Improving Density Forecasts and Value-at-Risk Estimates by Combining Densities

Anne Opschoor* Dick van Dijk Michel van der Wel

This version: March 14, 2014†

Abstract

We investigate the added value of combining density forecasts for asset return prediction in a specific region of support. We develop a new technique that takes into account model uncertainty by assigning weights to individual predictive densities using a scoring rule based on the censored likelihood. We apply this approach in the context of recently developed univariate volatility models (including HEAVY and Realized GARCH models), using daily returns from the S&P 500, DJIA, FTSE and Nikkei stock market indexes from 2000 until 2013. The results show that combined density forecasts based on the censored likelihood scoring rule significantly outperform pooling based on equal weights or the log scoring rule and individual density forecasts. In addition, using our technique improves VaR estimates at short horizons.

Keywords: Density forecast evaluation, Volatility modeling, Censored likelihood, Risk management

JEL: C53, C58, G17.

*Corresponding author, e-mail address: opschoor@ese.eur.nl. Postal address: Erasmus University Rotterdam, Erasmus School of Economics, P.O.Box 1738, 3000 DR, Rotterdam, The Netherlands. Phone number: +31(0)10-4088659.

†Anne Opschoor is from Erasmus University Rotterdam and the Tinbergen Institute. Van Dijk is from Erasmus University Rotterdam, Tinbergen Institute and ERIM. Van der Wel is from Erasmus University Rotterdam, CREATES, Tinbergen Institute and ERIM. Michel van der Wel is grateful to Netherlands Organisation for Scientific Research (NWO) for a Veni grant; and for support from CREATES, funded by the Danish National Research Foundation. We are responsible for all errors.

1 Introduction

Value-at-Risk (VaR) is a commonly used measure of downside risk for investments. Financial institutions are allowed by regulation (i.e. the Basel accords) to report VaR estimates for their asset portfolios obtained from their own "internal" model. An important related issue in this estimation is *model uncertainty*, as each model has its prespecified known form and takes no account of possible uncertainty regarding the model structure. In addition, given the availability of a considerable number of different risk-management methods, based on academic literature and/or his expertise, it is a difficult task for a decision-maker to choose the "best" model. Moreover, each model is an incomplete description of reality. Hence relying upon a single model is dangerous to construct a VaR, i.e. a density forecast in the left tail, as any model is "wrong" in some sense.

In this paper, we investigate the usefulness of combining density forecasts with the focus on a particular region of the density. This is motivated in the first place by well known advantages of combining point or density forecasts.[1] We aim to obtain more realistic and more accurate VaR estimates and density forecasts in the left tail. This motivates the investigation of combining density forecasts based on their behavior in the left tail as using the whole density does not necessarily lead to the same quality of forecasts as when we focus purely on the left tail of a density. Therefore, we develop a density forecast combination method that extends the method of Geweke and Amisano (2011), which uses the *whole* density, by considering the censored likelihood (*csl*) scoring rule of Diks *et al.* (2011) that focuses on a region of the densities' support of particular interest, such as the left tail.

We use our novel methodology in an empirical application involving several recently developed univariate volatility models. Hence, as a second contribution to the literature, we make a comparison between these models with respect to their predictive ability in terms of density forecasts. In particular, beyond the traditional GARCH model (Bollerslev, 1986), we consider the Heavy model (Shephard and Sheppard, 2010) and the Realized GARCH model (Hansen *et al.*, 2012) that include realized measures, as well as the GAS model (Creal *et al.*, 2013). All models are applied to daily returns on the S&P 500, DJIA, FTSE and Nikkei stock market indexes from 2000 until 2013.

[1] We discuss this literature in more detail below.

We evaluate the added value of combining density forecasts both statistically and economically. First, we test equal predictive accuracy in the left tail of a combined density forecast based on our new method and three alternatives: (i) the method based on the whole density, (ii) a benchmark that consists of equal weights, and (iii) the density forecast of each individual model. Second, we compare 1- and 5-day VaR estimates based on these methods using the Unconditional Coverage test and the Independence test of Christoffersen (1998). In addition we test on equal accuracy based on an asymmetric tick-loss function using the test procedure of Giacomini and White (2006).

Our results show statistically that density forecasts in the tail are more accurate if one pools density forecasts using the *csl* scoring rule than using the aforementioned method, using equal weights or using the density forecast of any individual volatility model. 90% and 95% 1-day VaR estimates improve significantly compared to the other pooling methods or the individual models, such that less violations are made and the unconditional coverage matches more closely to the nominal value. Moreover, the accuracy of the VaR estimates improves significantly upon using equal weights or any individual model according to the asymmetric tick-loss function. In addition, we show that the combination weights based on the *csl* scoring rule differ considerably from the weights obtained by using the whole density. Hence, a certain volatility model could get no or less weight in the method of Geweke and Amisano (2011), but may be useful in our new method.

We contribute to the literature on combining forecasts, see Timmermann (2006) for a survey. Starting with the seminal work of Bates and Granger (1969), combining point forecasts appears to be a successful forecasting strategy, improving upon individual forecasts. Timmermann (2006) shows from a theoretical point of view why forecast combinations could work well. This is confirmed by numerous empirical applications in different areas including macroeconomic and financial forecasting. For example, forecasting output growth using individual predictors typically delivers forecasts that are unstable over time. Combining forecasts offers more stable forecasts which improve upon autoregressive forecasts (Stock and Watson, 2004). Rapach *et al.* (2010) provide similar evidence in the context of equity premium prediction, by showing that combining forecasts leads to statistically and economically significant out-of-sample gains relative to the historical average return.

Although the literature shows the usefulness of combining point forecasts, point forecasts

themselves are not very informative if there is no indication of their uncertainty (see Granger and Pesaran, 2000; Garratt *et al.*, 2003). This finding has led to a growing interest in *density* forecasts, which represent a full predictive distribution of a random variable and hence provide the most complete measure of this uncertainty. It is a natural step forward to bring together the concepts of forecast combinations and density forecasts. The literature on combining density forecasts is yet scarce, although the interest in this topic of research grows with applications to for example macro-economics (Jore *et al.*, 2010; Aastveit *et al.*, 2011). Wallis (2005) considers a finite mixture distribution, which takes a weighted linear combination of multiple density forecasts. Hall and Mitchell (2007) address the issue how to choose the weights assigned to each competing density. They propose a methodology with the aim to obtain the most accurate density forecast from a statistical point of view. This boils down to using the logarithmic scoring rule, which takes the log of the predictive density evaluated at the observed value of the variable of interest. Closely related is the work of Geweke and Amisano (2011), who use the logarithmic scoring rule to obtain weights to form optimal linear combinations of predictive densities. We extend this approach, by substituting the log score rule by the censored likelihood scoring rule.

The remainder of this paper is organized as follows. Section 2 puts forward our methodology of combining density forecasts using the *csl* scoring rule. In Section 3, we provide an overview of the univariate volatility models and the related assumed conditional density functions, which are used in the empirical application (Section 4). Section 5 concludes.

2 Combining density forecasts

Suppose a decision maker has n different models for a variable of interest y. Conditional on information available up to and including time $t-1$, the predictive density corresponding with a particular model at time t is of the form $p_t(y_t|I_{t-1}, \theta_{A_i}, A_i)$, where I_{t-1} indicate the information set up to and including time $t-1$, A_i denotes the particular model i, $(i = 1, \ldots, n)$ and θ_{A_i} the estimated parameters of model A_i given I_{t-1}. Suppose further that the decision maker aims to choose the best predictive density at time $T+1$, given the available density forecasts from time $t = 1, \ldots, T$. An often used approach is to make use of scoring rules. A scoring rule measures the quality of density forecasts by assigning

a numerical score.[2] Typically, this rule is a objective function that depends on the density forecast and the actually observed value, such that a higher score is associated with a "better" density forecast. According to Gneiting and Raftery (2007), a scoring rule is *proper* if it satisfies the condition that incorrect density forecasts do not receive a higher average score than the true density. This property is important and a natural requirement for any rational decision maker.

A well founded scoring rule is the log score function (see Mitchell and Hall, 2005; Amisano and Giacomini, 2007). This function for a particular model A_i at a specific time t is defined as

$$S^l(y_t; A_i) = \log p_t(y_t|I_{t-1}, A_i), \tag{1}$$

with S^l the abbreviation of the log scoring rule, which simply takes the logarithm of the predictive density evaluated at y_t. This scoring rule is closely related to information theoretic goodness-of-fit measures such as the Kullback-Leibler Information Criterion (KLIC) associated with the density forecast $p_t(y_t|I_{t-1}, A_i)$. It can be shown that a higher value of the logarithmic score coincides with a lower value of the KLIC. Put differently, maximizing the logarithmic score is equivalent with minimizing the KLIC.

Geweke and Amisano (2011) argue that it is highly unlikely that one model is the true model for constructing a predictive density. They propose therefore to combine the predictive densities using the log score function of (1). In particular, they consider predictive densities of the form

$$\sum_{i=1}^{n} w_i p_t(y_t|I_{t-1}, A_i), \tag{2}$$

for $i = 1, \ldots, n$ and weights w_i, restricted such that they are positive and sum to one to ensure that (2) is a valid probability density function. It is natural to choose the weights in such a way that the log score function in (1) is maximized (and hence the KLIC is

[2]Note that we use the term 'score' twice: (i) in the GAS models to indicate the derivative of the logarithm of the density with respect to a certain parameter and (ii) a number that is assigned to measure density forecasts.

minimized):

$$S^l(Y_T, C) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{n} w_i p_t(y_t|I_{t-1}, A_i) \right], \tag{3}$$

with $Y_T = y_1, \dots y_T$, and C representing the fact that a combination of models is evaluated instead of a single model A_i. Following Bacharach (1974), linear combinations of (subjective) probability distributions are known as *linear opinion pools*. We use the term *pooling* and *linear opinion pools* interchangeably in this paper.

The main idea of this paper is to extend the approach of Geweke and Amisano (2011) by focusing on a particular region of interest of the predictive density. In order to do so, we consider a scoring rule based on the censored likelihood (*csl*), advocated by Diks *et al.* (2011). They prove that this scoring rule is proper and show the usefulness of this scoring rule if one is interested in the accuracy of density forecasts in a specific region. In this study, the focus is on the left tail, which is important for risk management purposes. The *csl* score function for a specific region B_t for model A_i at time t reads

$$\begin{aligned} S^{csl}(y_t|A_i) = I[y_t \in B_t] \log p_t(y_t|I_{t-1}, A_i) \\ + I[y_t \in B_t^c] \log \left(\int_{B_t^c} p_t(y|I_{t-1}, A_i) dy \right) \end{aligned} \tag{4}$$

with B_t^c the complement of B_t and $I[\cdot]$ an indicator function that takes the value 1 if the argument is true. The first part of this scoring rule focuses on the behavior of the density forecast in the region of interest B_t. The second part computes the cdf of the density in the region outside B_t.[3] Hence any observation outside B_t ignores the shape of $p_t(y_t|I_{t-1}, A_i)$ outside B_t. Note that (4) simplifies to the log scoring rule of (1) if B_t represents the full sample space.

The next step is combine the predictive densities based on the *csl* scoring rule. That is, we consider again predictive densities as defined in (2), however with the weights obtained by optimizing the corresponding censored likelihood score function over the values $Y_T =$

[3]To interpret this second part, if B_t is the left tail $y_{t+1} < r$ (with r a certain quantile of the cdf of y_t), the second part of (4) ensures that the tail probability implied by $p_t(y_t|I_{t-1}, A_i)$ matches with the frequency at which tail observations actually occur.

$y_1, \dots y_T$:

$$S^{csl}(Y_T, C) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{n} w_i \left(I[y_t \in B_t] p_t(y_t | I_{t-1}, A_i) + I[y_t \in B_t^c] \int_{B_t^c} p_t(y | I_{t-1}, A_i) dy \right) \right]. \tag{5}$$

We end this section by a brief comment about the optimization of the weights w_t in (3) and (5). Although (numerical) constrained optimization techniques may be used, we consider the algorithm of Conflitti *et al.* (2012). This iterative algorithm is easy to implement works well even when the number of forecasts to combine gets large. See Appendix A for more details.

3 Models and distributions

This study focuses on density forecasting in the context of univariate volatility models. We consider several classes of models, including the standard GARCH model of Bollerslev (1986), the HEAVY model of Shephard and Sheppard (2010), the Realized GARCH model of Hansen *et al.* (2012) and the GAS model of Creal *et al.* (2013). All models are based on the following general specification for y_t, the return for a financial asset at day t:

$$y_t = \mu + \sqrt{h_t} z_t, \quad \text{with } z_t | I_{t-1} \sim D(0, 1), \tag{6}$$

where μ denotes the conditional mean of the returns, h_t the conditional variance and z_t the standardized unexpected return following a certain conditional distribution $D(\cdot)$ with mean zero and unit variance. Further, I_t denotes the information set up to and including time t.[4] The following subsections differentiate between various specifications for the dynamics of h_t and possible choices for the conditional return density function $D(\cdot)$.

[4]For ease of exposition, we assume the conditional mean fixed, although it could easily be extended to a time-varying mean μ_t.

