paper is as follows. Section 2 develops the theory behind the generalised density combinations or pools. It proposes the use of sieve estimation (cf. Chen & Shen (1998)) as a means of optimising the score of the generalised density combination over a tractable, approximating space of weight functions on the component densities. We consider, in particular, the use of piecewise linear weight functions that have the advantage of explicitly letting the combination weights depend on the region, or specific quantiles, of the density. This means prediction models can be weighted according to their respective abilities to forecast across different regions of the distribution. We also discuss implementation and estimation of the generalised pool in practice, given the extra parameters involved in a generalised pool. We consider cross-validation as a data-dependent means of determining the degree of flexibility of the generalised pool. Importantly to mitigate the risks of over-fitting in-sample we suggest that cross-validation is undertaken over an out-of-sample period. Section 3 draws out the flexibility afforded by generalised density combinations by undertaking a range of Monte Carlo simulations. These show that the generalised combinations are more flexible than their linear counterparts and in general can better mimic a range of true but unknown densities, irrespective of their forms. But this additional flexibility does come at the expense of the introduction of additional parameters and the simulations indicate that the benefits of generalised combinations mostly survive the extra parameter estimation uncertainty; and increasingly so for larger sample sizes and more distinct component densities. Section 4 then shows how the generalised combinations can work better in practice, finding that they deliver more accurate density forecasts of the S&P500 daily return than optimal linear combinations of the sort used in Hall & Mitchell (2007) and Geweke & Amisano (2011). Section 5 concludes.

2 Generalised density combinations: theory

We provide a general scheme for combining density forecasts. Consider a covariance stationary stochastic process of interest y_t, $t = 1, ..., T$ and a vector of covariance stationary predictor variables x_t, $t = 1, ..., T$. Our aim is to forecast the density of y_{t+1} conditional on $\mathfrak{F}_t = \sigma\left(x_{t+1}, (y_t, x_t')', ..., (y_1, x_1')'\right)$, where σ denotes a sigma field.

We assume the existence of a set of N density forecasts, denoted $q_i\left(y|\mathfrak{F}_t\right) \equiv q_{it}\left(y\right)$, $i = 1, ..., N$, $t = 1, ...T$. We suggest a generalised combined density forecast, given by the generalised "opinion pool"

$$p_t\left(y\right) = \sum_{i=1}^{N} w_{it}\left(y\right) q_{it}\left(y\right), \tag{1}$$

such that

$$\int p_t\left(y\right) dy = 1, \tag{2}$$

where $w_{it}\left(y\right)$ are the weights on the individual or component density forecasts which themselves depend on y. This generalises existing work on optimal linear density forecast combinations, where $w_{it}\left(y\right) = w_i$ and w_i are scalars; see Hall & Mitchell (2007) and Geweke & Amisano (2011). Note the dependence in (1) of w on t. This arises out of the need to satisfy (2); when $w_{it}\left(y\right) = w_i$ only $\sum_{i=1}^{N} w_i = 1$ is required. In the next few paragraphs we abstract from this time dependence to discuss the general problem of determining w. We will revisit this issue after that discussion. As a result, for notational ease only, we temporarily drop the subscript t on w. Note that, unlike Billio et al. (2013), we do not explicitly parameterise the time-variation. Our time variation arises due to the need to normalise the combined density to integrate to one; and this normalisation is by construction time-varying.

We need to provide a mechanism for deriving $w_i\left(y\right)$. Accordingly, we define a predictive loss function given by

$$L_T = \sum_{t=1}^{T} l\left(p_t\left(y_t\right); y_t\right). \tag{3}$$

We assume that the true weights, $w_i^0\left(y\right)$, exist in the space of q_i-integrable functions Ψ_{q_i} where

$$\Psi_{q_i} = \left\{ w\left(.\right) : \int w\left(y\right) q_i\left(y\right) dy < \infty \right\}, i = 1, .., N, \tag{4}$$

such that

$$E\left(l\left(p_t\left(y_t\right); y_t\right)\right) \equiv E\left(l\left(p_t\left(y_t; w_1^0, .., w_N^0\right); y_t\right)\right) \leq E\left(l\left(p_t\left(y_t; w_1, .., w_N\right); y_t\right)\right), \tag{5}$$

for all weight functions $(w_1, .., w_N) \in \prod_i \Psi_{q_i}$. We suggest an extremum estimator for the weight

function $w_i(y)$ which involves minimising L_T, i.e.,

$$\{\hat{w}_{1T}, ..., \hat{w}_{NT}\} = \arg \min_{w_i, i=1,...,N} L_T. \tag{6}$$

But the minimisation problem in (6) is impossible to solve unless one restricts the space over which one searches from Ψ_{q_i} to a more tractable space. A general way forward à la Chen & Shen (1998) is to minimise over

$$\Phi_{q_i} = \left\{ w_{\eta i}(.) : w_{\eta i}(y) = \tilde{v}_{i0} + \sum_{s=1}^{\infty} \tilde{\nu}_{is} \eta_s(y, \boldsymbol{\theta}_s), \quad \tilde{\nu}_{is} \geq 0, \ \eta_s(y, \boldsymbol{\theta}_s) \geq 0 \right\}, i = 1, .., N, \tag{7}$$

where $\{\eta_s(y, \boldsymbol{\theta}_s)\}_{s=1}^{\infty}$ is a known basis, up to a finite dimensional parameter vector $\boldsymbol{\theta}_s$, such that Φ_{q_i} is *dense* in Ψ_{q_i}, and $\{\tilde{v}_{is}\}_{s=0}^{\infty}$ are a sequence of constants.[1] Such a basis can be made of any of a variety of functions including trigonometric functions, indicator functions, neural networks and splines.

Φ_{q_i}, in turn, can be approximated through sieve methods by

$$\Phi_{q_i}^T = \left\{ w_{T\eta i}(.) : w_{T\eta i}(.) = \tilde{v}_{i0} + \sum_{s=1}^{p_T} \tilde{\nu}_{is} \eta_s(y, \boldsymbol{\theta}_s) \quad \tilde{\nu}_{is} \geq 0, \ \eta_s(y, \boldsymbol{\theta}_s) \geq 0 \right\}, i = 1, .., N, \tag{8}$$

where $p_T \to \infty$ is either a deterministic or data-dependent sequence, and $\left\{\Phi_{q_i}^T\right\}$ is dense in Φ_{q_i} as $p_T \to \infty$.

Note that a sufficient condition for (2) is

$$\begin{aligned} \int_Y \sum_{i=1}^{N} \left(\tilde{v}_{it0} + \sum_{s=1}^{p_T} \tilde{\nu}_{its} \eta_s(y, \boldsymbol{\theta}_s) q_{it}(y) \right) dy &= \sum_{i=1}^{N} \left(\tilde{v}_{it0} + \sum_{s=1}^{p_T} \tilde{\nu}_{its} \int_Y \eta_s(y, \boldsymbol{\theta}_s) q_{it}(y) \, dy \right) \\ &= \sum_{i=1}^{N} \left(\tilde{v}_{it0} + \sum_{s=1}^{p_T} \tilde{\nu}_{its} \kappa_{its} \right) = 1, \end{aligned} \tag{9}$$

where $\kappa_{its} = \int_Y \eta_s(y, \boldsymbol{\theta}_s) q_{it}(y) \, dy$. It is clear that $\tilde{v}_{it0}$ and $\tilde{\nu}_{its}$ depend on t given that κ_{its} depends on $q_{it}(y)$ which is a function of t. A natural way to impose this normalisation, (9), is to define

$$\tilde{v}_{it0} = \tilde{v}_t(v_{i0}; \boldsymbol{v}_i^{(-0)}) = \tilde{v}_t(v_{i0}) = \frac{v_{i0}}{\sum_{i=1}^{N} (v_{i0} + \sum_{s=1}^{p_T} v_{is} \kappa_{its})}, \tag{10}$$

$$\tilde{\nu}_{its} = \tilde{v}_t(v_{is}; \boldsymbol{v}_i^{(-s)}) = \tilde{v}_t(v_{is}) = \frac{v_{is}}{\sum_{i=1}^{N} (v_{i0} + \sum_{s=1}^{p_T} v_{is} \kappa_{its})}, \tag{11}$$

[1] Informally, a space is *dense* in another larger space if every element of the larger space can be approximated arbitrarily well by a sequence of elements of the smaller space.

where $\boldsymbol{\nu}_i = \left(v_{i0}, ..., \nu'_{ip_T}\right)'$, $\boldsymbol{v}_i^{(-s)}$ is $\boldsymbol{\nu}_i$ with element v_{is} removed and the new transformed weights v_{i0} and v_{is} are fixed over time and need only be positive. Further, it is assumed throughout that the $\boldsymbol{\nu}_i$ satisfy $\inf_t \sum_{i=1}^N \left(v_{i0} + \sum_{s=1}^{p_T} v_{is}\kappa_{its}\right) > 0$. The notation $\tilde{v}_t(.)$ is used to denote the function between the time-varying weights $\tilde{v}_{its}$, $s = 0, ..., p_T$, and $\boldsymbol{\nu}_i$, given in (10)-(11), and it is readily seen to be continuous and twice-differentiable for all t.

Defining $\boldsymbol{\vartheta}_T = \left(\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_N, \boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_{p_T}\right)'$,

$$\hat{\boldsymbol{\vartheta}}_T = \arg\min_{\vartheta_T} L_T\left(\boldsymbol{\vartheta}_T\right), \tag{12}$$

we can use Corollary 1 of Chen & Shen (1998) to show the following convergence result for the loss function evaluated at the extremum estimator $\hat{\boldsymbol{\vartheta}}_T$ and the true parameters

$$\frac{1}{T}\left(L_T\left(\hat{\boldsymbol{\vartheta}}_T\right) - L_T\left(\boldsymbol{\vartheta}^0\right)\right) = o_p\left(1\right), \tag{13}$$

where $\boldsymbol{\vartheta}^0$ denotes the true parameter value defined formally in Assumption 1 below.

In fact, Theorem 1 of Chen & Shen (1998) also provides rates for the convergence of $\hat{\boldsymbol{\vartheta}}_T$ to $\boldsymbol{\vartheta}^0$. However, the proofs depend crucially on the choice of the basis $\{\eta_s\left(y, \boldsymbol{\theta}_s\right)\}_{s=1}^{\infty}$. Chen & Shen (1998) discuss many possible bases. We will derive a rate for one such basis when p_T continues to depend on time, and potentially tends to infinity, in Section 2.1 below.

Alternatively, rather than focus on specific bases, we can limit our analysis to less general spaces and then derive convergence and normality results. In particular, we can search over

$$\Phi_{q_i}^p = \left\{ w_{T\eta}\left(.\right) : w_{p\eta}\left(.\right) = \tilde{v}_t(v_{i0}) + \sum_{s=1}^{p} \tilde{v}_t(v_{is})\eta_s\left(y, \boldsymbol{\theta}_s\right) \right\}, i = 1, .., N, \tag{14}$$

for some finite p. (Below we consider how, in practice, to choose p *via* cross-validation.) Let $\boldsymbol{\vartheta}_p = \left(\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_N, \boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_p\right)'$.

We make the following assumptions:

Assumption 1 *For all p, there exists a unique $\boldsymbol{\vartheta}_p^0 \in$ int Θ_p that minimises $E\left(L_T\left(\mathbf{w}_{T\eta}\left(\boldsymbol{\vartheta}_T\right)\right)\right)$, for all p.*

Assumption 2 *$l\left(p_t\left(., \boldsymbol{\vartheta}_p\right); .\right)$ has bounded derivatives with respect to $\boldsymbol{\vartheta}_p$ uniformly over Θ_p, t and p.*

Assumption 3 *y_t is a L_r-bounded ($r > 2$), $L_2 - NED$ (near epoque dependent) process of size $-a$, on an α-mixing process, V_t, of size $-r/(r-2)$ such that $a \geq (r-1)/(r-2)$.*

Assumption 4 *Let $l^{(i)}\left(p_t\left(y_t, \boldsymbol{\vartheta}_p\right); y_t\right)$ denote the i-th derivative of l with respect to y_t, where*

$l^{(0)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)=l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$. *Let*

$$\left|l^{(i)}\left(p_t\left(y^{(1)},\boldsymbol{\vartheta}_p^0\right);y^{(1)}\right)-l^{(i)}\left(p_t\left(y^{(2)},\boldsymbol{\vartheta}_p^0\right);y^{(2)}\right)\right|\leq B_t^{(i)}\left(y^{(1)},y^{(2)}\right)\left|y^{(1)}-y^{(2)}\right|,\quad i=0,1,2 \tag{15}$$

where $B_t^{(i)}(.,.)$ *are nonnegative measurable functions and* $y^{(1)},y^{(2)}$ *denote the arguments of the relevant functions. Then, for all lags* m *(*$m=1,2,...$*), and some* $r>2$,

$$\left\|B_t^{(i)}\left(y_t,E\left(y_t|V_{t-m},...,V_{t+m}\right)\right)\right\|_r<\infty,\quad i=0,1,2, \tag{16}$$

$$\left\|l^{(i)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p^0\right);y^{(1)}\right)\right\|_r<\infty,\quad i=0,1,2, \tag{17}$$

where $\|.\|_r$ *denotes* L_r *norm.*

Assumption 5 $q_{it}(y)$ *are bounded functions for* $i=1,...,N$.

It is worth commenting on the use of NED processes in Assumption 3. NED processes can accommodate a wider variety of dependence than mixing and as such provide a useful broad framework of analysis. They are discussed in detail in a number of sources including Davidson (1994, Ch. 17).

It is then straightforward to show that:

Theorem 1 *Under Assumptions 1-5, and for a finite value of* p,

$$\sqrt{T}\left(\hat{\boldsymbol{\vartheta}}_T-\boldsymbol{\vartheta}_p^0\right)\to^p N(0,V) \tag{18}$$

where

$$V=V_1^{-1}V_2V_1^{-1}, \tag{19}$$

$$V_1=\frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial^2 l\left(p_t\left(y_t\right);y_t\right)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\right|_{\hat{\boldsymbol{\vartheta}}_T}, \tag{20}$$

$$V_2=\frac{1}{T}\sum_{t=1}^{T}\left(\left.\frac{\partial l\left(p_t\left(y_t\right);y_t\right)}{\partial\boldsymbol{\vartheta}}\right|_{\hat{\boldsymbol{\vartheta}}_T}\right)\left(\left.\frac{\partial l\left(p_t\left(y_t\right);y_t\right)}{\partial\boldsymbol{\vartheta}}\right|_{\hat{\boldsymbol{\vartheta}}_T}\right)'. \tag{21}$$

Using this result one can test the null hypothesis of fixed weights, $v_{i0}=w_i$, versus our proposed generalised weights:

$$H_0:\boldsymbol{\nu}_i^0=(v_{i0},0,...,0)',\ \forall i. \tag{22}$$

Rejection of (22) implies that it does help to let the weights depend on y. It is clear from Theorem 1 that the generalised pool reduces the value of the objective (loss) function relative to the linear pool except when $\boldsymbol{\nu}_i^0$, under H_0, minimises the population objective function. If

$\boldsymbol{\nu}_i^0$ does minimise the loss function then both pools achieve this minimum. And if one of the component densities is, in fact, the true density, then the loss function is minimised when the true density is given a unit weight, and all other densities are given a zero weight. However, there is no guarantee, theoretically, that this will be the case in practice, as other sets of weights potentially can also achieve the minimum objective function.

Note that in Theorem 1, and the test in (22), the basis function is assumed fully known and does not depend on unknown parameters. In practice, in many applications, this basis is not known down to the number of terms, p, and there are unknown parameters. This means that under H_0 these unknown parameters $\left(v_{i1}, ..., \nu'_{ip}\right)'$, $\forall i$, and $\boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_p$ have no significance and become nuisance parameters unidentified under the null (e.g., see Hansen (1996)). Below in Section 2.1 we suggest, in the specific context of indicator basis functions, but the discussion is general, an estimation and inferential procedure when there are these unknown parameters.

In practice, p can also be estimated to minimise the loss, $l\left(p_t\left(., \boldsymbol{\vartheta}_p\right); .\right)$. Since it is clear that increases in p lead to lower loss this cannot be undertaken over the whole in-sample estimation period. We suggest the use of cross-validation (CV) to determine p. In our forecasting context a natural variant of CV, which mitigates the risk of over-fitting in-sample, is to choose p, over the range $1, ..., p^{\max}$, to minimise the average loss associated with the series of recursively computed generalised combination density forecasts over an out-of-sample period $t_0, ..., T$

$$\widehat{p} = \arg \min_{1 \leq p \leq p^{\max}} \sum_{t=t_0}^{T} l\left(p_t\left(y_{t+1}, \hat{\boldsymbol{\vartheta}}_{t,p}\right); y_t\right), \tag{23}$$

where $\hat{\boldsymbol{\vartheta}}_{t,p}$ denotes the (recursively computed) estimate of $\boldsymbol{\vartheta}_p$ for a given value of p for the generalised density forecast, made at time t; and the loss function for this generalised density forecast is evaluated at the value for y that subsequently materialises, here assumed without loss of generality to be at time $t+1$. Although we are not aware of any formal theoretical results for such an estimator $\widehat{p}$, it is well known that CV has desirable properties in a number of contexts (see, e.g., Arlot & Celisse (2010) for a review).

Furthermore, in general, there may not be a unique set of parameters in the generalised combination which minimise the loss. We may then wish to modify the loss function in a variety of ways. We examine a couple of possibilities here. First, we can require that w_i that are far away from a constant function are penalised. This would lead to a loss function of the form

$$L_T = \sum_{t=1}^{T} l\left(p_t\left(y_t\right); y_t\right) + T^{\gamma} \sum_{i=1}^{N} \int_C \left|w_i\left(y\right) - f_C(y)\right|^{\delta} dy, \;\; \delta > 0, \, 0 < \gamma < 1, \tag{24}$$

where we impose the restriction $\int \left|w_i\left(y\right) - f_C(y)\right|^{\delta} dy < \infty$ and $f_C(y)$ is the uniform density over C. An alternative way to guarantee uniqueness for the solution of (6) is to impose restrictions

that apply to the functions belonging in Φ_{q_i}. Such a loss function could take the form

$$L_T = \sum_{t=1}^{T} l\left(p_t\left(y_t\right); y_t\right) + T^{\gamma} \sum_{i=1}^{N} \sum_{s=1}^{\infty} \left|\nu_s\right|^{\delta}, \;\; \delta > 0,\, 0 < \gamma < 1, \tag{25}$$

where we assume that $\sum_{s=1}^{\infty} \left|\nu_s\right|^{\delta} < \infty$. This sort of modification relates to the penalisations involved in penalised likelihood type methods. Given this, it is worth noting that the above general method for combining densities can easily cross the bridge between density forecast combination and outright density estimation. For example, simply setting $N = 1$ and $q_{1t}\left(y\right)$ to the uniform density and using some penalised loss function such as (24) or (25), reduces the density combination method to density estimation akin to penalised maximum likelihood.

2.1 Piecewise linear weight functions

In what follows we suggest, in particular, the use of indicator functions for $\eta_s\left(y\right)$, i.e. $\eta_s = I\left(r_{s-1} \leq y < r_s\right)$, $s = 1, ..., p$, which defines p $(p \geq 2)$ intervals or regions of the density. The $(p-1)$ inner thresholds $r_0 < r_1 < ... < r_p$, given that $r_0 = -\infty$ and $r_p = \infty$, are either known *a priori* or estimated from the data. For example, these thresholds might be assumed known on economic grounds. In a macroeconomic application, say, some models might be deemed to forecast better in recessionary than expansionary times (suggesting a threshold boundary of zero), or when inflation is inside its target range (suggesting threshold boundaries determined by the inflation target). Otherwise, the thresholds must be estimated, as Section 3 below considers further.

With piecewise linear weight functions, the generalised combined density forecast is[2]

$$p_t\left(y\right) = \sum_{i=1}^{N} \sum_{s=1}^{p_T} \tilde{v}_t(v_{is}) q_{it}\left(y\right) I\left(r_{s-1} \leq y < r_s\right), \tag{27}$$

[2]We note that this combination scheme can be equivalently parameterised with rather than without an intercept. Both parameterisations are the same; e.g., we can rewrite as

$$p_t\left(y\right) = \sum_{i=1}^{N} \left(\tilde{v}_t'(v_{i1}) + \sum_{s=2}^{p_T} \left(\tilde{v}_t'(v_{is}) - \tilde{v}_t'(v_{i1})\right) q_{it}\left(y\right) I\left(r_{s-1} \leq y < r_s\right) \right), \tag{26}$$

where

$$\tilde{v}_t'(v_{i1}) = \frac{v_{i1}}{\sum_{i=1}^{N} \left(v_{i1} + \sum_{s=2}^{p_T} \left(v_{is} - v_{i1}\right) \kappa_{its}\right)}$$

and

$$\tilde{v}_t'(v_{is}) = \frac{v_{is} - v_{i1}}{\sum_{i=1}^{N} \left(v_{i1} + \sum_{s=2}^{p_T} \left(v_{is} - v_{i1}\right) \kappa_{its}\right)}, \;\; s = 2, .., p_T.$$

where ν_{is} are constants (to be estimated) and

$$\kappa_{is} = \int_Y I\left(r_{s-1} \leq y < r_s\right) q_{it}\left(y\right) dy = \int_{r_{s-1}}^{r_s} q_{it}\left(y\right) dy. \tag{28}$$

Note that in (27), as anticipated before Theorem 1 above, we revert to the more general space by allowing $p = p_T$ to depend on T and potentially tend to infinity. Then in Theorem 2 below we present a convergence result for our estimators of ν_{is}.

The piecewise linear weights allow for considerable flexibility, and have the advantage that they let the combination weights vary by region of the density; e.g. as intimated it may be helpful to weight some models more highly when y is negative, say there is a bear market, than when y is high and there is a bull market. This flexibility increases as p_T, the number of regions, tends to infinity.

We now consider estimation of the threshold boundary parameters and inference about the weights when the threshold boundary parameters are estimated.

2.1.1 Estimation of the thresholds

The thresholds, r_s $(s = 1, ..., p)$, need to be estimated if they are not assumed known *a priori*.

Similarly to threshold time-series models, we suggest constructing a grid of possible values for these parameters, optimising the objective (loss) function for every value in the grid and then choosing the value in the grid that yields the overall optimum. The design of the grid will naturally depend on the likely values for y; so some knowledge of these is required and assumed. Quantiles of (historical) y values might be used; or the anticipated range of y could be divided into equally spaced intervals. Increasing the number of points in the grid carries a computational cost, although when estimating a single generalised pool we found this cost not to be prohibitive but also not to affect empirical results, for example, materially.

Of course, inference about these threshold boundary estimates is likely to be non-standard, as with threshold models. For example, it is well known that for threshold autoregressive (TAR) models the estimator of these parameters is super-consistent and has a non-standard asymptotic distribution (see Chan (1992)). A way forward has been proposed by Gonzalo & Wolf (2005) who use subsampling to carry out inference for the boundary parameters of a TAR model. Kapetanios et al. (2013) have extended the use of subsampling for threshold models to more complex panel data settings. Subsampling enables the determination of the rate at which boundary parameters converge to their probability limits thus removing another problem with the associated inference. In the appendix we show that subsampling can provide asymptotically valid inference for the estimated threshold parameters.

2.1.2 Inference about the weights

A second issue relates to the ability to carry out inference on the weights when the boundary parameters, the r's, are estimated. Again, it is expected that if estimators of the boundary parameters, r_s, are superconsistent, as is the case for threshold models, then inference about the remaining parameters does not depend on whether one knows or estimates these boundary parameters. However, this result does not extend to tests of the null hypothesis in (22). In this case it is well known that, under the null hypothesis, the boundary parameters are unidentified. This problem is widely discussed in the literature. A review can be found in Hansen (1999) where simulation based solutions to the problem are suggested (p. 564). These solutions are expected to be applicable to the current setting, although we defer further analysis to future research. An alternative solution is to use subsampling as discussed above.