3.1 Univariate volatility models

The first model we consider is the traditional GARCH(1,1) model (Bollerslev, 1986) for the conditional variance h_t:

$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{7}$$

with $\omega > 0$, $\alpha > 0$ and $\beta > 0$ to ensure a positive variance. The past squared demeaned return in this model is the innovation for the conditional variance. Many extensions of the GARCH model are proposed (e.g. the EGARCH and GJR GARCH models Nelson, 1991; Glosten *et al.*, 1993), however we stick to the basis specification as given in (7). We restrict also the other considered model classes in this study to the basis specification, although many variants/extensions are possible. The reason is that the aim is to compare model *classes* combined with distributions, and not models *within* a specific class.

Creal *et al.* (2013) develop a broader set of models which also includes the GARCH model of (7), namely the Generalized Autoregressive Score (GAS) models. The key property of these models is that innovations for time-varying parameters are based on the score of the probability density function at time t. In terms of our univariate volatility models, the time-varying parameters are the conditional variances h_t. The GAS(1,1) model proposes the following structure for h_t:

$$\begin{aligned} h_t &= \omega + \alpha s_{t-1} + \beta h_{t-1}, \\ s_t &= Q_t \nabla_t, \\ \nabla_t &= \frac{\partial \log p(y_t|h_t, I_{t-1}\theta)}{\partial h_t}, \end{aligned} \tag{8}$$

with $p(y_t|h_t, I_{t-1}, \theta)$ the conditional return density, θ the parameter vector, ∇_t the score and Q_t a scale factor. We follow Creal *et al.* (2013) and define the scale factor as $1/\mathsf{E}_{t-1}[\nabla_t^2]$, where E_t denotes the expectation with respect to the return density $p(y_t|h_t, I_{t-1}, \theta)$. For example, when the returns y_t follow a conditional Normal distribution, the GAS model corresponds exactly to the GARCH(1,1) model of (7).[5] In case of a fat-tailed Student-t

[5]When $y_t \sim N(0, h_t)$, $\nabla_t = -0.5h_t^{-1} + 0.5h_t^{-2}y_t^2$ and $Q_t = 2h_t^2$. Hence the GAS model becomes $h_t = \omega + \alpha(y_t^2 - h_t) + \beta h_t$, which is equivalent with the GARCH model of (7).

distribution for y_t, the score based volatility model reads

$$h_t = \omega + \alpha(1 + 3/\nu)\frac{\nu + 1}{(\nu - 2) + \frac{(y_{t-1}-\mu)^2}{h_{t-1}}}(y_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{9}$$

and will be labeled as the GAS-t model. The specification downweights the more extreme observations, in the sense that if the distribution is more heavy tailed, it is less likely that an extreme observation is due to an increase in volatility. Note that this is a function of ν; when $\nu \to \infty$, (9) converges to the GARCH(1,1) model of (7). We again impose $\omega > 0$,$\alpha > 0$ and $\beta > 0$ in the estimation of the parameters.

The third and fourth model classes in this study include realised measures to describe the dynamics of daily volatility. A realised measure is a high-frequency estimator of the variance of a particular asset return during the times the asset is trade on an exchange. For example, the realised variance (RV) for a particular day sums the squared returns during a specific intra-day period. The intuition is that realised measures are a more accurate estimate of daily volatility than the squared daily return, as used in the GARCH models (see Andersen *et al.*, 2003).

A recently developed model that explicitly introduces high-frequency estimators in daily volatility models is the HEAVY model of Shephard and Sheppard (2010). In particular, this model assumes the following structure for the conditional variance h_t and the expectation of the realised measure $\xi_t = \mathsf{E}[RM_t|I_{t-1}]$:

$$h_t = \omega + \alpha RM_{t-1} + \beta h_{t-1}, \tag{10}$$

$$\xi_t = \omega_R + \alpha_R RM_{t-1} + \beta_R \xi_{t-1}. \tag{11}$$

All parameters should be positive to avoid negative values of h_t and ξ_t. The Heavy model is seen to consist of a GARCH structure for both h_t and ξ_t, with RM_t as innovation term. One may also include the squared (demeaned) daily return in (10), however in practice the estimate of the corresponding parameter is generally close to zero and insignificant, as noted by Shephard and Sheppard (2010). Equation (11) "completes" the system, in the sense that without this equation one can only perform one-step ahead forecasts of the conditional variance h from (10) since future values of the realised measure are unknown at time t.

A second model that relates conditional volatility with realised measures is the Realized GARCH model (RGARCH) of Hansen *et al.* (2012). The basic specification is given by:

$$h_t = \omega + \alpha RM_{t-1} + \beta h_{t-1}, \tag{12}$$

$$RM_t = \delta + \phi h_t + \tau(z_t) + u_t, \tag{13}$$

with $\tau(z_t)$ the leverage function, defined in the basic form as $\tau_1 z_t + \tau_2(z_t^2 - 1)$. This function allows for the empirical finding that negative and positive shocks may have a different impact on the volatility. Except τ_1, which is typically negative, all parameters are restricted to be positive. The dynamics for h_t are similar for both the HEAVY and RGARCH model, however the difference arises in the specification of (the expectation of) RM_t. The HEAVY model proposes a GARCH structure for $\mathsf{E}[RM_t|I_{t-1}]$, while the RGARCH model explicitly relates RM_t to the conditional variance at time t and introduces additionally a leverage component.

3.2 Conditional distributions

We consider four possible distributions $D(\cdot)$ of z_t in (6), which corresponds with the conditional density of the returns y_t . The starting point is the conditional Normal distribution, since this distribution is simple and often used. However, to take into account possible conditional non-normality, skewness, and excess kurtosis, we also allow the return y_t to follow a Student-t distribution with mean μ, variance h_t and ν degrees of freedom. That is,

$$f(y_t|\mu, h_t, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{h_t(\nu-2)\pi}} \left(1 + \frac{(y_t-\mu)^2}{h_t(\nu-2)}\right)^{-\frac{\nu+1}{2}}. \tag{14}$$

The degrees of freedom ν is treated as an unknown parameter and is estimated together with the volatility parameters. In addition, $\nu > 2$ is required to ensure a existing variance. The excess kurtosis of the Student-t distribution is equal to $6/(\nu-4)$, hence it is only defined if $\nu > 4$. In general, a lower value of ν implies a more fat-tailed distribution. Third, we consider the Laplace distribution, which also exhibits fatter tails than the Normal distribution, but

does not involve additional parameters:

$$f(y_t|\mu, h_t) = \frac{1}{\sqrt{2h_t}} \exp\left(-\sqrt{2}\frac{|y_t - \mu|}{\sqrt{h_t}}\right) \tag{15}$$

with again mean μ and variance h_t. Finally the Skewed-t distribution of Hansen (1994) enables returns to be distributed asymmetrically, in contrast to the three symmetric distributions discussed above. For a zero mean and unit variance variable $z_t = (y_t - \mu)/\sqrt{h_t}$, the distribution reads

$$f(z_t; \lambda, \nu) = \begin{cases} bc\left(1 + \frac{1}{\nu-2}(\frac{bz_t+a}{1-\lambda})^2\right)^{-\frac{\nu+1}{2}} & \text{if} \quad z_t < -\frac{a}{b} \\ bc\left(1 + \frac{1}{\nu-2}(\frac{bz_t+a}{1+\lambda})^2\right)^{-\frac{\nu+1}{2}} & \text{if} \quad z_t \geq -\frac{a}{b} \end{cases} \tag{16}$$

with

$$a = 4\lambda c\frac{\nu-2}{\nu-1}, \quad b^2 = 1 + 3\lambda^2 - a^2, \text{and} \quad c = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}$$

such that $f(y_t|\mu, h_t, \nu, \lambda) = 1/h_t f(z_t; \lambda, \nu)$. Further, λ is the skewness parameter and ν again represents the degrees of freedom. A (positive) negative value of λ indicates (positive) negative skewness.

Table 1 summarizes the various choices for the dynamics of the conditional variance h_t and a conditional distribution $D(\cdot)$, both defined in the general specification for the daily return of (6). For the GARCH(1,1), HEAVY and RGARCH models, we estimate their parameters in combination with the assumption of the four described conditional distributions of 3.2 (i.e. Normal, Student-t, Laplace and Skewed-t). Further, we assume a Student-t and Laplace distribution for the GAS models. This delivers 14 models in total. We estimate all models by Maximum Likelihood. This is not a computationally involved step, since we are dealing with univariate models with a maximum of 8 parameters (RGARCH models) to be estimated. In addition, we can estimate the HEAVY parameters of (10) and (11) separately, see Shephard and Sheppard (2010) for more details.

Table 1: Overview of volatility models and conditional distributions
This table reports the various choices for the dynamics of the conditional variance h_t and the possible conditional distributions $D(\cdot)$, both apparent in the general specification of (6). An "x" ("-") denotes that an particular specification together with a conditional distribution is (not) chosen.

	Normal	Student-t	Laplace	Skewed-t
GARCH(1,1)	x	x	x	x
GAS(1,1)	x[a]	x	x	-[b]
Heavy	x	x	x	x
RGARCH	x	x	x	x

[a] The GAS(1,1) model with Normal distributed errors is the same as the GARCH(1,1) model with Normal errors.
[b] We leave the GAS(1,1) with Skewed-t distributed errors as a topic of further research.

4 Application

This section contains an application of our new method of combining density forecasts, applied in the context of univariate volatility models. In the following subsections, we discuss the data and implementation details, the evaluation of the density forecasts and finally the results.

4.1 Data and implementation details

We apply the volatility models of Section 3 to daily returns from four major stock market indexes: S&P 500, DJIA, Nikkei and the FTSE. The sample period goes from January 3, 2000 until June 28, 2013. Daily returns as well as their corresponding realized measures are obtained from the Oxford-Man Institute's "realised library".[6] We follow Shephard and Sheppard (2010) and use the realised kernel (see Barndorff-Nielsen *et al.*, 2008) as the realised measure at time t (RM_t). When the exchange is closed, days are deleted from the sample.[7] Figure 1 shows the dynamics of the S&P 500 index and Japanese equity index, together with the square root of the realised kernel estimate of the daily variance. The dynamics of both indexes are quite similar, however the Nikkei index contains more downward spikes (e.g. the 2011 Tohoku earthquake). Nevertheless, both return graphs

[6] See http://realized.oxford-man.ox.ac.uk/
[7] We have to delete 1-1.5% of the daily returns on the S&P500, DJIA and FTSE index. The Nikkei index loses 3% of its daily returns.

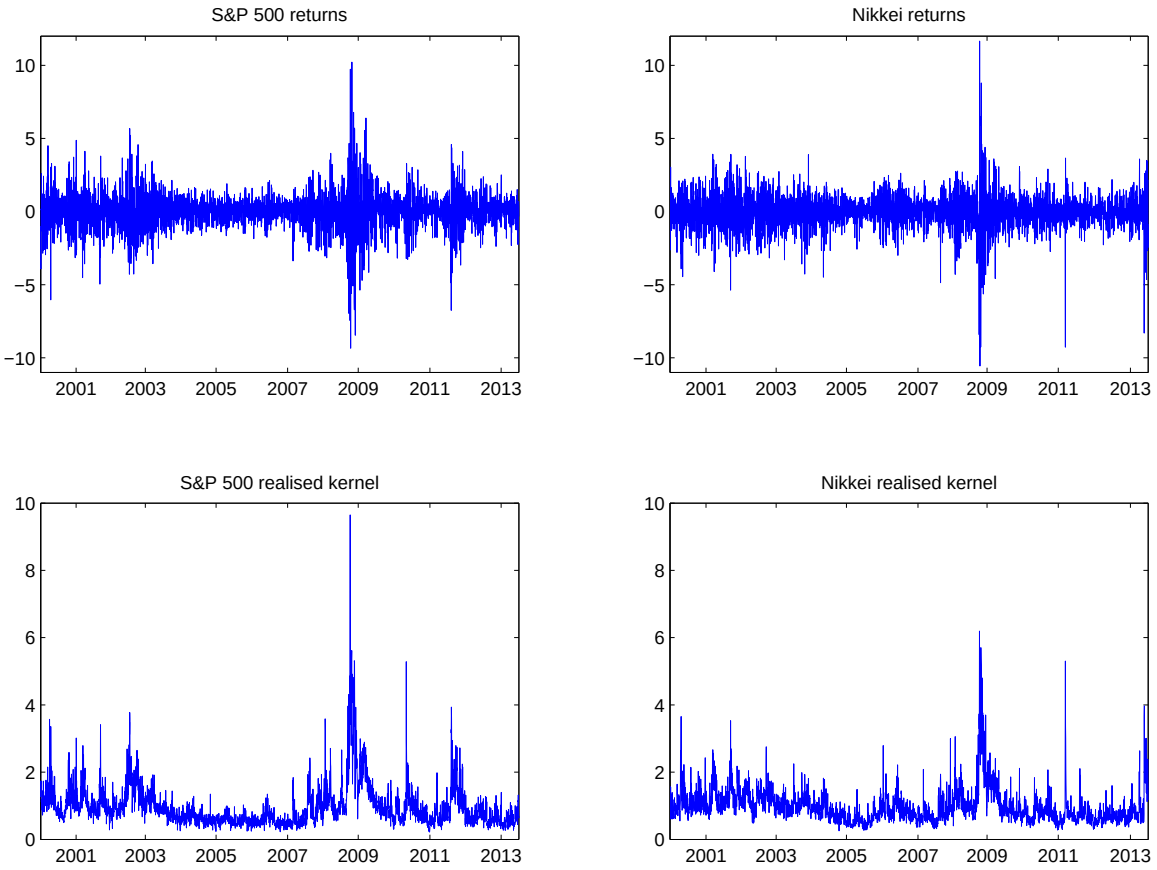
clearly show the presence of conditional heteroskedasticity, since calm periods and periods of high volatility occur in an alternating pattern.

We apply a rolling window scheme to estimate the model parameters and construct density forecasts. More specifically, we use an estimation window of approximately 3 years ($T_{est} = 750$ observations), estimate the model parameters and construct 1- until 5-step ahead forecasts of h_t at each time t ($t = T_{est}, T_{est}+1, \ldots, T-5$). Given these forecasts, we also construct the corresponding 1- and 5-step ahead density forecasts. After 750 subsequent density forecasts ($T_w = 750$) have been obtained for each model, we optimize (3) and (5) to obtain w_t. In case of the *csl* score function, we define the region B_t as the left tail $y_t < \hat{r}_t^{\kappa}$ with $\hat{r}_t^{\kappa}$ the κth quantile of the empirical CDF of the 750 returns corresponding with the estimation window T_{est}. We repeat also this optimization by means of a rolling window scheme with a window of T_w density forecasts evaluations at each time t ($t = T_{est} + T_w + 1, T_{est} + T_w + 2, \ldots, T-5$). We choose κ equal to 0.15 and 0.25 respectively. We choose $T_{est} = 750$ such that there is a sufficient number of observations for parameter estimation of the models. Further, we emphasize the trade-off in the choice of κ. Given our interest in the left tail, we should take a small value of κ. However, the corresponding number of observations in the region of interest becomes very low, such that the variation in the *csl* scores of the different models declines.[8] Similarly, there is a trade-off in the choice of T_w. On the one hand, one would choose T_w as high as possible in order to use the largest amount of available observations to compute the weights w_t. But on the other hand, if the relative performance of different models varies through time, one should take this into account and choose a smaller value of T_w. In addition, T_w and κ are related in the sense that a low value of κ combined with a small window results in a small amount of observations within the region B_t. Hence given these trade-offs and the relation between those two variables, we choose T_w and κ as 750 and 0.15 (0.25) such that there are 112 (187) observations in the left tail.

[8]Recall that if y_t is outside the region B_t with B_t the left tail $y_t < \hat{r}_t^{\kappa}$, the *csl* score is the cdf of y_t in the complement of the region.

Figure 1: Daily returns and realised measures

This figure depicts the daily (close-to-close) returns on the S&P 500 index and the Nikkei index (upper part) and a realised kernel estimate of the corresponding daily (open-to-close) volatility (bottom part) from January 3, 2000, through June 27, 2013 (3,364 and 3,206 observations respectively). Both daily returns and volatilities are given in percentages.



4.2 Evaluation

We assess the accuracy of our (combined) density forecasts in two ways. First, we focus purely on the predictive density in the left tail and investigate statistically whether pooling based on censored densities adds any value. Following Diks *et al.* (2011), we test the null hypothesis of equal performance of two density forecasts $p_t(y_t; I_{t-1}, A_i)$ and $p_t(y_t; I_{t-1}, A_j)$ based on the scoring rule of (4).[9] That is, given a sample of density forecasts and corresponding realizations for m periods, define the relative score d_t as

$$d_t = S^{csl}(y_t; A_i) - S^{csl}(y_t; A_j) \tag{17}$$

with corresponding null-hypothesis $H_0 : \mathsf{E}[d_t] = 0$ for all m periods. The resulting Diebold and Mariano (1995) test-statistic is then given by

$$t_m = \frac{\bar{d}_m}{\sqrt{\hat{\sigma}_m^2/m}}, \tag{18}$$

with $\bar{d}_m$ the sample average of the score differences and $\hat{\sigma}_m^2$ a HAC-consistent variance estimator of the true variance σ_m^2 of d_t. A positive value means that the density forecasts in the tail of model A_i are more accurate than the corresponding density forecasts of model A_j. This test allows for parameter estimation uncertainty and fits the framework of Giacomini and White (2006), who show that the use of a rolling window of m past observations for parameter estimation simplifies the asymptotic theory of tests of equal predictive accuracy. Moreover, the test allows to compare density forecasts of both nested and non-nested models.

The second way to explore the additional value of using censored densities in this study is based on 1- and 5-day Value-at-Risk (VaR) estimates. For the individual models considered in this study, the 1-day VaR estimate reads

$$VaR_t^{1-q} = \mu + z_q\sqrt{h_t}, \tag{19}$$

with μ the estimated conditional mean return, h_t the (forecasted) conditional variance, and z_q represents the q-th quantile of the assumed cdf. However, we cannot apply (19) when

[9] In case we consider density forecasts using combinations, the density forecast is given by $\sum_{i=1}^{n} w_{it} p_t(y_t; I_{t-1}, A_i)$.

our predictive distribution is a combination of individual distributions.[10] This also holds for the h-day ($h \geq 2$) VaR estimates if the assumed distribution is non-Normal. We use simulation techniques to overcome this issue. That is, we simulate daily returns from each individual model/distribution according to the assigned weight (and conditional variance) to obtain the required quantile of the total distribution to compute the $(1-q)\%$ VaR.

Finally, we test the accuracy of the VaR estimates by focusing on two aspects. First, we assess the frequency of the VaR violations with the unconditional coverage (UC) of Kupiec (1995) and Christoffersen (1998). These tests compare the actual with the expected number of violations. In addition, we test whether the violations occur in clusters by means of the Independence test (Ind) of Christoffersen (1998). In order to apply both tests on the estimated 5-day VaRs, we create first 5 different sub-series to avoid any overlap. Thus, sub-series j contains the estimates $\{VaR_j^{1-q}, VaR_{j+5}^{1-q}, VaR_{j+10}^{1-q}, \ldots\}$ for $j = 1, \ldots, 5$. According to the suggestion of Diebold *et al.* (1998), we use Bonferroni bounds for the 5 sub-series. That is, we assume that the VaR series has autocorrelation up to and including lag 4, whereas each sub-series should have correct coverage and independent VaR violations. Hence we therefore backtest each sub-series separately with a size of $\alpha/5$, with α the used significance level. Rejecting the null hypothesis of unconditional coverage/independence occurs when the null is rejected for *any* of the 5 sub-series. Second, we compare the 1-day VaR estimates of two different methods/models using the following asymmetric linear (tick) loss function of order q, which is also used in the CPA test of Giacomini and White (2006):

$$L_{A_i}^q(e_t) = (q - I[e_t < 0])e_t, \tag{20}$$

where $q = 5\%$ and 10% and $e_t = y_t - VaR_t^{1-q}$. The loss function is asymmetric in the sense that if there occurs a violation (i.e. $e_t < 0$)) the negative number $q-1$ is multiplied by the magnitude of the violation e_t, resulting in a penalization of $(1-q) \times e_t$. In contrast to this, if there is no violation, the loss is equal to $q \times e_t$, which is considerable lower.[11] Hence a model A_i is more penalized when a VaR violation is observed. The larger the magnitude of this violation, the larger the penalization. Similar to the density forecasts, we define the

[10] The VaR of a mixture of densities is not equal to the weighted average of each individual VaR.

[11] Suppose the 95% 1-day VaR of model A and B are equal to -5% and -8% respectively, while the actual return is -6%. The loss associated with model A is equal to $(0.05-1)(-1) = 0.95$, while the loss of model B is equal to $(0.05-0)(-2) = 0.10$.

relative loss as

$$d_t^q = L_{A_i}^q(e_t) - L_{A_j}^q(e_t) \tag{21}$$

and consider again a Diebold and Mariano (1995) type statistic as given in (18). A negative value of the unconditional mean of d_t^q means that on average the VaR estimates of model A_i are better than the corresponding estimates of model A_j.

4.3 Results

In this subsection, we present both the statistical and economic results. In order to understand these results, we first present the weights which are obtained by optimizing the log score function of (3) and the *csl* score function of (5). Figure 2 shows the result of the iterative process of optimizing weights according to both score functions. The sub-graphs depict the dynamics of the weights using daily returns from the DJIA index according to the 14 models listed in Table 1.[12] The top part of the figure corresponds with the log score function, while the bottom part corresponds with the *csl* score function with $\kappa = 0.25$. The top part shows that using the log score function results in a large weight for the Heavy model with Skewed-t distributed errors until 2009. Moreover, it gets the full weight until 2008. Subsequently, the weight of the Heavy Skewed-t model declines to zero and there is room for the Heavy N model and the Heavy model with Laplace distributed returns. The latter gets almost a weight of 0.6 in 2011. Nevertheless, the Heavy Skewed-t model appears again and dominates from 2012 onwards.

A rather different dynamic pattern arises from the lower part of Figure 2, i.e. when the *csl* score function is optimized. Although the graph is similar in the sense that (i) the Heavy Skewed-t model dominates the other models during 2006-2007 and since 2012 and (ii) the Heavy model with Laplace distributed errors dominates during 2009-2012, the years 2008-2012 show two main differences. First, the GARCH Skewed-t model has more impact in case of the *csl* score function, reaching a maximum weight of 0.65 at the end of 2008. Second, the RGARCH model class gets considerably more weight, either combined with the Laplace distribution (2009) or the Normal distribution (2010-2011). Hence the Heavy N

[12] Figure B.1 in Appendix B provides weights corresponding to the other three stock market indexes.

model is replaced by the RGARCH N model during the period 2010-2012.

To ease the interpretation of this finding, Figure 3 sums up the weights according to each model class (upper part) and distribution (lower part), for both types of scoring rules. It seems that in case of the log score function, the Heavy model with Skewed-t distributed errors dominates all the remaining models for most years. Only during 2009-2010, the RGARCH model has some influence, while the Skewed-t distribution is replaced by the Normal and Laplace distribution. In contrast to this, focusing of the left tail of the distribution does lead to more influence of the GARCH and RGARCH class of models. Furthermore, the Laplace distribution is more apparent during the years 2009-2012, with a climax at the start of 2012. Finally, both the Laplace and Normal distribution characterize 2012, while the log score function allocates the most weight to the Skewed-t distribution in that year.

4.3.1 Statistical results

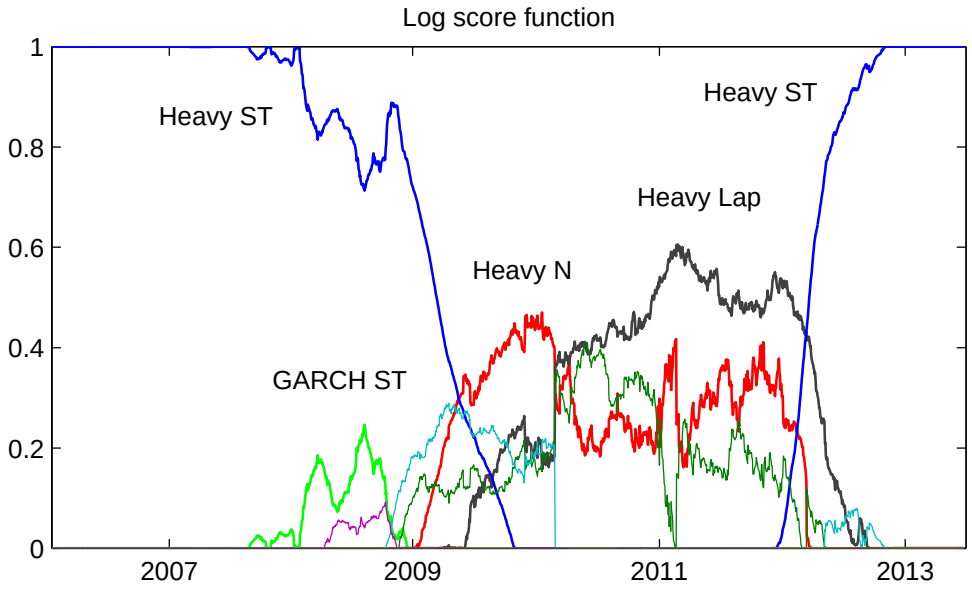
Table 2 provides the importance of pooling of censored densities by showing results of the t-test on equal predictive accuracy of (18). In more detail, we test equal accuracy of the combined density forecasts based on the *csl* score function and based on the log score function. In addition, we test the accuracy of the individual censored density of each competing model. Panel A reports HAC-based t-statistics of the test of equal accuracy of density forecasts made by means of combination, using the *csl* or log score function of Section 2. As a benchmark, we consider also the case of equal weights assigned to each competing density. A positive number corresponds with more accurate density forecasts of the forecast method based on the *csl* score function. The table suggests that (except for the FTSE returns), combined forecasts based on the *csl* score function statistically outperform the density forecasts based on the log scoring rule. In addition, the benchmark forecasts are also improved, especially when $\kappa = 0.25$, as indicated by the t-statistics 1.65 and 1.77 (DJIA), 3.85 and 4.62 (FTSE) and 1.66 (Nikkei). However for the S&P 500 returns this improvement is not statistically significant.