2.2 Scoring rules

Gneiting & Raftery (2007) discuss a general class of proper scoring rules to evaluate density forecast accuracy, whereby a numerical score is assigned based on the predictive density at time t and the value of y that subsequently materialises, here assumed without loss of generality to be at time $t+1$. A common choice for the loss function L_T, within the 'proper' class (cf. Gneiting & Raftery (2007)), is the logarithmic scoring rule. More specific loss functions that might be appropriate in some economic applications can readily be used instead (e.g. see Gneiting & Raftery (2007)) and again minimised *via* our combination scheme. But an attraction of the logarithmic scoring rule is that, absent knowledge of the loss function of the user of the forecast, by maximising the logarithmic score one is simultaneously minimising the Kulback-Leibler Information Criterion relative to the true but unknown density; and when this distance measure is zero, we know from Diebold et al. (1998) that all loss functions are, in fact, being minimised.

Using the logarithmic scoring rule, with piecewise linear weights, the loss function L_T is given by

$$L_T = \sum_{t=1}^{T} -\log p_t\left(y_{t+1}\right) = \sum_{t=1}^{T} -\log\left(\sum_{i=1}^{N}\sum_{s=1}^{p_T} \tilde{v}_t(v_{is}) q_{it}\left(y_{t+1}\right) I\left(r_{s-1} \leq y_{t+1} < r_s\right)\right), \tag{29}$$

where the restriction

$$\sum_{i=1}^{N}\sum_{s=1}^{p_T} \tilde{v}_t(v_{is})\kappa_{is} = 1, \tag{30}$$

is satisfied automatically for any value of v_{is}. As a result our normalisation in (10) and (11) removes the need for constrained optimisation *via* Lagrangeans.

In practice, in a time-series application, without knowledge of the full sample ($t = 1, ..., T$)

this minimisation would be undertaken recursively at each period t based on information through $(t-1)$. In a related context, albeit for evaluation, Diks et al. (2011) discuss the weighted logarithmic scoring rule, $w_t(y_{t+1}) \log q_{it}(y_{t+1})$, where the weight function $w_t(y_{t+1})$ emphasises regions of the density of interest; one possibility, as in (29), is that $w_t(y_{t+1}) = I(r_{s-1} \leq y_{t+1} < r_s)$. But, as Diks et al. (2011) show and we discuss below, the weighted logarithmic score rule is 'improper' and can systematically favour misspecified densities even when the candidate densities include the true density.

We can then prove the following rate of convergence result, as $p_T \to \infty$, for the estimators of the sub-vector of parameters $\boldsymbol{\nu}_i$, $i = 1, ..., N$, with the remaining parameters $\{r_s\}_{s=1}^{p_T}$ assumed known, using Theorem 1 of Chen & Shen (1998).

Theorem 2 *Let Assumptions 1, 3 and 4 hold. Let $p_T = T^{1/2}$, and*

$$L_T = \sum_{t=1}^{T} -\log p_t(y_{t+1}) \tag{31}$$

and

$$\eta_s = I(r_{s-1} \leq y < r_s) \tag{32}$$

where $\{r_s\}_{s=1}^{p_T}$ are known constants. Then,

$$\left\| \boldsymbol{\nu}_i^0 - \hat{\boldsymbol{\nu}}_i \right\| = o_p\left(T^{-\varphi}\right) \tag{33}$$

for all $\varphi < 1/2$.

Note that since we restrict our analysis to specific basis and loss functions we only need a subset of our assumptions given in the previous Section. Unfortunately, the rate of convergence in the Theorem is not fast enough to satisfy the condition associated with (4.2) in Theorem 2 of Chen & Shen (1998) and, therefore, one cannot prove asymptotic normality for the parameter estimates when $p_T \to \infty$.

But we also have the following Corollary of Theorem 1 for the leading case of using the logarithmic score as a loss function, piecewise linear sieves and component densities from the exponential family, when p is finite.

Corollary 3 *Let Assumptions 1 and 3 hold. Let $q_i(y)$ be bounded functions such that $q_i(y) \sim \exp\left(-y^2\right)$ as $y \to \pm\infty$ for all i. Let*

$$L_T = \sum_{t=1}^{T} -\log p_t(y_{t+1}) \tag{34}$$

and

$$\eta_s = I(r_{s-1} \leq y < r_s) \tag{35}$$

where $\{r_s\}_{s=1}^{p}$ *are known constants and* p *is finite. Then, the asymptotic normality result of Theorem 1 holds.*

We thereby establish consistency and asymptotic normality for the estimated sub-vector of parameters $\widehat{\boldsymbol{\nu}}_i$, $i = 1, ..., N$. When the thresholds are unknown but estimated, as discussed above, it is reasonable to re-consider the analogy with threshold models. Threshold parameter estimates are superconsistent and their estimation does not affect the asymptotic properties of the remaining model parameters. Therefore, in our case it is reasonable to expect that the conclusions of Corollary 3 hold when the threshold parameters are estimated; but we defer detailed analysis to future research.

2.2.1 Extensions: interpreting the weights

The weights ν_{is} cannot easily be interpreted. They do not convey the superiority of fit for density i for region s. This inability to interpret the weights arises due to the fact that, *via* (30), restrictions are placed on the weights across regions.

In order to facilitate interpretation of the weights, which might be helpful in some applications, we draw on Amisano & Giacomini (2007) and Diks et al. (2011) who consider weighted scoring rules and suggest the following restricted variant of our method.[3]

Define the sequence of weighted logarithmic score loss functions

$$L_{s,T} = \sum_{t=1}^{T} I\left(r_{s-1} \leq y_{t+1} < r_s\right) \log\left(\sum_{i=1}^{N} \nu_{is} q_{it}\left(y_{t+1}\right)\right), \quad s = 1, ..., p_T, \tag{36}$$

where $I\left(r_{s-1} \leq y_{t+1} < r_s\right) = \eta_s$ emphasises the region(s) of interest.

We could then minimise each $L_{s,T}$, $s = 1, ..., p_T$ with respect to ν_{is}, and thereby maximise the logarithmic score over each region s by assigning a higher weight to *better* individual densities i. This would provide a clearer link between our estimated weights and the best performing density in a given region than the unrestricted approach we have been discussing thus far. Then, the proper combined density, $p_t^w\left(y_{t+1}\right)$, can be defined, *via* normalisation, as

$$p_t^w\left(y_{t+1}\right) = \frac{\sum_{i=1}^{N}\sum_{s=1}^{p_T} \hat{\nu}_{is} q_{it}\left(y_{t+1}\right) I\left(r_{s-1} \leq y_{t+1} < r_s\right)}{\sum_{i=1}^{N}\sum_{s=1}^{p_T} \hat{\nu}_{is}\kappa_{is}}, \tag{37}$$

where $\hat{\nu}_s = (\hat{\nu}_{1s}, ..., \hat{\nu}_{p_T})'$ is the minimiser of $L_{s,T}$. The sum of these weighted logarithmic scores, $L_T^w = \sum_{s=1}^{p_T} L_{s,T}$, is such that $L_T^w > L_T$ is a likely outcome although not guaranteed as we use different normalisations for the weights in the two cases.

[3]While we propose - in theory - use of these restricted weights, in practice we defer detailed analysis to future work. However, as a start we did augment the set of Monte Carlo experiments reported below to consider use of these restricted weights and found, without exception but as expected, use of the restricted weights to involve a considerable loss in accuracy as measured by the average logarithmic score. Whether this loss is offset by the additional interpretation benefits we again defer to future discussion.

As discussed by Diks et al. (2011), the weighted logarithmic scoring rule, $w_t(y_{t+1}) \log q_{it}(y_{t+1})$, used in (36) is not proper (see also Gneiting & Ranjan (2011)); i.e., there can exist incorrect density forecasts that would receive a higher average score than the actual (true) conditional density. Therefore, following Diks et al. (2011), one might modify (36) and consider use of the conditional likelihood score function, given by

$$\tilde{L}_{s,T} = \sum_{t=1}^{T} I(r_{s-1} \leq y_{t+1} < r_s) \log \left(\frac{\sum_{i=1}^{N} \nu_{is} q_{it}(y_{t+1})}{\sum_{i=1}^{N} v_{is} \kappa_{is}} \right), \quad s = 1, ..., p_T, \tag{38}$$

where the division by $\sum_{i=1}^{N} v_{is}\kappa_{is}$ normalises the density $\sum_{i=1}^{N} \nu_{is} q_{it}(y_{t+1})$ on the region s of interest. Another possibility is to use the censored likelihood of Diks et al. (2011) rather than the conditional likelihood to define region-specific loss functions.

3 Monte Carlo study

We undertake four sets of experiments to investigate the performance of the generalised pool relative to the standard (optimised) linear pool. These four experiments differ according to the assumed true (but in practice, if not reality, unknown) density and the nature of the (misspecified) component densities which are subsequently combined. Thereby we seek to provide some robustness to our results; and some empirical relevance by considering some widely used nonlinear and stochastic volatility models.

In the first Data Generating Process (DGP1) a non-Gaussian true density is considered, and we compare the ability of linear and generalised combinations of misspecified Gaussian densities to capture the non-Gaussianity. In DGP2 we instead assume the true density is Gaussian, but again consider combinations of two misspecified Gaussian densities. In this case we know that linear combinations will, in general, incorrectly yield non-Gaussian densities, given that they generate mixture distributions. It is therefore important to establish if and how the generalised combinations improve upon this. In DGP3 we consider a more realistic scenario in economics where the component conditional densities change over time. Specifically, we consider a Threshold Auto-Regressive (TAR) nonlinear model, and assess the ability of combinations of linear Gaussian models with different autoregressive parameters to approximate the nonlinear process. For robustness, we consider a variety of parameter settings to explore whether the performance of the method is sensitive to characteristics like the persistence of the data. DGP4 assumes the true density evolves over time according to an unobserved components trend-cycle model with stochastic volatility. This model has been found to mimic successfully the changing behaviour of US inflation and its transition from high and volatile values (in the so-called Great Inflation period) to lower and more stable inflation (in the so-called Great Moderation period); see Stock & Watson (2007). We then investigate the density forecasting ability of combinations of two

widely used models without stochastic volatility, which are known to fit the Great Inflation and Great Moderation sub-samples respectively.

The performance of the generalised pool relative to the linear pool is assessed by tests for equal predictive accuracy on the basis of the logarithmic scoring rule, (29). A general test for equal performance is provided by Giacomini & White (2006); a Wald-type test statistic is given as

$$T\left(T^{-1}\sum_{t=1}^{T}\Delta L_t\right)'\Sigma^{-1}\left(T^{-1}\sum_{t=1}^{T}\Delta L_t\right), \tag{39}$$

where ΔL_t is the difference in the logarithmic scores of the generalised and linear pools at time t and equals their Kullback Leibler Information Criterion or distance measure; and Σ is an appropriate autocorrelation robust, estimate of the asymptotic covariance matrix. Under the null hypothesis of equal accuracy $E(\Delta L_t) = 0$, Giacomini & White (2006) show that the test statistic tends to χ^2_1 as $T \to \infty$. We undertake two-sided tests of the null of equal accuracy at a nominal size of 10% and in the Tables below report the proportion of rejections in favour of both the generalised and linear pools.

For DGP1 as the true density in fact characterises a specific instance of what we call a generalised combination we report some additional results. To gauge absolute density forecasting performance we compute the Integrated Mean Squared Error (IMSE) of the density forecasts relative to the true density. We also examine the size and power properties of the test of the null hypothesis in (27); and consider the properties of the CV estimator for p.

Throughout we implement the generalised pool using piecewise linear weight functions, as in (27); and we focus on its use in the realistic situation that p and the r's have to be estimated. We consider samples sizes, T, of 100, 200, 400, and 1000 observations and carry out 1000 Monte Carlo replications in each experiment. For a simulated T-sample, taking the component densities as fixed, we estimate the parameters in the generalised and linear pools over the first $T/2$ observations. When using CV, as discussed above (see (23)) based on a series of recursively computed 1-step ahead generalised combination density forecasts over an out-of-sample period, this involves using observations $T/4+1, ..., T/2$ to select p and estimate the r's. Then, keeping these parameters fixed, we generate the generalised and linear pools for the last $T/2$ simulated observations and evaluate them either relative to the simulated outturn (to calculate the average logarithmic score and conduct the Giacomini & White (2006) tests) or the true density (to calculate the IMSE).