Panel B shows test results of using combined forecasts based on the left tail and the individual censored density forecasts. In general, using the *csl* score function results statistically in better density forecasts. This holds in particular when $\kappa = 0.25$.[13] Even if the

[13]Table B.1 in Appendix B provides results where the weights are based on the log score function. The

Figure 2: Pooling weights DJIA index

This figure depicts the evolution of weights based on optimizing the logarithmic score function (upper part) of (3) or the *csl* score function (bottom part) of (5) with a moving window of $T = 750$ one-step ahead evaluated density forecasts using daily returns of the DJIA Index. In case of the *csl* score function, B_t represents the left tail $y_t < \hat{r}^{0.25}$ with $\hat{r}^{0.25}$ the 0.25th quantile of the empirical CDF of the moving estimation window of 750 returns. The labels refer to the models that have the highest weight at a given period. The abbreviations "ST", "Lap" and "N" stand for Skewed-t, Laplace and Normal respectively.



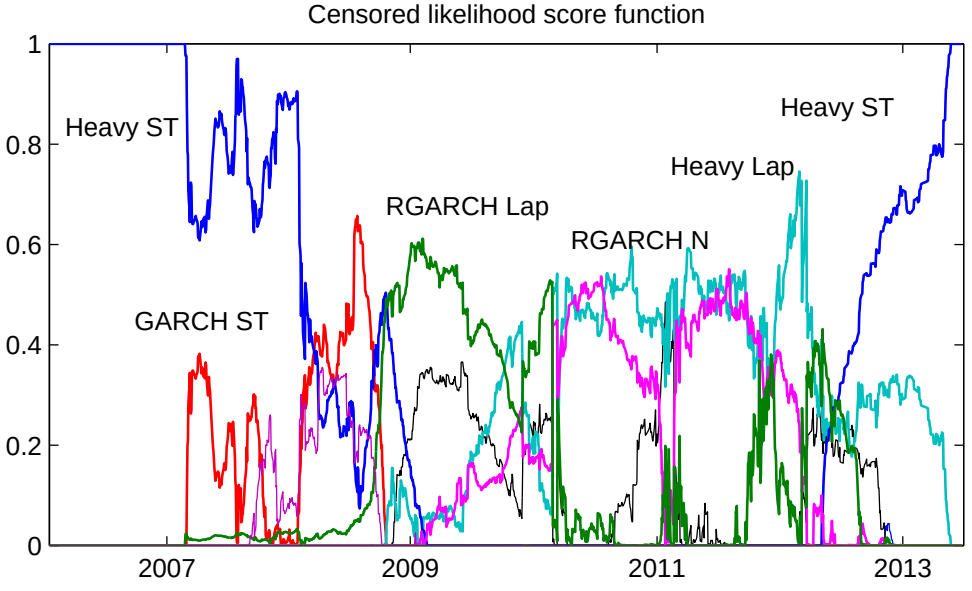


Figure 3: Pooling weights per model and distribution

This figure sums up the optimized weights per model class (top panels) and distribution (bottom panels) based on optimizing the logarithmic score function (left part) of (3) or the *csl* score function of (5) (right part) with a moving window of $T = 750$ one-step ahead evaluated density forecasts using daily returns of the DJIA Index. In case of the *csl* scoring function, B_t represents the left tail $y_t < \hat{r}^{0.25}$ with $\hat{r}^{0.25}$ the 0.25th quantile of the empirical CDF of the moving estimation window of 750 returns.

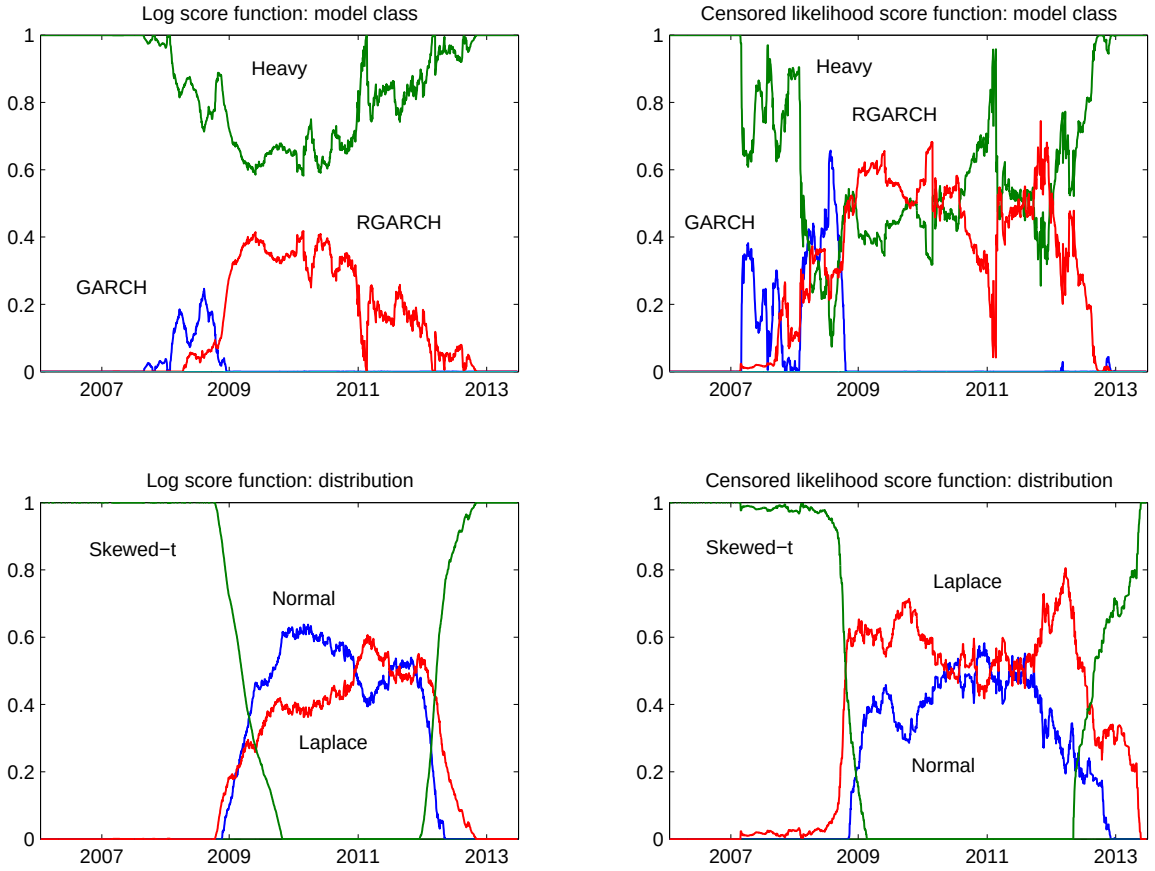


Table 2: Evaluation of 1- and 5-day ahead censored density forecasts

This table reports results of testing equal predictive accuracy using the censored likelihood scoring rule of (4), with B_t the left tail $y_t < \hat{r}^{\kappa}$ with $\hat{r}^{\kappa}$ the κth quantile of the empirical CDF of the in-sample returns. We set κ equal to 0.15 and 0.25 respectively. The weights are repeatedly optimized based on a moving window of 750 evaluated density forecasts. We focus on 1- and 5-step ahead density forecasts. The test statistic is given in (18). Panel A compares combined density forecasts where the weights are based on the *csl* score function with (i) weights based on the log score function and (ii) each competing model gets the same weight. In Panel B, we test equal predictive accuracy of combined density forecasts based on the *csl* score function and density forecasts of each competing model, which are listed in Table 1. All models are estimated with a moving window of 750 daily returns from the S&P500, DJIA, FTSE and Nikkei index through the period January, 2000 - June, 2013. The test statistics are based on HAC-based standard errors and 1864 (S&P 500), 1866 (DJIA), 1882(FTSE) and 1766 (Nikkei) out-of-sample observations respectively.

Panel A: Csl score function vs. log score function and equal weighted								
	S&P500	DJIA	FTSE	Nikkei	S&P500	DJIA	FTSE	Nikkei
	1-step ahead forecasts				5-step ahead forecasts			
	$\kappa = 0.15$							
csl vs log	3.54^{***}	3.60^{***}	-0.09	2.69^{**}	3.76^{***}	3.78^{***}	-1.15	2.36^{**}
csl vs eqw	1.14	1.08	2.36^{**}	0.89	0.74	1.83^{*}	3.17^{***}	1.03
	$\kappa = 0.25$							
csl vs log	3.06^{***}	3.34^{***}	-0.75	2.06^{**}	3.34^{***}	3.36^{***}	-0.96	1.93^{*}
csl vs eqw	1.32	1.65^{*}	3.85^{***}	1.66^{*}	0.73	1.77^{*}	4.62^{***}	1.31
Panel B: Pooled (csl score function) vs. individual								
	$\kappa = 0.15$							
GARCH N	3.22^{***}	3.30^{***}	4.17^{***}	2.27^{**}	2.34^{**}	2.43^{**}	3.74^{***}	2.00^{**}
GARCH T	3.24^{***}	2.96^{***}	4.25^{***}	2.26^{**}	1.56	1.84^{*}	4.41^{***}	1.60
GARCH Lap	1.59	1.44	4.46^{***}	2.65^{***}	0.15	0.80	4.42^{***}	1.34
GARCH ST	6.09^{***}	5.64^{***}	2.92^{***}	3.71^{***}	6.12^{***}	5.96^{***}	2.58^{***}	3.50^{***}
HEAVY N	1.41	1.50	3.13^{***}	1.34	1.58	1.64	3.64^{***}	1.57
HEAVY T	0.92	1.16	2.95^{***}	0.29	0.26	0.75	3.96^{***}	1.25
HEAVY Lap	-0.24	-0.13	3.42^{***}	0.24	-0.99	-0.34	3.78^{***}	0.16
HEAVY ST	5.41^{***}	4.88^{***}	-0.23	3.32^{***}	5.50^{***}	5.01^{***}	0.23	2.88^{***}
RGARCH N	1.44	1.53	3.57^{***}	2.53^{**}	3.09^{***}	3.33^{***}	5.62^{***}	3.58^{***}
RGARCH T	1.22	1.07	3.35^{***}	2.04^{**}	2.82^{***}	3.77^{***}	5.49^{***}	3.44^{***}
RGARCH Lap	0.21	0.26	3.62^{***}	3.10^{***}	1.36	2.38^{**}	5.06^{***}	3.63^{***}
RGARCH ST	5.51^{***}	5.14^{***}	-0.60	5.64^{***}	6.76^{***}	7.32^{***}	1.66^{*}	6.31^{***}
GAS T	2.95^{***}	2.71^{***}	4.27^{***}	2.07^{**}	1.53	1.65^{*}	4.47^{***}	1.78^{*}
GAS Lap	1.45	1.42	4.43^{***}	2.52^{**}	0.09	0.70	4.45^{***}	1.30
	$\kappa = 0.25$							
GARCH N	3.70^{***}	3.89^{***}	5.80^{***}	2.53^{**}	2.53^{**}	2.64^{***}	4.79^{***}	2.09^{**}
GARCH T	3.65^{***}	3.72^{***}	5.88^{***}	3.07^{***}	1.58	1.96^{*}	5.78^{***}	2.00^{**}
GARCH Lap	1.89^{*}	2.11^{**}	5.52^{***}	3.33^{***}	0.11	0.82	5.39^{***}	1.68^{*}
GARCH ST	5.98^{***}	5.60^{***}	2.58^{***}	3.26^{***}	5.63^{***}	5.58^{***}	2.40^{**}	3.13^{***}
HEAVY N	2.12^{**}	2.39^{**}	5.16^{***}	1.52	2.13^{**}	2.26^{**}	5.07^{***}	1.65^{*}
HEAVY T	1.66^{*}	2.43^{**}	4.86^{***}	1.29	0.94	1.65^{*}	5.62^{***}	1.54
HEAVY Lap	0.18	0.81	4.59^{***}	1.40	-0.80	0.20	4.94^{***}	0.85
HEAVY ST	5.00^{***}	4.50^{***}	-1.25	2.35^{**}	5.34^{***}	4.55^{***}	-0.35	2.52^{**}
RGARCH N	2.52^{**}	2.81^{***}	5.62^{***}	2.93^{***}	3.98^{***}	4.10^{***}	7.05^{***}	3.83^{***}
RGARCH T	2.33^{**}	2.31^{**}	5.25^{***}	3.05^{***}	3.68^{***}	4.27^{***}	7.09^{***}	3.93^{***}
RGARCH Lap	0.66	0.86	4.78^{***}	3.86^{***}	1.30	2.33^{**}	6.08^{***}	4.12^{***}
RGARCH ST	6.00^{***}	5.88^{***}	-1.19	7.56^{***}	8.71^{***}	9.16^{***}	2.32^{**}	8.57^{***}
GAS T	3.36^{***}	3.46^{***}	5.89^{***}	2.83^{***}	1.57	1.82^{*}	5.87^{***}	2.19^{**}
GAS Lap	1.75^{*}	2.05^{**}	5.48^{***}	3.30^{***}	0.02	0.69	5.44^{***}	1.74^{*}

null hypothesis cannot be rejected for a particular model, this result is not consistent for all data sets. For example, the Heavy Lap model performs well in case of the US stock market indexes, but is statistically beaten in case of the FTSE returns. In general, there is no striking difference between the 1-step and the 5-step ahead density forecasts. Interestingly, considering the S&P 500 and DJIA indexes in the upper part of Panel B, the RGARCH model with Normal or Student-t distributed errors produces accurate 1-step ahead density forecasts, while forecasting 5 steps ahead results in inaccurate forecasts compared to pooled density forecasts with weights based on the 15th quantile of the individual densities.

Table 3 reports additional evidence of the added value of pooling using the *csl* score function, by providing the *csl* score over the out-of-sample period:

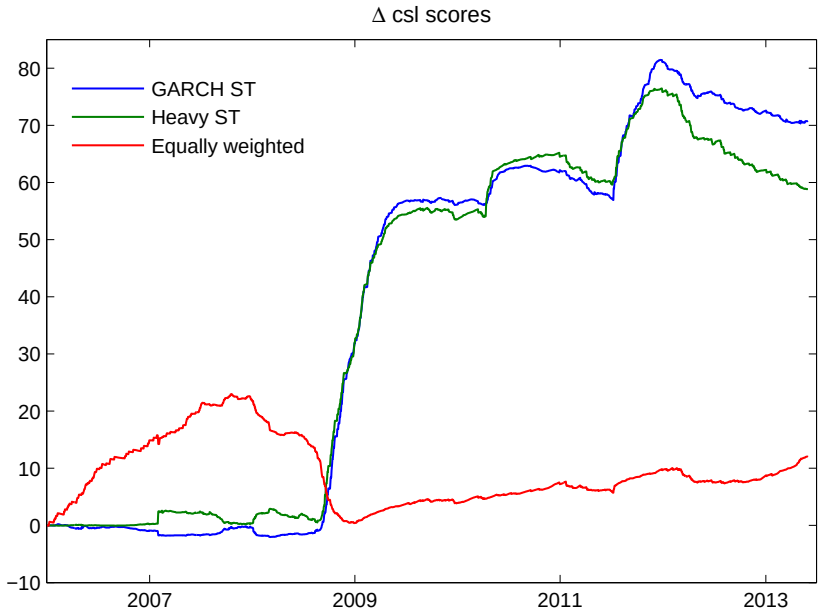
$$\sum_{t=1}^{T} \log \left[\sum_{i=1}^{n} w_{i,t-1}^{*} \left(I[y_t \in B_t] \log p_t(y_t; I_{t-1}, A_i) + I[y_t \in B_t^c] \int_{B_t^c} p_t(y; I_{t-1}, A_i) dy \right) \right] \tag{22}$$

where $w_{i,t-1}^{*}$ is the optimized weight for model A_i at the end of trading day $t-1$, based on the evaluated density forecasts at time $t - T_w$ through $t-1$. In addition, we provide corresponding values of the individual models with bold numbers representing the maximum *csl* score over the competing models. The pooled *csl* scores are higher than the *csl* scores of most of the individual models. If this is not the case, the differences are small, with a maximum difference of 9 points (S&P 500, 5-step ahead forecasts with $\kappa = 0.25$). Further, the *csl* score of our pooling method are higher than the scores of the remaining 13 models (12 in case of the FTSE index), with differences that can be quite substantial. For example, if one favours the best performing individual model, i.e. the Heavy Lap model, this results in a loss of 43 or 46 points with respect to the pooled *csl* score based with $\kappa = 0.15$ in case of the FTSE data set. Finally, there is quite some positive difference between the *csl* scores of pooling with weights based on the *csl* score function and simply using equal weights. For the US and Japanese indexes, the difference is on average around 8 points, however in case of the FTSE index, the difference increases to 20 ($\kappa = 0.15$) or 45 ($\kappa = 0.25$) points

results indicate that the pooled density forecasts do not add any value in case of the S&P 500, DJIA and Nikkei indexes. Only in case of the FTSE index, there is evidence that the combined density forecasts statistically outperform the individual density forecasts.

Figure 4: Censored likelihood scores w.r.t. individual models

This figure depicts the cumulative sum of the difference of the censored likelihood score corresponding with one-step ahead density forecasts of the pooled densities and the *csl* score of the three best competing individual models according to Table 3. The weights of the pooled densities are based on maximizing the *csl* score function of (5) with a moving window of 750 evaluated density forecasts, using daily returns of the DJIA index. Further, B_t the left tail $y_t < \hat{r}^{0.25}$ with $\hat{r}^{0.25}$ the κth quantile of the empirical CDF of the in-sample returns.



respectively. Note that the table relates to Table 2, in the sense that a negative t-stat of a particular model corresponds with a higher *csl* score of that model than the pooled *csl* score. We refer to Table B.2 in Appendix B for similar type of results regarding the log scores.

Figure 4 illustrates the evolution of the cumulative gain in the *csl* scores of Table 3 through time. In particular, it shows the cumulative difference of the *csl* scores corresponding with the combined density forecasts relative to the *csl* scores of the GARCH and Heavy models with a Skewed-t distribution and the benchmark (i.e. pooling based on equal weights). The figure shows two different patterns. First, the gain of pooling with respect to the benchmark occurs mainly during the first years, decreases during the crisis period and increases slowly from 2009 onwards. Second, pooling does not add much value with respect to the Skewed-t distribution during the first years, regardless whether the GARCH or Heavy model class is used. However, the gain becomes striking at the end of 2008 and

Table 3: Censored likelihood scores

This table reports censored likelihood scores corresponding with individual models and combined models, where the weights are based on optimizing the *csl* score function of (5), with B_t the left tail $y_t < \hat{r}^{\kappa}$ with $\hat{r}^{\kappa}$ the κth quantile of the empirical CDF of the in-sample returns. We set κ equal to 0.15 and 0.25. The weights are repeatedly optimized based on a moving window of 750 evaluated density forecasts. In addition, *csl* scores are reported of combined models using equal weights (eqw). The bold numbers represent the maximum of all models per data set. All models are estimated with a moving window of 750 daily returns from the S&P500, DJIA, FTSE and Nikkei index through the period January, 2000 - June, 2013. The number of out-of-sample observations are equal to 1864 (S&P 500), 1866 (DJIA), 1882(FTSE) and 1766 (Nikkei)respectively.

	S&P500	DJIA	FTSE	Nikkei	S&P500	DJIA	FTSE	Nikkei
	1-step ahead forecasts				5-step ahead forecasts			
	$\kappa = 0.15$							
GARCH N	-1052	-1014	-1020	-1002	-1081	-1031	-1050	-1083
GARCH T	-1030	-991	-1002	-937	-1039	-993	-1013	-956
GARCH Lap	-1019	-980	-1001	-934	-1029	-985	-1009	-952
GARCH ST	-1091	-1043	-964	-961	-1105	-1049	-973	-986
HEAVY N	-1026	-990	-988	-979	-1055	-1005	-1027	-1060
HEAVY T	-1013	-976	-983	-919	-1030	-985	-1007	-962
HEAVY Lap	**-1004**	**-967**	-990	**-919**	**-1019**	**-977**	-1004	**-944**
HEAVY ST	-1075	-1028	-946	-950	-1093	-1037	**-959**	-995
RGARCH N	-1022	-983	-992	-1001	-1079	-1027	-1051	-1096
RGARCH T	-1014	-975	-987	-936	-1050	-1010	-1023	-994
RGARCH Lap	-1008	-970	-991	-953	-1040	-998	-1017	-999
RGARCH ST	-1080	-1039	**-944**	-985	-1114	-1074	-969	-1055
GAS T	-1031	-990	-1005	-941	-1039	-992	-1015	-964
GAS Lap	-1019	-980	-1001	-935	-1028	-984	-1010	-951
pooled csl	-1006	-968	-947	-917	-1027	-979	-958	-943
eqw	-1013	-974	-968	-922	-1032	-988	-984	-950
	$\kappa = 0.25$							
GARCH N	-1480	-1443	-1364	-1338	-1508	-1458	-1398	-1421
GARCH T	-1454	-1419	-1343	-1274	-1461	-1417	-1358	-1295
GARCH Lap	-1440	-1404	-1336	-1273	-1447	-1405	-1348	-1291
GARCH ST	-1506	-1460	-1270	-1281	-1521	-1468	-1286	-1312
HEAVY N	-1453	-1421	-1332	-1313	-1486	-1438	-1376	-1398
HEAVY T	-1436	-1405	-1324	**-1253**	-1455	-1414	-1353	-1300
HEAVY Lap	**-1422**	**-1390**	-1324	-1256	**-1437**	**-1399**	-1342	**-1283**
HEAVY ST	-1485	-1439	-1244	-1265	-1512	-1456	**-1265**	-1318
RGARCH N	-1454	-1416	-1337	-1340	-1516	-1463	-1402	-1435
RGARCH T	-1441	-1403	-1328	-1274	-1481	-1441	-1370	-1334
RGARCH Lap	-1427	-1391	-1325	-1290	-1460	-1420	-1356	-1338
RGARCH ST	-1502	-1466	**-1244**	-1337	-1555	-1518	-1285	-1427
GAS T	-1455	-1418	-1346	-1278	-1462	-1416	-1360	-1303
GAS Lap	-1439	-1404	-1336	-1274	-1446	-1404	-1349	-1291
pooled csl	-1420	-1381	-1251	-1243	-1446	-1397	-1268	-1276
eqw	-1430	-1394	-1295	-1255	-1451	-1409	-1317	-1287

in 2011. This result is related with Figure 2, as the Heavy model with Skewed-t distributed returns dominates all the remaining models until the end of 2009.

4.3.2 Economic results

Tables 4 and 5 shed light on the economic impact of pooling (censored) density forecasts by shedding VaR estimates. For each data set, we first compare the frequency and independence of the VaR violations corresponding with the combined densities based on pooling, either using the *csl* or log scoring rule or using equally weights. The latter can be seen as a benchmark. We report results of the approach using equal weights by using simulation corresponding with the actual weight (eqw(1)), and the approach that simply takes the average of all individual VaR's (eqw(2)), as done in Giacomini and Komunjer (2005). Second, we show results of each individual model per data set. Furthermore, we compare the accuracy of the 1-day VaR estimates based on the *csl* scoring rule with VaR estimates from any other pooling method or from any individual model by applying a t-test based on the asymmetric tick loss function of (21). A negative number indicates that the pooled *csl* based VaR estimates are more accurate. Apart from this test, both tables contain the same type of results, although Table 4 focuses on the 1-day VaR estimates, while Table 5 provides results of the 5-day VaR estimates.

Three main conclusions are apparent from Table 4. First, the VaR estimates corresponding with our new proposed technique outperform the benchmark of equal weights, both regarding the frequency of violations and the test on equal accuracy. Using equal weights leads to rejection of the nominal frequency of 5% for the S&P 500 and FTSE indexes, while this is not the case for VaR estimates based on the *csl* score function. Furthermore, the 90% VaR estimates using the *csl* scoring rule are closer to its nominal values using optimized weights. According to the t-statistics of equal accuracy of the VaR estimates, using the *csl* score function produces significantly better VaR estimates than the benchmark in case of the S&P 500 index (both 90% and 95%) and the Nikkei index (90%).

Second, pooling based on the *csl* scoring rule generally outperforms pooling based on the log scoring rule, but only from the perspective of the nominal frequency of the VaR violations. The differences between the *csl* and log scoring rules arise mainly in the case of US stock market indexes. For example, considering the S&P 500, using the whole density

Table 4: Evaluation of 1-day Value-at-Risk estimates

This table provides the accuracy of 1-day VaR estimates. For each data set, the table reports results based on combined density forecasts using the log scoring rule of (1), the *csl* scoring rule of (4) and using equal weights applied on 14 volatility models using daily returns from the S&P 500, DJIA, FTSE and Nikkei index over the period January, 2000 - June, 2013. In case of using equal weights, we report the approach by means of simulation (eqw(1)) and the approach that takes simply the average of all individual VaR estimates (eqw(2)). Further, we report results based on VaR estimates of the individual models. The columns represent for both 95% and 90% VaRs the number of violations, the percentage of violations with respect to the total number of VaR estimates in parentheses, the p-values of the Unconditional Coverage (UC) and Independence (Ind) test of Christoffersen (1998) and finally HAC-based t-statistics of the unconditional test on predictive ability of the combination method/individual model and the combined density forecasts with weights based on the *csl* score function, using the tick-loss function of (21). Bold numbers represent those models which have p-values for the UC and Ind test above 5% for both horizons. The number of estimated VaRs for each series is equal to 1864 (S&P 500), 1866 (DJIA), 1882(FTSE) and 1766 (Nikkei) respectively.