Below we provide details of and results for each of the four DGPs in turn.

3.1 DGP1

The true density for the random variable Y has a two part normal density given by

$$f(Y) = \begin{cases} -A\exp(y-\mu)^2/(2\sigma_1^2) \text{ if } y < \mu \\ -A\exp(y-\mu)^2/(2\sigma_2^2) \text{ if } y \geq \mu \end{cases} \tag{40}$$
$$A = \left(\sqrt{2\pi}(\sigma_1+\sigma_2)/2\right)^{-1}.$$

We assume that the practitioner combines two Gaussian densities with the same mean μ and variances given by σ_1^2 and σ_2^2, respectively. Combination is *via* (a) the generalised pool and (b) the optimised (with respect to the logarithmic score) linear pool with fixed weights, as in Hall & Mitchell (2007) and Geweke & Amisano (2011). We set $\mu = 0$, $\sigma_1^2 = 1$ and consider various values for $\sigma_2^2 = 1.5, 2, 4$ and 8.

Given that (40) characterises a generalised combination of $N = 2$ Gaussian component densities with piecewise linear weights, where $p = 2$ and $r_1 = 0$, we use this experiment to draw out three facts. Firstly, to assess, in the unrealistic situation that we know both $p = 2$ and $r_1 = 0$, the accuracy of the generalised combination as a function of T. Secondly, to quantify the costs associated with having to estimate r_1 but still assuming $p = 2$. Thirdly, to quantify the costs of having both to estimate p and the r's. In practice, both p and the r's are typically unknown and so this third case is of particular relevance for applied work.

Implementation of the generalised combination when the r's are to be estimated, as discussed above, requires the practitioner to select the estimation grid. With the mode of the two part normal set at $\mu = 0$, we experimented with grids for r in the range -1 to 1 with an interval of 0.1.[4] When estimating p we consider values from $p = 2, ..., 4$ which is a reasonable spread of values for this parameter trading off the bias inherent in small p with the variance inherent in larger values of this parameter.

Table 1, in the columns labelled G/L (L/G), presents the rejection proportions (across the simulations) in favour of the generalised (linear) pool using the Giacomini & White (2006) test for equal density forecast performance, (39).

It is clear from Table 1 that, irrespective of whether p and r_s are assumed known or estimated, the generalised combination is preferred to the linear combination with rejection proportions clearly in its favour. These proportions approach 1 as σ_2^2 and T increase. Even for the smallest values of σ_2^2, which permit less skew in the true density, the generalised combination is still preferred with rejection proportions above 0.9 for the larger sample sizes, T. In turn, across σ_2^2 and T the linear combination is never preferred over the generalised combination with rejection proportions below 0.03 and decreasing to 0 as σ_2^2 and T increase. Estimation of both p and r_s (the "Unknown p" column in Table 1) does, in general, involve a loss of relative performance

[4]Inference was not found to be particularly sensitive either to widening these outer limits or to finer intervals.

Table 1: DGP1: Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

σ_2^2	T	$r_1 = 0$		r_1 estimated ($p = 2$)		Unknown p	
		G/L	L/G	G/L	L/G	G/L	L/G
1.5	100	0.302	0.012	0.188	0.008	0.138	0.022
	200	0.514	0.000	0.380	0.000	0.254	0.008
	400	0.718	0.000	0.616	0.000	0.536	0.000
	1000	0.968	0.000	0.966	0.000	0.936	0.000
2	100	0.620	0.000	0.508	0.002	0.342	0.010
	200	0.872	0.000	0.752	0.000	0.688	0.004
	400	0.976	0.000	0.968	0.000	0.964	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000
4	100	0.966	0.000	0.920	0.004	0.748	0.012
	200	1.000	0.000	0.996	0.002	0.932	0.004
	400	1.000	0.000	1.000	0.000	0.996	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000
8	100	0.988	0.006	0.874	0.002	0.690	0.006
	200	1.000	0.000	0.986	0.000	0.872	0.004
	400	1.000	0.000	1.000	0.000	0.972	0.000
	1000	1.000	0.000	1.000	0.000	1.000	0.000

for the generalised pool. But despite this loss the generalised pool still offers clear advantages relative to the linear pool. Moreover this loss again deteriorates both with increases in T and increases in the skewness of the underlying DGP, σ_2^2.

To gauge absolute performance IMSE estimates are presented in Table 2 for the generalised and linear pools as well as the two component Gaussian density forecasts. Table 2 shows that the linear combination always delivers more accurate densities than either component density. And when the threshold, r_1, is assumed known and set to 0 the generalised combination scheme dominates the linear scheme for all sample sizes and values of σ_2^2 - with lower IMSE estimates as expected. The gains in accuracy are clear, and increase with T and σ_2^2. Continuing to assume $p = 2$ but now estimating r_1 as expected we find the accuracy of the generalised combination scheme to deteriorate, with the IMSE estimates at least tripling and often quadrupling. For small sample sizes and low values of σ_2^2 this loss in accuracy is large enough for the linear scheme to be preferred.[5] But for larger T and larger variances, σ_2^2, the generalised pool is again the better performing pool; and by a considerable margin. Reassuringly, when both p and the r's are estimated although accuracy is again lost, the incremental losses associated with estimation

[5]But this superiority on the basis of IMSE does not translate into improved rejection proportions in Table 1. This is explained by the fact that IMSE and the logarithmic score are different measures of density forecast "fit". However, comparison of Tables 1 and 2 indicates that in general these different measures point in the same direction.

Table 2: DGP1: IMSE estimates for the Generalised and Linear Combinations

σ_2^2	T	Generalised			Linear	Component Densities	
		$r_1 = 0$	r_1 estimated ($p = 2$)	Unknown p		Component 1	Component 2
1.5	100	0.343	2.293	1.910	0.773	1.654	1.103
	200	0.167	0.883	0.926	0.715	1.654	1.103
	400	0.084	0.389	0.456	0.691	1.654	1.103
	1000	0.032	0.149	0.197	0.673	1.654	1.103
2	100	0.297	1.520	1.562	1.592	4.421	2.211
	200	0.139	0.679	0.772	1.531	4.421	2.211
	400	0.072	0.293	0.445	1.505	4.421	2.211
	1000	0.030	0.115	0.203	1.486	4.421	2.211
4	100	0.189	0.619	0.872	2.649	12.728	3.182
	200	0.090	0.310	0.414	2.595	12.728	3.182
	400	0.045	0.139	0.218	2.569	12.728	3.182
	1000	0.019	0.062	0.103	2.556	12.728	3.182
8	100	0.101	0.291	0.373	2.221	19.413	2.427
	200	0.053	0.139	0.196	2.190	19.413	2.427
	400	0.026	0.073	0.100	2.174	19.413	2.427
	1000	0.011	0.033	0.048	2.164	19.413	2.427

Table 3: Rejection probabilities for linearity test under the null hypothesis of linearity

T/σ_2^2	1.1	1.25	1.5	2
100	0.021	0.021	0.014	0.014
200	0.017	0.014	0.016	0.012
400	0.042	0.019	0.014	0.023
1000	0.070	0.042	0.021	0.019

are confined to an order of $20-30\%$. Again larger values for T and σ_2^2 help the generalised pool, with the IMSE estimates approaching 0 as T and σ_2^2 increase.

Finally, to study the properties of the linearity test, (22), Tables 3 and 4 report the test's rejection probabilities at the nominal 5% level under both the null of linearity and the alternative. We focus on a narrower range of values for σ_2^2 reflecting the finding that results were unchanged for $\sigma_2^2 > 2$; and deferring to future work analysis and implementation of the test in the unknown parameters case when we know there are unidentified parameters (as discussed in Section 2.1.2 above) we estimate the generalised pools correctly assuming $p = 2$ and $r_1 = 0$.

Table 3 shows that the test is, in general, slightly under-sized but Table 4 indicates that power increases strongly with both T and σ_2^2. Even for relatively modest T (e.g., $T = 100$) the test has power above 0.6 even when σ_2^2 is only 1.5. Note that we use much smaller values of

Table 4: Rejection probabilities for linearity test under the alternative

T/σ_2^2	1.1	1.25	1.5	2
100	0.024	0.134	0.611	0.998
200	0.028	0.267	0.915	1.000
400	0.081	0.552	0.997	1.000
1000	0.420	0.950	1.000	1.000

σ_2^2 than in the previous Monte Carlo experiment. If we had used those values the rejections probabilities would have been invariably equal to one. This simply illustrates that the test is very powerful and is able to reject the linear combination in favour of the generalised one even for small deviations from the case when the linear combination is optimal.

Figure 1 then investigates the properties of the CV estimator for p, by plotting across T and σ_2^2 a histogram indicating the number of times (out of the 1000 replications) a given value for p was selected. We note that we continue to confine the CV search to values of p in the range 2 to 4. Figure 1 shows that encouragingly $p = 2$ is the modal estimate, although higher values are often selected. This is a reasonable outcome since CV provides some protection against the risk of over-fitting, as without it, the maximum value of p would invariably be selected. Further Tables 1 and 2 show that despite this estimation uncertainty over p the generalised pool remains preferable to the linear pool.

3.2 DGP2

In contrast to DGP1 the true density is assumed to be the standard normal. We then entertain two Gaussian component densities, each of which is misspecified as the mean is incorrect. We fix the variances of the component densities at unity but consider different values for their means, μ_1 and μ_2. We then take a combination of these two component normal densities *via* the generalised and linear pools.[6] In contrast to DGP1 there is neither a *true* value for p nor the r's and we therefore proceed to examine the generalised combination when both p and the r's are estimated. As in DGP1, we use an identical grid search design to estimate the thresholds, r_s; and when using CV to select $\widehat{p}$ consider values from $p = 2, ..., 10$. We consider higher values for p than in DGP1 given that there is now no finite p value that would give the same loss function, for any of the combinations we consider, as the true density.

Table 5 presents the rejection proportions in favour of the generalised and linear pools, using the Giacomini-White test for equal density forecast performance. Results are presented for values of $(\mu_1, \mu_2) = (-0.25, 0.25), (-0.5, 0.5), (-1, 1)$ and $(-2, 2)$. In general, both as μ_1 and μ_2

[6] Another possibility would have been to consider misspecified variances for the Gaussian components rather than misspecified means. We feel that mean misspecification is usually a more influential misspecification (especially in explaining forecast failure; cf. Clements & Hendry (1999)) and, in any case, we consider volatility misspecification in DGP4.

Table 5: DGP2: Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

(μ_1, μ_2)	T	Unknown p	
		G/L	L/G
$(-0.25, 0.25)$	100	0.016	0.184
	200	0.010	0.214
	400	0.008	0.180
	1000	0.010	0.158
$(-0.5, 0.5)$	100	0.008	0.114
	200	0.028	0.112
	400	0.016	0.062
	1000	0.108	0.016
$(-1, 1)$	100	0.130	0.020
	200	0.372	0.016
	400	0.650	0.014
	1000	0.954	0.000
$(-2, 2)$	100	0.748	0.000
	200	0.900	0.000
	400	0.986	0.000
	1000	1.000	0.000

increase in absolute value (such that the component densities become less similar) and as T rises, we see increasing gains to the use of the generalised pool rather than the linear pool.[7] However, when the component densities are more similar, for values of $(\mu_1, \mu_2) = (-0.25, 0.25)$, the linear pool does delivers modest gain. While the generalised pool nests the linear pool, especially for smaller samples T, it requires extra parameters to be estimated and Table 5 reveals that this pays off only for larger T and when the component densities become more distinct because μ_1 and μ_2 increase in absolute value. We also note (detailed results available upon request) that the modal CV estimate of $\widehat{p}$ is around 4. While there is not a monotonic relationship - and the pattern varies across both values of μ_1, μ_2 and T - increases in p lead to a deterioration in the performance of the generalised combination beyond a changing threshold. This hints at a nonlinear trade-off between the complexity or flexibility of the generalised combination and estimation error.

[7]Both pools (results not reported) confer statistically significant advantages (with Giacomini-White rejection rates close to unity) relative to use of either component density alone.

3.3 DGP3

DGP3 generalises the two part normal density used in DGP1 to the more realistic time-series case by assuming the true model is the TAR model

$$y_t = \begin{cases} \rho_1 y_{t-1} + \sigma_1 \epsilon_t, \text{ if } y_{t-1} < q \\ \rho_2 y_{t-1} + \sigma_2 \epsilon_t, \text{ if } y_{t-1} \geq q \end{cases}, \tag{41}$$

where ϵ_t is assumed to be standard normal, and ρ_1 and ρ_2 control the degree of persistence. We consider values $(\rho_1, \rho_2) = \{(0.1, 0.7), (0.1, 0.9), (0.3, 0.7), (0.3, 0.9), (0.5, 0.7), (0.5, 0.9)\}$, where $(\sigma_1^2, \sigma_2^2) = (1, 4)$ and $q = 0$. The two component density forecasts are $N(\rho_1 y_{t-1}, \sigma_1^2)$ and $N(\rho_2 y_{t-2}, \sigma_2^2)$.

Following the same implementation of the generalised pool as in DGP2, Table 6 reports the rejection proportions. These demonstrate that the generalised pool is preferred to the linear pool, except on two occasions when T is only 100 and when ρ_1 and ρ_2 are relatively close together. The superiority of the generalised pool increases with T, increases with the size of $\rho_2 - \rho_1$ and, for a given sized difference $\rho_2 - \rho_1$, increases in the values of ρ_1 and ρ_2.

3.4 DGP4

DGP4 is the Unobserved Components (UC) model with Stochastic Volatility proposed by Stock & Watson (2007) to model US inflation. This model allows the variances of both the permanent and transitory component of inflation to evolve randomly over time. The UC-SV model is

$$\begin{aligned} \pi_t &= \tau_t + \eta_t, \text{ where } \eta_t = \sigma_{\eta,t}\zeta_{\eta,t} \\ \tau_t &= \tau_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t = \sigma_{\varepsilon,t}\zeta_{\varepsilon,t} \\ \ln \sigma_{\eta,t}^2 &= \ln \sigma_{\eta,t-1}^2 + \upsilon_{\eta,t} \\ \ln \sigma_{\varepsilon,t}^2 &= \ln \sigma_{\varepsilon,t-1}^2 + \upsilon_{\varepsilon,t} \end{aligned}$$

where $\zeta_t = (\zeta_{\eta,t}, \zeta_{\varepsilon,t})$ is i.i.d. $N(0, I_2)$, $\upsilon_t = (\upsilon_{\eta,t}, \upsilon_{\varepsilon,t})$ is i.i.d. $N(0, \gamma I_2)$, ζ_t and υ_t are independently distributed and γ is a scalar parameter set equal to 0.01.[8]

The two component density forecasts are UC models but without stochastic volatility. The first, in fact found by Stock & Watson (2007) to offer a good fit for high inflation values, sets $\sigma_\eta = 0.66$ and $\sigma_\varepsilon = 0.91$. The second component model was found to offer a good fit for the lower and more stable inflation values during the Great Moderation and sets $\sigma_\eta = 0.61$ and $\sigma_\varepsilon = 0.26$. So by combining these two models we are seeing whether it helps to let the combination weights

[8]Stock & Watson (2007) found a value of $\gamma = 0.2$ best fit US inflation. We experimented with a range of γ values; and as in Table 7 below found the generalised pool tended to be preferred over the linear pool. But higher γ values did induce explosive behaviour in many Monte Carlo replications explaining why Stock & Watson (2007) focus on forecasting a first difference of inflation.

Table 6: DGP3: Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

(ρ_1, ρ_2)	T	Unknown p	
		G/L	L/G
(0.1,0.7)	100	0.121	0.064
	200	0.236	0.007
	400	0.304	0.000
	1000	0.371	0.000
(0.1,0.9)	100	0.326	0.042
	200	0.509	0.002
	400	0.606	0.000
	1000	0.600	0.000
(0.3,0.7)	100	0.065	0.097
	200	0.114	0.017
	400	0.210	0.014
	1000	0.223	0.005
(0.3,0.9)	100	0.279	0.079
	200	0.435	0.007
	400	0.494	0.001
	1000	0.482	0.001
(0.5,0.7)	100	0.026	0.122
	200	0.045	0.039
	400	0.071	0.024
	1000	0.100	0.010
(0.5,0.9)	100	0.207	0.092
	200	0.303	0.011
	400	0.345	0.003
	1000	0.364	0.003

Table 7: DGP4: Rejection probabilities in favour of the Generalised (G) and Linear (L) pools using the Giacomini-White test for equal density forecast performance

T	Unknown p	
	G/L	L/G
100	0.124	0.168
200	0.270	0.130
400	0.354	0.126
1000	0.476	0.180

vary according to the variable of interest.

Table 7 again indicates that the relative performance of the generalised pool is dependent on the sample size, T. For $T = 100$ the linear pool is preferred more frequently than the generalised pool. However, as T increases the generalised pool is preferred two to three time more frequently.

4 Empirical Application

We consider S&P 500 daily percent logarithmic returns data from 3 January 1972 to 9 September 2013, an extension of the dataset used by Geweke & Amisano (2010, 2011) in their analysis of optimal linear pools. Following Geweke & Amisano (2010, 2011) we then estimate a Gaussian GARCH(1,1) model, a Student T-GARCH(1,1) model and a Gaussian exponential GARCH(1,1) *via* maximum likelihood; and the stochastic volatility model of Kim et al. (1998) using an integration sampler. These four models are estimated using rolling samples of 1250 trading days (about five years). One day ahead density forecasts are then produced recursively from each model for the return on 15 December 1976 through to the return on 9 September 2013 giving a total of 9268 observations. The predictive densities are formed by substituting the ML estimates for the unknown parameters.

These component densities are then combined using either a linear or generalised combination scheme in two, three and four model pools. We evaluate the combination schemes in two ways.

Firstly, we fit the generalised and linear combinations *ex post* (so effectively we treat the forecast data as an in-sample dataset). This involves extending the empirical analysis in Geweke & Amisano (2011) who analysed the optimised linear combinations of similar marginal (component) densities, over a shorter sample, and found gains to linear combination relative to use of the component densities alone. In Table 8 we provide the average logarithmic scores of the four component models as well as the scores of two, three and four model pools of these component densities. While we are able to replicate the results of Geweke & Amisano (2011) over their sample period ending in 16 December 2005, Table 8 reveals that inference is in fact sensitive to the sample period. Across the columns in Table 8 we see that over our longer sample period the

optimised linear pool is at best able to match the performance of the best component density; this means that the optimised weight vector in the linear pool often involves a unit element, i.e., one of the component densities receives all the weight.

But Table 8 does indicate clear gains to fitting the generalised pools. Indeed the reported linearity test, (22), rejects linearity with p-values of 0.000 across all the columns in Table 8. The generalised pools involve, in each variant, using CV to estimate p (with values from 2 to 10 considered) with the thresholds estimated *via* a grid search of width 0.5 in the interval -2.5% to 2.5%. This interval was chosen on the basis of our historical judgment over the 1957-1976 period about the likely range of values for daily stock returns. Over this pre-estimation sample the -2.5% to 2.5% interval amounts to more than a 99% confidence interval.

We see from Table 8 that the generalised pools, whether two or three model pools, involving the T-GARCH density yield the highest scores; and importantly a much higher score than use of the T-GARCH density alone. Interestingly, looking at the time-invariant weights (v_{is} in the notation of (10) but normalised to add to unity) on the different component densities, we see that 5 of the 6 generalised pools involving the T-GARCH all yield identical scores of 2.762. Although this does involve weighting the component models in different ways across the regions of the density, Table 8 shows that in these pools the T-GARCH does always receive a high weight, approaching and reaching unity for the central regions of its forecast density. The sixth generalised pool involving the T-GARCH is in fact the four-component density pool; and this pool performs slightly worse than the two and three component pools. The prominence of the T-GARCH component density seems reasonable given the volatility observed over the sample period, which includes the turbulent 2007-8 crisis period; see Figure 2. But despite this prominence Table 8 clearly shows how one can improve upon use of the T-GARCH component density alone by taking a generalised combination. This adjusts, in particular, the T-GARCH forecast density in its tails.[9]

Table 8 also lists the value for p chosen by CV, $\widehat{p}$, on the basis of the first 7000 observations as the estimated values of p are used in the forecasting exercise discussed in the next paragraph. For all the generalised pools except for the four-component pool $\widehat{p} = 8$. But for the more complex four-component pool $\widehat{p} = 2$, reminding us that for more complicated pools there can be benefits to use of a more parsimonious weighting function.

[9]Plots of the generalised densities over time (not reported) reveal that they can indeed capture fat-tails. While their precise shape changes both over time and according to the component models considered the generalised densities often experience modest spikes at the thresholds. As p increases the severity of these spikes decreases, and the generalised densities become smoother. Table 8 reveals that accommodating these spikes does not prejudice performance on the basis of the logarithmic score rule. But under alternative loss functions (e.g., those that penalise lack of smoothness) these spikes may not deliver an improved score for the generalised combination and may therefore be deemed unattractive. For these loss functions it may prove beneficial to fit generalised combinations where p is higher, despite the extra parameters, to ensure smoothness. We defer analysis of the generalised combinations under alternative loss functions to future work; but we do remark in defence of the current analysis that the logarithmic scoring rule, as discussed above, is used widely because of its attractive theoretical properties.

The clear risk in using these piecewise functions is that because of their more flexible forms they fit well in-sample, but provide disappointing performance out of sample because of additional parameter estimation error. Secondly, therefore, we estimate the weights and boundaries in the pools recursively from 3 September 2004 (observation 7000) and form out-of-sample linear and generalised pools over the remaining sample of 2268 observations (i.e., through to 9 September 2013), using the previously obtained estimates of p. To provide an indication of how robust results are to the chosen out-of-sample window we also present results over two sub-samples. These correspond to the period before and after August 2007 when the banking system began to freeze up due to problems in US mortgage debt. These real-time exercises mean we are only using past data for optimisation. This is an important test given the extra parameters involved when estimating the generalised rather than linear combination. As before, we consider a range of generalised pools and compare their average logarithmic score with the optimal linear combination. We also test equal predictive accuracy between the linear and generalised pools using the Giacomini-White test, (39), and report the p-values in Table 9 below.

Table 9 shows that real-time generalised combinations do deliver higher scores than the linear pools over all sub-periods; and these differences are statistically significant. While accuracy is higher over the pre 2007 sample than the post 2007 sample, as we might expect given the heightened uncertainty and volatility in the aftermath of the global financial crisis, the generalised pool remains superior even in this more volatile period. Table 9 also indicates that despite the potential for the four-model generalised pool to offer a more flexible fit it does not work as well as the more parsimonious generalised pools. Over the pre-2007 evaluation period the preferred generalised pool is in fact a pool of just two models.

We also find that unlike the generalised pools the (optimised) linear pools, on an out–of-sample basis, can but often do not beat the four component densities individually. Thus while, as Geweke & Amisano (2011) show on this same dataset, optimal linear combinations will at least match the performance of the best component density when estimated over the full sample ($t = 1, ..., T$), there is no guarantee that the "optimal" linear combination will help out-of-sample when the combination weights are computed recursively. By way of example, over the 3 Sept 2004 - 9 Sept 2013 evaluation period as a whole, the average logarithmic scores of the GARCH, EGARCH, SV and TGARCH forecast densities are -0.885, -0.773, -0.954 and -0.204, respectively. Thus, the T-GARCH is the preferred component density and always delivers a higher average logarithmic score than any linear (but not generalised) pool containing it. In contrast, demonstrating that some linear pools are helpful, Table 9 shows that the linear pool of the GARCH and SV densities is preferred to either component alone. But again the generalised pool of these two densities confers further gains in forecast accuracy; and these gains are statistically significant.