S&P500								
Model/Sc. rule	V(%)	p_{uc}	p_{ind}	t-stat	V(%)	p_{uc}	p_{ind}	t-stat
	95% VaR				90% VaR			
csl	111 (5.97)	**0.063**	**0.484**		200 (10.75)	**0.284**	**0.056**	
log	115 (6.18)	0.024	0.375	0.44	199 (10.70)	0.320	0.032	1.19
eqw(1)	123 (6.61)	0.002	0.089	-2.50^{**}	213 (11.45)	0.041	0.043	-1.90^{*}
eqw(2)	120 (6.45)	0.006	0.117	-2.38^{**}	208 (11.18)	0.095	0.042	-1.39
GARCH N	124 (6.67)	0.002	0.192	-3.50^{***}	192 (10.32)	0.644	0.127	-2.41^{**}
GARCH T	132 (7.10)	0.000	0.385	-3.58^{***}	229 (12.31)	0.001	0.021	-2.98^{***}
GARCH Lap	125 (6.72)	0.001	0.177	-3.35^{***}	229 (12.31)	0.001	0.021	-2.81^{***}
GARCH ST	123 (6.61)	0.002	0.209	-3.50^{***}	207 (11.13)	0.110	0.084	-2.57^{**}
HEAVY N	125 (6.72)	0.001	0.177	-1.92^{*}	191 (10.27)	0.700	0.020	1.21
HEAVY T	133 (7.15)	0.000	0.360	-1.93^{*}	215 (11.56)	0.028	0.371	-0.56
HEAVY Lap	116 (6.24)	0.018	0.165	-0.69	220 (11.83)	0.010	0.166	0.29
HEAVY ST	113 (6.08)	0.039	0.210	-0.47	194 (10.43)	0.539	0.058	0.92
RGARCH N	113 (6.08)	0.039	0.427	-1.92^{*}	185 (9.95)	0.938	0.041	-0.53
RGARCH T	123 (6.61)	0.002	0.209	-2.15^{**}	208 (11.18)	0.095	0.042	-2.40^{**}
RGARCH Lap	108 (5.81)	0.119	0.577	-1.86^{*}	214 (11.51)	0.034	0.038	-2.00^{**}
RGARCH ST	103 (5.54)	**0.295**	**0.750**	-0.96	189 (10.16)	**0.817**	**0.168**	-1.23
GAS T	130 (6.99)	0.000	0.044	-3.28^{***}	223 (11.99)	0.005	0.024	-2.67^{***}
GAS Lap	122 (6.56)	0.003	0.032	-3.03^{***}	223 (11.99)	0.005	0.012	-2.51^{**}
DJIA								
csl	116 (6.23)	0.019	0.744		202 (10.85)	0.228	0.024	
log	120 (6.44)	0.006	0.900	1.19	211 (11.33)	0.060	0.031	0.65
eqw(1)	120 (6.44)	0.006	0.792	-1.06	210 (11.28)	0.071	0.112	-0.79
eqw(2)	118 (6.34)	0.011	0.567	-0.81	209 (11.22)	0.083	0.123	-0.73
GARCH N	130 (6.98)	0.000	0.456	-3.08^{***}	198 (10.63)	0.366	0.074	-1.92^{*}
GARCH T	137 (7.36)	0.000	0.479	-2.95^{***}	221 (11.87)	0.009	0.059	-2.43^{**}
GARCH Lap	116 (6.23)	0.019	0.364	-2.47^{**}	223 (11.98)	0.006	0.204	-2.35^{**}
GARCH ST	124 (6.66)	0.002	0.647	-2.62^{***}	212 (11.39)	0.051	0.150	-2.10^{**}
HEAVY N	130 (6.98)	0.000	0.713	0.12	201 (10.79)	0.258	0.161	1.81^{*}
HEAVY T	138 (7.41)	0.000	0.690	-0.47	214 (11.49)	0.035	0.074	-0.08
HEAVY Lap	103 (5.53)	0.300	0.876	1.18	215 (11.55)	0.029	0.037	0.55
HEAVY ST	112 (6.02)	0.051	0.597	1.39	203 (10.90)	0.200	0.042	1.15
RGARCH N	117 (6.28)	0.014	0.783	-1.15	187 (10.04)	0.951	0.069	0.46
RGARCH T	124 (6.66)	0.002	0.767	-1.69^{*}	202 (10.85)	0.228	0.048	-1.02
RGARCH Lap	103 (5.53)	**0.300**	**0.876**	-1.03	199 (10.69)	**0.328**	**0.117**	-0.49
RGARCH ST	101 (5.42)	0.407	0.798	-0.33	191 (10.26)	0.712	0.045	-0.67
GAS T	133 (7.14)	0.000	0.201	-2.79^{***}	216 (11.60)	0.024	0.059	-1.97^{**}
GAS Lap	113 (6.07)	0.040	0.220	-2.40^{**}	219 (11.76)	0.013	0.074	-2.00^{**}

(continued from previous page)

FTSE								
Sc. rule/Model	V(%)	p_{uc}	p_{ind}	t-stat	V(%)	p_{uc}	p_{ind}	t-stat
	95% VaR				90% VaR			
csl	112 (5.96)	**0.063**	**0.246**		208 (11.08)	**0.126**	**0.483**	
log	108 (5.75)	**0.144**	**0.328**	0.94	206 (10.97)	**0.167**	**0.398**	1.44
eqw(1)	115 (6.12)	0.031	0.195	−1.49	208 (11.08)	0.126	0.232	−1.24
eqw(2)	114 (6.07)	0.039	0.211	−1.54	209 (11.13)	0.109	0.213	−1.34
GARCH N	122 (6.50)	0.004	0.737	−2.52**	203 (10.81)	0.248	0.145	−1.98**
GARCH T	129 (6.87)	0.000	0.502	−2.61***	222 (11.82)	0.010	0.975	−2.39**
GARCH Lap	105 (5.59)	0.248	0.714	−2.26**	224 (11.93)	0.007	0.721	−2.48**
GARCH ST	106 (5.64)	**0.209**	**0.678**	−2.40**	210 (11.18)	**0.093**	**0.422**	−1.93*
HEAVY N	124 (6.60)	0.002	0.405	−0.11	202 (10.76)	0.280	0.691	1.25
HEAVY T	135 (7.19)	0.000	0.179	−0.76	219 (11.66)	0.019	0.433	0.19
HEAVY Lap	97 (5.17)	0.744	0.985	−0.05	227 (12.09)	0.003	0.336	−0.56
HEAVY ST	103 (5.48)	**0.342**	**0.786**	0.75	199 (10.60)	**0.393**	**0.808**	0.99
RGARCH N	128 (6.82)	0.001	0.308	−1.98**	201 (10.70)	0.315	0.105	−0.98
RGARCH T	136 (7.24)	0.000	0.316	−2.22**	220 (11.71)	0.016	0.285	−1.56
RGARCH Lap	101 (5.38)	0.457	0.861	−0.93	225 (11.98)	0.005	0.186	−1.96*
RGARCH ST	112 (5.96)	**0.063**	**0.246**	−0.53	200 (10.65)	**0.353**	**0.190**	−0.53
GAS T	126 (6.71)	0.001	0.354	−2.60***	218 (11.61)	0.023	0.963	−2.34**
GAS Lap	93 (4.95)	0.924	0.416	−2.53**	224 (11.93)	0.007	0.721	−2.26**
Nikkei								
csl	90 (5.11)	0.836	0.161		178 (10.10)	0.887	0.001	
log	89 (5.05)	0.922	0.175	−0.03	172 (9.76)	0.738	0.004	−0.72
eqw(1)	84 (4.77)	0.652	0.248	−0.62	167 (9.48)	0.462	0.007	−1.51
eqw(2)	86 (4.88)	0.818	0.221	−1.05	164 (9.31)	0.328	0.011	−1.78*
GARCH N	104 (5.90)	0.091	0.042	−0.85	166 (9.42)	0.414	0.047	−1.69*
GARCH T	108 (6.13)	0.035	0.027	−1.04	189 (10.73)	0.314	0.030	−1.51
GARCH Lap	88 (4.99)	0.991	0.182	−0.66	181 (10.27)	0.704	0.038	−1.64
GARCH ST	100 (5.68)	0.202	0.062	−0.72	178 (10.10)	0.887	0.026	−1.50
HEAVY N	91 (5.16)	0.752	0.142	1.73*	162 (9.19)	0.254	0.014	1.42
HEAVY T	99 (5.62)	0.242	0.072	1.70*	176 (9.99)	0.987	0.002	2.46**
HEAVY Lap	74 (4.20)	0.113	0.150	1.15	175 (9.93)	0.924	0.002	2.02**
HEAVY ST	88 (4.99)	0.991	0.189	1.93*	172 (9.76)	0.738	0.004	1.90*
RGARCH N	85 (4.82)	0.733	0.566	−2.51**	144 (8.17)	0.008	0.110	−3.56***
RGARCH T	90 (5.11)	**0.836**	**0.417**	−2.71***	160 (9.08)	**0.192**	**0.094**	−2.81***
RGARCH Lap	73 (4.14)	0.089	0.993	−2.88***	151 (8.57)	0.041	0.217	−3.32***
RGARCH ST	76 (4.31)	**0.176**	**0.889**	−3.11***	152 (8.63)	**0.050**	**0.200**	−3.48***
GAS T	106 (6.02)	0.058	0.034	−1.48	191 (10.84)	0.246	0.011	−1.52
GAS Lap	86 (4.88)	0.818	0.213	−1.60	183 (10.39)	0.591	0.002	−2.19**

Table 5: Evaluation of 5-day Value-at-Risk estimates

This table provides the accuracy of 5-day VaR estimates. For each data set, the table reports results based on combined density forecasts using the log scoring rule of (1), the *csl* scoring rule of (4) and using equal weights applied on 14 volatility models using daily returns from the S&P 500, DJIA, FTSE and Nikkei index over the period January, 2000 - June, 2013. I In case of using equal weights, we report the approach by means of simulation (eqw(1)) and the approach that takes simply the average of all individual VaR estimates (eqw(2)). Further, we report results based on VaR estimates of the individual models. For each combination method/model, we have 5 different sub-series of VaRs. The table reports for both 95% and 90% VaRs the sub-series that has the lowest p-value of the test on unconditional coverage of the VaR estimates. The columns represents the corresponding number of violations, the percentage of violations with respect to the total number of VaR estimates in parentheses and the p-values of the Unconditional Coverage (UC) and Independence (Ind) test of Christoffersen (1998). The number of estimated VaRs for each series is equal to 372 (S&P 500), 372 (DJIA), 375(FTSE) and 352 (Nikkei) respectively.

S&P500						
Model/sc. rule	V(%)	p_{uc}	$p_{m,ind}$	V(%)	p_{uc}	p_{ind}
	95% VaR			90% VaR		
csl	24 (6.45)	0.218	0.712	43 (11.56)	0.327	0.615
log	24 (6.45)	0.218	0.712	44 (11.83)	0.252	0.703
eqw(1)	24 (6.45)	0.218	0.615	46 (12.37)	0.141	0.888
eqw(2)	24 (6.45)	0.218	0.615	44 (11.83)	0.252	0.913
GARCH N	26 (6.99)	0.096	0.479	45 (12.10)	0.190	0.461
GARCH T	28 (7.53)	0.037	0.390	55 (14.78)	0.004	0.360
GARCH Lap	25 (6.72)	0.147	0.544	49 (13.17)	0.051	0.237
GARCH ST	24 (6.45)	0.218	0.615	46 (12.37)	0.141	0.396
HEAVY N	27 (7.26)	0.060	0.455	43 (11.56)	0.327	0.993
HEAVY T	28 (7.53)	0.037	0.531	45 (12.10)	0.190	0.821
HEAVY Lap	26 (6.99)	0.096	0.889	47 (12.63)	0.102	0.648
HEAVY ST	15 (4.03)	0.376	0.278	43 (11.56)	0.327	0.993
RGARCH N	23 (6.18)	0.312	0.729	30 (8.06)	0.199	0.802
RGARCH T	24 (6.45)	0.218	0.712	43 (11.56)	0.327	0.993
RGARCH Lap	23 (6.18)	0.312	0.729	33 (8.89)	0.470	0.971
RGARCH ST	13 (3.49)	0.160	0.427	29 (7.80)	0.142	0.337
GAS T	28 (7.53)	0.037	0.363	51 (13.71)	0.023	0.652
GAS Lap	25 (6.72)	0.147	0.320	49 (13.17)	0.051	0.237
DJIA						
csl	25 (6.72)	0.147	0.579	47 (12.63)	0.102	0.648
log	25 (6.72)	0.147	0.579	47 (12.63)	0.102	0.648
eqw(1)	26 (6.99)	0.096	0.844	48 (12.90)	0.073	0.567
eqw(2)	28 (7.53)	0.037	0.492	48 (12.90)	0.073	0.567
GARCH N	32 (8.60)	0.004	0.830	48 (12.90)	0.073	0.567
GARCH T	29 (7.80)	0.022	0.570	50 (13.44)	0.035	0.196
GARCH Lap	28 (7.53)	0.037	0.492	48 (12.90)	0.073	0.284
GARCH ST	28 (7.53)	0.037	0.492	49 (13.17)	0.051	0.237
HEAVY N	26 (6.99)	0.096	0.511	49 (13.17)	0.051	0.492
HEAVY T	27 (7.26)	0.060	0.933	53 (14.25)	0.010	0.494
HEAVY Lap	23 (6.18)	0.312	0.729	51 (13.71)	0.023	0.360
HEAVY ST	24 (6.45)	0.218	0.652	48 (12.90)	0.073	0.567
RGARCH N	23 (6.18)	0.312	0.729	30 (8.04)	0.194	0.763
RGARCH T	25 (6.72)	0.147	0.579	31 (8.31)	0.264	0.782
RGARCH Lap	15 (4.03)	0.376	0.132	30 (8.04)	0.194	0.763
RGARCH ST	14 (3.76)	0.253	0.099	30 (8.04)	0.194	0.763
GAS T	28 (7.53)	0.037	0.492	46 (12.37)	0.141	0.733
GAS Lap	27 (7.26)	0.060	0.418	46 (12.37)	0.141	0.733