Table 8: In-sample (15 December 1976 to 9 September 2013) average logarithmic scores for the Generalised pool, the Linear pool and the four component densities indicated 1 to 4. $\widehat{p}$ is the CV estimator for p. H_0 p-value refers to the p-value of the test that tests the null hypothesis that the linear pool is the appropriate combination to use. The weights v_{is} from (10) are normalised to sum to one.

	Component Densities (1: GARCH, 2: EGARCH, 3: SV, 4: TGARCH									
	1,2	1,3	1,4	2,3	2,4	3,4	1,2,3	1,2,4	2,3,4	1,2,3,4
G	-0.288	-0.313	2.762	-0.284	2.762	2.762	-0.288	2.762	2.762	2.572
L	-1.169	-1.169	-1.169	-1.187	-1.183	-1.183	-1.169	-1.169	-1.183	-1.169
Component	-1.169	-1.187	-1.580	-1.183	-	-	-	-	-	-
$\widehat{p}$	8	8	8	8	8	8	8	8	8	2
Lin. Test p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GARCH Weight 1	0.500	0.501	0.403				0.334	0.333		0.001
GARCH Weight 2	0.000	0.994	0.000				0.956	0.000		0.000
GARCH Weight 3	0.001	0.439	0.007				0.134	0.000		
GARCH Weight 4	0.936	0.853	0.162				0.034	0.104		
GARCH Weight 5	0.000	1.000	0.000				0.000	0.000		
GARCH Weight 6	0.367	0.460	0.007				0.460	0.008		
GARCH Weight 7	0.458	1.000	0.001				0.857	0.001		
GARCH Weight 8	0.500	0.551	0.500				0.334	0.333		
EGARCH Weight 1	0.500			0.501	0.500		0.334	0.333	0.338	0.193
EGARCH Weight 2	1.000			0.779	0.007		0.040	0.000	0.000	0.000
EGARCH Weight 3	0.999			0.742	0.161		0.000	0.013	0.022	
EGARCH Weight 4	0.064			0.934	0.020		0.000	0.092	0.092	
EGARCH Weight 5	1.000			0.000	0.000		1.000	0.000	0.000	
EGARCH Weight 6	0.633			0.162	0.007		0.105	0.000	0.000	
EGARCH Weight 7	0.542			0.620	0.019		0.123	0.002	0.000	
EGARCH Weight 8	0.500			0.501	0.500		0.334	0.333	0.339	
SV Weight 1		0.499		0.499		0.476	0.333		0.323	0.102
SV Weight 2		0.006		0.221		0.000	0.003		0.000	0.000
SV Weight 3		0.561		0.258		0.002	0.865		0.033	
SV Weight 4		0.147		0.066		0.002	0.966		0.022	
SV Weight 5		0.000		1.000		0.000	0.000		0.000	
SV Weight 6		0.540		0.838		0.000	0.435		0.002	
SV Weight 7		0.000		0.380		0.000	0.020		0.000	
SV Weight 8		0.449		0.499		0.475	0.333		0.323	
TGARCH Weight 1			0.397		0.500	0.524		0.333	0.339	0.704
TGARCH Weight 2			1.000		0.993	1.000		1.000	1.000	1.000
TGARCH Weight 3			0.993		0.839	0.998		0.987	0.945	
TGARCH Weight 4			0.838		0.980	0.998		0.805	0.886	
TGARCH Weight 5			1.000		1.000	1.000		1.000	1.000	
TGARCH Weight 6			0.993		0.993	1.000		0.992	0.998	
TGARCH Weight 7			0.999		0.981	1.000		0.997	1.000	
TGARCH Weight 8			0.500		0.500	0.525		0.333	0.339	

Table 9: Out-of-sample average logarithmic scores of the Generalised and Linear combinations over selected evaluation periods

	Component Densities (1: GARCH, 2: EGARCH, 3: SV, 4: TGARCH									
	1,2	1,3	1,4	2,3	2,4	3,4	1,2,3	1,2,4	2,3,4	1,2,3,4
	3 Sept 2004: 9 Sept 2013									
G	0.742	0.775	0.659	0.803	0.690	0.724	0.808	0.723	0.735	-0.296
L	-0.786	-0.585	-0.893	-0.531	-0.810	-0.610	-0.548	-0.793	-0.558	-0.554
	3 Sept 2004: 31 Aug 2007									
G	0.917	0.864	0.860	0.914	0.861	0.861	0.910	0.862	0.861	0.406
L	-0.137	-0.183	-0.218	-0.117	-0.146	-0.193	-0.131	-0.143	-0.137	-0.135
	4 Sept 2007: 9 Sept 2013									
G	0.652	0.730	0.555	0.746	0.603	0.654	0.757	0.652	0.670	-0.655
L	-1.119	-0.791	-1.238	-0.743	-1.150	-0.823	-0.761	-1.126	-0.773	-0.769

5 Conclusion

With the growing recognition that point forecasts are best seen as the central points of ranges of uncertainty more attention is now paid to density forecasts. Coupled with uncertainty about the best means of producing these density forecasts, and practical experience that combination can render forecasts more accurate, density forecast combinations are being used increasingly in macroeconomics and finance.

This paper extends this existing literature by letting the combination weights follow more general schemes. It introduces generalised density forecast combinations or pools, where the combination weights depend on the variable one is trying to forecast. Specific attention is paid to the use of piecewise linear weight functions that let the weights vary by region of the density. These weighting schemes are examined theoretically, with sieve estimation used to optimise the score of the generalised density combination. The paper then shows both in simulations and in an application to S&P500 returns that the generalised combinations can deliver more accurate forecasts than linear combinations with optimised but fixed weights as in Hall & Mitchell (2007) and Geweke & Amisano (2011). Their use therefore seems to offer the promise of more effective forecasts in the presence of a changing economic climate.

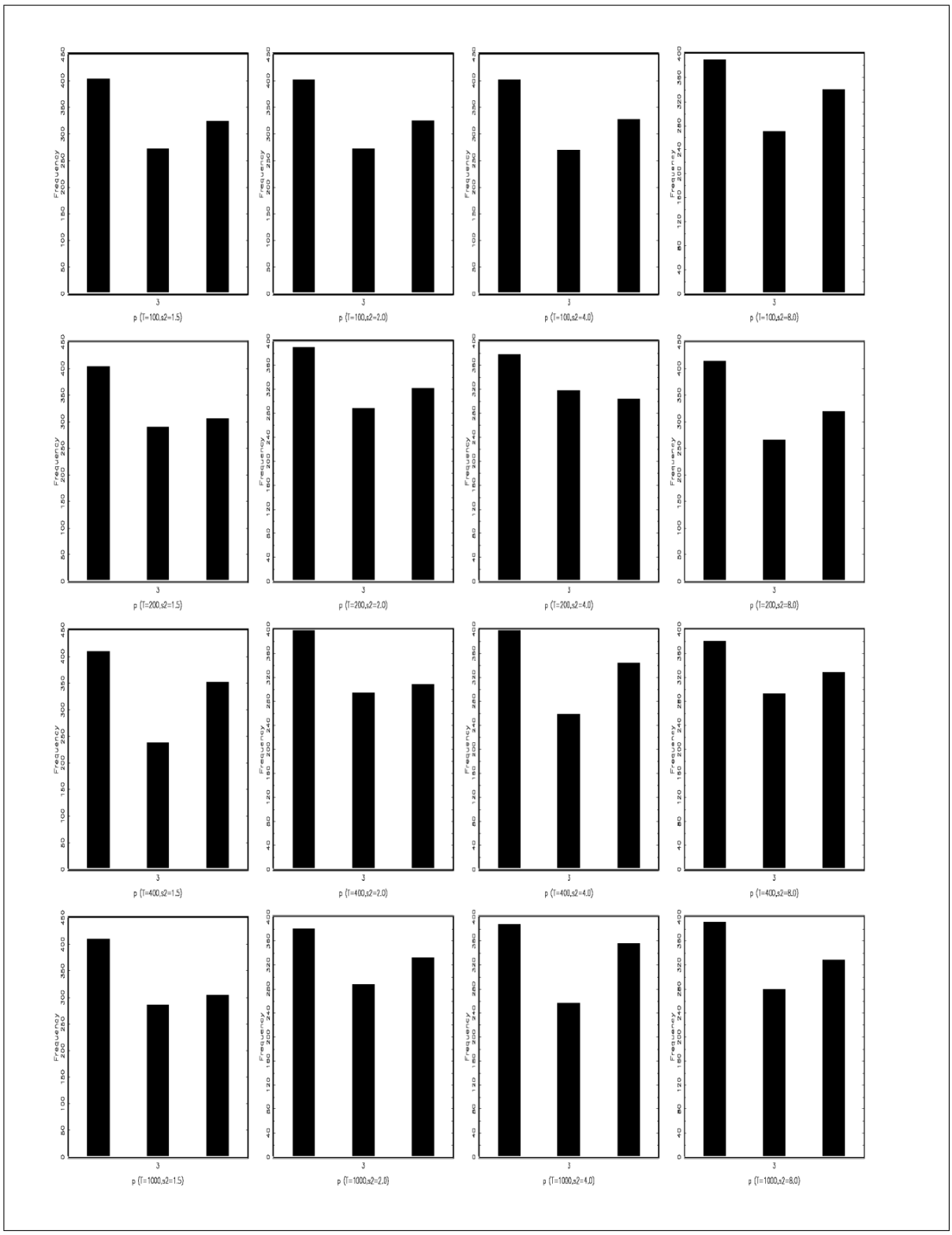

Figure 1: DGP1: Number of times a given p value was selected by CV

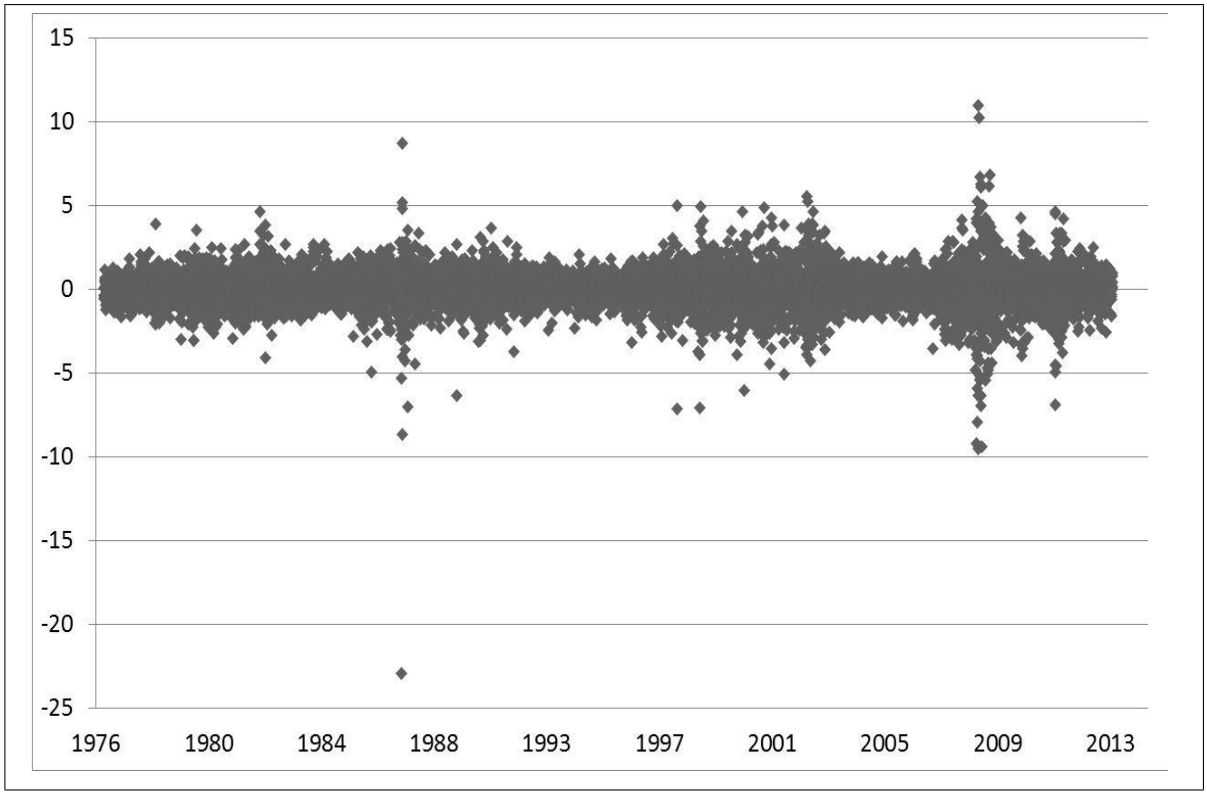

Figure 2: S&P 500 daily percent logarithmic returns data from 15 December 1976 to 9 September 2013

References

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press.

Amisano, G. & Giacomini, R. (2007), 'Comparing density forecasts via weighted likelihood ratio tests', *Journal of Business and Economic Statistics* **25**(2), 177–190.

Arlot, S. & Celisse, A. (2010), 'A survey of cross-validation procedures for model selection', *Statistics Surveys* **4**, 40–79.

Billio, M., Casarin, R., Ravazzolo, F. & van Dijk, H. K. (2013), 'Time-varying combinations of predictive densities using nonlinear filtering', *Journal of Econometrics* . Forthcoming.

Chan, K. S. (1992), 'Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model', *Annals of Statistics* **21**, 520–533.

Chen, X. & Shen, X. (1998), 'Sieve extremum estimates for weakly dependent data', *Econometrica* **66**(2), 289–314.

Clements, M. P. & Hendry, D. F. (1999), *Forecasting Non-Stationary Economic Time Series*, MIT Press.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press.

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

Diks, C., Panchenko, V. & van Dijk, D. (2011), 'Likelihood-based scoring rules for comparing density forecasts in tails', *Journal of Econometrics* **163**(2), 215–230.

Geweke, J. & Amisano, G. (2010), 'Comparing and evaluating Bayesian predictive distributions of asset returns', *International Journal of Forecasting* **26**(2), 216–230.

Geweke, J. & Amisano, G. (2011), 'Optimal prediction pools', *Journal of Econometrics* **164**(1), 130–141.

Geweke, J. & Amisano, G. (2012), 'Prediction with misspecified models', *American Economic Review: Papers and Proceedings* **102**(3), 482–486.

Giacomini, R. & White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**, 1545–1578.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**, 359–378.

Gneiting, T. & Ranjan, R. (2011), 'Comparing density forecasts using threshold- and quantile-weighted scoring rules', *Journal of Business and Economic Statistics* **29**(3), 411–422.

Gneiting, T. & Ranjan, R. (2013), 'Combining predictive distributions', *Electronic Journal of Statistics* **7**, 1747 – 1782.

Gonzalo, J. & Wolf, M. (2005), 'Subsampling inference in threshold autoregressive models', *Journal of Econometrics* **127**, 201–224.

Hall, S. G. & Mitchell, J. (2007), 'Combining density forecasts', *International Journal of Forecasting* **23**, 1–13.

Hansen, B. E. (1996), 'Inference when a nuisance parameter is not identified under the null hypothesis', *Econometrica* **64**(2), 413–30.

Hansen, B. E. (1999), 'Testing for linearity', *Journal of Economic Surveys* **13**(551–576), 413–30.

Jong, R. D. (1997), 'Central limit theorems for dependent heterogeneous random variables', *Econometric Theory* **13**, 353–367.

Jore, A. S., Mitchell, J. & Vahey, S. P. (2010), 'Combining forecast densities from VARs with uncertain instabilities', *Journal of Applied Econometrics* **25**, 621–634.

Kapetanios, G., Mitchell, J. & Shin, Y. (2013), 'A nonlinear panel data model of cross-sectional dependence', *Working Paper No. 12/01, University of Leicester* .

Kim, S., Shephard, N. & Chib, S. (1998), 'Stochastic volatility: Likelihood inference and comparison with arch models', *Review of Economic Studies* **65**(3), 361–93.

Politis, D., Romano, J. & Wolf, M. (1996), *Subsampling*, Springer.

Rossi, B. (2013), Advances in Forecasting under Instabilities, *in* G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting, Volume 2', Elsevier-North Holland Publications.

Stock, J. H. & Watson, M. W. (2007), 'Why has u.s. inflation become harder to forecast?', *Journal of Money, Credit and Banking* **39**(s1), 3–33.

Waggoner, D. F. & Zha, T. (2012), 'Confronting model misspecification in macroeconomics', *Journal of Econometrics* **171**(2), 167 – 184.

Appendix

Proof of Theorem 1

C or C_i where i takes integer values, denote generic finite positive constants.

We wish to prove that minimisation of

$$L_T\left(\boldsymbol{\vartheta}_p\right)=\sum_{t=1}^{T} l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right),$$

where

$$p_t\left(y,\boldsymbol{\vartheta}_p\right)=\sum_{i=1}^{N} w_{p\eta,i}\left(y,\boldsymbol{\vartheta}_p\right)q_{it}\left(y\right),$$

$$w_{p\eta,i}\left(y,\boldsymbol{\vartheta}_p\right)=\tilde{v}_t(v_{i0})+\sum_{s=1}^{p}\tilde{v}_t(v_{is})\eta_s\left(y,\boldsymbol{\theta}_{is}\right),$$

produces an estimate, denoted by $\hat{\boldsymbol{\vartheta}}_T$, of the value of $\boldsymbol{\vartheta}_p=\left(\nu_1,...,\nu_p,\boldsymbol{\theta}_1',...,\boldsymbol{\theta}_p'\right)'$ that minimises $\lim_{T\to\infty}E(L_T\left(\boldsymbol{\vartheta}_p\right))=L\left(\boldsymbol{\vartheta}_p\right)$, denoted by $\boldsymbol{\vartheta}_p^0$, that is asymptotically normal with an asymptotic variance given in the statement of the Theorem. To prove this we use Theorem 4.1.3 of Amemiya (1985). The conditions of this Theorem are satisfied if the following hold:

$$L_T\left(\boldsymbol{\vartheta}_p\right)\to^p L\left(\boldsymbol{\vartheta}_p\right),\quad\text{uniformly over }\boldsymbol{\vartheta}_p, \tag{42}$$

$$\frac{1}{T}\left.\frac{\partial L_T\left(\boldsymbol{\vartheta}_p\right)}{\partial\boldsymbol{\vartheta}_p}\right|_{\boldsymbol{\vartheta}_p^0}\to^d N\left(0,V_2\right) \tag{43}$$

$$\frac{1}{T}\left.\frac{\partial^2 L_T\left(\boldsymbol{\vartheta}_p\right)}{\partial\boldsymbol{\vartheta}_p\partial\boldsymbol{\vartheta}_p'}\right|_{\hat{\boldsymbol{\vartheta}}_T}\to^p V_1 \tag{44}$$

To prove (42), we note that by Theorems 21.9, 21.10 and (21.55)-(21.57) of Davidson (1994), (42) holds if

$$L_T\left(\boldsymbol{\vartheta}_p\right)\to^p L\left(\boldsymbol{\vartheta}_p\right), \tag{45}$$

and

$$\sup_{\boldsymbol{\vartheta}_p\in\Theta_p}\left\|\frac{\partial L_T\left(\boldsymbol{\vartheta}_p\right)}{\partial\boldsymbol{\vartheta}_p}\right\|<\infty, \tag{46}$$

(46) holds by the fact that $l\left(p_t\left(.,\boldsymbol{\vartheta}_p\right);.\right)$ has uniformly bounded derivatives with respect to $\boldsymbol{\vartheta}_p$ over Θ_p and t, by Assumption 2. (44)-(45) and (43) follow by Theorems 19.11 of Davidson (1994) and Jong (1997), respectively given Assumption 4 on the boundedness of the relevant moments and if the processes involved are NED processes on some α-mixing processes satisfying the required size restrictions. To show the latter we need to show that (A) $l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ and

$l^{(2)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ are L_1-NED process on an α- mixing process, where $l^{(i)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ denotes the i-th derivative of l with respect to y_t, and (B) $l^{(1)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ is an L_r-bounded, L_2-NED process, of size $-1/2$, on an α-mixing process of size $-r/(r-2)$. These conditions are satisfied by Assumption 4, given Theorem 17.16 of Davidson (1994).

Proof of Theorem 2

We need to show that the conditions A of Chen & Shen (1998) hold. Condition A.1 is satisfied by Assumption 3. Conditions A.2-A.4 have to be confirmed for the particular instance of the loss function and the approximating function basis given in the Theorem. We show A.2 for

$$L_T=\sum_{t=1}^{T}-\log p_t\left(y_{j+1}\right),$$

where

$$p_t\left(y\right)=\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(v_{is})q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right).$$

Let $\nu^0=\left(\nu_1^0,...,\nu_N^0\right)$, $\nu_i^0=\left(\nu_{i1}^0,...\right)'$ denote the set of coefficients that maximise $E\left(L_T\right)$ and ν_T a generic point of the space of coefficients $\{\nu_{T,is}\}_{i,s=1}^{N,p_T}$. We need to show that

$$\sup_{\{\|\nu^0-\nu_T\|\leq\varepsilon\}}Var\left(\log p_t\left(y,\nu^0\right)-\log p_t\left(y,\nu_T\right)\right)\leq C\varepsilon^2. \tag{47}$$

We have

$$\log p_t\left(y,\nu^0\right)-\log p_t\left(y,\nu_T\right)=\log\left(\frac{p_t\left(y,\nu^0\right)}{p_t\left(y,\nu_T\right)}\right)=\log\left(\frac{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(v_{is}^0)q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(\nu_{T,is})q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}\right)=$$

$$\log\left(1+\frac{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\left(\tilde{v}_t(\nu_{is}^0)-\tilde{v}_t(\nu_{T,is})\right)q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(\nu_{T,is})q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}\right).$$

But,

$$\left|\log\left(1+\frac{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\left(\tilde{v}_t(\nu_{is}^0)-\tilde{v}_t(\nu_{T,is})\right)q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(\nu_{T,is})q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}\right)\right|\leq$$

$$C_1\varepsilon+\left|\frac{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\left(\tilde{v}_t(\nu_{is}^0)-\tilde{v}_t(\nu_{T,is})\right)q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}{\sum_{i=1}^{N}\sum_{s=1}^{p_T}\tilde{v}_t(\nu_{T,is})q_{it}\left(y\right)I\left(r_{s-1}\leq y<r_s\right)}\right|.$$

Then (47) follows from Assumption 5 and the uniform continuity of the mapping $\tilde{v}_t(.)$ over t. Condition A.4, requiring that

$$\sup_{\{\|\nu^0-\nu_T\|\leq\varepsilon\}}\left|\log p_t\left(y,\nu^0\right)-\log p_t\left(y,\nu_T\right)\right|\leq\varepsilon^s CU_T\left(y\right), \tag{48}$$

where $\sup_T E\left(U_T\left(y_t\right)\right)^\gamma < \infty$, for some $\gamma > 2$, follows similarly.