(continued from previous page)

FTSE						
Sc. rule/Model	V(%)	p_{uc}	p_{ind}	V(%)	p_{uc}	p_{ind}
	95% VaR			90% VaR		
csl	24 (6.38)	0.237	0.701	33 (8.80)	0.430	0.214
log	23 (6.12)	0.336	0.616	32 (8.53)	0.333	0.171
eqw(1)	24 (6.38)	0.237	0.059	34 (9.07)	0.541	0.085
eqw(2)	23 (6.12)	0.336	0.043	32 (8.53)	0.333	0.047
GARCH N	24 (6.40)	0.232	0.232	34 (9.07)	0.541	0.098
GARCH T	25 (6.65)	0.162	0.080	41 (10.90)	0.564	0.029
GARCH Lap	22 (5.87)	0.453	0.145	33 (8.80)	0.430	0.074
GARCH ST	24 (6.40)	0.232	0.232	33 (8.80)	0.430	0.074
HEAVY N	24 (6.38)	0.237	0.059	33 (8.80)	0.430	0.214
HEAVY T	25 (6.65)	0.162	0.788	41 (10.90)	0.564	0.073
HEAVY Lap	22 (5.85)	0.460	0.144	33 (8.80)	0.430	0.214
HEAVY ST	23 (6.12)	0.336	0.184	33 (8.80)	0.430	0.909
RGARCH N	26 (6.91)	0.106	0.375	32 (8.53)	0.333	0.430
RGARCH T	27 (7.18)	0.068	0.409	33 (8.80)	0.430	0.214
RGARCH Lap	25 (6.65)	0.162	0.312	31 (8.27)	0.250	0.774
RGARCH ST	26 (6.91)	0.106	0.375	31 (8.27)	0.250	0.774
GAS T	24 (6.40)	0.232	0.232	41 (10.90)	0.564	0.029
GAS Lap	21 (5.60)	0.601	0.901	33 (8.80)	0.430	0.192
Nikkei						
csl	13 (3.69)	0.239	0.317	21 (5.95)	0.006	0.804
log	14 (3.98)	0.362	0.281	20 (5.67)	0.003	0.890
eqw(1)	12 (3.40)	0.144	0.414	20 (5.67)	0.003	0.432
eqw(2)	11 (3.12)	0.082	0.341	19 (5.38)	0.002	0.979
GARCH N	14 (3.97)	0.356	0.575	24 (6.80)	0.034	0.767
GARCH T	13 (3.68)	0.234	0.492	27 (7.65)	0.126	0.957
GARCH Lap	12 (3.40)	0.144	0.414	24 (6.80)	0.034	0.767
GARCH ST	13 (3.68)	0.234	0.492	26 (7.37)	0.085	0.951
HEAVY N	13 (3.69)	0.239	0.317	21 (5.95)	0.006	0.804
HEAVY T	14 (3.98)	0.362	0.576	24 (6.80)	0.034	0.767
HEAVY Lap	9 (2.56)	0.021	0.491	19 (5.38)	0.002	0.979
HEAVY ST	14 (3.98)	0.362	0.281	21 (5.95)	0.006	0.804
RGARCH N	12 (3.41)	0.147	0.357	18 (5.10)	0.001	0.009
RGARCH T	12 (3.41)	0.147	0.415	19 (5.38)	0.002	0.002
RGARCH Lap	10 (2.84)	0.044	0.275	17 (4.82)	0.000	0.006
RGARCH ST	12 (3.40)	0.144	0.005	17 (4.82)	0.000	0.006
GAS T	21 (5.97)	0.419	0.802	28 (7.93)	0.181	0.590
GAS Lap	11 (3.12)	0.082	0.341	26 (7.37)	0.085	0.431

implies a violation frequency that is significantly different from 5% (using a significance level of 5%), while it is not significant using the *csl* scoring rule. A similar view arises in case of the DJIA returns, although the *csl* scoring rule produces still too many violations of the 95% VaR. Nevertheless, the unconditional coverage corresponding with the *csl* scoring rule is closer to its nominal value, especially for the 90% VaR estimates (10.85% vs. 11.33%). Finally, according to the t-statistics, there is no significant difference between the accuracy of the VaR of both pooling methods, although all numbers are negative.

Third, there is no single model that consistently outperforms our method of combining density forecasts. Each model fails at least once in the frequency of violations or in the test of equal accuracy. The best competitors are the HEAVY and RGARCH model classes. The RGARCH model with Skewed-t distributed errors outperforms the combined approach in case of the DJIA returns with regards to the frequency of violations, however the unconditional coverage of 10% is borderline significant in case of the Nikkei data set. Moreover, using the same data set, the corresponding t-statistics on equal accuracy favour significantly our combined method.

The differences between our various methods to estimate a VaR become much smaller if we put attention to the 5-day estimated VaRs, as indicated by Table 5. Using the Bonferroni bound corresponding with a 5% significance level, we conclude that using the *csl* scoring rule or the log scoring rule to obtain weights does not make a clear difference in the VaR estimates. In addition, the individual models perform also well. This could be explained partly by the decreasing power of the tests when the number of exceptions decreases. Although there is a gain of using the *csl* scoring rule over the other pooling methods in case of 90% 5-day VaR estimates of the daily Nikkei returns, the difference boils down to one exception.

To summarize, short-horizon VaR estimates improve when using combined density forecasts based on the *csl* score function, either with respect the nominal size and/or with respect to the statistical accuracy using the asymmetric tick-loss function of (21).

5 Conclusion

We investigate the benefits of combining density forecasts based on a specific region of interest. We develop a new density forecast method that combines density forecasts of different models based on the censored likelihood scoring rule (Diks *et al.*, 2011). Using daily returns from the S&P 500, DJIA, FTSE and Nikkei stock market indexes from 2000 until 2013, we apply our technique on recently developed univariate volatility models, including the HEAVY, GAS and Realized GARCH models.

Our results show that density forecasts in the tail are statistically more accurate if one pools density forecasts using the censored likelihood scoring rule than using density forecasts based on the log score rule, using the benchmark of equal weights or density forecasts of any individual volatility model. Second, we show that the 1-day 95% and 90% VaR estimates improve significantly compared to the benchmark forecasting method or the method based on the log scoring rule. Moreover, the VaR estimates of each individual is beaten, either with respect to the nominal frequency of the VaR violations, or with respect to a statistical test on equal accuracy of the VaR estimates. Our results imply that risk managers and portfolio managers should not rely on one single model if they are interested in the left tail. Instead, they should make combinations of density forecasts using the *csl* scoring rule.

References

Aastveit, K.A., K.R. Gerdrup, A.S. Jore and L.A. Thorsrud (2011), Nowcasting GDP in real-time: A density combination approach, Working Paper.

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

Andersen, T., T. Bollerslev, F.X. Diebold and P. Labys (2003), Modeling and forecasting realized volatility, *Econometrica* **71**, 529–626.

Bacharach, J. (1974), Bayesian dialogues, Working Paper.

Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde and N. Shephard (2008), Designing re-

alised kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica* **76**, 1481–1536.

Bates, J.M. and C.W.J. Granger (1969), The combination of forecasts, *Operational Research* **20**, 451–468.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.

Christoffersen, P. (1998), Evaluating interval forecasts, *International Economic Review* **39**, 841–862.

Conflitti, C., C. De Mol and D. Giannone (2012), Optimal combination of survey forecasts, Working Paper.

Creal, D., S.J. Koopman and A. Lucas (2013), Generalized autoregressive score models with applications, *Journal of Applied Econometrics* **28**, 777–795.

Diebold, F.X. and R.S. Mariano (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**, 253–263.

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

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

Garratt, A., K. Lee, M.H. Peseran and Y. Shin (2003), Forecast uncertainties in macroeconomic modelling: an application to the UK economy, *Journal of the American Statistical Association* **98**, 829–838.

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

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

Giacomini, R. and I. Komunjer (2005), Evaluation and combination of conditional quantile forecasts, *Journal of Business and Economic Statistics* **23**, 416–431.

Glosten, L., R. Jagannathan and D. Runkle (1993), On the relationship between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* **48**, 1779–1801.

Gneiting, T. and A.E. Raftery (2007), Strictly proper scoring rules, prediction and estimation, *Journal of the American Statistical Association* **102**, 359–378.

Granger, C.W.J. and M.H. Pesaran (2000), Economic and statistical measures of forecast accuracy, *Journal of Forecasting* **19**, 537–560.

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

Hansen, B.E. (1994), Autoregressive conditional density estimation, *International Economic Review* **35**, 705–730.

Hansen, P.R., Z. Huang and H.H. Shek (2012), Realized GARCH: A joint model for returns and realized measures of volatility, *Journal of Applied Econometrics* **27**, 877–906.

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

Kupiec, P.H. (1995), Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **3**, 73–82.

Mitchell, J. and S.G. Hall (2005), Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR fan charts of inflation, *Oxford Bulletin of Economics and Statistics* **67**, 995–1033.

Nelson, D.B. (1991), Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.

Rapach, D.E., J.K. Strauss and G. Zhou (2010), Out-of-sample equity premium prediction: Combination forecasts and links to the real economy, *Review of Financial Studies* **23**, 821–862.

Shephard, N. and K. Sheppard (2010), Realising the future: forecasting with high-frequency-based volatility (heavy) models, *Journal of Applied Econometrics* **25**, 197–231.

Stock, J. and M. Watson (2004), Combination forecasts of output growth in a seven-country data set, *Journal of Forecasting* **23**, 405–430.

Timmermann, A. (2006), Forecast combinations, *Handbook of economic forecasting* **1**, 135–196.

Wallis, K.F. (2005), Combining density and invertal forecasts: A modest appraoch, *Oxford Bulletin of Economics and Statistics* **67**, 983–994.

Appendix

A Optimizing weights

We follow Conflitti *et al.* (2012) to optimize the weights according tot the log or *csl* score function of (3) and (5) respectively. We provide here only an outline of the algorithm.

Define $\mathbf{p}(y_{t+1})$ as the vector of n density forecasts $p_i(y_{t+1}) = p_{t+1}(y_{t+1}; Y_t, A_i)$ $(i = 1, \ldots n)$ of the variable y_{t+1} at time t over a one-day horizon. The combined density is then equal to:

$$p(y_{t+1}) = \mathbf{w}^{'}\mathbf{p}(y_{t+1}) = \sum_{i=1}^{n} w_i p_i(y_{t+1}), \tag{A.1}$$

with the assumption that the weights are positive and sum to one. For both scoring rules, we have to maximize the logarithm of the combined (censored) density over a given time period:[14]

$$\Phi(\mathbf{w}) = \frac{1}{T-1}\sum_{t=1}^{T-1} \log p(y_{t+1}). \tag{A.2}$$

Note that we omitted the factor $\frac{1}{T-1}$ in this paper. This does not change the result as it is a constant. Define the $(T-1) \times n$ matrix $\hat{P}$ with non-negative elements $P_{ti} = p_i(y_{t+1})$. Now, (A.2) can be rewritten as $\frac{1}{T-1}\sum_{t=1}^{T-1} \log(P\mathbf{w}_t)$. Denote $\mathbf{w}_{opt}$ as the maximum of $\Phi(\mathbf{w})$ subject to the weight constraints. Further, the Lagrange multiplier is introduced to take into account these constraints:

$$\Phi_\lambda(\mathbf{w}) = \frac{1}{T-1}\sum_{t=1}^{T-1} \log(P\mathbf{w}_t) - \lambda\sum_{i=1}^{N} w_i. \tag{A.3}$$

Instead of optimizing (A.3), Conflitti *et al.* (2012) consider the following "surrogate" func-

[14] For the log scoring rule, it is indeed the log of the combined density. For the *csl* scoring rule, it is the log of the first part (corresponding with the region B_t) or the second part (corresponding with the region outside B_t) of (5).

tion, which depends on a vector $\mathbf{a}$ of arbitrary weights:

$$\Psi_\lambda(\mathbf{w};\mathbf{a}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{i=1}^{n} b_{ti}\log\left(\frac{w_i}{a_i}\sum_{l=1}^{n}\log P_{tl}a_l\right) - \lambda\sum_{i=1}^{n} w_i. \tag{A.4}$$

with $b_{ti} = \frac{P_{ti}a_i}{\sum_{l=1}^{n} P_{tl}a_l}$. Further, the function has the properties $\Psi_\lambda(\mathbf{a};\mathbf{a}) = \Psi_\lambda(\mathbf{a})$ for any $\mathbf{a}$ and $\Psi_\lambda(\mathbf{w};\mathbf{a}) \leq \Psi_\lambda(\mathbf{w})$ for any $\mathbf{a}$ and $\mathbf{w}$.