Next we focus on Condition A.3. First, we need to determine the rate at which a function in $\Phi_{q_i}^T$, where $\eta_s = I\left(r_{s-1} \leq y < r_s\right)$ can approximate a continuous function, f, with finite L_2-norm defined on a compact interval of the real line denoted by $M = [M_1, M_2]$. To simplify notation and without loss of generality we denote the piecewise linear approximating function by $\sum_{s=0}^{p_T} \nu_s I\left(r_{s-1} \leq y < r_s\right)$ and would like to determine the rate at which $\left(\int_M \left(\sum_{s=1}^{p_T} \tilde{v}_t(\nu_s) I\left(r_{s-1} \leq y < r_s\right) - f\left(y\right)\right)^2 dy\right)^{1/2}$ converges to zero as $p_T \to \infty$. We assume that the triangular array $\{\{r_s\}_{s=0}^{p_T}\}_{T=1}^{\infty} = \{\{r_{Ts}\}_{s=0}^{p_T}\}_{T=1}^{\infty}$ defines an equidistant grid in the sense that $M_1 \leq r_{T0} < r_{Tp_T} \leq M_2$ and

$$\sup_s \left(r_s - r_{s-1}\right) = O\left(\frac{1}{p_T}\right),$$

and

$$\inf_s \left(r_s - r_{s-1}\right) = O\left(\frac{1}{p_T}\right).$$

For simplicity, we set $M_1 = r_{T0}$ and $r_{Tp_T} = M_2$. We have

$$\int_M \left(\sum_{s=1}^{p_T} \tilde{v}_t(\nu_s) I\left(r_{s-1} \leq y < r_s\right) - f\left(y\right)\right)^2 dy = \sum_{s=0}^{p_T} \int_{r_{s-1}}^{r_s} \left(\tilde{v}_t(\nu_s) - f\left(y\right)\right)^2 dy.$$

By continuity of f and uniform continuity of $\tilde{v}_t(.)$, we have that there exist ν_s such that

$$\sup_s \sup_{y \in [r_{s-1}, r_s]} \left|\tilde{v}_t(\nu_s) - f\left(y\right)\right| = O\left(\frac{1}{p_T}\right).$$

This implies that

$$\sup_s \int_{r_{s-1}}^{r_s} \left(\tilde{v}_t(\nu_s) - f\left(y\right)\right)^2 dy \leq \frac{C_1}{p_T^2} \sup_s \int_{r_{s-1}}^{r_s} dy \leq \frac{C_1}{p_T^3},$$

uniformly over t, which implies that

$$\sum_{s=0}^{p_T} \int_{r_{s-1}}^{r_s} \left(\tilde{v}_t(\nu_s) - f\left(y\right)\right)^2 dy \leq p_T \sup_s \int_{r_{s-1}}^{r_s} \left(\tilde{v}_t(\nu_s) - f\left(y\right)\right)^2 dy \leq \frac{C_1}{p_T^2},$$

uniformly over t, giving

$$\left(\int_M \left(\sum_{s=1}^{p_T} \tilde{v}_t(\nu_s) I\left(r_{s-1} \leq y < r_s\right) - f\left(y\right)\right)^2 dy\right)^{1/2} = O\left(p_T^{-1}\right).$$

The next step involves determining $\mathcal{H}_{\Phi_{q_i}}\left(\epsilon\right)$ which denotes the bracketing L_2 metric entropy

of Φ_{q_i}. $\mathcal{H}_{\Phi_{q_i}}(\epsilon)$ is defined as the logarithm of the cardinality of the ϵ-bracketing set of Φ_{q_i} that has the smallest cardinality among all ϵ-bracketing sets. An ϵ-bracketing set for Φ_{q_i}, with cardinality Q, is defined as a set of L_2 bounded functions $\left\{h_1^l, h_1^u, ..., h_Q^l, h_Q^u\right\}$ such that $\max_j \left\|h_j^u - h_j^l\right\| \leq \epsilon$ and for any function h in Φ_{q_i}, defined on $M = [M_1, M_2]$, there exists j such that $h_j^l \leq h \leq h_j^u$ almost everywhere. We determine the bracketing L_2 metric entropy of Φ_{q_i} where $\eta_s = I(r_{s-1} \leq y < r_s)$, from first principles. By the definition of Φ_{q_i}, any function in Φ_{q_i} is bounded. We set

$$\sup_T \sup_{h \in \Phi_{q_i}} \sup_y |h(y)| = B < \infty.$$

Then, it is easy to see that an ϵ-bracketing set for Φ_{q_i} is given by $\left\{h_i^l, h_i^u\right\}_{i=1}^Q$, where

$$h_i^l = \inf_t \sum_{s=0}^{p_T} \tilde{v}_t(\nu_{is}^l) I(r_{s-1} \leq y < r_s),$$

$$h_i^u = \sup_t \sum_{s=0}^{p_T} \tilde{v}_t(\nu_{is}^u) I(r_{s-1} \leq y < r_s),$$

$\nu_{is}^l = \nu_{Tis}^l$ takes values in $\{-B, -B + \epsilon/p_T, -B + 2\epsilon/p_T, ..., B - \epsilon/p_T\}$ and $\nu_{is}^u = \nu_{Tis}^u$ takes values in $\{-B + \epsilon/p_T, -B + 2\epsilon/p_T, ..., B\}$. Clearly for $\epsilon > C > 0$, $Q = Q_T = O\left(p_T^2\right)$ and so $\mathcal{H}_{\Phi_{q_i}}(\epsilon) = \ln\left(\frac{2Bp_T^2}{\epsilon}\right) = O(\ln(p_T))$. For some δ, such that $0 < \delta < 1$, Condition A.3 of Chen & Shen (1998) involves

$$\delta^{-2} \int_{\delta^2}^{\delta} \mathcal{H}_{\Phi_{q_i}}^{1/2}(\epsilon)\, d\epsilon = \delta^{-2} \int_{\delta^2}^{\delta} \ln\left(\frac{2Bp_T^2}{\epsilon}\right)^{1/2} d\epsilon \leq \delta^{-2} \int_{\delta^2}^{\delta} \ln\left(\frac{2Bp_T^2}{\epsilon}\right) d\epsilon =$$

$$\delta^{-2}\left(\delta \ln\left(\frac{2Bp_T^2}{\delta}\right) - \delta^2 \ln\left(\frac{2Bp_T^2}{\delta^2}\right)\right) < \delta^{-1} \ln\left(\frac{2Bp_T^2}{\delta}\right),$$

which must be less than $T^{1/2}$. We have that

$$\delta^{-1} \ln\left(\frac{2Bp_T^2}{\delta}\right) \leq T^{1/2},$$

or

$$\delta^{-1} \ln\left(\frac{p_T^2}{\delta}\right) = o\left(T^{1/2}\right).$$

Setting $\delta = \delta_T$ and parameterising $\delta_T = T^{-\varphi}$ and $p_T = T^{\phi}$ gives $\varphi < 1/2$. So using the result of Theorem 1 of Chen & Shen (1998) gives

$$\left\|v^0 - \hat{\nu}_T\right\| = O_p\left(\max\left(T^{-\varphi}, T^{-\phi}\right)\right) = O_p\left(T^{-\min(\varphi,\phi)}\right).$$

Proof of Corollary 3

We need to show that the conditions of Theorem 1 hold for the choices made in the statement of the Corollary. We have that

$$l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)=\log\left(\sum_{i=1}^{N}\sum_{s=1}^{p}\tilde{v}_t(\nu_{is})q_{it}\left(y_t\right)I\left(r_{s-1}\leq y_t<r_s\right)\right).$$

Without loss of generality we can focus on a special case given by

$$l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)=\log\left(\tilde{v}_t(\nu_1)q_1\left(y_t\right)+\tilde{v}_t(\nu_2)q_2\left(y_t\right)\right).$$

It is clear that both $l^{(1)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ and $l^{(2)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ are bounded functions of y_t so, by Theorem 17.13 of Davidson (1994), the NED properties of y_t given in Assumption 3 are inherited by $l^{(1)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ and $l^{(2)}\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$. So we focus on $l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ and note that, since $q(y)\sim\exp\left(-y^2\right)$ as $y\to\pm\infty$,

$$\log\left(\exp\left(-y^2\right)\right)=-y^2.$$

Then, using Example 17.17 of Davidson (1994) we get that if y_t is L_2-NED of size $-a$ and L_r-bounded ($r>2$), then y_t^2 is L_2-NED of size $-a(r-2)/2(r-1)$. Since we need that the NED size of $l\left(p_t\left(y_t,\boldsymbol{\vartheta}_p\right);y_t\right)$ to be greater than $1/2$ the minimum acceptable value for r is $a\geq(r-1)/(r-2)$.

Subsampling inference on threshold parameters

In this appendix we show that subsampling provides valid inference for estimated threshold parameters. Subsampling provides valid inference for estimators under extremely weak conditions. As a result it is easy to show the validity of subsampling even when other properties of the estimator are difficult to obtain. This is the case for threshold parameter estimates where consistency and a rate of convergence are difficult to derive. Without loss of generality, we carry out the analysis for the case of a single threshold parameter. We start by assuming that the estimator of r, $\hat{r}$, has a probability limit, r^0, and there exists some sequence c_T such that the distribution of $c_T\left(\hat{r}-r^0\right)$ converges weakly to a non-degenerate limit. Subsampling can be used to determine c_T, as well, if it is unknown; but we refer the reader to (Politis et al. 1996) for further discussion. For the remainder we will assume a known c_T which we consider to be equal to T, as is the case for threshold parameter estimates for threshold models.

Following (Politis et al. 1996), we suggest the following algorithm. Set the subsample sizes to $b_T=T^{\zeta}$, for some $0<\zeta<1$. Construct subsamples by sampling blocks of data temporally. These are given by $\{\tilde{y}_{1,b_T},\tilde{y}_{2,b_T+1},...,\tilde{y}_{T-b_T+1,T}\}$ where $\tilde{y}_{t_1,t_2}=(y_{t_1},...,y_{t_2})'$. ζ is a tuning

parameter related to block size. There exists no theory on its determination, but usual values range between 0.7 and 0.8. Then, threshold parameters are estimated for each subsample created. The empirical distribution of the set of estimates, denoted by $\hat{r}^{*,(i)}$, $i = 1, ...B$, $B = T - b_T + 1$, can be used for inference, and is given by

$$L_{b_T}(x) = \frac{1}{B}\sum_{s=1}^{B} 1\left\{c_{b_T}\left(\hat{r}^{*,(s)} - \hat{r}\right) \leq x\right\}. \tag{49}$$

Below we show that this empirical distribution is valid for inference asymptotically.

Define

$$J_T(x, P) = \Pr_P\left\{c_T\left(\hat{r} - r^0\right) \leq x\right\}. \tag{50}$$

Denote by $J(x, P)$ the limit of $J_T(x, P)$ as $T \to \infty$. We have assumed above that this limit exists and is non-degenerate. The subsampling approximation to $J(x, P)$ is given by $L_{b_T}(x)$. For x_α, where $J(x_\alpha, P) = \alpha$, we need to prove that

$$L_{b_T}(x_\alpha) \to J(x_\alpha, P),$$

for the result to hold. But,

$$E(L_{b_T}(x_\alpha)) = J_T(x, P),$$

because as discussed in Section 2.1.1 the subsample is a sample from the true model, retaining the temporal ordering of the original sample. Hence, it suffices to show that $Var(L_{b_T}(x_\alpha)) \to 0$ as $T \to \infty$. Let

$$1_{b_T,s} = 1\left\{c_{b_T}\left(\hat{r}^{*,(s)} - \hat{r}\right) \leq x_\alpha\right\}, \tag{51}$$

$$v_{B,h} = \frac{1}{B}\sum_{s=1}^{B} Cov\left(1_{b_T,s}, 1_{b_T,s+h}\right). \tag{52}$$

Then

$$Var\left(L_{b_T}(x_\alpha)\right) = \frac{1}{B}\left(v_{B,0} + 2\sum_{h=1}^{B} v_{B,h}\right) = \tag{53}$$

$$\frac{1}{B}\left(v_{B,0} + 2\sum_{h=1}^{Cb_T-1} v_{B,h}\right) + \frac{2}{B}\sum_{h=Cb_T}^{B} v_{B,h} = V_1 + V_2,$$

for some $C > 1$. We first determine the order of magnitude of V_1. By the boundedness of $1_{b_T,s}$, it follows that $v_{B,h}$ is uniformly bounded across h. Hence, $|V_1| \leq \frac{Cb_T}{B}\max_h |v_{B,h}|$, from which

it follows that $V_1 = O(Cb_T/B) = o(1)$. Examining V_2 we notice that

$$|V_2| \leq \frac{2}{B} \sum_{h=Cb_T}^{B-1} |v_{B,h}|. \tag{54}$$

But

$$v_{B,h} = o(1), \text{ uniformly across } h. \tag{55}$$

This follows from the β-mixing of the process which we have assumed above. Hence,

$$\frac{2}{B} \sum_{h=Cb_T}^{B-1} |v_{B,h}| = o(1),$$

proving the convergence of $L_{b_T}(x_\alpha)$ to $J(x_\alpha, P)$ and the overall result.