The iterative algorithm is now defined as

$$\mathbf{w}_\lambda^{(k+1)} = \arg\max_{w} \Psi_\lambda(\mathbf{w};\mathbf{w}_\lambda^{(k)}) \tag{A.5}$$

which yields a monotonic increase of Ψ_λ, according to the two aforementioned properties. Setting the derivatives of $\Psi_\lambda(\mathbf{w};\mathbf{w}_\lambda^{(k)})$ with respect to w_i equal to zero leads to the maximum $w_{\lambda,i} = (1/\lambda)\sum_{t=1}^{T-1} b_{ti}$. Using the constraint that the weights should sum up to one, it holds that $\lambda = T-1$. This changes (A.5) into

$$w_i^{(k+1)} = w_i^{(k)}\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{P_{ti}}{\sum_{l=1}^{n} P_{tl}w_l^{(k)}} \tag{A.6}$$

where we replace a_i by $w_i^{(k)}$ in the expression of b_{ti}. We start the algorithm with equal weights, that is $w_i 0 = 1/n$ and use as a stopping criterion a tolerance of $1e^{-6}$ of the sum of the absolute deviation of two successive iterates.

B Pooling results of the log score function

Figure B.1: Pooling weights of the S&P 500, FTSE and Nikkei index
This figure depicts the evolution of weights based on optimizing the logarithmic score function (left part) of (3) or the *csl* score function (right part) of (5) with a moving window of $T = 750$ one-step ahead evaluated density forecasts using daily returns of the S&P500, FTSE and Nikkei indexes. In case of the *csl* score function, B_t the left tail $y_t < \hat{r}^{0.25}$ with $\hat{r}^{0.25}$ the 0.25th quantile of the empirical CDF of the in-sample returns. The labels refer to the models that have the highest weight at a given period. The abbreviations "ST", "Lap" and "N" stand for Skewed-t, Laplace and Normal respectively.

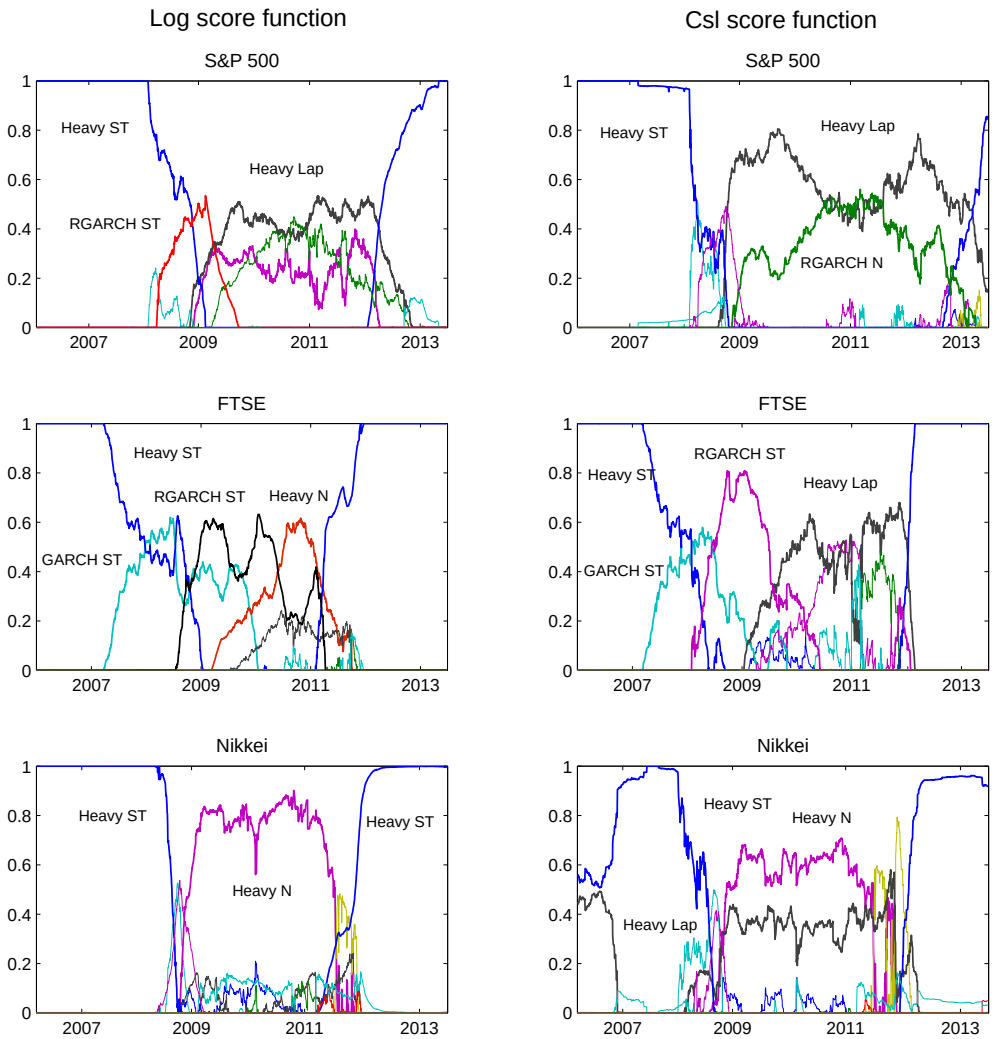


Table B.1: Evaluation of 1- and 5-day ahead censored density forecasts based on the log scoring rule

This table reports results of testing equal predictive accuracy using the censored likelihood scoring rule of (4), with B_t the left tail $y_t < \hat{r}^{\kappa}$ with $\hat{r}^{\kappa}$ the κth quantile of the empirical CDF of the in-sample returns. We set κ equal to 0.15 and 0.25 respectively. The weights are repeatedly optimized based on a the log score function of (3), using a moving window of 750 evaluated density forecasts. We focus on 1- and 5-step ahead density forecasts. The test statistic is given in (18) and compares censored density forecast with weights based on the log score function and density forecasts of each competing model, which are listed in Table 1. All models are estimated with a moving window of 750 daily returns from the S&P500, DJIA, FTSE and Nikkei index through the period January, 2000 - June, 2013. The test statistics are based on HAC-based standard errors and 1864 (S&P 500), 1866 (DJIA), 1882(FTSE) and 1766 (Nikkei) out-of-sample observations respectively.

Pooled (log score function) vs. individual								
	S&P500	DJIA	FTSE	Nikkei	S&P500	DJIA	FTSE	Nikkei
	1-step ahead forecasts				5-step ahead forecasts			
	$\kappa = 0.15$							
GARCH N	0.70	0.64	4.12^{***}	1.86^{*}	0.80	0.73	3.71^{***}	1.82^{*}
GARCH T	-0.78	-0.86	4.15^{***}	0.35	-1.63	-1.50	4.44^{***}	-0.63
GARCH Lap	-1.59	-1.71^{*}	4.38^{***}	0.11	-2.42^{**}	-2.20^{**}	4.64^{***}	-0.92
GARCH ST	6.34^{***}	5.12^{***}	3.37^{***}	3.37^{***}	5.85^{***}	4.98^{***}	3.18^{***}	3.31^{***}
HEAVY N	-0.78	-0.63	3.03^{***}	0.99	-0.19	-0.40	3.62^{***}	1.38
HEAVY T	-2.05^{**}	-2.00^{**}	2.85^{***}	-1.80^{*}	-2.27^{**}	-2.12^{**}	3.95^{***}	0.07
HEAVY Lap	-2.75^{***}	-2.73^{***}	3.34^{***}	-1.50	-3.25^{***}	-2.90^{***}	3.98^{***}	-1.64
HEAVY ST	5.92^{***}	4.90^{***}	-0.31	2.86^{***}	5.39^{***}	4.65^{***}	0.79	2.62^{***}
RGARCH N	-1.11	-1.30	3.45^{***}	2.07^{**}	1.01	0.84	5.54^{***}	3.45^{***}
RGARCH T	-2.06^{**}	-2.08^{**}	3.23^{***}	0.22	-0.77	-0.21	5.53^{***}	2.32^{**}
RGARCH Lap	-2.42^{**}	-2.37^{**}	3.54^{***}	1.64	-1.60	-1.17	5.35^{***}	2.41^{**}
RGARCH ST	5.39^{***}	4.27^{***}	-0.64	5.08^{***}	7.18^{***}	8.19^{***}	2.08^{**}	6.01^{***}
GAS T	-0.66	-0.90	4.20^{***}	0.64	-1.50	-1.47	4.50^{***}	0.29
GAS Lap	-1.59	-1.67^{*}	4.37^{***}	0.19	-2.42^{**}	-2.21^{**}	4.67^{***}	-1.03
	$\kappa = 0.25$							
GARCH N	1.84^{*}	1.76^{*}	5.97^{***}	2.33^{**}	1.44	1.34	5.00^{***}	2.02^{**}
GARCH T	0.72	0.72	5.95^{***}	2.04^{**}	-0.51	-0.45	5.98^{***}	0.86
GARCH Lap	-0.26	-0.24	5.49^{***}	1.96^{*}	-1.48	-1.22	5.46^{***}	0.38
GARCH ST	6.06^{***}	5.21^{***}	3.52^{***}	3.11^{***}	5.36^{***}	4.78^{***}	3.32^{***}	2.87^{***}
HEAVY N	0.47	0.62	5.21^{***}	1.35	0.81	0.69	5.21^{***}	1.56
HEAVY T	-0.56	-0.22	4.87^{***}	0.23	-0.91	-0.62	5.74^{***}	0.97
HEAVY Lap	-1.46	-1.16	4.54^{***}	0.38	-2.21^{**}	-1.66^{*}	4.93^{***}	-0.27
HEAVY ST	5.11^{***}	4.17^{***}	-1.46	2.03^{**}	5.03^{***}	3.92^{***}	0.07	2.44^{**}
RGARCH N	0.56	0.45	5.69^{***}	2.70^{***}	2.37^{**}	2.12^{**}	7.26^{***}	3.82^{***}
RGARCH T	-0.22	-0.33	5.26^{***}	2.02^{**}	0.96	1.18	7.25^{***}	3.18^{***}
RGARCH Lap	-1.11	-1.08	4.73^{***}	2.99^{***}	-0.58	-0.25	6.08^{***}	3.12^{***}
RGARCH ST	6.18^{***}	5.58^{***}	-1.05	7.26^{***}	9.38^{***}	9.89^{***}	2.82^{***}	8.25^{***}
GAS T	0.74	0.63	6.00^{***}	2.13^{**}	-0.43	-0.47	6.06^{***}	1.50
GAS Lap	-0.32	-0.25	5.46^{***}	2.07^{**}	-1.52	-1.28	5.48^{***}	0.41

Table B.2: Log scores

This table reports log scores corresponding with individual models and combined models, where the weights are based on optimizing the log score function of (3). The weights are repeatedly optimized based on a moving window of 750 evaluated density forecasts. The bold numbers represent the maximum of all models per data set. All models are estimated with a moving window of 750 daily returns from the S&P500, DJIA, FTSE and Nikkei index through the period January, 2000 - June, 2013. The number of out-of-sample observations are equal to 1864 (S&P 500), 1866 (DJIA), 1882(FTSE) and 1766 (Nikkei)respectively.

	S&P500	DJIA	FTSE	Nikkei	S&P500	DJIA	FTSE	Nikkei
	1-step ahead forecasts				5-step ahead forecasts			
GARCH N	-2687	-2609	-2338	-2500	-2735	-2648	-2400	-2642
GARCH T	-2651	-2574	-2314	-2434	-2673	-2588	-2348	-2491
GARCH Lap	-2642	-2568	-2338	-2454	-2662	-2584	-2364	-2499
GARCH ST	-2653	-2507	-1948	-2377	-2711	-2564	-2017	-2480
HEAVY N	-2622	-2552	-2269	-2459	-2698	-2611	-2350	-2603
HEAVY T	-2603	-2533	-2263	-2399	-2660	-2578	-2321	-2490
HEAVY Lap	-2603	-2535	-2309	-2430	**-2646**	-2572	-2348	-2483
HEAVY ST	**-2542**	**-2402**	**-1812**	**-2301**	-2683	**-2518**	**-1907**	**-2466**
RGARCH N	-2636	-2564	-2281	-2517	-2780	-2691	-2420	-2742
RGARCH T	-2622	-2549	-2276	-2453	-2728	-2647	-2373	-2584
RGARCH Lap	-2619	-2546	-2315	-2493	-2703	-2627	-2384	-2598
RGARCH ST	-2634	-2541	-1853	-2671	-2898	-2802	-2065	-2987
GAS T	-2655	-2573	-2317	-2439	-2678	-2591	-2353	-2495
GAS Lap	-2643	-2567	-2339	-2457	-2662	-2582	-2366	-2494
pooled log	-2488	-2374	-1867	-2310	-2601	-2472	-1959	-2412